

**PREDIKSI *TWEET NETIZEN* MENGGUNAKAN ALGORITMA
RANDOM FOREST, DECISION TREE, NAÏVE BAYES, DAN
ENSEMBLE (STUDI KASUS GUBERNUR DKI JAKARTA)**



TESIS

Antonius Yadi Kuntoro
14002185

PROGRAM STUDI MAGISTER ILMU KOMPUTER SEKOLAH TINGGI
MANAJEMEN INFORMATIKA DAN KOMPUTER NUSA MANDIRI
JAKARTA
2020

**PREDIKSI *TWEET NETIZEN* MENGGUNAKAN ALGORITMA
RANDOM FOREST, DECISION TREE, NAÏVE BAYES, DAN
ENSEMBLE (STUDI KASUS GUBERNUR DKI JAKARTA)**



TESIS

Diajukan sebagai salah satu syarat untuk memperoleh gelar
Magister Ilmu Komputer (M.Kom)

Antonius Yadi Kuntoro
14002185

PROGRAM STUDI MAGISTER ILMU KOMPUTER SEKOLAH TINGGI
MANAJEMEN INFORMATIKA DAN KOMPUTER NUSA MANDIRI
JAKARTA
2020

SURAT PERNYATAAN ORISINALITAS DAN BEBAS PLAGIARISM

Yang bertanda tangan di bawah ini :

Nama : Antonius Yadi Kuntoro
NIM : 14002185
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : *Data Mining*

Dengan ini menyatakan bahwa tesis yang telah saya buat dengan judul: **“Prediksi *Tweet Netizen Menggunakan Algoritma Random Forest, Decision Tree, Naïve Bayes, dan Ensembles (Studi Kasus Gubernur DKI Jakarta)*”** adalah hasil karya sendiri, dan semua sumber baik yang kutip maupun yang dirujuk telah saya nyatakan dengan benar dan tesis belum pernah diterbitkan atau dipublikasikan dimanapun dan dalam bentuk apapun.

Demikianlah surat pernyataan ini saya buat dengan sebenar-benarnya Apabila dikemudian hari ternyata saya memberikan keterangan palsu dan atau ada pihak lain yang mengklaim bahwa tesis yang telah saya buat adalah hasil karya milik seseorang atau badan tertentu, saya bersedia diproses baik secara pidana maupun perdata dan kelulusan saya dari Program Studi Ilmu Komputer (S2) Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri dicabut/dibatalkan.

Jakarta, 23 January 2020
Yang Menyatakan



Antonius Yadi Kuntoro

PERSETUJUAN DAN PENGESAHAN

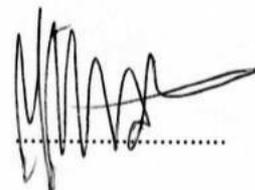
Tesis ini diajukan oleh :
Nama : Antonius Yadi Kuntoro
NIM : 14002185
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : *Data Mining*
Judul Tesis : Prediksi Tweet Netizen Menggunakan Algoritma
Random Forest, Decision Tree, Ensemble, dan Naïve
Bayes (Studi Kasus Gubernur DKI)

Telah dipertahankan pada periode 2019 - 2020 dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Ilmu Komputer (M.Kom) pada Program Studi Ilmu Komputer (S2) Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri)

Jakarta, 23 Januari 2020

PEMBIMBING TESIS

Dosen Pembimbing : Dr. Yan Rianto, M.Eng



DEWAN PENGUJI

Penguji I : Dr. Windu Gata, M.Kom



Penguji II : Dr. Tr. Lindung Parningotan, M.T.I



Penguji III / Dosen Pembimbing : Dr. Yan Rianto, M.Eng



SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Yang bertanda tangan di bawah ini, saya :

Nama : Antonius Yadi Kuntoro
NIM : 14002185
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : *Data Mining*
Jenis Karya : Tesis

Demi pengembangan ilmu pengetahuan, dengan ini menyetujui untuk memberikan izin kepada pihak Program Studi Ilmu Komputer (S2) Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri (STMIK Nusa Mandiri) **Hak Bebas Royalti Non-Eksklusif (*Non-exclusive Royalti-Free Right*)** atas karya ilmiah kami yang berjudul : **“Prediksi Tweet Netizen Menggunakan Algoritma *Random Forest, Decision Tree, Naïve Bayes, dan Ensemble* (Studi Kasus Gubernur DKI Jakarta)”** beserta perangkat yang diperlukan (apabila ada).

Dengan **Hak Bebas Royalti Non-Eksklusif** ini pihak STMIK Nusa Mandiri berhak menyimpan, mengalih-media atau *bentuk*-kan, mengelolanya dalam pangkalan data (*database*), mendistribusikannya dan menampilkan atau mempublikasikannya di *internet* atau media lain untuk kepentingan akademis tanpa perlu meminta izin dari kami selama tetap mencantumkan nama kami sebagai penulis/pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak STMIK Nusa Mandiri, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini .

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 23 January 2020
Yang Menyatakan



Antonius Yadi Kuntoro

KATA PENGANTAR

Puji syukur kepada Tuhan Yang Maha Esa atas berkat dan anugerah-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul **PREDIKSI *TWEET NETIZEN MENGGUNAKAN ALGORITMA RANDOM FOREST, DECISION TREE, NAÏVE BAYES, DAN ENSEMBLE (STUDI KASUS GUBERNUR DKI JAKARTA)***.

Tujuan penulisan tesis ini dibuat sebagai salah satu syarat untuk mendapatkan gelar Magister Ilmu Komputer (M.Kom) pada Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri. Penulis dalam menyusun tesis ini tidak lepas dari bantuan banyak pihak. Oleh karenanya dan dari hati yang paling dalam, penulis ingin menyampaikan rasa terima kasih dan penghargaan penulis kepada:

1. Ibu Dr. Dwiza Riana, S.Si, MM, M.Kom, selaku ketua Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri.
2. Bapak Dr. Yan Rianto, M.Eng, selaku pembimbing tesis yang telah menyediakan waktu, pikiran dan tenaga dalam membimbing penulis dalam menyelesaikan tesis ini.
3. Dosen / Staff / Karyawan di lingkungan Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri yang telah memberikan pelajaran yang berarti bagi penulis selama menempuh studi.
4. Keluarga besar kami yang telah memberikan dukungan moral maupun spiritual.
5. Sahabat Penulis atas keceriaan, dukungan dan semangat yang diberikan.
6. Tak lupa terima kasih penulis ucapkan bagi semua pihak yang tidak dapat penulis ungkapkan satu per satu.

Akhir kata semoga tesis ini dapat berguna bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Jakarta, 23 January 2020



Antonius Yadi Kuntoro
Penulis

ABSTRAK

Nama : Antonius Yadi K
NIM : 14002185
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : *Data Mining*
Judul : “Prediksi *Tweet Netizen* Menggunakan Algoritma *Random Forest*,
Decision Tree, *Naïve Bayes*, dan *Ensemble* (Studi Kasus Gubernur
DKI Jakarta)”

Gubernur DKI Jakarta saat ini, meskipun sudah terpilih sejak tahun 2017 selalu menarik untuk di bicarakan atau bahkan dikomentari. Komentar yang muncul bisa dari media langsung ataupun melalui media social. *Twitter* menjadi salah satu media sosial yang kerap kali digunakan sebagai media untuk memberi komentar terhadap gubernur terpilih bahkan bisa menjadi topik trending di media sosial twitter. *Netizen* yang berkomentar pun beragam, ada yang selalu men-*tweet* kritikan, ada yang berkomentar *positive*, bahkan ada yang hanya melakukan *re-tweet*. Dalam penelitian ini akan dilakukan prediksi apakah *netizen* yang *active* terdapat kecenderungan selalu mengarah pada komentar *positive* atau *negative*. Model algoritma yang digunakan adalah *Decision Tree*, *Naïve Bayes*, *Random Forest* dan juga *Ensemble*. Data *twitter* yang di olah harus melalui *preprocessing* terlebih dahulu sebelum di olah lebih lanjut menggunakan *Rapidminer*. Dalam uji coba menggunakan *Rapidminer* dilakukan dalam empat kali percobaan dengan membagi dalam dua bagian yaitu data *testing* dan *data training*. Perbandingan yang dilakukan adalah data *testing* 10% : data *training* 90%, kemudian data *testing* 20% : data *training* 80%, lalu data *testing* 30% : data *training* 70% dan yang terakhir adalah data *testing* 35% : data *training* 65%. Hasil akurasi rata-rata untuk algoritma *Decision Tree* adalah 93,15%, sedangkan untuk algoritman *Naïve Bayes* akurasinya sebesar 91,55%, kemudian untuk algoritma *Random Forest* adalah sebesar 93,41, dan yang terakhir adalah algoritma *Ensemble* dengan akurasi sebesar 93,42%.

Kata kunci: *Decision Tree*, *Naive Bayes*, *Random Forest*, *Ensemble*, *twitter*.

ABSTRACT

Nama : Antonius Yadi K
NIM : 14002185
Program Studi : Ilmu Komputer
Jenjang : Strata Dua (S2)
Konsentrasi : Data Mining
Judul : “Prediksi *Tweet Netizen* Menggunakan Algoritma *Random Forest*, *Decision Tree*, Naïve Bayes, dan Ensemble (Studi Kasus Gubernur DKI Jakarta)”

The current Governor of DKI Jakarta, even though he has been elected since 2017 is always interesting to talk about or even comment on. Comments that appear come from the media directly or through social media. Twitter has become one of the social media that is often used as a media to comment on elected governors and can even become a trending topic on Twitter social media. Netizens who comment are also varied, some are always tweeting criticism, some are commenting positively, and some are only re-tweeting. In this research, a prediction of whether active netizens will tend to always lead to positive or negative comments will be carried out in this study. Model algorithms used are Decision Tree, Naïve Bayes, Random Forest and also Ensemble. Twitter data that is processed must go through preprocessing first before proceeding using Rapidminer. In trials using Rapidminer conducted in four trials by dividing into two parts, namely testing data and training data. Comparisons made are 10% testing data: 90% Training data, then 20% testing data: 80% training data, then 30% testing data: 70% training data and the last is 35% testing data: 65% training data. The average accuracy for the Decision Tree algorithm is 93.15%, while for the Naïve Bayes algorithm the accuracy is 91.55%, then for the Random Forest algorithm is 93.41, and the last is the Ensemble algorithm with an accuracy of 93, 42%.

Key Words: Decision Tree, Naive Bayes, Random Forest, Ensemble, Twitter.

DAFTAR ISI

	Halaman
HALAMAN SAMPUL	i
HALAMAN JUDUL	ii
HALAMAN PERNYATAAN ORISINALITAS	iii
HALAMAN PERSETUJUAN DAN PENGESAHAN	iv
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS	v
KATA PENGANTAR	vi
ABSTRAK	vii
ABSTRACK	viii
DAFTAR ISI	ix
DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN	xiv
BAB I PENDAHULUAN	1
1.1. Latar Belakang Masalah	1
1.2. Identifikasi Masalah	3
1.3. Tujuan Penelitian	4
1.4. Ruang Lingkup	4
1.5. Rumusan Masalah	4
1.6. Sistematika Penulisan	4
BAB II LANDASAN TEORI	6
2.1. Tinjauan Pustaka	6
2.1.1. Pengertian <i>Data Mining</i>	6
2.1.2. Fungsi <i>Data Mining</i>	6
2.2. Klasifikasi	10
2.2.1. Algoritma <i>Random Forest</i>	11
2.2.2. Algoritma <i>Decision Tree</i>	12
2.2.3. Algoritma <i>Naïve Bayes</i>	13
2.2.4. Algoritma <i>Ensemble Classification</i>	15
2.2.5. <i>Twitter</i>	16
2.2.6. <i>Netizen</i>	16
2.3. Pengujian <i>K-Fold Cross Validation</i>	16
2.4. Evaluasi dan Validasi Hasil	17
2.5. <i>Confusion Matrix</i>	18
2.6. Kurva ROC	19
2.7. Pareto	20
2.8. Tinjauan Studi	21
2.9. Kerangka Pemikiran	24

BAB III METODOLOGI PENELITIAN	26
3.1. Metode Penelitian	26
3.2. Kerangka Kerja Penelitian	26
3.2.1 <i>Business / Research Understanding Phase</i>	27
3.2.2. <i>Data Understanding Phase</i>	27
3.2.3. <i>Data Preparation Phase</i>	28
3.2.4. <i>Modeling Phase</i>	28
3.2.5. <i>Evaluation Phase</i>	29
3.2.6. <i>Deployment Phase</i>	30
BAB IV HASIL PENELITIAN DAN PEMBAHASAN	31
4.1 Hasil Penelitian	31
4.1.1. <i>Business / Research Understanding Phase</i>	31
4.1.2. <i>Data Understanding Phase</i>	31
4.2. <i>Data Preparation and modeling</i>	32
4.3 Hasil Eksperimen dan Pengujian Model	39
4.3.1 Hasil Eksperimen dan Pengujian Model <i>Decision Tree</i> ...	40
4.3.2 Hasil Eksperimen dan Pengujian Model <i>Naive Bayes</i>	45
4.3.3 Hasil Eksperimen dan Pengujian Model <i>Ensemble Vote</i> ...	51
4.3.4 Hasil Eksperimen dan Pengujian Model <i>Random Forest</i> ...	57
4.4 Perbandingan Keseluruhan Model	62
4.5 Analisa Prediksi <i>Tweet Netizen</i>	65
BAB V PENUTUP	66
5.1 Kesimpulan	66
5.2 Saran	66
DAFTAR PUSTAKA	68
DAFTAR RIWAYAT HIDUP	71
HALAMAN KONSULTASI BIMBINGAN	72
LAMPIRAN-LAMPIRAN	74

DAFTAR TABEL

		Halaman
Tabel 2.1	<i>Model Confusion Matrix</i> (Gorunescu, 2011)	18
Tabel 3.1	Spesifikasi <i>Hardware</i> dan <i>Software</i>	30
Tabel 4.1	Atribut tambahan dari data olah	33
Tabel 4.2	Perbandingan Text sebelum dan setelah proses <i>@Anonation Removal</i>	35
Tabel 4.3	Perbandingan Text sebelum dan setelah proses <i>@ Transformation: Remove URL</i>	36
Tabel 4.4	Perbandingan Text sebelum dan setelah proses <i>@ Tokenization: Regexp</i>	37
Tabel 4.5	Perbandingan Text sebelum dan setelah proses <i>Indonesian Stemming</i>	38
Tabel 4.6	Perbandingan Text sebelum dan setelah proses <i>Transformation: Not (Negative)</i>	39
Tabel 4.7	Perbandingan Text sebelum dan setelah proses <i>Indonesian Stop word removal</i>	39
Tabel 4.8	Rekapitulasi dataset percobaan <i>Rapidminer</i>	63
Tabel 4.9	Hasil Rata-rata dataset percobaan <i>Rapidminer</i>	64

DAFTAR GAMBAR

	Halaman
Gambar 1.1	Contoh berita mengenai Gubernur DKI Jakarta 1
Gambar 1.2	Contoh <i>tweet</i> dari <i>netizen</i> 2
Gambar 2.1	Proses CRISP-DM (Larose, 2005) 7
Gambar 2.2	Taksonomi Peranan <i>Data Mining</i> (Gorunescu, 2011) 10
Gambar 2.3	Metode <i>Ensemble learning</i> 15
Gambar 2.4	Ilustrasi <i>10-Fold Cross Validation</i> (Paprotny, 2014) 17
Gambar 2.5	Grafik ROC (<i>discrete</i> dan <i>continuous</i>) (Gorunescu, 2011) 20
Gambar 2.6	Kerangka pemikiran 25
Gambar 3.1	Tahapan Penelitian 26
Gambar 4.1	Contoh data hasil <i>download twitter</i> 32
Gambar 4.2	Contoh lembar kerja <i>preprocessing</i> 33
Gambar 4.3	Contoh data hasil proses <i>Gataframework</i> 34
Gambar 4.4	Tampilan <i>Gataframework</i> 34
Gambar 4.5	Tampilan <i>Rapidminer</i> model <i>Decision Tree</i> 40
Gambar 4.6	<i>Confussion Matrix Decision Tree</i> data <i>testing</i> 10% dan data <i>training</i> 90% 41
Gambar 4.7	Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Decision Tree</i> , data <i>testing</i> 10% dan data <i>training</i> 90% 41
Gambar 4.8	<i>Confussion Matrix Decision Tree</i> data <i>testing</i> 20% dan data <i>training</i> 80% 41
Gambar 4.9	Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Decision Tree</i> , data <i>testing</i> 20% dan data <i>training</i> 80% 42
Gambar 4.10	<i>Confussion Matrix Decision Tree</i> data <i>testing</i> 30% dan data <i>training</i> 70% 43
Gambar 4.11	Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Decision Tree</i> , data <i>testing</i> 30% dan data <i>training</i> 70% 43
Gambar 4.12	<i>Confussion Matrix Decision Tree</i> data <i>testing</i> 35% dan data <i>training</i> 65% 44
Gambar 4.13	Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Decision Tree</i> , data <i>testing</i> 30% dan data <i>training</i> 70% 45
Gambar 4.14	Tampilan <i>Rapidminer</i> model <i>Decision Tree</i> 45
Gambar 4.15	<i>Confussion Matrix Naïve Bayes</i> data <i>testing</i> 10% dan data <i>training</i> 90% 46
Gambar 4.16	Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Naïve Bayes</i> , data <i>testing</i> 10% dan data <i>training</i> 90% 46
Gambar 4.17	Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Decision Tree</i> , data <i>testing</i> 20% dan data <i>training</i> 80% 47
Gambar 4.19	Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Naïve Bayes</i> , data <i>testing</i> 20% dan data <i>training</i> 80% 48
Gambar 4.19	Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Naïve Bayes</i> , data <i>testing</i> 30% dan data <i>training</i> 70% 48
Gambar 4.20	Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Naïve Bayes</i> , data <i>testing</i> 30% dan data <i>training</i> 70% 49

DAFTAR GAMBAR

	Halaman
Gambar 4.21 Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Naïve Bayes</i> , data <i>testing</i> 35% dan data <i>training</i> 65%	50
Gambar 4.22 Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Naïve Bayes</i> , data <i>testing</i> 35% dan data <i>training</i> 65%	50
Gambar 4.23 Tampilan <i>Rapidminer</i> model <i>Ensemble</i>	51
Gambar 4.24 Tampilan <i>Performance</i> pada <i>Rapidminer</i> model <i>Ensemble future Vote</i>	51
Gambar 4.25 <i>Confussion Matrix Ensemble</i> , data <i>testing</i> 10% dan data <i>training</i> 90%	52
Gambar 4.26 Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Ensemble</i> , data <i>testing</i> 10% dan data <i>training</i> 90%	52
Gambar 4.27 <i>Confussion Matrix Ensemble</i> , data <i>testing</i> 20% dan data <i>training</i> 80%	53
Gambar 4.28 Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Ensemble</i> , data <i>testing</i> 20% dan data <i>training</i> 80%	54
Gambar 4.29 <i>Confussion Matrix Ensemble</i> , data <i>testing</i> 30% dan data <i>training</i> 70%	54
Gambar 4.30 Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Ensemble</i> , data <i>testing</i> 30% dan data <i>training</i> 70%	55
Gambar 4.31 <i>Confussion Matrix Ensemble</i> , data <i>testing</i> 35% dan data <i>training</i> 65%	56
Gambar 4.32 Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Ensemble</i> , data <i>testing</i> 35% dan data <i>training</i> 65%	56
Gambar 4.33 Tampilan <i>Rapidminer</i> model <i>Random Forest</i>	57
Gambar 4.34 <i>Confussion Matrix Random Forest</i> , data <i>testing</i> 10% dan data <i>training</i> 90%	57
Gambar 4.35 Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Ensemble</i> , data <i>testing</i> 10% dan data <i>training</i> 90%	58
Gambar 4.36 <i>Confussion Matrix Random Forest</i> , data <i>testing</i> 20% dan data <i>training</i> 80%	58
Gambar 4.37 Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Ensemble</i> , data <i>testing</i> 20% dan data <i>training</i> 80%	59
Gambar 4.38 <i>Confussion Matrix Random Forest</i> , data <i>testing</i> 30% dan data <i>training</i> 70%	60
Gambar 4.39 Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Ensemble</i> , data <i>testing</i> 30% dan data <i>training</i> 70%	60
Gambar 4.40 <i>Confussion Matrix Random Forest</i> , data <i>testing</i> 35% dan data <i>training</i> 65%	61
Gambar 4.41 Grafik <i>Area Under Curve (AUC)</i> Algoritma <i>Ensemble</i> , data <i>testing</i> 35% dan data <i>training</i> 65%	62
Gambar 4.42 Grafik Perbandingan hasil percobaan <i>Rapidminer</i>	63
Gambar 4.43 Grafik Perbandingan Rata-rata hasil percobaan <i>Rapidminer</i> ...	64
Gambar 4.44 Diagram <i>Tree</i> dari model <i>Random Forest</i> , dengan dataset 2000	65

DAFTAR LAMPIRAN

	Halaman
Lampiran 1	Data <i>Twitter</i> 74
Lampiran 2	Atribut Tambahan 79
Lampiran 3	Hasil Proses <i>GataFramework</i> 84

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Twitter merupakan layanan jejaring sosial dan mikroblog daring yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks (Wikipedia, 2019). Demikian halnya dengan *official account Twitter* milik gubernur DKI Jakarta yaitu @aniesbaswedan. Bapak Anies Rasyid Baswedan dan Sandiaga Salahuddin Uno merupakan Gubernur dan Wakil Gubernur Propinsi DKI Jakarta untuk periode 2017-2022. Banyaknya berita perihal gubernur termasuk *tweet* yang *diposting* pada akun bapak Anies Rasyid Baswedan, baik yang bermakna positif, *Negative*, maupun Netral, seperti pada Gambar 1.1 dan Gambar 1.2 berikut :



Sumber: Media elektronik

Gambar 1.1 Contoh berita mengenai Gubernur DKI Jakarta



Sumber: akun *Twitter* @aniesbaswedan

Gambar 1.2 Contoh tweet dari netizen

Oleh karenanya penulis akan melakukan penelitian atas berita-berita tersebut diatas dengan pengambilan data yang dilakukan menggunakan hasil *posting* masyarakat pada *official account Twitter* Bapak Anies Rasyid Baswedan yaitu @aniesbaswedan. Pengambilan data dilakukan dengan efektif dan efisien yang selanjutnya dilakukan koreksi label *Positive* dan *Negative* oleh pihak ketiga yaitu oleh 100 responden. Pemberian label berguna untuk menganalisis pendapat seseorang, evaluasi seseorang, sikap seseorang dan emosi seseorang ke dalam bahasa tertulis, dalam hal ini bisa disebut sebagai sentimen. Salah satu disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar adalah *data mining*. *Data mining* adalah proses melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data (Witten et al., 2011). *Huge of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data* (Jiawei Han & Kamber, 2013). *Data mining* sering dianggap sebagai bagian dari *Knowledge Discovery in Database* (KDD) yaitu sebuah proses mencari pengetahuan yang bermanfaat dari data. Selain itu *data mining* juga dikenal dengan nama *knowledge extraction*, *pattern analysis*, *information harvesting*, dan *Business intelligence*.

Ada 5 peranan utama *data mining*, yaitu: Estimasi, Prediksi, Klasifikasi, Klustering, dan Asosiasi. Algoritma *data mining* yang sering digunakan dalam klasifikasi diantaranya adalah *Naïve Bayes*, *K-Nearest Neighbors*, *Decision Tree*,

ID3, CART, Linear Discriminant Analysis, Logistic Regression, Ensembles dan lain-lain. Namun, pada penelitian ini penulis hanya akan menggunakan algoritma *Random Forest, Decision Tree, Naïve Bayes, dan Ensemble* untuk mengolah, mengklasifikasikan, serta *me-mining knowledge* dari *dataset Twitter* pada akun @aniesbaswedan.

Dalam *data mining*, penelitian mengenai klasifikasi data *posting-an Twitter* sudah pernah dilakukan oleh peneliti lain. Sebagian besar penelitian tersebut difokuskan pada identifikasi variabel prediktor. Ada banyak penelitian dalam literatur terdahulu yang menjelaskan faktor-faktor apa saja yang dapat memaksimalkan proses klasifikasi data *posting-an Twitter*. Faktor-faktor tersebut secara umum dibagi menjadi dua, yaitu faktor *pre-processing* pengolahan data *Twitter* dan *text mining* data *Twitter*.

Data *posting-an* pada *Twitter* atau yang disebut *tweet* (kicauan) sangat rentan terhadap *noisy* data, data hilang atau tidak lengkap, dan data yang tidak konsisten karena biasanya data *tweet* setiap *posting-an* akan berbeda-beda bentuk penulisannya dan sangat heterogen. Untuk itu, harus disiapkan data yang baik, memadai dan representatif merupakan langkah awal yang tidak dapat diabaikan begitu saja. Keandalan informasi yang akan di *mining* dari sebuah *database* yang ada bergantung pada kualitas data yang nantinya akan diproses. Ada beberapa teknik data preprocessing yang dapat digunakan untuk menghasilkan data yang berkualitas. *Data cleaning* dapat diterapkan untuk menghilangkan *noise* dan data yang tidak konsisten. *Data reduction* dapat mengurangi ukuran data. *Data transformation* untuk meningkatkan akurasi dan efisiensi algoritma *data mining* yang melibatkan pengukuran jarak (Jiawei Han & Kamber, 2013) .

1.2. Identifikasi Masalah

Banyak penelitian yang menjadikan pemimpin daerah atau pemimpin negara dan juga *public figure* menjadi *object* penelitian, dan sebagian besar membahas mengenai analisa sentimen atau *sentiment analysis*, menggunakan teori *text mining*. Sedangkan pada penelitian ini, penulis mencoba untuk melakukan penelitian menggunakan data *Twitter* pada *official account Twitter* milik @aniesbaswedan untuk diprediksi apakah *netizen* memiliki kecenderungan untuk *men-tweet* berita *Positive* atau *Negative*.

Pada penelitian ini penulis akan membandingkan algoritma *Random Forest*, *Decision Tree*, *Naïve Bayes*, dan *Ensemble* untuk mendapatkan akurasi terbaik dari prediksi tersebut diatas.

1.3. Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah untuk mendapatkan asumsi terhadap *user id Twitter* apakah akan mengarah ke sentimen *Positive* atau *Negative* dari opini yang di *posting*-nya pada *official account Twitter @aniesbaswedan*, menggunakan algoritma *Random Forest*, *Decision Tree*, *Naïve Bayes*, dan *Ensemble*.

1.4. Ruang Lingkup

Peneliti akan membatasi penelian pada ruang lingkup mengolah, melakukan komparasi algoritma klasifikasi algoritma *Random Forest*, *Decision Tree*, *Naïve Bayes*, dan *Ensemble*. Data yang digunakan dalam penelitian ini adalah data komentar masyarakat pada *official Twitter @aniesbaswedan*.

1.5. Rumusan Masalah

Berdasarkan identifikasi masalah di atas, maka pada penelitian ini berusaha menjawab pertanyaan model mana yang lebih akurat dan tepat antara algoritma *Random Forest*, *Decision Tree*, *Naïve Bayes*, dan *Ensemble* dalam prediksi *user ID* atau *netizen* yang *men-tweet* hal *positive* dan *negative*.

1.6. Sistematika Penulisan

Sistematika penulisan tesis ini terdiri dari lima bab dan tiap bab terdiri dari beberapa sub bab atau sub bahasan, yaitu:

BAB I PENDAHULUAN

Bab ini akan menjelaskan latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian dan manfaat penelitian, dan sistematika penulisan.

BAB II KERANGKA PEMIKIRAN

Bab ini akan berisi landasan-landasan teoritis yang digunakan untuk melaksanakan penelitian dan uraian sistematis dari penelitian-penelitian terkait.

BAB III METODOLOGI PENELITIAN

Bab ini akan berisi metode-metode penelitian yang digunakan yang secara garis besar terbagi menjadi dua, yaitu preprocessing dan evaluasi hasil penelitian.

BAB IV PEMBAHASAN

Pada Bab ini berisi pemaparan hasil penelitian yang digambarkan dalam bentuk uraian, tabel, grafik beserta pembahasannya.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi tentang kesimpulan dari hasil penelitian dan saran dari penulis mengenai penelitian ini selanjutnya.

BAB II

LANDASAN TEORI

2.1. Tinjauan Pustaka

Tinjauan pustaka dalam penulisan tesis ini dilakukan dengan menggunakan literature yang berhubungan dengan judul yang dipilih. Untuk lebih detail, tinjauan pustaka dalam penulisan tesis ini dapat dijelaskan sebagai berikut:

2.1.1 Pengertian *Data Mining*

Data mining adalah proses menemukan pola yang menarik dan pengetahuan dari data yang berjumlah besar (Jiawei Han & Kamber, 2013). Menurut Linoff dan Berry, 2011 *data mining* merupakan suatu pencarian dan analisis dari jumlah data yang sangat besar dan bertujuan untuk mencari arti dari pola dan aturan. Sedangkan menurut (Lee et al., 2013) *data mining* adalah kajian yang meliputi kegiatan pengumpulan, pembersihan, pemrosesan dan analisa sekumpulan data sehingga dengan kegiatan tersebut dapat diperoleh pemahaman yang dalam akan data.

2.1.2 Fungsi *Data Mining*

Salah satu disiplin ilmu yang dapat digunakan untuk menemukan pola atau *mining knowledge* dari suatu *big data* yang ada adalah *data mining*. *Data mining* adalah teknik untuk menemukan dan mendeskripsikan pola-pola yang ada dalam data tersebut dan membuat prakiraan dari data itu (Witten et al., 2011). *Data mining* sering dianggap sebagai bagian dari *Knowledge Discovery in Database (KDD)* yaitu sebuah proses mencari pengetahuan yang bermanfaat dari data atau ekstrasi pola secara otomatis mewakili pengetahuan yang disimpan atau ditangkap secara tersembunyi di dalam sebuah *database* besar, gudang data, *web*, *repositori* informasi lainnya, atau *data stream* (Han et al., 2012).

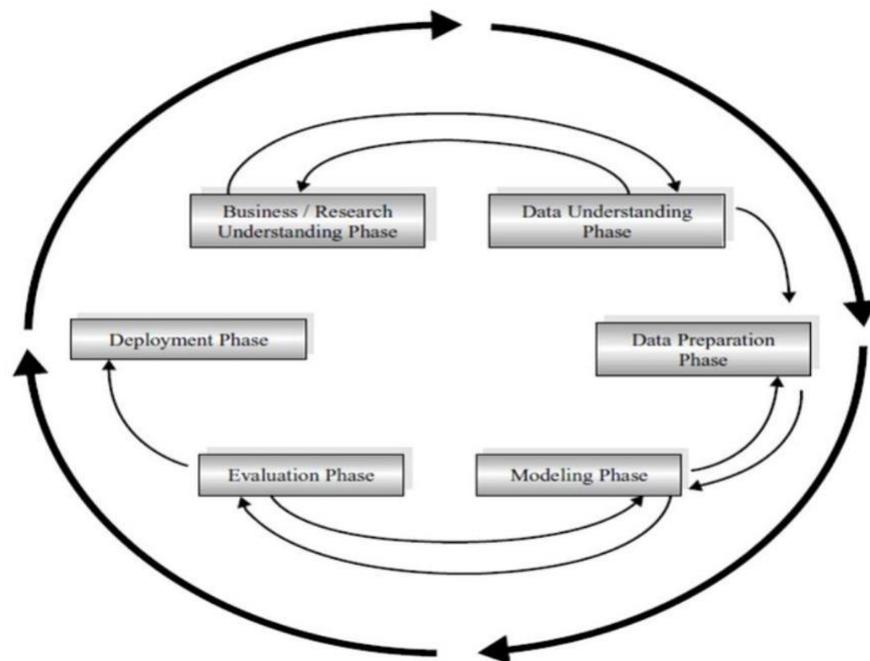
Pada dasarnya, *data mining* dapat dilihat sebagai ilmu yang mengeksplorasi *dataset* dalam jumlah besar untuk penggalian informasi yang tersirat, yang sebelumnya tidak diketahui dan berpotensi menghasilkan informasi yang berguna (Gorunescu, 2011) . *Data mining* adalah suatu proses menemukan hubungan yang berarti, ber-pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Larose, 2005). *Data mining* adalah

proses terorganisir untuk mengidentifikasi pola yang valid, baru, berguna, dan dapat dimengerti dari sebuah *dataset* yang besar dan kompleks (Maimon & Rokach, 2010) .

Kemajuan dalam bidang *data mining* didorong oleh beberapa faktor, antara lain (Larose, 2005):

1. Pertumbuhan yang cepat dalam kumpulan data.
2. Penyimpanan data dalam *data warehouse*, sehingga seluruh perusahaan memiliki akses ke dalam *database* yang andal.
3. Adanya peningkatan akses data melalui *navigasi web* dan *intranet*.
4. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
5. Perkembangan teknologi perangkat lunak untuk *data mining* (ketersediaan teknologi).
6. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

Data mining adalah sebuah proses, sehingga dalam melakukan prosesnya harus sesuai dengan prosedur proses CRISP-DM (*Cross-Industry Standard Process for Data Mining*) yaitu sebagai keseluruhan proses, *preprocessing data*, pembentukan model, model evaluasi, dan akhirnya penyebaran model (Larose, 2005). CRISP-DM menyediakan standar proses *data mining* sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian. Dapat dilihat pada Gambar 2.1 sebagai berikut:



Gambar 2.1 Proses CRISP-DM (Larose, 2005)

1. Fase Pemahaman Bisnis (*Business / Research Understanding Phase*).
 - a. Penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan.
 - b. Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan *data mining*.
 - c. Menyiapkan strategi awal untuk mencapai tujuan.
2. Fase Pemahaman Data (*Data Understanding Phase*)
 - a. Mengumpulkan data.
 - b. Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal.
 - c. Mengevaluasi kualitas data.
 - d. Jika diinginkan, pilih sebagian kecil group data yang mungkin mengandung pola dari permasalahan.
3. Fase Pengolahan Data (*Data Preparation Phase*)
 - a. Siapkan dari data awal, kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu di laksanakan secara intensif.
 - b. Pilih kasus dan variabel yang ingin dianalisis dan yang sesuai analisis yang akan dilakukan.
 - c. Lakukan perubahan pada beberapa variable jika di butuhkan.
 - d. Siapkan data awal sehingga siap untuk perangkat permodelan.
4. Fase Pemodelan (*Modeling Phase*)
 - a. Pilih dan aplikasikan teknik permodelan yang sesuai.
 - b. Kalibrasi aturan model untuk mengoptimalkan hasil.
 - c. Perlu diperhatikan bahwa beberapa teknik mungkin untuk digunakan pada permasalahan *data mining* yang sama.
 - d. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data ke dalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik *data mining* tertentu.
5. Fase Evaluasi (*Evaluation Phase*)
 - a. Mengevaluasi satu atau lebih model yang di gunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektifitas sebelum disebarkan untuk

digunakan.

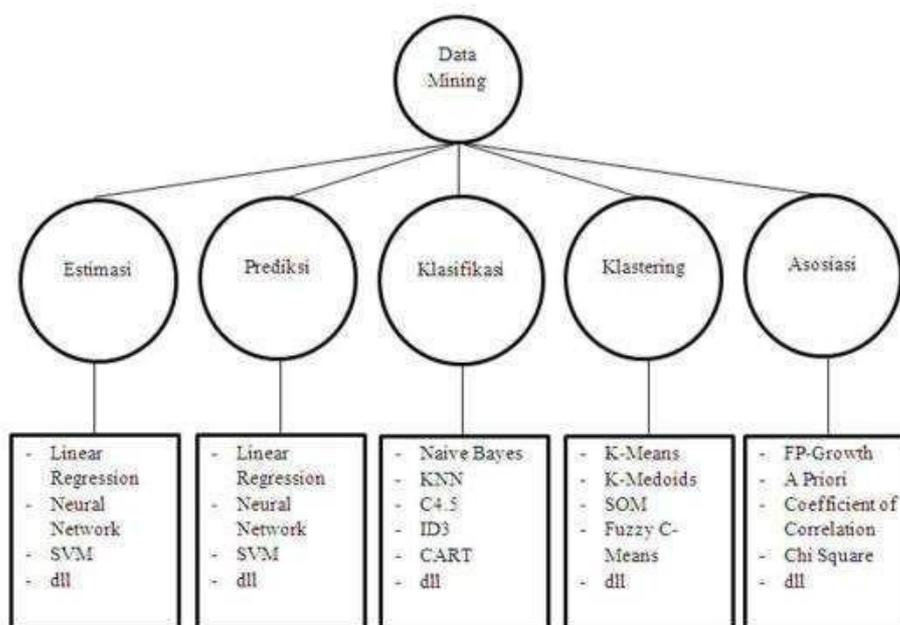
- b. Menetapkan apakah terdapat model yang memenuhi tujuan pada fase awal.
 - c. Menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik.
 - d. Mengambil keputusan berkaitan dengan penggunaan hasil dari *data mining*.
6. Fase Penempatan (*Deployment Phase*)
- a. Menggunakan model yang dihasilkan. Terbentuknya model tidak menandakan telah terselesaikannya proyek.
 - b. Contoh sederhana: Pembuatan laporan.
 - c. Contoh kompleks: Penerapan proses *data mining* secara paralel pada departemen lain.

Terdapat lima peranan utama *data mining* mengacu pada (Larose, 2005), lima peranan tersebut yaitu:

1. Fungsi estimasi (*estimation*)
Fungsi estimasi adalah fungsi untuk memperkirakan suatu hal yang sudah ada datanya. Fungsi estimasi terdiri dari dua cara yaitu Estimasi Titik dan Estimasi Selang Kepercayaan.
2. Fungsi prediksi (*prediction*)
Fungsi prediksi adalah memperkirakan hasil dari hal yang belum diketahui, untuk mendapatkan hal baru yang akan muncul selanjutnya. Cara memprediksi dalam fungsi ini adalah *Regresi Linier*.
3. Fungsi klasifikasi (*classification*)
Fungsi klasifikasi atau menggolongkan suatu data. Cara yang digunakan terdiri dari *algoritma Mean Vector*, *algoritma K-nearest Neighbor*, *algoritma ID3*, *algoritma Decision Tree*, dan *algoritma C5.0*
4. Fungsi pengelompokan (*cluster*)
Fungsi pengelompokan, data yang dikelompokkan disebut objek catatan yang memiliki kemiripan atribut kemudian dikelompokkan pada kelompok yang berbeda. Algoritma yang digunakan adalah *algoritma Hirarchical Clustering*, *algoritma Partitional Clustering*, *algoritma Single Linkage*, *algoritma Complete Linkage*, *algoritma Average Linkage*, *algoritma K-Means* dan lain- lain
5. Fungsi asosiasi (*association*)

Fungsi asosiasi adalah untuk menemukan aturan asosiasi (*association rule*) yang mampu mengidentifikasi item-item yang menjadi objek. Algoritma yang digunakan adalah algoritma *Generalized Association Rule*, *Quantitative Association Rule*, *asynchronous Parallel Mining*.

Kelima peranan di atas terlihat seperti pada gambar 2.2 berikut :



Gambar 2.2 Taksonomi Peranan *Data Mining* (Gorunescu, 2011)

2.2. Klasifikasi

Klasifikasi adalah proses menempatkan obyek atau konsep tertentu kedalam satu set kategori, berdasarkan sifat obyek atau konsep yang bersangkutan (Gorunescu, 2011). Dalam klasifikasi terdapat dua pekerjaan utama yang dilakukan: pertama, pembangunan model sebagai *prototype* untuk disimpan sebagai memori. Kedua, penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut berada. Proses klasifikasi didasarkan pada komponen (Gorunescu, 2011):

1. Kelas (*Class*)

Variabel dependen dari model yang merupakan kategori variabel yang mewakili label-label yang diletakkan pada obyek setelah pengklasifikasian. Contoh: kelas bintang, kelas gempa bumi.

2. Prediktor (*Predictor*)

Variabel independen dari model yang diwakili oleh karakteristik atau atribut dari

data yang diklasifikasikan berdasarkan klasifikasi yang dibuat. Contoh: tekanan darah, status perkawinan, musim.

3. *Dataset* Pelatihan (*Training Dataset*)

Merupakan *dataset* yang berisi dua komponen nilai yang digunakan untuk pelatihan mengenali model yang sesuai dengan kelasnya, berdasarkan prediktor yang ada. Contoh: *database* penelitian gempa, *database* badai, *database* pelanggan supermarket.

4. *Database* Pengujian (*Testing Database*)

Merupakan *dataset* baru yang akan diklasifikasikan oleh model yang dibangun sehingga dapat dievaluasi hasil akurasi klasifikasi tersebut.

2.2.1. Algoritma *Random Forest*

Random Forest merupakan pengembangan dari *Decision Tree*, dimana setiap *Decision Tree* telah dilakukan training menggunakan sampel individu dan setiap atribut dipecah pada *tree* yang dipilih antara atribut subset yang bersifat acak. Dan pada proses klasifikasi, individunya didasarkan pada *vote* dari suara terbanyak pada kumpulan populasi *tree*.

Random Forest juga pengembangan dari metode CART, yaitu dengan menerapkan metode *bootstrap aggregating (bagging)* dan *random feature selection* (Breiman, 2001). Dalam penelitiannya, Breiman telah menunjukkan beberapa kelebihan *Random Forest* antara lain dapat menghasilkan error yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi *data training* dalam jumlah sangat besar secara efisien, dan metode yang efektif untuk mengestimasi missing data. Dalam *Random Forest*, banyak pohon ditumbuhkan sehingga terbentuk hutan (*forest*), kemudian analisis dilakukan pada kumpulan pohon tersebut.

Random Forest dikembangkan dengan ide bahwa perlu ada penambahan layer dari proses *resampling* acak pada *Bagging*. Selain data sampel yang diambil secara acak untuk membentuk pohon klasifikasi, variabel prediktor juga diambil sebagian secara acak dan baru dipilih sebagai pemilah terbaik saat penentuan pemilah pohon (Liaw, 2002). Pada gugus data yang terdiri dari n amatan dan p peubah penjelas, *Random Forest* dilakukan dengan cara (Breiman, 2001):

1. Lakukan penarikan contoh acak berukuran n dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan *bootstrap*.
2. Dengan menggunakan contoh *bootstrap*, pohon / *tree* dibangun sampai mencapai

ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilahan-pemilahan dilakukan dengan memilih m peubah penjelas secara acak, dimana $m \ll p$. Pemilah terbaik dipilih dari m peubah penjelas tersebut. Tahapan ini adalah tahapan *random feature selection*.

3. Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon.

Error klasifikasi diprediksi melalui error OOB (*Out of Bag*) yang diperoleh dengan cara (Breiman, 2001; Liaw, 2002; Breiman & Cutler, 2003):

1. Lakukan prediksi pada setiap data OOB pada pohon yang bersesuaian. Data OOB (*Out of Bag*) adalah data yang tidak termuat dalam contoh *bootstrap*.
2. Secara rata-rata, setiap amatan gugus data asli akan menjadi data OOB sebanyak sekitar 36% dari banyaknya pohon. Oleh karena itu, pada langkah 1, masing-masing amatan gugus data asli mengalami prediksi sebanyak sekitar sepertiga kali dari banyaknya pohon. Jika a adalah sebuah amatan dari gugus data asli, maka hasil prediksi *Random Forest* terhadap a adalah gabungan dari hasil prediksi setiap kali a menjadi data OOB.
3. Error OOB dihitung dari proporsi misklasifikasi hasil prediksi *Random Forest* dari seluruh amatan gugus data asli.

2.2.2. Algoritma *Decision Tree*

Algoritma *Decision Tree* diperkenalkan oleh J. Ross Quinlan yang merupakan pengembang dari algoritma ID3, algoritma tersebut digunakan untuk membentuk pohon keputusan. Pohon keputusan dibuat dengan membagi nilai-nilai atribut menjadi cabang untuk setiap kemungkinan. Cara kerja pohon keputusan yaitu dengan melakukan penelusuran dari akar hingga ke cabang sampai *class* suatu objek ditemukan. Instance diklasifikasikan dengan mengarahkan dari akar pohon sampai ke daun sesuai dengan hasil tes melalui *node internal* (Alfisahrin, 2013). Beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma *Decision Tree* (Ridwan et al., 2013) sebagai berikut :

1. Mempersiapkan *data training*, dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masing atribut atau berdasarkan nilai *index entropy* terendah. Sebelum menghitung nilai *gain* terlebih dahulu mencari nilai *entropy* dengan rumus sebagai

berikut :

$$Entropy(i) = - \sum_{j=1}^m f(i,j) \cdot \log_2 f(i,j) \dots \dots \dots (2.1)$$

Keterangan :

i : himpunan kasus

m : jumlah partisi i

$f(i, j)$: proposi j terhadap i

3. Kemudian menghitung nilai *gain* dengan rumus sebagai berikut :

$$Entropy(i) = - \sum_{i=1}^p \frac{n_i}{n} \cdot IE(i) \dots \dots \dots (2.2)$$

Keterangan :

P : jumlah partisi atribut

n_i : proporsi n_i terhadap i

n : jumlah kasus dalam n

4. Ulangi langkah ke 2 hingga semua nilai *record* terpartisi proses partisi akan berhenti jika :

- a. Semua record dalam node N mendapat kelas yang sama.
- b. Tidak ada atribut dalam record yang terpartisi kembali.
- c. Tidak ada record dalam cabang yang kosong.

2.2.3. Algoritma *Naïve Bayes*

Klasifikasi adalah suatu bentuk analisis data yang ekstrak model menggambarkan kelas data penting. Klasifikasi memiliki berbagai aplikasi termasuk deteksi penipuan, pemasaran target, kinerja prediksi, manufaktur dan diagnosis media karena akurasi yang tinggi, tingkat prediksi dan metode otomatis untuk mencari hipotesis. Pada fase ini, *Naïve Bayes* (NB) *classifier* digunakan untuk mengklasifikasikan Sinyal ECG sebagai normal dan abnormal diimplementasikan dalam alat cepat *Miner. Naïve Bayes* pengklasifikasi adalah statistik pengklasifikasi dapat memprediksi probabilitas keanggotaan kelas seperti probabilitas bahwa *tupel* diberikan milik kelas tertentu (Padmavathi et al., 2015).

Algoritma *Naïve Bayes* merupakan suatu bentuk klasifikasi data dengan menggunakan metode probabilitas dan statistik. Metode ini pertama kali dikenalkan oleh ilmuwan Inggris Thomas Bayes, yaitu digunakan untuk memprediksi peluang yang terjadi dimasa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *teorema bayes* (Nuraeni, 2017).

Metode *teorema bayes* kemudian dikombinasikan dengan *naive* yang diasumsikan dengan kondisi antar atribut yang saling bebas. Algoritma *Naïve Bayes* dapat diartikan sebagai sebuah metode yang tidak memiliki aturan, *Naïve Bayes* menggunakan cabang matematika yang dikenal dengan *teori probabilitas* untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada *data training*.

Persamaan dari *teorema bayes* (Mukminin & Riana, 2017) adalah :

$$P(H|X) = \frac{P(H \cap X) \cdot P(X)}{P(X)} \dots\dots\dots(2.3)$$

Keterangan :

- X : Data dengan *class* yang belum diketahui Y: Hipotesis data yaitu suatu *class* spesifik
- P(H|X) : Probabilitas hipotesis H berdasar kondisi X (*posteriori probability*)
- P(H) : Probabilitas hipotesis H (*prior probability*)
- P(X|Y) : Probabilitas X berdasarkan kondisi pada hipotesis.
- H P (X) : Probabilitas X Penjabaran.

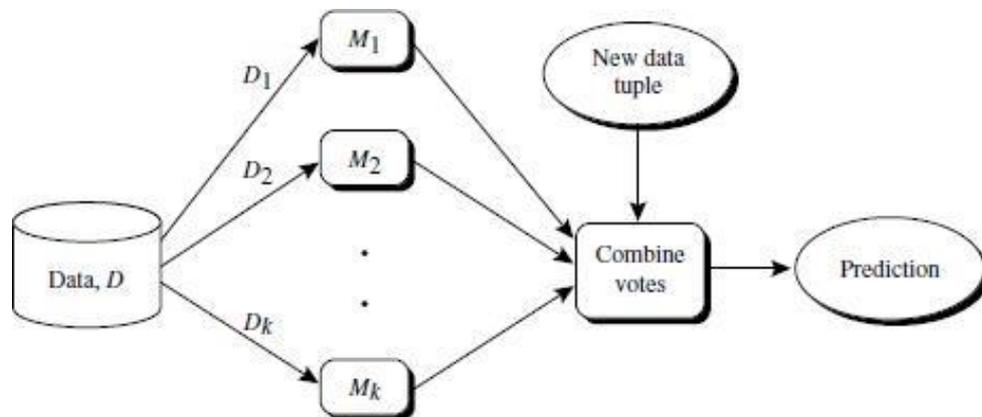
Algoritma *Naïve Bayes* merupakan penyederhanaan metode *bayes*. Untuk mempermudah pemahaman, maka *Theorema Bayes* disederhanakan menjadi :

$$P(H \cap X) = P(X \cap H) P(X) \dots\dots\dots(2.4)$$

Bayes rule diterapkan untuk menghitung *posterior* dan *probabilitas* dari data sebelumnya. Dalam analisis *bayesian*, klasifikasi akhir dihasilkan dengan menggabungkan kedua sumber informasi (*prior* dan *posterior*) untuk menghasilkan probabilitas menggunakan aturan *bayes*.

2.2.4. Algoritma *Ensemble Classification*

Ensemble learning pada klasifikasi adalah metode yang berdasar kepada pendekatan *machine learning* dimana beberapa algoritma pembelajaran dapat digunakan untuk memecahkan sebuah masalah.



Sumber: Han, Kamber, dan Pei (2012)

Gambar 2.3 Metode *ensemble learning*

Pada Gambar 2.3 Han, Kamber dan Pei (2012) menjelaskan dalam memecahkan masalah dapat menggunakan kombinasi sejumlah k model atau *base classifier*, M_1, M_2, \dots, M_k , dengan tujuan yaitu mendapatkan model terbaik dari gabungan sejumlah model. Sumber data D digunakan sebagai *data training* sejumlah k set *data training* D_1, D_2, \dots, D_k dimana D_i ($1 \leq i \leq k$) digunakan untuk menghasilkan *classifier* M_i . Beberapa contoh metode *ensemble learning* adalah *bagging*, *boosting* dan *Stacking*.

Metode *Bagging* pada dasarnya menganut konsep mayoritas suara, dimana *subset* dari *data training* yang berbeda digunakan secara acak dalam melatih *base learner* atau model yang berbeda dengan cara yang sama (Wang, Hao, Ma, & Jiang, 2011). Pembentukan model pada metode *bagging* dilakukan sejumlah k iterasi. Setiap iterasi, model yang dibentuk melakukan prediksi terhadap setiap subset data. Pada setiap iterasi metode *bagging*, model yang dibentuk memiliki bobot suara yang sama. Metode *bagging* akan memilih model klasifikasi dengan suara terbanyak. Model klasifikasi yang dihasilkan oleh metode *bagging* memiliki akurasi yang lebih baik dan cukup signifikan dibandingkan dengan model klasifikasi tunggal (*base*). Peningkatan akurasi tersebut terjadi karena penggabungan model dapat mengurangi variansi dari pengelompokan tunggal (Han, Kamber, & Pei, 2012).

2.2.5. *Twitter*

Merupakan layanan *microblogging* yang populer, pengguna dapat melakukan *posting* status atau pesan yang di sebut sebagai *tweet* yang tidak lebih dari 140 karakter. Sebagian besar kasus, para pengguna menuliskan pesan mereka jauh lebih sedikit dari batasan karakter yang telah ditentukan. *Twitter* merupakan salah satu media sosial penyumbang data terbesar dan dinamis berdasarkan konten buatan pengguna. Diperkirakan sekitar 200 juta pengguna *Twitter* melakukan *posting* 400 juta *tweet* per hari. *Tweet* ini dapat berupa ekspresi pendapat terhadap berbagai macam topik yang dapat membantu agar dapat berbagi pendapat terhadap para pengguna, identifikasi penyimpangan, kejadian yang menimbulkan ketidaknyamanan, prediksi perilaku politik dan olahraga, penerimaan atau penolakan politik dan semua komunikasi yang diutarakan dari mulut ke mulut (Da Silva et al., 2014).

2.2.6 *Netizen*

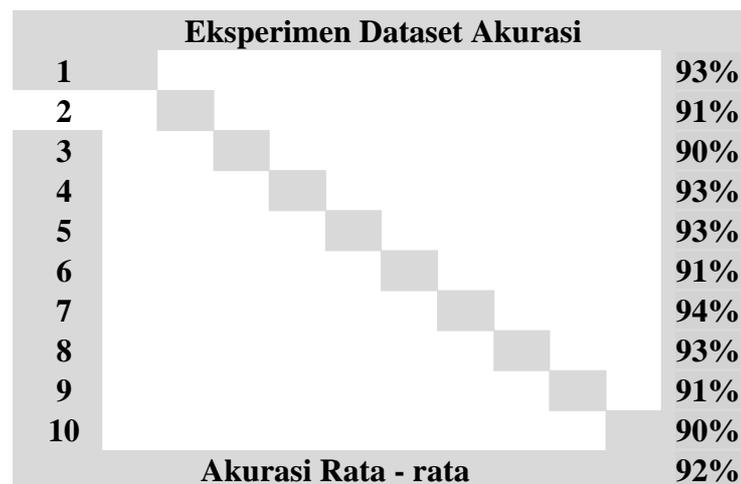
Netizen adalah gabungan dari kata *internet* dan *citizen* yang berarti warga internet atau personal yang menggunakan internet pada umumnya dan aktif dalam media sosial.

2.3. Pengujian *K-Fold Cross Validation*

K-Fold Cross Validation adalah teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi (Han & Kamber, 2006). *Cross validation* adalah metode statistik yang digunakan untuk mengevaluasi dan membandingkan algoritma dengan membagi data menjadi dua segmen, Segmen pertama digunakan sebagai *data training* dan segmen kedua sebagai *data testing* dalam validasi model (Witten, Frank, 2011). *Data training* adalah data yang akan dipakai dalam melakukan pembelajaran sedangkan *data testing* adalah data yang belum pernah dipakai sebagai pembelajaran dan akan berfungsi sebagai data pengujian kebenaran atau keakurasian hasil pembelajaran (Witten & Frank, 2011).

Dengan menggunakan *K-Fold Cross Validation* atau disebut juga dengan *10-Fold Cross Validation* akan dilakukan percobaan sebanyak k . Hasil dari berbagai

percobaan yang ekstensif dan pembuktian teoritis, menunjukkan bahwa *10-Fold Cross Validation* adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat. Oleh karena itu, secara umum pengujian nilai k dilakukan sebanyak 10 kali. *10-Fold Cross Validation* akan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian. Setiap percobaan akan menggunakan satu *data testing* dan $k-1$ bagian akan menjadi *data training*, kemudian *data testing* itu akan ditukar dengan satu buah *data training* sehingga untuk tiap percobaan akan didapatkan *data testing* yang berbeda-beda, seperti yang terlihat pada gambar 2.4 berikut:



k -subset (*data testing*)

Gambar 2.4 Ilustrasi *10-Fold Cross Validation* (Paprotny et al., 2014)

2.4. Evaluasi dan Validasi Hasil

Secara umum pengukuran model *data mining* mengacu kepada tiga kriteria:

a. Akurasi (*Accuracy*)

- Ukuran dari seberapa baik model mengkorelasikan antara hasil dengan atribut dalam data yang telah disediakan.
- Terdapat berbagai model akurasi, tetapi semua model akurasi bergantung pada data yang digunakan.

b. Keandalan (*Reliability*)

- Ukuran dimana model *data mining* diterapkan pada *dataset* yang berbeda.
- Model *data mining* dapat diandalkan jika menghasilkan pola umum yang sama

terlepas dari *data testing* yang disediakan.

c. Kegunaan (*Usefulness*)

- Mencakup berbagai metrik yang mengukur apakah model tersebut memberikan informasi yang berguna.

Keseimbangan diantara ketiganya diperlukan karena belum tentu model yang akurat adalah handal, dan yang handal atau akurat belum tentu berguna.

2.5. *Confusion Matrix*

Confusion Matrix adalah alat (*tools*) visualisasi yang biasa digunakan pada *supervised learning*. Tiap kolom pada matriks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya (Gorunescu, 2011). Sedangkan (Han et al., 2012) menjelaskan bahwa *confusion matrix* adalah tabel yang digunakan untuk menganalisis seberapa baik kualitas pengklasifikasi dapat mengenali data dari kelas yang berbeda. *Confusion matrix* merupakan *matrix* 2 dimensi yang menggambarkan perbandingan antara hasil prediksi dengan kenyataan. Tabel 2.1 adalah contoh tabel *confusion matrix* yang menunjukkan klasifikasi dua kelas.

Tabel 2.1 Model *Confusion Matrix*
(Gorunescu, 2011)

<i>Classification</i>		<i>Predicted Class</i>	
		<i>Class = Yes</i>	<i>Class = No</i>
<i>Observed</i>	<i>Class = Yes</i>	A (<i>true Positive</i> –TP)	B (<i>false Negative</i> –FN)
<i>Class</i>	<i>Class = No</i>	C (<i>false Positive</i> –FP)	D (<i>true Negative</i> –TN)

Keterangan:

TP = proporsi *Positive* dalam data set yang diklasifikasikan *Positive*.

TN = proporsi *Negative* dalam data set yang diklasifikasikan *Negative*.

FP = proporsi *Negative* dalam data set yang diklasifikasikan positif.

FN = proporsi *Negative* dalam *dataset* yang diklasifikasikan *Negative*.

Berikut adalah persamaan model *confusion matrix*:

- a. Nilai *Accuracy* adalah proporsi jumlah prediksi yang benar. Dapat dihitung dengan menggunakan persamaan:

$$\frac{TP+TN}{TP+FP+FN+TN}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots (2.5)$$

- b. *Sensitivity* digunakan untuk membandingkan proporsi TP terhadap tupel yang *Positive*, yang dihitung dengan menggunakan persamaan:

$$Sensitivity = \frac{TP}{TP+FN} \dots\dots\dots (2.6)$$

- c. *Specificity* digunakan untuk membandingkan proporsi TN terhadap tupel yang *Negative*, yang dihitung dengan menggunakan persamaan:

$$Specificity = \frac{TN}{TN+FP} \dots\dots\dots (2.7)$$

- d. PPV (*Positive Predictive Value*) adalah proporsi kasus dengan hasil diagnosa *Positive*, yang dihitung dengan menggunakan persamaan:

$$PPV = \frac{TP}{TP+FP} \dots\dots\dots (2.9)$$

- e. NPV (*Negative Predictive Value*) adalah proporsi kasus dengan hasil diagnosa *Negative*, yang dihitung dengan menggunakan persamaan:

$$NPV = \frac{TN}{TN+FN} \dots\dots\dots (2.10)$$

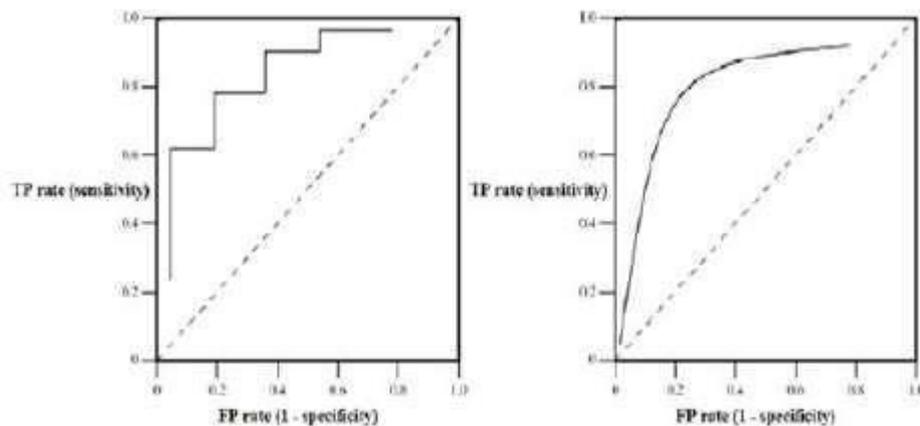
2.6. Kurva ROC

Kurva ROC (*Receiver Operating Characteristic*) adalah alat visual yang berguna untuk membandingkan dua model klasifikasi hasil ekspresi dari *confusion matrix*. ROC adalah grafik dua dimensi dengan *false Positives* sebagai garis horisontal dan *true Positives* sebagai garis vertikal (Vecellis, 2009). Sedangkan (Attenberg & Ertekin, 2013) menjelaskan bahwa ROC adalah ukuran numerik untuk membedakan kinerja model, dan menunjukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan *Positive* dan *Negative*.

Dengan kurva ROC, kita dapat melihat *trade off* antara tingkat dimana suatu model dapat mengenali tupel *Positive* secara akurat dan tingkat dimana model

tersebut salah mengenali tupel *Negative* sebagai tupel *Positive*. Sebuah grafik ROC adalah plot dua dimensi dengan proporsi *Positive* salah (FP) pada sumbu X dan proporsi *Positive* benar (TP) pada sumbu Y. Titik (0,1) merupakan klasifikasi yang sempurna terhadap semua kasus *Positive* dan kasus *Negative*. nilai *Positive* salah adalah tidak ada (FP = 0) dan nilai *Positive* benar adalah tinggi (TP = 1). Titik (0,0) adalah klasifikasi yang memprediksi setiap kasus menjadi *Negative* {-1}, dan titik (1,1) adalah klasifikasi yang memprediksi setiap kasus menjadi *Positive* {1}.

Grafik ROC pada gambar 2.5 dibawah ini, menggambarkan *trade off* antara manfaat (*true Positive*) dan biaya (*false Positives*). Berikut tampilan dua jenis kurva ROC (*discrete* dan *continuous*).



Gambar 2.5 Grafik ROC (*discrete* dan *continuous*) (Gorunescu, 2011)

Poin di atas garis diagonal merupakan hasil klasifikasi yang baik, sedangkan poin di bawah garis diagonal merupakan hasil klasifikasi yang buruk. Dapat disimpulkan bahwa satu poin pada kurva ROC adalah lebih baik daripada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas di dalam grafik. Untuk tingkat akurasi nilai AUC (*Area Under Curve*) dalam klasifikasi *data mining* dibagi menjadi lima kelompok (Gorunescu, 2011):

- a. 0.90 – 1.00 = *Excellent Classification*
- b. 0.80 – 0.90 = *Good Classification*
- c. 0.70 – 0.80 = *Fair Classification*
- d. 0.60 – 0.70 = *Poor Classification*
- e. 0.50 – 0.60 = *Failure*

2.7. Pareto

Prinsip Pareto juga dikenal sebagai aturan 80-20, menyatakan bahwa untuk banyak kejadian, sekitar 80% daripada efeknya disebabkan oleh 20% dari penyebabnya. Prinsip ini diajukan oleh pemikir manajemen bisnis Joseph M. Juran, yang menamakannya berdasarkan ekonom Italia Vilfredo Pareto (Flux & Pareto, 1897). Dengan pendekatan Pareto ini, penulis akan menggunakannya pada proses *preprocessing* pada *dataset* di excel untuk menentukan beberapa atribut baru yang berisi *Active* atau *Passive*.

2.8. Tinjauan Studi

1. Penelitian yang dilakukan oleh (Cureg et al., 2019) yang berjudul “*Sentiment Analysis on Tweets with Punctuations, Emoticons, and Negations*”. Dalam penelitian ini, *tweet* yang menyertakan parameter berbeda (emotikon, negasi, dan tanda baca) dikumpulkan dan dijelaskan oleh para ahli untuk memberi label pada sentimen mereka. Pembelajaran mesin diterapkan dalam penelitian ini untuk merumuskan model yang optimal. Hasil percobaan menunjukkan bahwa fitur yang disertakan memberikan kinerja yang signifikan untuk mengidentifikasi sentimen pernyataan microblog yang diberikan. Algoritma yang digunakan untuk membangun model adalah KNN dan *Naïve Bayes* dengan data bahasa Inggris dan Filipina. Dengan hasil penelitian pada *Calculating Kappa* adalah algoritma k-NN dengan akurasi 40%, algoritma *Naïve Bayes* **32.5%** dan algoritma SVM adalah 38.77%.
2. Penelitian yang dilakukan oleh (Kartiko & Sfenrianto, 2019) yang berjudul “*Accuracy for Sentiment Analysis of Twitter Students on E- Learning in Indonesia using Naïve Bayes Algorithm Based on Particle Swarm Optimization*”. Penelitian ini bertujuan untuk menganalisis akurasi bagi sentimen siswa tentang *E-Learning* yang menggunakan bahasa Indonesia di media sosial *Twitter* baik pendapat *Positive* maupun *Negative*. Algoritma yang digunakan adalah *Naïve Bayes (NB)*. Kemudian untuk mengoptimalkan keakuratan hasil perhitungan, pendekatan *Particle Swarm Optimization (NB-PSO)* digunakan. Untuk mengoptimalkan hasil yang akurat, penelitian ini menggunakan tiga urutan eksperimental (skenario 1, skenario 2, dan skenario 2)

untuk algoritma NB dan NB-PSO. Setiap skenario menggunakan komentar *Positive* dan *Negative* yang berbeda. Hasil percobaan menunjukkan bahwa dalam skenario 1 peningkatan akurasi adalah 10,00% untuk NB-PSO. Skenario 2 ada peningkatan akurasi 13,33% pada NB-PSO. Sementara itu, dalam skenario 3 peningkatan akurasi adalah 27,22% untuk NB-PSO. Hasil ini membuktikan bahwa akurasi NB-PSO lebih baik daripada NB untuk semua skenario.

3. Penelitian yang dilakukan oleh (Al-Rubaiee et al., 2016) yang berjudul “*Analysis of the Relationship Between Saudi Twitter Posts and the Saudi Stock Market*”. Dalam tulisan ini, *Twitter* telah dipilih sebagai platform untuk penambangan pendapat dalam strategi perdagangan dengan pasar saham Saudi untuk melaksanakan dan menggambarkan hubungan antara *tweet* Saudi (yaitu dialek standar dan Teluk Arab) dan indeks pasar Saudi. Sepengetahuan kami, ini adalah studi pertama yang dilakukan di *tweet* Saudi dan pasar saham Saudi. Dengan akurasi pada algoritma *Naïve Bayes* sebesar **69.86%** dan akurasi pada algoritma SVM sebesar 96.60%
4. Penelitian yang dilakukan oleh (Blatnik et al., 2014) yang berjudul “*Movie sentiment analysis based on public tweets*”. Dalam penelitian tersebut, menggunakan bahasa pemrograman Python dengan perpustakaan NLTK dan membandingkan hasil yang diperoleh dengan teknik pembelajaran mesin tradisional menggunakan *Rapidminer*. Fokus kami adalah di *Twitter - platform microblogging* dengan maksimum 140 karakter per *posting (tweet)*, lebih khusus lagi, pada pengumpulan sentimen untuk film tertentu. Dalam penelitian ini hasil yang dicapai memiliki akurasi tepat waktu sebesar 93% dan akurasi terlambat sebesar 71%. Dengan hasil akurasi algoritma *Naïve Bayes* **74.14%** dan algoritma k-NN adalah 77.59%.
5. Penelitian yang dilakukan oleh (Buntoro, 2017) yang berjudul “*Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter*”. Pada penelitian ini diharapkan dapat bermanfaat membantu untuk melakukan riset atas opini masyarakat yang mengandung sentimen *Positive*, Netral atau *Negative*. Metode yang digunakan dalam penelitian ini, untuk *preprocessing* data menggunakan tokenisasi, *cleansing* dan *filtering*, untuk menentukan class sentimen dengan metode *Lexicon Based*. Untuk proses klasifikasinya menggunakan metode *Naïve Bayes Classifier (NBC)* dan *Support Vector Machine (SVM)*. Data yang

digunakan adalah *tweet* dalam bahasa Indonesia dengan kata kunci AHY, Ahok, Anies, dengan jumlah *dataset* sebanyak 300 *tweet*. Hasil dari penelitian ini adalah analisis sentimen terhadap calon gubernur DKI Jakarta 2017. Akurasi tertinggi didapat saat menggunakan metode klasifikasi *Naïve Bayes Classifier* (NBC), dengan nilai rata-rata akurasi mencapai **95%**, nilai presisi 95%, nilai recall 95% nilai TP rate 96,8% dan nilai TN rate 84,6%.

6. Prediksi Penyakit Jantung Menggunakan Metode-Metode *Machine Learning* Berbasis *Ensemble – Weighted Vote* yang di tulis oleh (Alhamad et al., 2019) menyatakan bahwa penerapan metode-metode *machine learning* pada *dataset public* (Cleveland, Hungary, Switzerland, VA Long Beach, & Statlog) yang umumnya digunakan oleh para peneliti untuk prediksi penyakit jantung, termasuk pengembangan alat bantu, masih belum menangani *missing value*, *noisy data*, *unbalanced class*, dan bahkan data *validation* secara efisien. Oleh karena itu, pendekatan imputasi *mean/mode* diusulkan untuk menangani *missing value replacement*, Min-Max Normalization untuk menangani *smoothing noisy data*, K-Fold *Cross Validation* untuk menangani data *validation*, dan pendekatan *ensemble* menggunakan metode *Weighted Vote* (WV) yang dapat menyatukan kinerja tiap-tiap metode *machine learning* untuk mengambil keputusan klasifikasi sekaligus untuk mereduksi *unbalanced class*. Hasil penelitian ini menunjukkan bahwa metode yang diusulkan tersebut memberikan akurasi sebesar 85,21%, sehingga mampu meningkatkan kinerja akurasi metode-metode *machine learning*, selisih 7,14% dengan *Artificial Neural Network*, 2,77% dengan *Support Vector Machine*, 0,34% dengan *Decision Tree*, 2,94% dengan *Naïve Bayes*, dan 3,95% dengan *k-Nearest Neighbor*.
7. Demikian juga dengan penelitian yang berjudul Analisis Sentimen Kurikulum 2013 pada *Twitter* menggunakan *Ensemble Feature* dan Metode *K-Nearest Neighbor*, oleh (Mentari et al., 2018). Dalam penelitian tersebut, dilakukan analisis sentimen untuk mengetahui opini yang berkembang tersebut yang dibagi ke dalam opini *Positive* atau opini *Negative*. Fitur dan metode yang digunakan adalah *ensemble feature* dan metode klasifikasi *K-Nearest Neighbor* (K-NN). *Ensemble feature* merupakan fitur gabungan, berupa fitur statistik *Bag of Words* (BoW) dan semantik (*Twitter specific*, *textual features*, *PoS features*, *lexicon based features*). Berdasarkan serangkaian pengujian, kombinasi fitur berdampak dalam meningkatkan akurasi metode *K-Nearest Neighbor* (K-NN) untuk menentukan opini *Positive* atau *Negative*. Penggabungan fitur ini dapat melengkapi kelemahan masing-masing fitur, sehingga hasil akhir akurasi yang didapatkan dengan menggabungkan kedua fitur tersebut mencapai **96%**. Berbeda hal jika

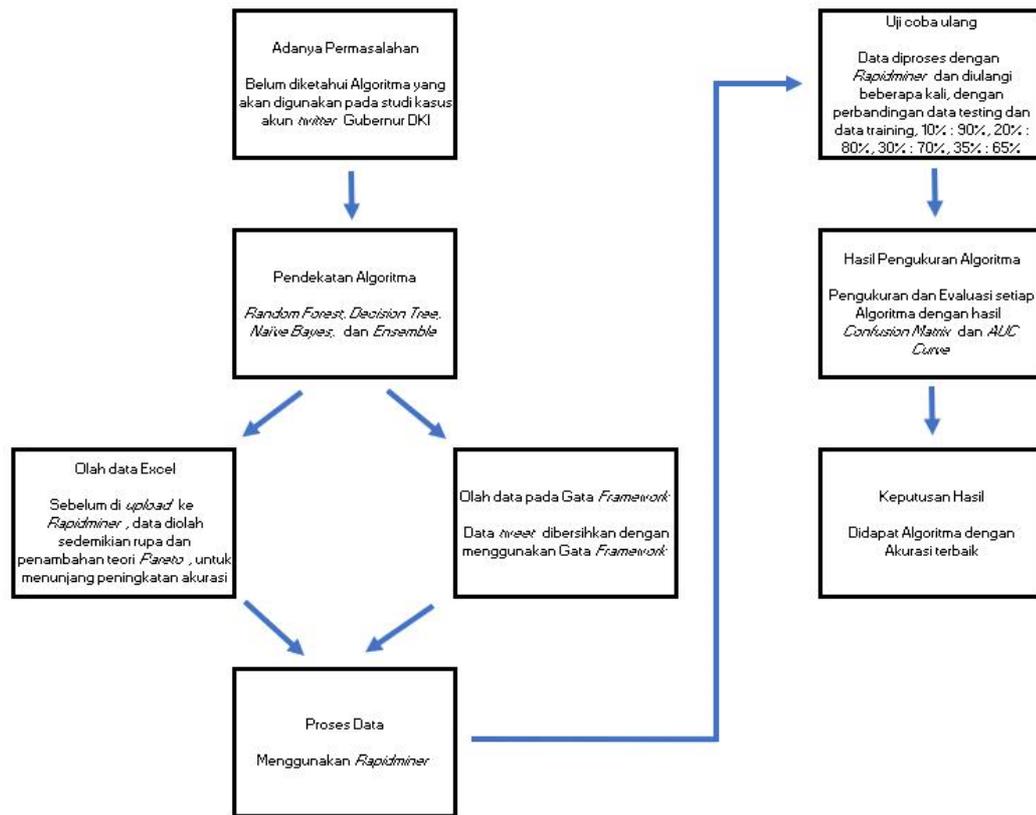
hanya menggunakan fitur secara independen saja, akurasi yang didapatkan hanya mencapai 80% pada fitur *Bag of Words* (BoW) dan 82% pada fitur *ensemble* tanpa *Bag of Words* (BoW).

8. Penelitian yang berjudul *A Comparative Study on Crime in Denver City Based on Machine Learning and data mining* oleh (Ratul & Engineering, n.d.) tahun 2020 adalah untuk memastikan keamanan secara umum termasuk pencegahan kejahatan kota dengan menerapkan beberapa algoritma klasifikasi seperti *Random Forest*, *Decision Tree*, *AdaBoost*, *Extra Tree Classification*, dan Klasifikasi *K-Neighbors*, dan 4 Model Ensemble. Hasil yang didapat. Algoritma *Random Forest*, *Decision Tree*, dan Ensemble Model 1, 3, dan 4 menghasilkan akurasi 90%.
9. Penelitian yang berjudul *Comparison of Performance of Various Data Classification Algorithms with Ensemble Methods Using Rapidminer* oleh (Puyalnithi et al., 2016), yang mempelajari dampak dari berbagai algoritma klasifikasi dalam prediksi yang tidak diketahui atribut labelnya. Prediksi yang digunakan dalam penelitian adalah *Naïve Bayes*, *Decision Tree* dan *Random Forest* menggunakan *Rapidminer*. Hasil akurasi yang didapat adalah **Naïve Bayes 84,34%, Random Forest 89,96% dan Decision Tree 89,97%**.
10. Penelitian lainnya, yang dilakukan pada tahun 2019 dengan judul *Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM And NB Methods* oleh (Pratama et al., 2019) menjelaskan algoritma SVM unggul dengan akurasi 86,96% sedangkan algoritma **Naïve Bayes** dengan akurasi **86.48%**. Nilai akurasi ini di dapat dari penelitian dengan obyek data adalah Korps Brimob pada media sosial twitter untuk dianalisa apakah tweet yang di posting bersifat positive atau negative.

2.9. Kerangka Pemikiran

Dalam menyelesaikan penelitian, penulis membuat sebuah kerangka pemikiran yang berguna sebagai pedoman penelitian ini sehingga penelitian dapat dilakukan secara konsisten. Permasalahan pada penelitian ini adalah belum diketahuinya metode yang paling akurat dalam klasifikasi pengguna *Twitter* kecenderungannya selalu *Positive* atau *Negative*. Atas dasar permasalahan tersebut penulis melakukan perbandingan terhadap metode klasifikasi *algoritma Random Forest*, *Decision Tree*, *Naïve Bayes*, dan *Ensemble* untuk memecahkan masalah penelitian ini. Berikut ini

adalah gambar 2.6 mengenai kerangka pemikiran yang penulis buat untuk penelitian ini:



Gambar 2.6 Kerangka pemikiran

BAB III

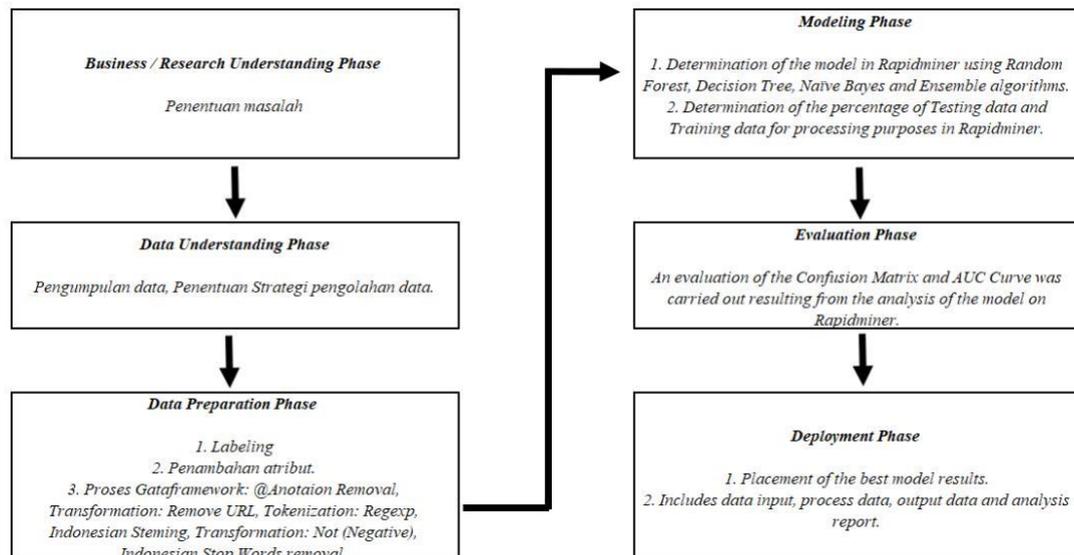
METODOLOGI PENELITIAN

3.1. Metode Penelitian

Penelitian adalah usaha mencari melalui proses yang metodis untuk menambahkan pengetahuan itu sendiri dan dengan yang lainnya, oleh penemuan fakta dan wawasan tidak biasa (Dawson, 2009). Untuk dapat menemukan fakta atau pengetahuan dari data, dibutuhkan suatu usaha ekstraksi yang disebut dengan *data mining*. Ekstraksi dilakukan untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui dari suatu data (Witten et al., 2001).

3.2. Kerangka Kerja Penelitian

Penulis mencoba untuk menterjemahkan metodologi standar dalam model *Cross-Standard Industry for data mining* (CRISP-DM) pada kerangka kerja penelitian pada Gambar 3.1 berikut:



Gambar 3.1 Tahapan Penelitian

Dalam model yang diusulkan pada Gambar 3.1 di atas, dijelaskan bahwa penelitian ini ditujukan untuk mencari satu algoritma yang terbaik diantara *Random Forest, Decision Tree, Naïve Bayes dan Ensemble*. Akurasi algoritma akan diukur dengan menggunakan *confusion matrix*. Sedangkan AUC akan diukur dengan

menggunakan *ROC Curve*. Hasil pengujian dengan akurasi yang paling tinggi adalah metode yang akan dilakukan untuk penentuan prediksi *netizen*. Berikut gambaran karakteristik dari masing-masing metode:

1. Algoritma *Random Forest* merupakan suatu metode hasil dari pengembangan *Decision Tree*, dimana setiap *Decision Tree* telah dilakukan *training* menggunakan sampel individu dan setiap atribut dipecah pada *tree* yang dipilih antara subset yang bersifat acak.
2. Algoritma *Decision Tree* merupakan algoritma dalam metode *Decision Tree* yang mengubah data menjadi pohon keputusan dengan menggunakan rumus perhitungan *entropy*.
3. Algoritma *Naïve Bayes* adalah salah satu metode *machine learning* yang memanfaatkan perhitungan probabilitas dengan cara memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya.
4. Algoritma *Ensemble* menggunakan beberapa algoritma pembelajaran untuk pencapaian solusi prediksi yang lebih baik daripada algoritma yang bisa diperoleh dari salah satu pembelajaran algoritma konsituen saja dan biasanya memungkinkan untuk menjadi lebih banyak lagi struktur fleksibel yang ada diantara alternatif model itu sendiri. Fitur *Vote* akan mengambil prediksi dengan suara / rekomendasi maksimum dari berbagai model prediksi sambil memprediksi hasil dari masalah klasifikasi.

3.2.1. Business / Research Understanding Phase

Pada fase ini, akan diamati tweet pada akun @aniesbaswedan yang berasal dari pada *netizen*. Banyaknya *tweet* dan pola yang ada, menjadikan materi ini layak untuk dijadikan bahan penelitian. Pada fase ini juga di pikirkan strategi untuk pengolahan data *tweet* tersebut, apakah akan menggunakan *Microsoft excell* untuk mengolahnya dan juga perlu dilakukan proses labeling pada text dari *tweet*. Sehingga bisa didapati makna dari *tweet* tersebut.

3.2.2. Data Understanding Phase

Proses pengumpulan data menjelaskan tentang bagaimana dan dari mana sumber data dikumpulkan dan kemudian diolah sehingga dapat digunakan untuk penelitian (Blaxter, Hughes, and Tight, 2010). Menurut (Kothari, 2004) berdasarkan sumbernya, data dibedakan menjadi dua, yaitu data primer dan data sekunder. Data primer adalah data yang dikumpulkan mula-mula untuk melihat apa yang

sebenarnya terjadi. Sedangkan data sekunder adalah data yang pernah dibuat oleh orang lain baik diterbitkan atau tidak. Data yang diperoleh untuk penelitian ini adalah data primer yang berupa sebuah *dataset* yang di *download* dari *Twitter* menggunakan Rapid Miner.

3.2.3. Data Preparation Phase

Data yang di *download* akan di proses terlebih dahulu secara parallel, yaitu:

a. Labeling

Pemberian label *Positive* dan *Negative* setiap text atau komentar yang diposting pada *official account* @aniesbaswedan. Pemberian label ini dilakukan pada responden sebanyak 100 orang.

b. Proses Excel

Data yang di *download* perlu di tambahkan dengan beberapa atribut lainnya agar data yang di proses ke dalam *Rapid Miner* didapat hasil terbaik. Dalam proses ini penulis menggunakan metode Pareto untuk mendapatkan user id yang *Active* atau tidak *Active*. Pareto merupakan salah satu dari tujuh alat gugus mutu yang sering digunakan dalam ilmu statistika. Urutannya mulai dari jumlah permasalahan yang paling banyak terjadi sampai yang paling sedikit terjadi.

3.2.4. Modeling Phase

Dalam tahap ini merupakan tahap yang dilalui setelah *document collection* adalah *preprocessing* bahasa indonesia melalui situs <http://www.GataFramework.com/textmining/>. Di dalamnya terdapat tahapan-tahapan sebagai berikut (Junianto & Riana, 2017) berikut:

a. @Anotation Removal

Langkah pertama dari ini adalah teks diurai berdasarkan *white space*, semua anotasi yang terkandung dalam *tweet* akan dihilangkan dan dilakukannya *lower case* atau mengubah huruf dalam text menjadi huruf kecil semua.

b. Transformation: Remove URL

Seringnya muncul sebuah url dari data *Twitter* membuat data tidak efektif dan tidak memiliki arti. Untuk itu perlu adanya penghapusan url tersebut atau bias juga untuk menghilangkan link *internet*.

c. Tokenization: Regexp

Proses tokenization dilakukan setelah *transform cases*. Semua karakter yang

tidak diperlukan akan dibuang. Termasuk *white space* yang berlebihan dan semua tanda baca. Proses ini akan dilakukan terhadap setiap dokumen yang dimasukkan dari *document collection*. Sehingga diperoleh kata yang unik dan dapat merepresentasikan dokumen.

d. Indonesian Stemming

Setelah hasil dari *transformation not Negative* akan dilanjutkan dengan proses *stemming* yaitu menghilangkan imbuhan yang terdapat pada masing-masing kata sehingga menjadi kata dasar dengan menggunakan *indonesian stemming* untuk *tweet* berbahasa Indoensia.

e. Transformation: Not (Negative)

Dari hasil *Tokenization (Regexp)* selanjutnya dilakukan proses *transformation not Negative*. Untuk contoh ini dalam text yang digunakan sebelumnya tidak ada perubahan dikarena tidak adanya kata yang dilakukan *Transformation Not Negative*. Namun untuk memperjelas maksud proses yang terjadi maka digunakan text yang lain dari *data local* yang sama.

f. Indonesian Stop Words removal

Tahap *stopwords* ini akan menyempurnakan tahap *filter token by length*. Kata yang terdiri lebih dari 3 huruf dan termasuk dalam *stopwords* akan dibuang.

Karena kata tersebut tidak mencerminkan isi dokumen walaupun sering muncul.

3.2.5. Evaluation Phase

Dalam penelitian ini akan dilakukan analisis komparasi menggunakan empat metode klasifikasi *data mining*. Algoritma yang akan digunakan adalah *Random Forest*, *Decision Tree*, *Naïve Bayes*, dan *Ensemble*. Setelah diolah dan menghasilkan model, selanjutnya terhadap model yang sudah dihasilkan tersebut dilakukan pengujian menggunakan *k-fold cross validation* dengan perbandingan antara *data testing* dan *data training* 10% : 90%, 20% : 80%, 30% : 70% dan 35% : 65%, kemudian mengulang pengujian tersebut beberapa kali.

Dalam penelitian ini, proses eksperimen dan pengujian model menggunakan beberapa spesifikasi *hardware* dan *software* sebagai alat bantu, spesifikasi *hardware* dan *software* tersebut dapat dilihat pada tabel 3.1:

Tabel 3.1 Spesifikasi *Hardware* dan *Software*

<i>Hardware</i>		<i>Software</i>	
CPU	: <i>Intel Pentium</i>	Sistem Operasi	: <i>Windows 10 Home</i>
Memory	: 16GB	<i>Data mining</i>	: <i>Rapid Miner 9.5</i>
Hardisk	: 500 GB SSD		

3.2.6. Deployment Phase

Tahap selanjutnya adalah melakukan evaluasi dan validasi hasil pengujian model tersebut dengan menggunakan *confussion matrix* dan kurva ROC. *Confussion matrix* adalah alat (*tools*) visualisasi yang biasa digunakan untuk menganalisis seberapa baik kualitas pengklasifikasi dapat mengenali data dari kelas yang berbeda (Han et al., 2012). Sedangkan kurva ROC menurut (Attenberg & Ertekin, 2013) adalah ukuran numerik untuk membedakan kinerja model, dan menunjukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan *Positive* dan *Negative*. Pada penelitian ini, penulis tidak melakukan proses *deployment*.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1 Hasil Penelitian

Berdasarkan metodologi penelitian yang telah di paparkan pada bab III, berikut implementasi metodologi yang dilakukan dalam penelitian ini.

4.1.1. Business / Research Understanding Phase

Pada tahapan ini dilakukan pemahaman terhadap objek penelitian. Dalam penelitian ini penulis menggunakan data *Twitter* dari *official account* Gubernur DKI Jakarta, yaitu @aniesbaswedan pada periode 28 Sept 2019 s/d 09 November 2019. Pengambilan data *Twitter* menggunakan aplikasi pada *Rapidminer*. Pada tahap ini juga dilakukan pemahaman untuk mencari label *Positive* dan *Negative* atas text yang diposting oleh user. Selain label text, juga bisa didapat *Active* dan *Passive* dari user *Twitter*.

4.1.2. Data Understanding Phase

Tahap ini adalah proses memahami data yang akan digunakan sebagai bahan yang akan diteliti untuk bisa dilakukan ke tahap setelahnya yaitu *Preprocessing*. Dibawah ini adalah langkah-langkah yang akan dikerjakan.

Menyiapkan total *dataset* dari *Twitter* @aniesbaswadana dan data yang berhasil diunduh sebanyak 29.340 *tweet*, data *tweet* di unduh menggunakan *tools* dari *Rapidminer*, kemudian *disave* dalam format excel. *Dataset* yang sudah disimpan dalam excel diproses lebih lanjut untuk mengidentifikasi adanya duplikasi *tweet* yang di post, dengan kata lain *dataset* dibersihkan dengan proses yang dinamakan *cleaning* data. Setelah dilakukan *cleaning* didapatkan data sebanyak 12.027 yang bisa digunakan. Pada penelitian ini hanya di ambil 10.000 data terdiri dari label *Positive* sebanyak 5.000 dan label *Negative* sebanyak 5.000 data. Pemberian label ini melibatkan 100 responden menggunakan metode *Crowdsourced labelling*, yaitu metode pelabelan data yang melibatkan partisipasi khalayak umum. Proses pemberian label untuk *dataset* yang tidak membutuhkan keahlian khusus atau melakukan pembelajaran kepada partisipan dalam melakukan pemberian label (Rachmat & Lukito, 2016). Dengan responden atau partisipan yang banyak, akan me mepercepat proses pemberian label dan juga akan lebih netral dalam perberian label tersebut. Hal lain yang juga menguntungkan penulis dari banyaknya responden atau

passitipan dalam proses ini, adalah tidak diperlukan biaya yang besar jika dibandingkan dengan menggunakan bantuan tenaga ahli untuk melakukannya.

4.2. Data Preparation and Modelling

Tahap selanjutnya adalah melakukan persiapan data sebelum data akan dilakukan modelling atau disebut dengan *Data Preparation*. Untuk tahap yang ke-2 ini yaitu mempersiapkan data untuk melakukan langkah-langkah yang disebut dengan *text preprocessing*, dengan menggunakan dua aplikasi *preprocessing*, pertama menggunakan *GataFramework* yang diakses melalui link <http://GataFramework.com/textmining> yang dapat digunakan secara gratis juga mudah dalam penggunaan dikarenakan tidak harus membuat *account* untuk memakai servicenya dan dilanjutkan *preprocessing* dari *Rapidminer*. Berikut adalah tampilan data yang di *download* dan juga tampilan lembar kerja pada excel, seperti pada Gambar 4.1 dan 4.2 berikut:

Created-At	From-User	From-User-id	To-User	To-User-id	Language	Source	Geo-Location-Latitude	Geo-Location-Longitude	Retweet-Count	Id	Text
2019-09-28 04:31:17	FaisalRT	214669644	smaya96	416381744	in	Twitter for Android			-	1177697242158792706	@smaya96 @mas_phyou @aniesbaswedan Astagfirullah.. pak anis dan para pemangkin daerah lain.. tolong bantu evakuasi.. lgn nunggu mrk mati?? @aniesbaswedan @sandiono @PEFESOfficial
2019-09-28 04:34:35	???????????? of FDI(AT)B of 常国	489850577	lmad_kun	2245061161	in	Twitter Web App			-	117769870840037377	@lmad_kun Itulah kenapa saya makin sengsi si jae bisa turun. Perlu kelengkapan jakarta agar mampu memaksa si jae turun. Tapi gubernur DKI itu mas @aniesbaswedan. saya paham benar kapasitas beliau sebagai gubernur. Disgusing in bless. ??
2019-09-28 04:37:53	Dkizamnow	963011589149356032	Coldice10911773	968953864798969857	in	Twitter for Android			-	1177698900586614785	@Coldice10911773 @Dennysiregar7 @pmidkiJakarta @aniesbaswedan Buka dong bang, penasaran neh. Sebenarnya batunya ada apa nggak seh
2019-09-28 04:38:56	jaiyy fath	324919350	dektampu	241909291	in	Twitter for Android			-	1177699167373647873	@dektampu @joe_Amstroeng @XPK_RI Tak bisakah bung @aniesbaswedan periksa kembali data pengadaan bus rombongan cina ini
2019-09-28 04:39:02	BODO.	849377349435092992	detektive88	395984870	in	Twitter for Android			1.0	1177699193512550400	@detektive88 @jokowi @aniesbaswedan Dimohon dengan amat sangat segera ke RSJ secepatnya. https://www.kemkes.go.id/

Gambar 4.1 Contoh data hasil *download* Twitter

#	Proses	Remark	Char Text	Isi Tweet	To-User ?	Jenis Tweet	Freq Re-Tweet	Active Re-Tweet	Freq Tweet	Active Tweet	Active Tweet/Re-Tweet	Freq All	Active Freq All	Freq Retweet-Count	Active Retweet-Count	From To-Active To	User Active/Passive Label User	Status	# Proses	# Label	Label		
3	Yes		93		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	6	Active	Passive	Positive	1	1	Positive
6	Yes		197		1	Tweet	0	Passive	3	Active	Active Tweet	4	Active	0	Passive	1	2	Active	Active	Positive	2	3	Positive
8	Yes		57		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	4	Active	Passive	Positive	3	5	Positive
9	Yes		71		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	4	Active	Passive	Positive	4	6	Positive
10	Yes		35		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	1	Passive	1	3	Active	Passive	Negative	1	3	Negative

Gambar 4.2 Contoh lembar kerja *preprocessing*

Tampilan diatas merupakan *paper work* dengan penambahan atribut pada kolom berwarna untuk mendapatkan informasi lainnya. Informasi tersebut antara lain, apakah user id dapat digolongkan menjadi user id yang *Active* atau *Passive*, kategori *tweet* dibedakan menjadi *tweet* dan *re-tweet*, seperti pada penjelasan pada Tabel 4.1 berikut:

Tabel 4.1 Atribut tambahan dari data olah

<i>Attribute</i>	<i>Remarks</i>	<i>Attribute</i>	<i>Remarks</i>
#	No Baris data	<i>Freq All</i>	Jml <i>Posting/User Id</i>
Proses	Proses ke <i>Rapidminer</i>	<i>Active Freq All</i>	Lht kolom <i>Freq All > 1</i> , <i>Active</i>
<i>Remark</i>	Keterangan Proses ke <i>Rapidminer</i>	<i>Freq Retweet-Count</i>	Jml <i>Freq Retweet-Count</i>
<i>Char Text</i>	Jml Character	<i>Active Retweet-Count</i>	Lht kolom <i>J</i> , <i>Active/Passive</i> dgn Pareto
Isi <i>Tweet</i>	Isi <i>Tweet</i> , Jelas/Tdk Jelas	From	Dari 1 <i>User Id</i>
<i>To-User ?</i>	Ada/Tdk Penerima <i>Tweet</i>	<i>To</i>	Ke berapa <i>User Id</i>
Jenis <i>Tweet</i>	Lht kolom Text	<i>Active To</i>	Lht kolom <i>To > 1</i> , <i>Active</i>
<i>Freq Re-Tweet</i>	Jml <i>Re-Tweet/User Id</i>	<i>User Active/Passive</i>	Lht kolom <i>Active Freq All</i> , <i>ActiveRetweet-Count</i> , <i>Active To</i>
<i>Active Re-Tweet</i>	Lht kolom <i>Freq Re-Tweet > 1</i> , <i>Active</i>	Status	Label terbanyak tiap
<i>Freq Tweet</i>	Jml <i>Tweet/User Id</i>	Label <i>User</i>	<i>User ID</i>
<i>Active Tweet</i>	Lht kolom <i>Freq Tweet > 1</i> , <i>Active</i>	# Proses	No Urut Proses vs Label
<i>Active Tweet/Re-Tweet</i>	Bandingkan kolom <i>Freq (Re-Tweet/ Twett)</i>	# Label	No Urut/Label
		Label	Label Text

Pada *paper work* ini, penulis akan menggunakan pendekatan teori Pareto untuk mentukan user id tersebut *Active* atau *Passive*. Total atribut yang digunakan dalam penelitian ini sebanyak 43 atribut yang terdiri dari 12 atribut berasal dari hasil *download Twitter*, 24 atribut berasal dari hasil penambahan dengan maksud untuk meningkatkan akurasi saat diproses dengan rapid miner. Sedangkan untuk 6 atribut terakhir merupakan hasil olah text yang *diposting* oleh user atau masyarakat menggunakan <http://www.GataFramework.com/textmining/>.

Berikut ini adalah total data yang di gunakan dalam penelitian ini seperti pada Gambar 4.3:

@Anotation Removal	Transformation: Remove URL	Tokenization: Regexp	Indonesian Stemming	Transformation: Not (Negative)	Indonesian Stop word removal
astaghfirullah.. pak anis dan para pemimpin daerah lain.. tolong bantu evakuasi.. jgn nunggu mrk mati??	astaghfirullah.. pak anis dan para pemimpin daerah lain.. tolong bantu evakuasi.. jgn nunggu mrk mati??	astaghfirullah pak anis dan para pemimpin daerah lain tolong bantu evakuasi jgn nunggu mrk mati	astaghfirullah pak anis dan para pimpin daerah lain tolong bantu evakuasi jgn nunggu mrk mati	astaghfirullah pak anis dan para pimpin daerah lain tolong bantu evakuasi jgn nunggu mrk mati	astaghfirullah pak anis dan para pimpin daerah lain tolong bantu evakuasi jgn nunggu mrk mati
itulah kenapa saya makin sangsi si jae bisa turun. perlu kelumpuhan jakarta agar mampu memaksa si jae turun. tapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur. disguising in bless. ??	itulah kenapa saya makin sangsi si jae bisa turun. perlu kelumpuhan jakarta agar mampu memaksa si jae turun. tapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur. disguising in bless. ??	itulah kenapa saya makin sangsi si jae bisa turun perlu kelumpuhan jakarta agar mampu memaksa si jae turun tapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur disguising in bless	itulah kenapa saya makin sangsi si jae bisa turun perlu lumpuh jakarta agar mampu paksa si jae turun tapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur disguising in bless	itulah kenapa saya makin sangsi si jae bisa turun perlu lumpuh jakarta agar mampu paksa si jae turun tapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur disguising in bless	itulah kenapa saya makin sangsi si jae bisa turun perlu lumpuh jakarta agar mampu paksa si jae turun tapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur disguising in bless
buka dong bang. penasaran neh. sebenarnya batunya ada apa nggak seh	buka dong bang. penasaran neh. sebenarnya batunya ada apa nggak seh	buka dong bang penasaran neh sebenarnya batunya ada apa nggak seh	buka dong bang penasaran neh benar batu ada apa nggak seh	buka dong bang penasaran neh benar batu ada apa nggak seh	buka dong bang penasaran neh benar batu ada apa nggak seh
tak bisakah bung periksa periksa kembali data pengadaan bus rongsokan cina ini	tak bisakah bung periksa periksa kembali data pengadaan bus rongsokan cina ini	tak bisakah bung periksa periksa kembali data pengadaan bus rongsokan cina ini	tak bisakah bung periksa periksa kembali data gada bus rongsok cina ini	tak bisakah bung periksa periksa kembali data gada bus rongsok cina ini	tak bisakah bung periksa periksa kembali data gada bus rongsok cina ini
dimohon dengan amat sangat segera ke rsj secepatnya. semoga lekas sembuh!	dimohon dengan amat sangat segera ke rsj secepatnya. semoga lekas sembuh!	dimohon dengan amat sangat segera ke rsj secepatnya semoga lekas sembuh.	mohon dengan amat sangat segera ke rsj cepat semoga lekas sembuh.	mohon dengan amat sangat segera ke rsj cepat semoga lekas sembuh.	mohon rsj cepat semoga lekas sembuh

Gambar 4.3 Contoh data hasil proses *GataFramework*

Tahapan berikutnya adalah *preprocessing* dari *Rapidminer* dengan urutan seperti Gambar 4.4 berikut :

The screenshot shows a web interface for 'GataFramework'. At the top, it says 'UPLOAD EXCELL 2003'. Below this, it specifies the format: 'Format: Coloumn 1 : No, Coloumn 2 : Text, Coloumn 3 : Status'. An example table is provided:

NO	Text	Status
1	Saya suka kamu	Positive
2	Saya tidak suka kamu alias benci	Negative

Below the table, there is a 'Choose File' button and a text input field containing 'Upload to Gata 121019.xls'. At the bottom, there are eight 'Techniques' listed with dropdown menus:

- Techniques 1 : @Anotation Removal
- Techniques 2 : Transformation: Remove URL
- Techniques 3 : Tokenization: Regexp
- Techniques 4 : Indonesian Stemming
- Techniques 5 : Transformation: Not (Negative)
- Techniques 6 : Indonesian Stop word removal
- Techniques 7 : None
- Techniques 8 : None

At the very bottom, there is a progress indicator 'Uploading Excell status'.

Gambar 4.4 Tampilan *GataFramework*

a. *@Anotation Removal*

Langkah pertama dari ini adalah teks diurai berdasarkan *white space*, semua anotasi yang terkandung dalam *tweet* akan dihilangkan dan dilakukannya *lower case* atau mengubah huruf dalam text menjadi huruf kecil semua, seperti contoh Tabel 4.2 berikut:

Tabel 4.2 Perbandingan Text sebelum dan setelah proses

@Anonation Removal

Text	@Anotation Removal
Tenangkan Massa @ganjarpranowo Turun ke Tengah Mahasiswa, kalau @aniesbaswedan? #AniesGaBener https://t.co/t9lqYpmGLx	tenangkan massa turun ke tengah mahasiswa, kalau #aniesgabener https://t.co/t9lqypmgLx
@aniesbaswedan @DKIJakarta @pln_123 @PT_TransJakarta @DishubDKI_JKT @dinaslhdkl Hati hati Pak @aniesbaswedan sampai saat ini saja wakilnya belum ada... kasihan @PKSejahtera di zholimi terus. https://t.co/K294APLorx	hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus. https://t.co/k294aplorx
@wong_sedeng @prahar_77 @PSI_Jakarta @rianernesto @psi_id @aniesbaswedan Siapapun presiden Indonesia... Pasti Ngutang .. gubernur nya aja ngutang ?????? https://t.co/wphXQRhFMw	siapapun presiden indonesia... pasti ngutang .. gubernur nya aja ngutang ?????? https://t.co/wphxqrhfmw
@sabar_mbok @IndoPluralitas @alvaro3_lee3_ @aniesbaswedan @DKIJakarta Alhamdulillah, mudah2an dosa2nya Pak Anies terhapus karena fitnah2 ini. https://t.co/Wk1KtLkTvJ	alhamdulillah, mudah2an dosa2nya pak anies terhapus karena fitnah2 ini. https://t.co/wk1kltkvtj
@addiems Konser lagi sepi ya mas...pantes WOW?????masih byk kerjanya dari pada cela-nya itulah @aniesbaswedan akui sajalah ?? https://t.co/pxmyn3uWJa	konser lagi sepi ya mas...pantes wow?????masih byk kerjanya dari pada cela-nya itulah akui sajalah ?? https://t.co/pxmyn3uwja

@AndriawanRandie @NgopiJahe_09 @aniesbaswedan @DivHumas_Polri Gw maklum lah lho, pantas aja lho mati2an nyinyir Pak @aniesbaswedan sahabtanya @Dennysiregar7 dan Buzzer2 lainnya, https://t.co/8USkFraaff	gw maklum lah lho, pantas aja lho mati2an nyinyir pak sahabtanya dan buzzer2 lainnya, https://t.co/8uskfraaff
--	--

b. *Transformation: Remove URL*

Seringnya muncul sebuah url dari data *Twitter* membuat data tidak efektif dan tidak memiliki arti. Untuk itu perlu adanya penghapusan url tersebut atau bias juga untuk menghilangkan *link internet*, seperti contoh Tabel 4.3 berikut:

Tabel 4.3 Perbandingan Text sebelum dan setelah proses

@ Transformation: Remove URL

Text	<i>Transformation: Remove URL</i>
tenangkan massa turun ke tengah mahasiswa, kalau #aniesgabener https://t.co/t9lqypmglx	tenangkan massa turun ke tengah mahasiswa, kalau #aniesgabener
hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus. https://t.co/k294aplorx	hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus.
siapapun presiden indonesia... pasti ngutang .. gubernur nya aja ngutang ?????? https://t.co/wphxqrhfmw	siapapun presiden indonesia... pasti ngutang .. gubernur nya aja ngutang ??????
alhamdulillah, mudah2an dosa2nya pak anies terhapus karena fitnah2 ini. https://t.co/wk1ktlktvj	alhamdulillah, mudah2an dosa2nya pak anies terhapus karena fitnah2 ini.
konser lagi sepi ya mas...pantes wow?????masih byk kerjanya dari pada cela-nya itulah akui sajalah ?? https://t.co/pxmyn3uwja	konser lagi sepi ya mas...pantes wow?????masih byk kerjanya dari pada cela-nya itulah akui sajalah ??
gw maklum lah lho, pantas aja lho mati2an nyinyir pak sahabtanya dan buzzer2 lainnya, https://t.co/8uskfraaff	gw maklum lah lho, pantas aja lho mati2an nyinyir pak sahabtanya dan buzzer2 lainnya,

c. *Tokenization: Regexp*

Proses tokenization dilakukan setelah *transform cases*. Semua karakter yang tidak diperlukan akan dibuang. Termasuk *white space* yang berlebihan dan semua tanda baca. Proses ini akan dilakukan terhadap setiap dokumen yang dimasukkan dari *document collection*. Sehingga diperoleh kata yang unik dan dapat merepresentasikan dokumen, seperti contoh Tabel 4.4 berikut:

Tabel 4.4 Perbandingan Text sebelum dan setelah proses**@ Tokenization: Regexp**

Text	Tokenization: Regexp
tenangkan massa turun ke tengah mahasiswa, kalau #aniesgabener	tenangkan massa turun ke tengah mahasiswa kalau aniesgabener
hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus.	hati hati pak sampai saat ini saja wakilnya belum ada kasihan di zholimi terus
siapapun presiden indonesia... pasti ngutang .. gubernur nya aja ngutang ??????	siapapun presiden indonesia pasti ngutang gubernur nya aja ngutang
alhamdulillah, mudah2an dosa2nya pak anies terhapus karena fitnah2 ini.	alhamdulillah mudahan dosanya pak anies terhapus karena fitnah ini
konser lagi sepi ya mas...pantes wow?????masih byk kerjanya dari pada cela-nya itulah akui sajalah ??	konser lagi sepi ya maspantes wow masih byk kerjanya dari pada celanya itulah akui sajalah
gw maklum lah lho, pantas aja lho mati2an nyinyir pak sahabtanya dan buzzer2 lainnya,	gw maklum lah lho pantas aja lho matian nyinyir pak sahabtanya dan buzzer lainnya

d. Indonesian Stemming

Setelah hasil dari *transformation not Negative* akan dilanjutkan dengan proses *stemming* yaitu menghilangkan imbuhan yang terdapat pada masing-masing kata sehingga menjadi kata dasar dengan menggunakan *indonesian stemming* untuk *tweet* berbahasa Indonesia, seperti contoh Tabel 4.5 berikut:

Tabel 4.5 Perbandingan Text sebelum dan setelah proses**Indonesian Stemming**

Text	Indonesian Stemming
tenangkan massa turun ke tengah mahasiswa kalau aniesgabener	tenang massa turun ke tengah mahasiswa kalau aniesgabener
hati hati pak sampai saat ini saja wakilnya belum ada kasihan di zholimi terus	hati hati pak sampai saat ini saja wakil belum ada kasihan di zholimi terus

siapapun presiden indonesia pasti ngutang gubernurnya aja ngutang	siapa presiden indonesia pasti ngutang gubernurnya aja ngutang
alhamdulillah mudah an dosanya pak anies terhapus karena fitnah ini	alhamdulillah mudah dosa pak anies hapus karena fitnah ini
konser lagi sepi ya maspantes wow masih byk kerjanya dari pada celanya itulah akui sajalah	konser lagi sepi ya maspantes wow masih byk kerja dari pada cela itulah akui saja
gw maklum lah lho pantas aja lho matian nyinyir pak sahabtanya dan buzzer lainnya	gw maklum lah lho pantas aja lho mati nyinyir pak sahabtanya dan buzzer lain

e. *Transformation: Not (Negative)*

Dari hasil *Tokenization (Regex)* selanjutnya dilakukan proses *transformation not Negative*. Untuk contoh ini dalam text yang digunakan sebelumnya tidak ada perubahan karena tidak adanya kata yang dilakukan *Transformation Not Negative*. Namun untuk memperjelas maksud proses yang terjadi maka digunakan text yang lain dari data local yang sama, seperti contoh Tabel 4.6 berikut:

**Tabel 4.6 Perbandingan Text sebelum dan setelah proses
*Transformation: Not (Negative)***

Text	<i>Transformation: Not (Negative)</i>
untuk erat tali saudara yg cerai berai oleh radikalisme pak gimana pak jawab saya udah keren belum	untuk erat tali saudara yg cerai berai oleh radikalisme pak gimana pak jawab saya udah keren belum_
tanya saya simple anda boleh orang dagang trotoar yang notabene buat untuk jalan kaki bukan untuk dagang iya atau tidak	tanya saya simple anda boleh orang dagang trotoar yang notabene buat untuk jalan kaki bukan_untuk dagang iya atau tidak_
lo me bicara harga juga semua dapat harga lo liat lapang sini lo orang jakarta bukan	lo me bicara harga juga semua dapat harga lo liat lapang sini lo orang jakarta bukan_
lha itu jpo sdh ada atap dul ngapain di lepas lain halnya kalo emang dr dolo ga ada kita trima dgn lapang dada buat bijak itu yg manfaat jgn buat bijak yg lebih tdk manfaat dr belum	lha itu jpo sdh ada atap dul ngapain di lepas lain halnya kalo emang dr dolo ga ada kita trima dgn lapang dada buat bijak itu yg manfaat jgn buat bijak yg lebih tdk manfaat dr belum_

omong kosong apa yg sedang anda bicara anda lupa kasus dugaan penyimpangan guna dana di pameran frankfurt book fair yg laku olh yg sdh lapor di namun apa wujud sampe skrg ada gk tindak dari nb kasus itu pernah buka apa belum	omong kosong apa yg sedang anda bicara anda lupa kasus dugaan penyimpangan guna dana di pameran frankfurt book fair yg laku olh yg sdh lapor di namun apa wujud sampe skrg ada gk tindak dari nb kasus itu pernah buka apa belum_
hati hati pak sampai saat ini saja wakil belum ada kasihan di zholimi terus	hati hati pak sampai saat ini saja wakil belum_ada kasihan di zholimi terus

f. *Indonesian Stop Words removal*

Tahap *stopwords* ini akan menyempurnakan tahap *filter token by length*. Kata yang terdiri lebih dari 3 huruf dan termasuk dalam *stopwords* akan dibuang. Karena kata tersebut tidak mencerminkan isi dokumen walaupun sering muncul, seperti contoh Tabel 4.6 berikut:

**Tabel 4.7 Perbandingan Text sebelum dan setelah proses
*Indonesian Stop word removal***

Text	<i>Indonesian Stop word removal</i>
tenang massa turun ke tengah mahasiswa kalau aniesgabener	tenang massa turun mahasiswa aniesgabener
hati hati pak sampai saat ini saja wakil belum_ada kasihan di zholimi terus	hati hati wakil belum_ada kasihan zholimi
siapa presiden indonesia pasti ngutang gubernur nya aja ngutang	presiden indonesia ngutang gubernur ngutang
alhamdulillah mudah dosa pak anies hapus karena fitnah ini	alhamdulillah mudah dosa anies hapus fitnah
konser lagi sepi ya maspantes wow masih byk kerja dari pada cela itulah akui saja	konser sepi maspantes wow byk kerja cela akui
gw maklum lah lho pantas aja lho mati nyinyir pak sahabtanya dan buzzer lain	maklum lho lho mati nyinyir sahabtanya buzzer

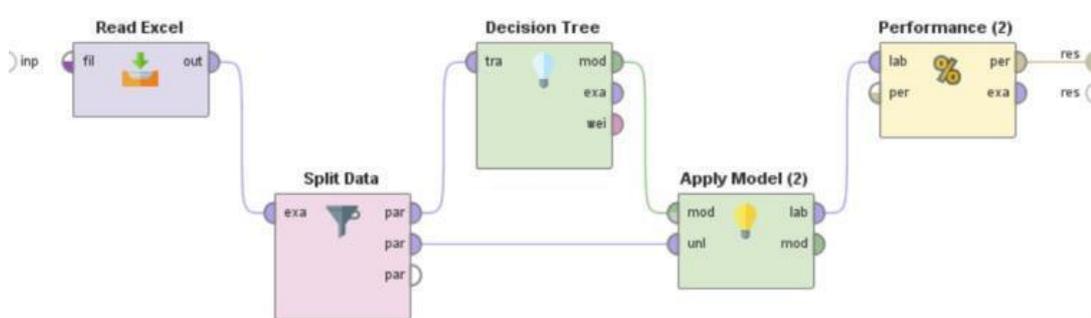
4.3 Hasil Eksperimen dan Pengujian Model

Metode klasifikasi bisa dievaluasi berdasarkan kriteria seperti tingkat akurasi, kecepatan, kehandalan, skalabilitas dan interpretabilitas (Vecellis, 2009). Pada penelitian ini, eksperimen yang dilakukan bertujuan untuk mengetahui tingkat

akurasi yang terbaik diantara algoritma *Random Forest*, *Decision Tree*, *Naïve Bayes*, dan *Ensemble* dengan membandingkan keempat algoritma tersebut. Total *dataset* yang digunakan adalah 10.000 data dan setelah diolah kemudian menghasilkan model, dilakukan pengujian dengan menggunakan *k-fold cross validation* dengan perbandingan antara *data testing* dan *data training* nya yaitu 10 : 90, 20 : 80, 30 : 70, dan 35% : 65% dengan 4 (tiga) algoritma, yaitu *Random Forest*, *Decision Tree*, *Naïve Bayes*, dan *Ensemble*.

4.3.1 Hasil Eksperimen dan Pengujian Model *Decision Tree*

Dari percobaan *dataset* sebanyak 10.000 menggunakan model *Decision Tree* seperti pada gambar 4.5 berikut:



Gambar 4.5 Tampilan Rapidminer model *Decision Tree*

Dari sebanyak 10.000 data *Text* yang diposting dan telah diolah menggunakan algoritma *Decision Tree* pada *Rapidminer* dengan perbandingan *data testing* dan *data training* 10 : 90 terdapat sebanyak 4152 data yang di prediksi *Positive* dan kenyataannya *Positive*, 4128 data diprediksi *Negative* dan kenyataannya *Negative*, 372 data diprediksi *Positive* tetapi kenyataannya *Negative* dan 348 data diprediksi *Negative* namun kenyataannya *Positive*, seperti pada Gambar 4.6 berikut:

accuracy: 92.00%

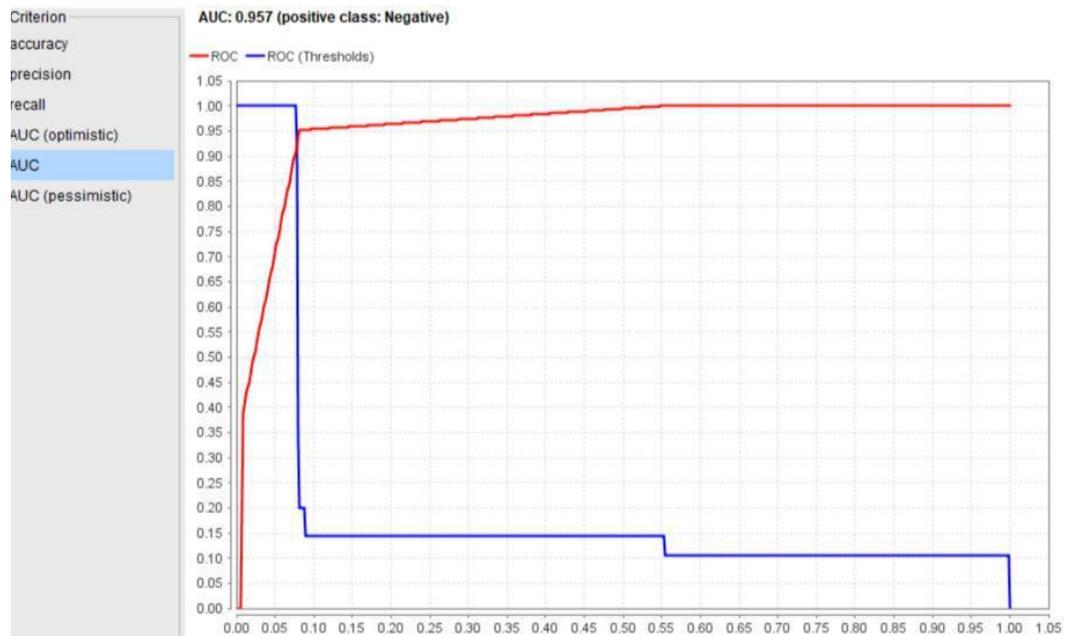
	true Positive	true Negative	class precision
pred. Positive	4152	372	91.78%
pred. Negative	348	4128	92.23%
class recall	92.27%	91.73%	

**Gambar 4.6 Confusion Matrix *Decision Tree*
data testing 10% dan data training 90%**

Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NVP* adalah sebagai berikut:

$$\begin{aligned}
 1. &= \frac{4152 + 4128}{4152 + 4128 + 372 + 348} = \frac{8280}{9000} = 0,9200 = 92,00\% \\
 2. &= \frac{4152}{4152 + 372} = \frac{4152}{4524} = 0,9227 = 92,27\% \\
 3. &= \frac{4128}{4128 + 348} = \frac{4128}{4476} = 0,9173 = 91,73\% \\
 4. & \text{PPV} = \frac{4152}{4152 + 372} = \frac{4152}{4524} = 0,9178 = 91,78\% \\
 5. & \text{NPV} = \frac{4128}{4128 + 348} = \frac{4128}{4476} = 0,9223 = 92,23\%
 \end{aligned}$$

Pengukuran ROC *curve* dengan menggunakan *Area Under Curve* (AUC) yang menghasilkan nilai AUC sebesar 0.957, seperti pada Gambar 4.7 berikut.



Gambar 4.7 Grafik Area Under Curve (AUC) Algoritma *Decision Tree*, data testing 10% dan data training 90%

Perbandingan berikutnya adalah 20% : 80%, dari *dataset* sejumlah 10000 didapat akurasi 93,4%. Dari hasil percobaan juga didapat sebanyak 3689 data yang di prediksi *Positive* dan kenyataannya *Positive*, 3783 data diprediksi *Negative* dan kenyataannya *Negative*, 217 data diprediksi *Positive* tetapi kenyataannya *Negative*, 311 data diprediksi *Negative* namun kenyataannya *Positive* seperti pada gambar 4.8

accuracy: 93.40%

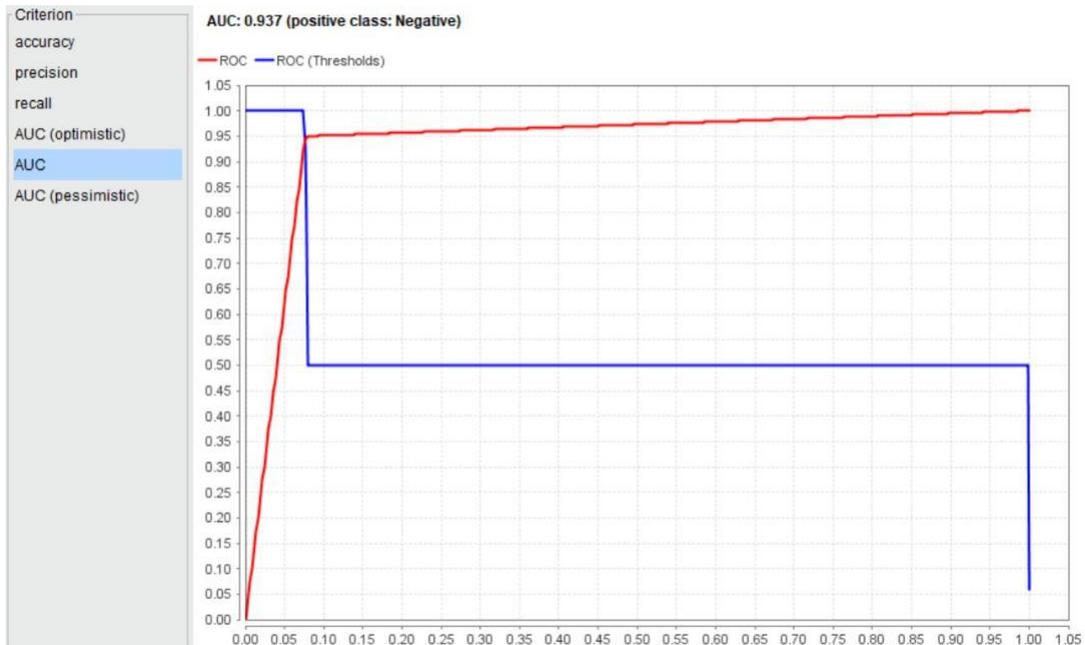
	true Positive	true Negative	class precision
pred. Positive	3689	217	94.44%
pred. Negative	311	3783	92.40%
class recall	92.22%	94.58%	

Gambar 4.8 *Confusion Matrix Decision Tree* data testing 20% dan data training 80%

Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NVP* adalah sebagai berikut:

$$\begin{aligned}
 1. &= \frac{3689 + 3783}{3689 + 311 + 3783 + 217} = \frac{7472}{8000} = 0,9340 = 93,40\% \\
 2. &= \frac{3689}{3689 + 311} = \frac{3689}{4000} = 0,92225 = 92,225\% \\
 3. &= \frac{3783}{3783 + 217} = \frac{3783}{4000} = 0,94575 = 94,575\% \\
 4. \text{ PPV} &= \frac{3689}{3689 + 273} = \frac{3689}{3962} = 0,9311 = 93,11\% \\
 5. \text{ NPV} &= \frac{3783}{3783 + 175} = \frac{3783}{3958} = 0,9558 = 95,58\%
 \end{aligned}$$

Pengukuran *ROC curve* dengan menggunakan *Area Under Curve* (AUC) yang menghasilkan nilai AUC sebesar 0.937 seperti pada gambar 4.9



Gambar 4.9 Grafik Area Under Curve (AUC) Algoritma *Decision Tree*, data testing 20% dan data training 80%

Perbandingan berikutnya adalah 30% : 70%, dengan akurasi 93,6% dan terdapat sebanyak 3227 data yang di prediksi *Positive* dan kenyataannya *Positive*, 3325 data diprediksi *Negative* dan kenyataannya *Negative*, 175 data diprediksi *Positive* tetapi kenyataannya *Negative*, 273 data diprediksi *Negative* namun kenyataannya *Positive* seperti pada gambar 4.10

accuracy: 93.60%

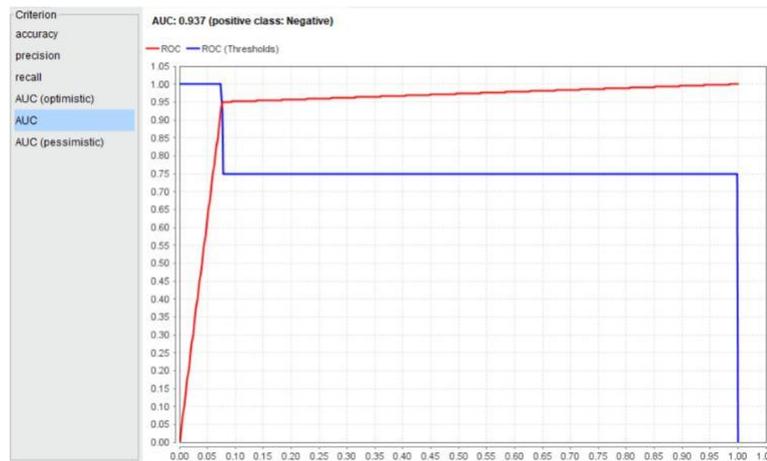
	true Positive	true Negative	class precision
pred. Positive	3227	175	94.86%
pred. Negative	273	3325	92.41%
class recall	92.20%	95.00%	

Gambar 4.10 Confussion Matrix Decision Tree data testing 30% dan data training 70%

Nilai Accuracy, Sensitivity, Specificity, PPV dan NVP adalah sebagai berikut:

1.
$$= \frac{3227 + 3325}{3227 + 175 + 273 + 3325} = \frac{6552}{7000} = 0,9360 = 93,60\%$$
2.
$$= \frac{3227}{3227 + 273} = \frac{3227}{3500} = 0,9220 = 92,20\%$$
3.
$$= \frac{3325}{3325 + 175} = \frac{3325}{3500} = 0,9500 = 95,00\%$$
4. PPV
$$= \frac{3227}{3227 + 175} = \frac{3227}{3402} = 0,9485 = 94,86\%$$
5. NPV
$$= \frac{3325}{3325 + 273} = \frac{3325}{3598} = 0,9241 = 92,41\%$$

Pengukuran ROC curve dengan menggunakan Area Under Curve (AUC) yang menghasilkan nilai AUC sebesar 0.937 seperti pada gambar 4.11



Gambar 4.11 Grafik Area Under Curve (AUC) Algoritma Decision Tree, data testing 30% dan data training 70%

Perbandingan berikutnya adalah 35% : 65%, dengan akurasi 93,58% dan terdapat sebanyak 2997 data yang di prediksi *Positive* dan kenyataannya *Positive*, 2086 data diprediksi *Negative* dan kenyataannya *Negative*, 164 data diprediksi *Positive* tetapi kenyataannya *Negative*, 253 data diprediksi *Negative* namun kenyataannya *Positive* seperti pada gambar 4.12

accuracy: 93.58%

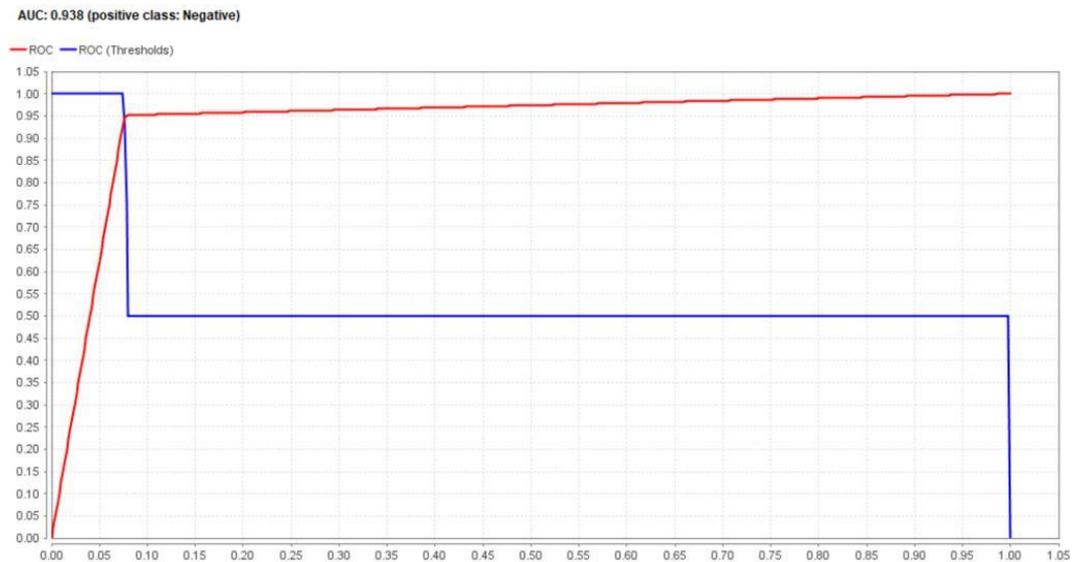
	true Positive	true Negative	class precision
pred. Positive	2997	164	94.81%
pred. Negative	253	3086	92.42%
class recall	92.22%	94.95%	

Gambar 4.12 Confussion Matrix Decision Tree
data testing 35% dan data training 65%

Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NVP* adalah sebagai berikut:

1.
$$= \frac{2997 + 3086}{2997 + 3086 + 164 + 253} = \frac{6083}{6500} = 0,9358 = 93,58\%$$
2.
$$= \frac{2997}{2997 + 253} = \frac{2997}{3250} = 0,9222 = 92,22\%$$
3.
$$= \frac{3086}{3086 + 164} = \frac{3086}{3250} = 0,9495 = 94,95\%$$
4. *PPV*
$$= \frac{2997}{2997 + 164} = \frac{2997}{3161} = 0,9481 = 94,81\%$$
5. *NPV*
$$= \frac{3086}{3086 + 253} = \frac{3086}{3339} = 0,9242 = 92,42\%$$

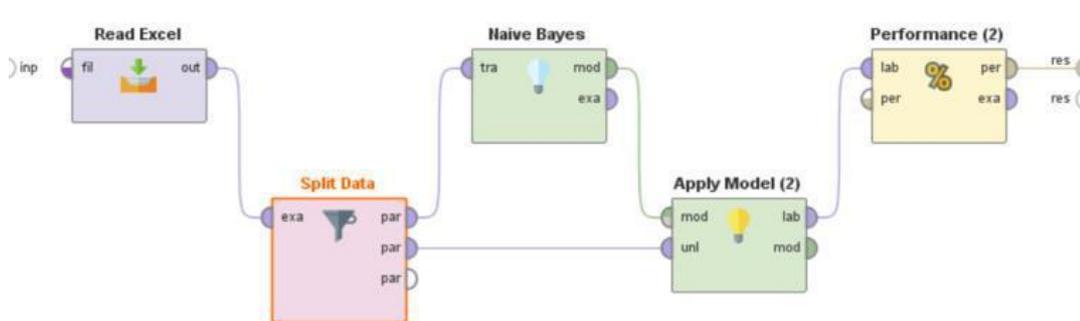
Pengukuran ROC *curve* dengan menggunakan *Area Under Curve* (AUC) yang menghasilkan nilai AUC sebesar 0.938 seperti pada gambar 4.13



Gambar 4.13 Grafik Area Under Curve (AUC) Algoritma *Decision Tree*, data testing 30% dan data training 70%

4.3.2 Hasil Eksperimen dan Pengujian Model *Naïve Bayes*

Algoritma ke dua yang digunakan adalah *Naïve Bayes* dengan model seperti pada gambar 4.14 berikut :



Gambar 4.14 Tampilan *Rapidminer* model *Decision Tree*

Nilai akurasi yang diperoleh dengan perbandingan *data testing* 10% : *data training*nya 90% adalah; $accuracy = 91.64\%$. Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 4024 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 4224 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 276 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 476 data diprediksi *Negative* tetapi kenyataannya *Positive* seperti pada gambar 4.15

accuracy: 91.64%

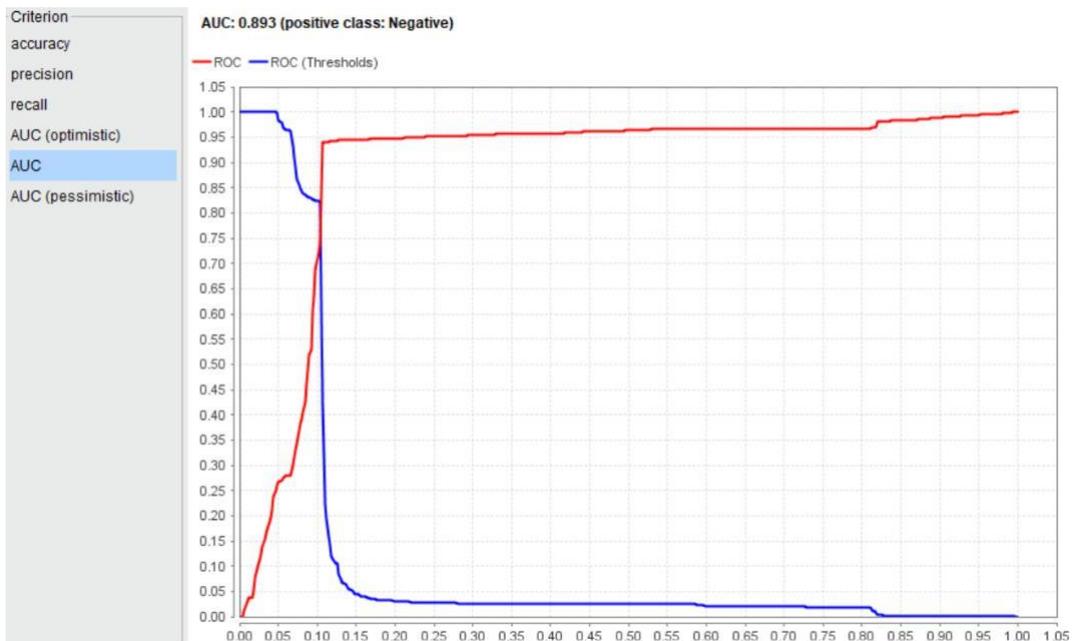
	true Positive	true Negative	class precision
pred. Positive	4024	276	93.58%
pred. Negative	476	4224	89.87%
class recall	89.42%	93.87%	

Gambar 4.15 Confussion Matrix Naïve Bayes data testing 10% dan data training 90%

Nilai Accuracy, Sensitivity, Specificity, PPV dan NVP adalah sebagai berikut:

$$\begin{aligned}
 1. \text{ Accuracy} &= \frac{4024 + 4224}{4024 + 276 + 476 + 4224} = \frac{8248}{8990} = 0,9160 = 91,60\% \\
 2. \text{ Sensitivity} &= \frac{4024}{4024 + 476} = \frac{4024}{4500} = 0,8942 = 89,42\% \\
 3. \text{ Specificity} &= \frac{4224}{4224 + 276} = \frac{4224}{4500} = 0,9387 = 93,87\% \\
 4. \text{ PPV} &= \frac{4024}{4024 + 276} = \frac{4024}{4300} = 0,9358 = 93,58\% \\
 5. \text{ NPV} &= \frac{4224}{4224 + 476} = \frac{4224}{4700} = 0,8987 = 89,87\%
 \end{aligned}$$

Pengukuran ROC curve dengan menggunakan Area Under Curve (AUC) yang menghasilkan nilai AUC sebesar 0.893 seperti pada gambar 4.16



Gambar 4.16 Grafik Area Under Curve (AUC) Algoritma Naïve Bayes, data testing 10% dan data training 90%

Perbandingan berikutnya adalah nilai akurasi yang diperoleh dengan perbandingan *data testing* 20% : *data training*nya 80% adalah; *accuracy* = 91.6% . Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 3609 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 3719 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 281 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 391 data diprediksi *Negative* tetapi kenyataannya *Positive* seperti pada gambar 4.17

accuracy: 91.60%

	true Positive	true Negative	class precision
pred. Positive	3609	281	92.78%
pred. Negative	391	3719	90.49%
class recall	90.22%	92.97%	

Gambar 4.17 Grafik Area Under Curve (AUC) Algoritma *Decision Tree*, *data testing* 20% dan *data training* 80%

Nilai *Accuracy*, *Sensitivity*, *Specificity*, PPV dan NPV adalah sebagai berikut:

$$\begin{aligned}
 1. \quad &= \frac{3609 + 3719}{3609 + 281 + 391 + 3719} = \frac{7328}{8000} = 0.916 = 91.60\% \\
 2. \quad &= \frac{3609}{3609 + 281} = \frac{3609}{4000} = 0,9022 = 90,22\% \\
 3. \quad &= \frac{3719}{3719 + 391} = \frac{3719}{4000} = 0,9297 = 92,97\% \\
 4. \quad \text{PPV} &= \frac{3609}{3609 + 281} = \frac{3609}{4000} = 0,9278 = 92,78\% \\
 5. \quad \text{NPV} &= \frac{3719}{3719 + 391} = \frac{3719}{4000} = 0,9049 = 90,49\%
 \end{aligned}$$

Pengukuran ROC *curve* dengan menggunakan *Area Under Curve* (AUC) yang menghasilkan nilai AUC sebesar 0.893 seperti pada gambar 4.18



Gambar 4.18 Grafik Area Under Curve (AUC) Algoritma Naïve Bayes, *data testing 20% dan data training 80%*

Perbandingan berikutnya adalah nilai akurasi yang diperoleh dengan perbandingan *data testing 30% : data trainingnya 70%* adalah; *accuracy = 91.3%* . Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 3146 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 3245 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 255 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 354 data diprediksi *Negative* tetapi kenyataannya *Positive*, seperti pada gambar 4.19

accuracy: 91.30%

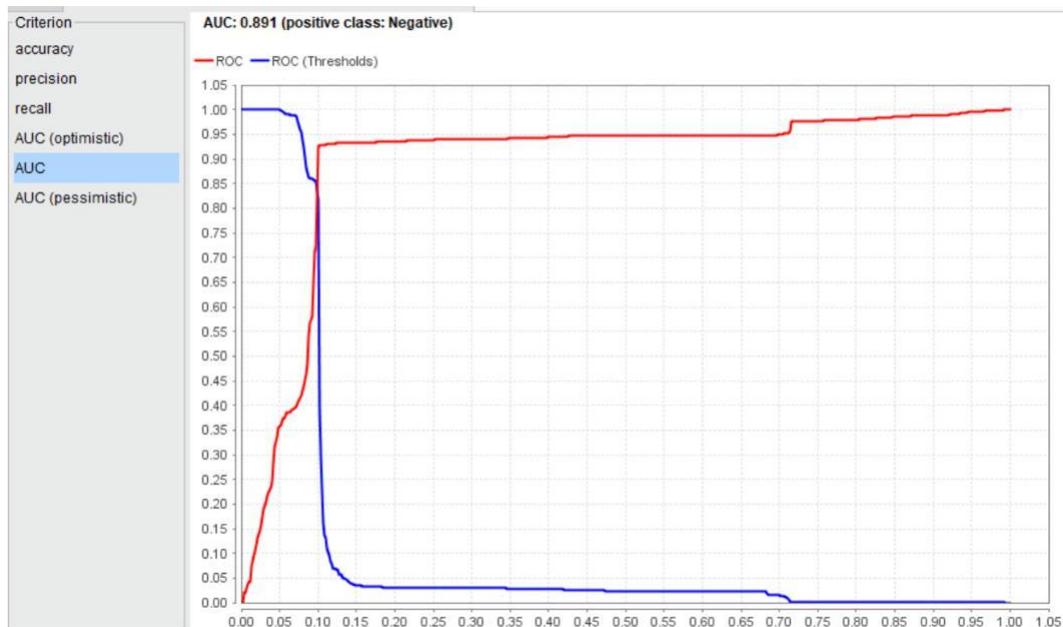
	true Positive	true Negative	class precision
pred. Positive	3146	255	92.50%
pred. Negative	354	3245	90.16%
class recall	89.89%	92.71%	

Gambar 4.19 Grafik Area Under Curve (AUC) Algoritma Naïve Bayes, *data testing 30% dan data training 70%*

Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NVP* adalah sebagai berikut:

$$\begin{aligned}
 1. \quad &= \frac{3146 + 3245}{3146 + 3245 + 354 + 255} = \frac{6391}{7000} = 0.913 = 91.30\% \\
 2. \quad &= \frac{3146}{3146 + 354} = \frac{3146}{3500} = 0,8988 = 89,88\% \\
 3. \quad &= \frac{3245}{3245 + 255} = \frac{3245}{3500} = 0,9271 = 92,71\% \\
 4. \quad \text{PPV} &= \frac{3146}{3146 + 255} = \frac{3146}{3401} = 0,9250 = 92,50\% \\
 5. \quad \text{NPV} &= \frac{3245}{3245 + 354} = \frac{3245}{3600} = 0,9016 = 90,16\%
 \end{aligned}$$

Pengukuran *ROC curve* dengan menggunakan *Area Under Curve (AUC)* yang menghasilkan nilai *AUC* sebesar 0.891, seperti pada gambar 4.20



Gambar 4.20 Grafik Area Under Curve (AUC) Algoritma Naïve Bayes, *data testing* 30% dan *data training* 70%

Perbandingan berikutnya adalah nilai akurasi yang diperoleh dengan perbandingan *data testing* 35% : *data training*nya 65% adalah; *accuracy* = 91.6% . Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 2939 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 3019 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 231 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 311 data diprediksi *Negative* tetapi kenyataannya *Positive*, seperti pada gambar 4.21

accuracy: 91.66%

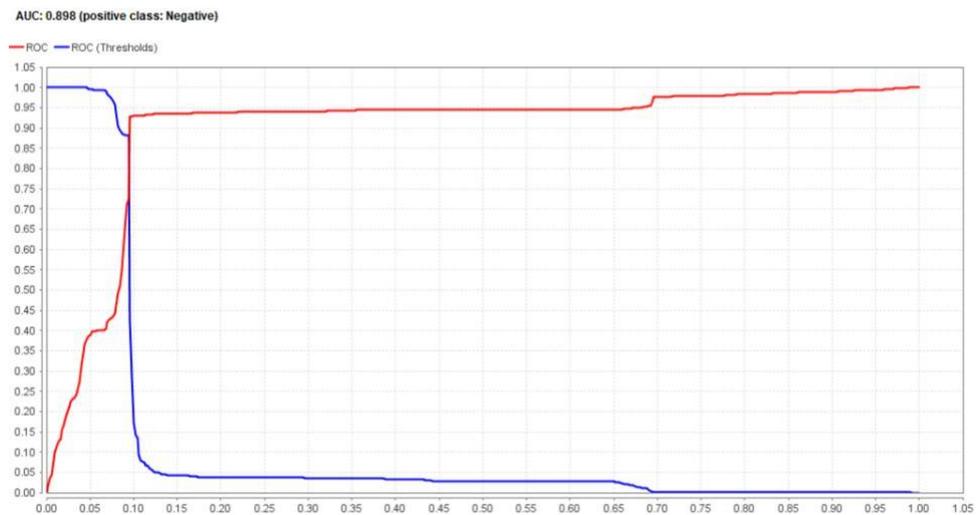
	true Positive	true Negative	class precision
pred. Positive	2939	231	92.71%
pred. Negative	311	3019	90.66%
class recall	90.43%	92.89%	

Gambar 4.21 Grafik Area Under Curve (AUC) Algoritma Naïve Bayes, data testing 35% dan data training 65%

Nilai Accuracy, Sensitivity, Specificity, PPV dan NVP adalah sebagai berikut:

- $$1. \text{ Accuracy} = \frac{2939 + 3019}{2939 + 231 + 311 + 3019} = \frac{5958}{6499} = 0.9166 = 91.66\%$$
- $$2. \text{ Sensitivity} = \frac{2939}{2939 + 311} = \frac{2939}{3250} = 0.9043 = 90.43\%$$
- $$3. \text{ Specificity} = \frac{3019}{3019 + 231} = \frac{3019}{3250} = 0.9289 = 92.89\%$$
- $$4. \text{ PPV} = \frac{2939}{2939 + 231} = \frac{2939}{3170} = 0.9271 = 92.71\%$$
- $$5. \text{ NPV} = \frac{3019}{3019 + 311} = \frac{3019}{3330} = 0.9066 = 90.66\%$$

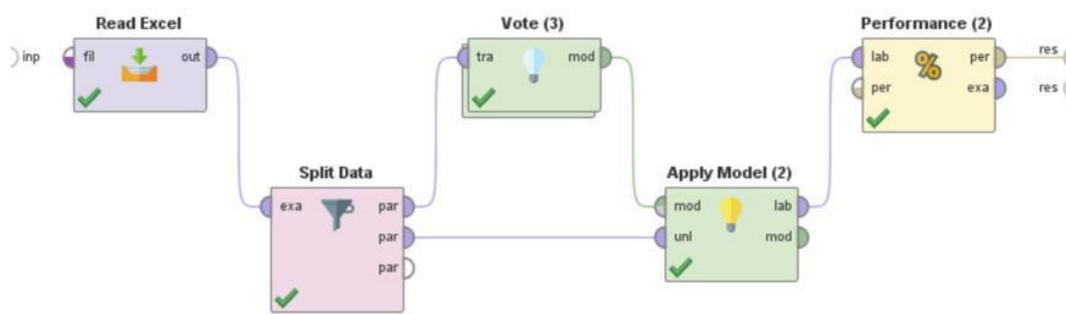
Pengukuran ROC curve dengan menggunakan Area Under Curve (AUC) yang menghasilkan nilai AUC sebesar 0.898, seperti pada gambar 4.22



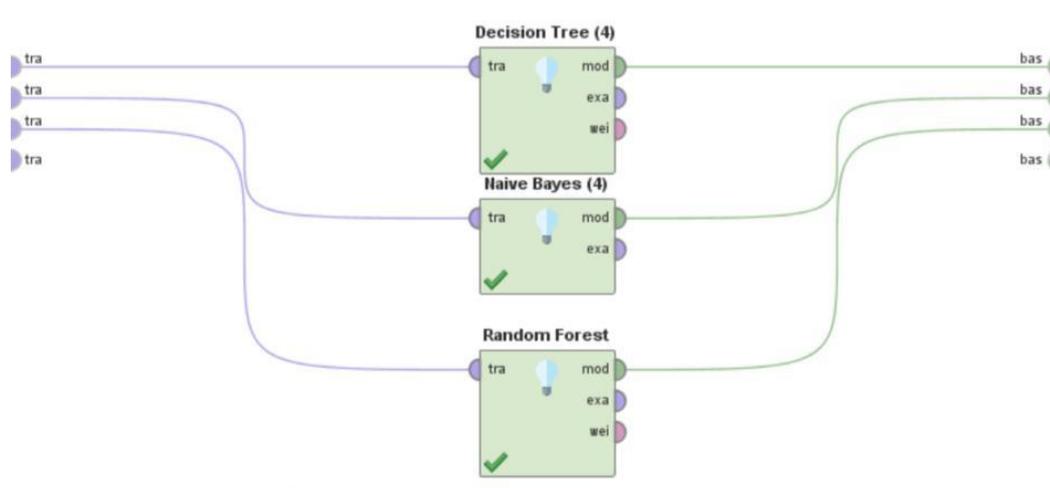
Gambar 4.22 Grafik Area Under Curve (AUC) Algoritma Naïve Bayes, data testing 35% dan data training 65%

4.3.3 Hasil Eksperimen dan Pengujian Model Ensemble Vote

Algoritma berikutnya atau yang ke tiga adalah *Ensemble* dengan fitur Vote, dengan model pada *Rapidminer* seperti pada gambar 4.23 dan 4.25 berikut :



Gambar 4.23 Tampilan *Rapidminer* model Ensemble



Gambar 4.24 Tampilan Performance pada *Rapidminer* model Ensemble fitur Vote

Dengan perbandingan *dataset* tester 10% dan *dataset* training 90%, maka dihasilkan akurasi sebesar . Hasil akurasi sebesar 91,44%. Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 4152 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 4223 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 227 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 348 data diprediksi *Negative* tetapi kenyataannya *Positive* seperti pada gambar 4.25

accuracy: 93.06%

	true Positive	true Negative	class precision
pred. Positive	4152	277	93.75%
pred. Negative	348	4223	92.39%
class recall	92.27%	93.84%	

Gambar 4.25 Confussion Matrix Ensemble, data testing 10% dan data training 90%

Nilai Accuracy, Sensitivity, Specificity, PPV dan NVP adalah sebagai berikut:

$$\begin{aligned}
 1. \quad & \text{Accuracy} = \frac{4152 + 4223}{4152 + 277 + 348 + 4223} = \frac{8375}{8999} = 0.9306 = 93,06\% \\
 2. \quad & \text{Sensitivity} = \frac{4152}{4152 + 348} = \frac{4152}{4500} = 0,9227 = 92,27\% \\
 3. \quad & \text{Specificity} = \frac{4223}{4223 + 277} = \frac{4223}{4500} = 0,9384 = 93,84\% \\
 4. \quad & \text{PPV} = \frac{4152}{4152 + 277} = \frac{4152}{4429} = 0,9375 = 93,75\% \\
 5. \quad & \text{NPV} = \frac{4223}{4223 + 348} = \frac{4223}{4571} = 0,9239 = 92,39\%
 \end{aligned}$$

Pengukuran ROC curve dengan menggunakan Area Under Curve (AUC) yang menghasilkan nilai AUC sebesar 0.905 seperti pada gambar 4.26



Gambar 4.26 Grafik Area Under Curve (AUC) Algoritma Ensemble, data testing 10% dan data training 90%

Percobaan berikutnya menggunakan perbandingan 20 : 80 antara data *dataset* tester dengan *dataset* training, maka dihasilkan akurasi sebesar . Hasil akurasi sebesar 93,44%. Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 3686 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 3786 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 211 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 314 data diprediksi *Negative* tetapi kenyataannya *Positive* seperti pada gambar 4.27

accuracy: 93.44%

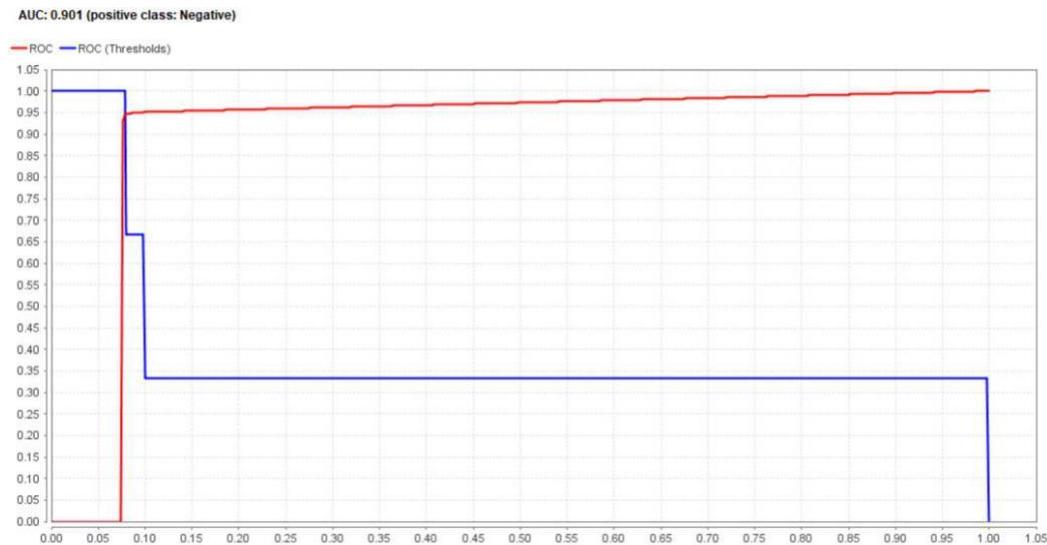
	true Positive	true Negative	class precision
pred. Positive	3686	211	94.59%
pred. Negative	314	3789	92.35%
class recall	92.15%	94.73%	

**Gambar 4.27 Confussion Matrix Ensemble,
data testing 20% dan data training 80%**

Nilai *Accuracy*, *Sensitivity*, *Specificity*, PPV dan NPV adalah sebagai berikut:

$$\begin{aligned}
 1. \quad & = \frac{3686 + 3789}{3686 + 211 + 314 + 3789} = \frac{7475}{8000} = 0.9344 = 93,44\% \\
 2. \quad & = \frac{3686}{3686 + 211} = \frac{3686}{4000} = 0,9215 = 92,15\% \\
 3. \quad & = \frac{3789}{3789 + 314} = \frac{3789}{4000} = 0,9473 = 94,73\% \\
 4. \quad \text{PPV} & = \frac{3686}{3686 + 211} = \frac{3686}{3907} = 0,9459 = 94,59\% \\
 5. \quad \text{NPV} & = \frac{3789}{3789 + 314} = \frac{3789}{4103} = 0,9235 = 92,35\%
 \end{aligned}$$

Pengukuran ROC *curve* dengan menggunakan *Area Under Curve* (AUC) yang menghasilkan nilai AUC sebesar 0.901 seperti pada gambar 4.28



Gambar 4.28 Grafik Area Under Curve (AUC) Algoritma Ensemble, *data testing 20% dan data training 80%*

Percobaan ke tiga menggunakan perbandingan 30 : 70 antara data *dataset* tester dengan *dataset* training, maka dihasilkan akurasi sebesar . Hasil akurasi sebesar 93,57%. Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 3227 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 3323 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 177 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 273 data diprediksi *Negative* tetapi kenyataannya *Positive* seperti pada gambar 4.29

accuracy: 93.57%

	true Positive	true Negative	class precision
pred. Positive	3227	177	94.80%
pred. Negative	273	3323	92.41%
class recall	92.20%	94.94%	

Gambar 4.29 *Confussion Matrix* Ensemble, *data testing 30% dan data training 70%*

Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NVP* adalah sebagai berikut:

$$\begin{array}{l}
 1. \quad = \frac{3227 + 3323}{3227 + 3323 + 177 + 273} = \frac{6650}{6999} = 0,9249 = 92,49\% \\
 2. \quad = \frac{3227}{3227 + 273} = \frac{3227}{3500} = 0,9220 = 92,20\% \\
 3. \quad = \frac{3323}{3323 + 177} = \frac{3323}{3500} = 0,9494 = 94,94\% \\
 4. \quad \text{PPV} = \frac{3227}{3227 + 177} = \frac{3227}{3404} = 0,9480 = 94,80\% \\
 5. \quad \text{NPV} = \frac{3323}{3323 + 273} = \frac{3323}{3596} = 0,9241 = 92,41\%
 \end{array}$$

Pengukuran *ROC curve* dengan menggunakan *Area Under Curve* (AUC)

yang menghasilkan nilai AUC sebesar 0.901 seperti pada gambar 4.30



Gambar 4.30 Grafik Area Under Curve (AUC) Algoritma Ensemble, data testing 30% dan data training 70%

Percobaan ke empat menggunakan perbandingan 35 : 65 antara data *dataset* tester dengan *dataset* training, maka dihasilkan akurasi sebesar . Hasil akurasi sebesar 93,60%. Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 2997 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 3087 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 163 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 253 data diprediksi *Negative* tetapi kenyataannya *Positive* seperti pada gambar 4.31

accuracy: 93.60%

	true Positive	true Negative	class precision
pred. Positive	2997	163	94.84%
pred. Negative	253	3087	92.43%
class recall	92.22%	94.98%	

**Gambar 4.31 Confussion Matrix Ensemble,
data testing 35% dan data training 65%**

Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NVP* adalah sebagai berikut:

$$\begin{aligned}
 1. \quad & = \frac{2997 + 3087}{2997 + 3087 + 253 + 163} = \frac{6084}{6540} = 0.9360 = 93,60\% \\
 2. \quad & = \frac{2997}{2997 + 253} = \frac{2997}{3250} = 0,9222 = 92,22\% \\
 3. \quad & = \frac{3087}{3087 + 163} = \frac{3087}{3250} = 0,9498 = 94,98\% \\
 4. \quad \text{PPV} & = \frac{2997}{2997 + 163} = \frac{2997}{3160} = 0,9484 = 94,84\% \\
 5. \quad \text{NPV} & = \frac{3087}{3087 + 253} = \frac{3087}{3340} = 0,9243 = 92,43\%
 \end{aligned}$$

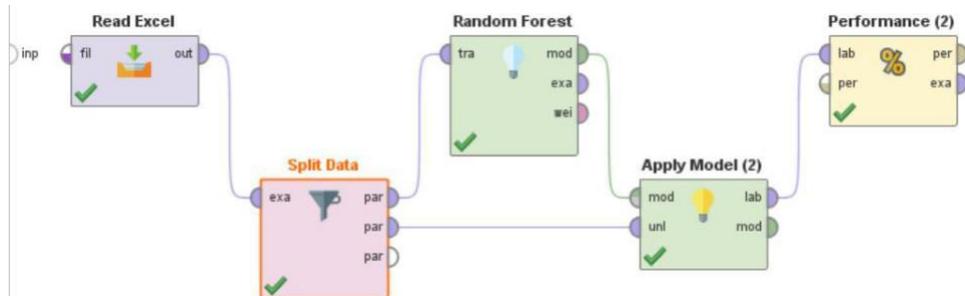
Pengukuran *ROC curve* dengan menggunakan *Area Under Curve* (AUC) yang menghasilkan nilai AUC sebesar 0.904 seperti pada gambar 4.33



**Gambar 4.32 Grafik Area Under Curve (AUC) Algoritma Ensemble,
data testing 35% dan data training 65%**

4.3.4. Hasil Eksperimen dan Pengujian Model *Random Forest*

Algoritma berikutnya atau yang ke empat adalah *Random Forest*, dengan model pada *Rapidminer* seperti pada gambar 4.33



Gambar 4.33 Tampilan *Rapidminer* model *Random Forest*

Dengan perbandingan *dataset* tester 10% dan *dataset* training 90%, maka dihasilkan akurasi sebesar . Hasil akurasi sebesar 93,08%. Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 4153 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 4224 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 226 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 347 data diprediksi *Negative* tetapi kenyataannya *Positive* seperti pada gambar 4.34

accuracy: 93.08%

	true Positive	true Negative	class precision
pred. Positive	4153	276	93.77%
pred. Negative	347	4224	92.41%
class recall	92.29%	93.87%	

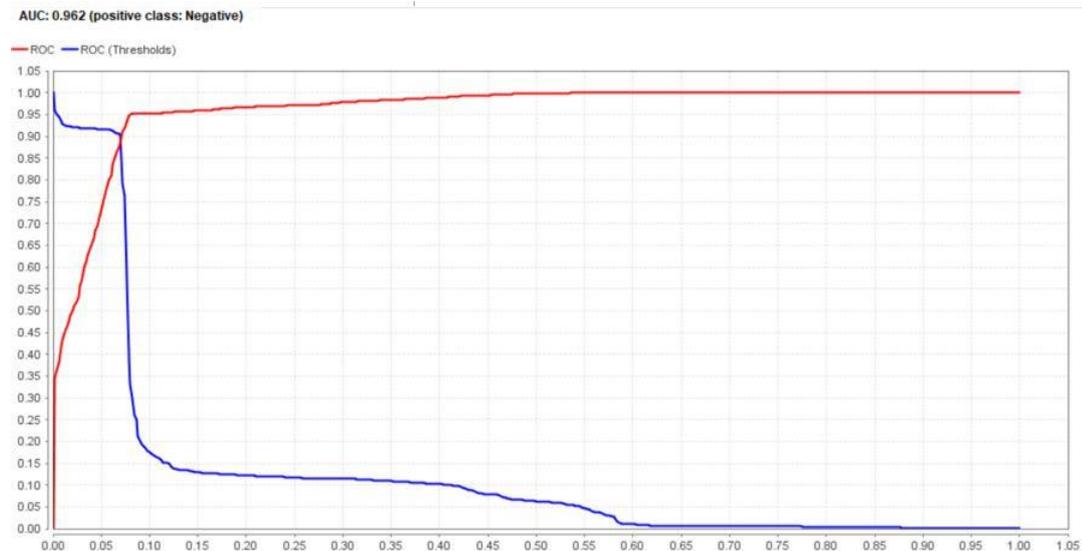
Gambar 4.34 *Confussion Matrix Random Forest*, data testing 10% dan data training 90%

Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NVP* adalah sebagai berikut:

1.
$$= \frac{4153 + 4224}{4153 + 4224 + 276 + 347} = \frac{8377}{10000} = 0.9306 = 93,06\%$$
2.
$$= \frac{4153}{4153 + 347} = \frac{4153}{4500} = 0,9229 = 92,29\%$$
3.
$$= \frac{4224}{4224 + 276} = \frac{4224}{4500} = 0,9387 = 93,87\%$$
4. *PPV*
$$= \frac{4153}{4153 + 276} = \frac{4153}{4429} = 0,9377 = 93,77\%$$

$$5. \text{ NPV} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{3789}{3789 + 211} = \frac{3789}{4000} = 0,94725 = 94,725\%$$

Pengukuran ROC *curve* dengan menggunakan *Area Under Curve* (AUC) yang menghasilkan nilai AUC sebesar 0.962 seperti pada gambar 4.35



Gambar 4.35 Grafik Area Under Curve (AUC) Algoritma Ensemble, data testing 10% dan data training 90%

Percobaan berikutnya menggunakan perbandingan 20 : 80 antara data *dataset* tester dengan *dataset* training, maka dihasilkan akurasi sebesar. Hasil akurasi sebesar 93,45%. Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 3687 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 3789 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 211 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 313 data diprediksi *Negative* tetapi kenyataannya *Positive* seperti pada gambar 4.36

accuracy: 93.45%

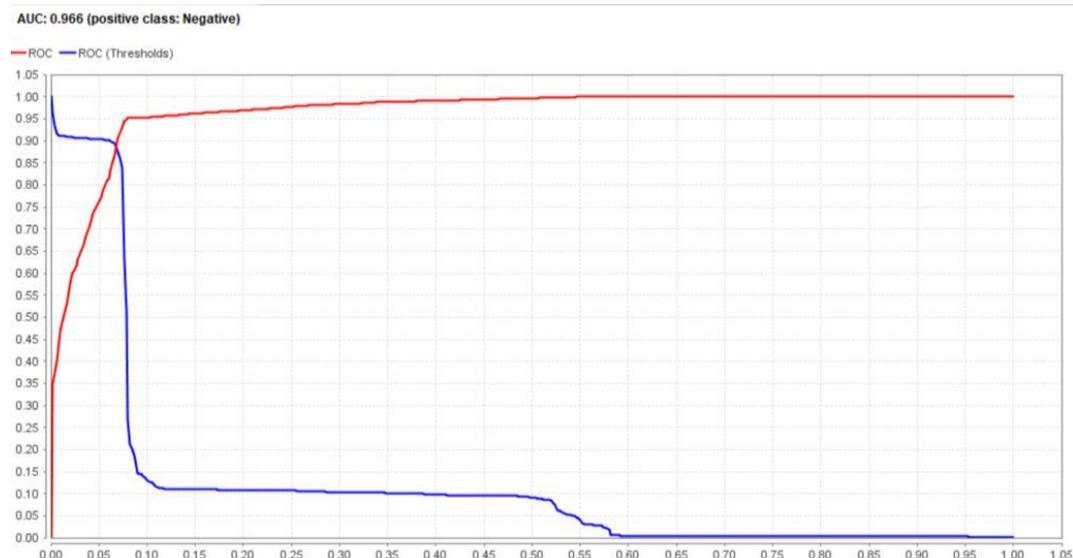
	true Positive	true Negative	class precision
pred. Positive	3687	211	94.59%
pred. Negative	313	3789	92.37%
class recall	92.17%	94.73%	

Gambar 4.36 Confussion Matrix Random Forest, data testing 20% dan data training 80%

Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NVP* adalah sebagai berikut:

$$\begin{aligned}
 1. & \quad = \frac{3687 + 3789}{8000} = \frac{7476}{8000} = 0.9345 = 93,45\% \\
 2. & \quad = \frac{3687}{3687 + 211} = \frac{3687}{3898} = 0,9217 = 92,17\% \\
 3. & \quad = \frac{3789}{3789 + 211} = \frac{3789}{4000} = 0,9473 = 94,73\% \\
 4. \text{ PPV} & \quad = \frac{3687}{3687 + 211} = \frac{3687}{3898} = 0,9459 = 94,59\% \\
 5. \text{ NPV} & \quad = \frac{3789}{3789 + 211} = \frac{3789}{4000} = 0,9237 = 92,37\%
 \end{aligned}$$

Pengukuran *ROC curve* dengan menggunakan *Area Under Curve* (AUC) yang menghasilkan nilai AUC sebesar 0.966 seperti pada gambar 4.37



Gambar 4.37 Grafik Area Under Curve (AUC) Algoritma Ensemble, data testing 20% dan data training 80%

Percobaan ke tiga menggunakan perbandingan 30 : 70 antara data *dataset* tester dengan *dataset* training, maka dihasilkan akurasi sebesar . Hasil akurasi sebesar 93,57%. Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 3227 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 3323 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 177 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 273 data diprediksi *Negative* tetapi kenyataannya *Positive* seperti pada gambar 4.38

accuracy: 93.57%

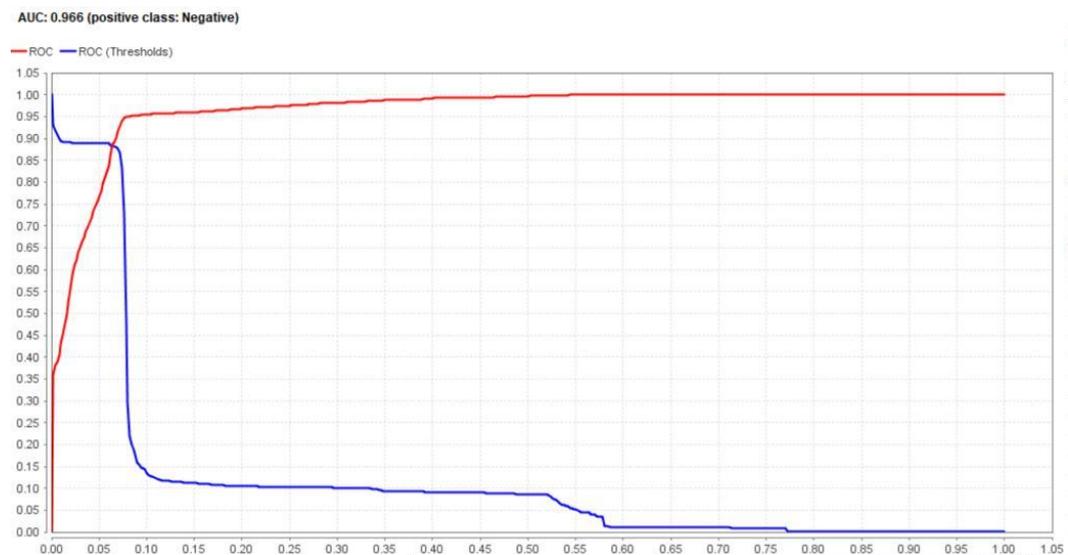
	true Positive	true Negative	class precision
pred. Positive	3227	177	94.80%
pred. Negative	273	3323	92.41%
class recall	92.20%	94.94%	

**Gambar 4.38 Confusion Matrix Random Forest,
data testing 30% dan data training 70%**

Nilai Accuracy, Sensitivity, Specificity, PPV dan NPV adalah sebagai berikut:

$$\begin{aligned}
 1. \quad & \text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} = \frac{3227 + 3323}{3227 + 177 + 273 + 3323} = \frac{6650}{7000} = 0.9249 = 92,49\% \\
 2. \quad & \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{3227}{3227 + 273} = \frac{3227}{3500} = 0,9220 = 92,20\% \\
 3. \quad & \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{3323}{3323 + 177} = \frac{3323}{3500} = 0,9494 = 94,94\% \\
 4. \quad & \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{3227}{3227 + 177} = \frac{3227}{3404} = 0,9480 = 94,80\% \\
 5. \quad & \text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} = \frac{3323}{3323 + 273} = \frac{3323}{3596} = 0,9241 = 92,41\%
 \end{aligned}$$

Pengukuran ROC curve dengan menggunakan Area Under Curve (AUC) yang menghasilkan nilai AUC sebesar 0.966 seperti pada gambar 4.39



**Gambar 4.39 Grafik Area Under Curve (AUC) Algoritma Ensemble,
data testing 30% dan data training 70%**

Percobaan ke empat menggunakan perbandingan 35 : 65 antara *dataset* tester dengan *dataset* training, maka dihasilkan akurasi sebesar . Hasil akurasi sebesar 93,60%. Dari keseluruhan 10.000 *dataset* yang diolah, sebanyak 2994 jumlah data yang diprediksi *Positive* dan pada kenyataannya memang *Positive*, 3087 data diprediksi *Negative* dan pada kenyataannya memang *Negative*, 163 data yang diprediksi *Positive* tetapi kenyataannya *Negative*, dan 256 data diprediksi *Negative* tetapi kenyataannya *Positive* seperti pada gambar 4.40

accuracy: 93.55%

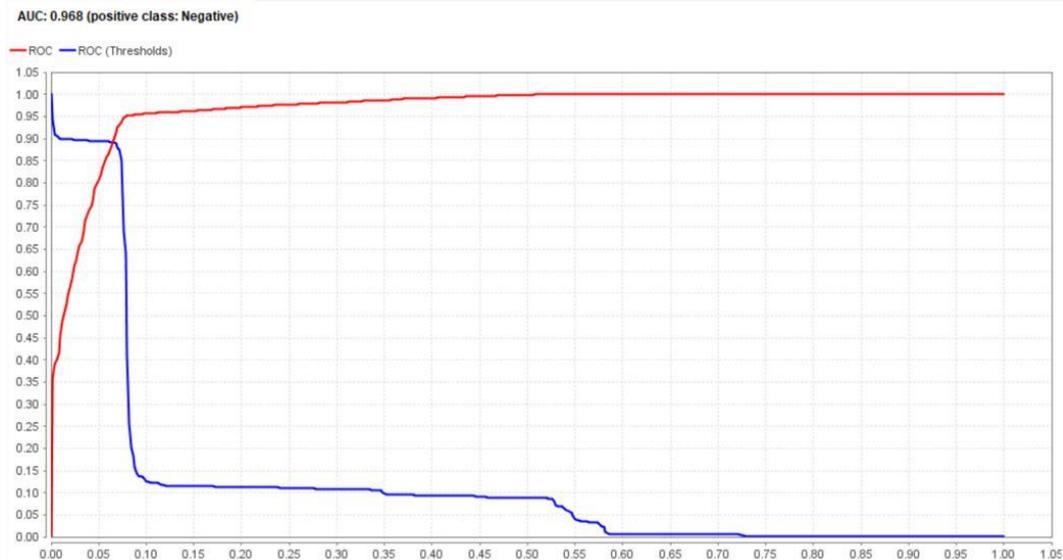
	true Positive	true Negative	class precision
pred. Positive	2994	163	94.84%
pred. Negative	256	3087	92.34%
class recall	92.12%	94.98%	

**Gambar 4.40 Confussion Matrix Random Forest,
data testing 35% dan data training 65%**

Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV* dan *NVP* adalah sebagai berikut:

$$\begin{aligned}
 1. \quad & = \frac{2994 + 3087}{2994 + 3087 + 163 + 256} = \frac{6081}{6510} = 0.9355 = 93,55\% \\
 2. \quad & = \frac{2994}{2994 + 256} = \frac{2994}{3250} = 0,9212 = 92,12\% \\
 3. \quad & = \frac{3087}{3087 + 163} = \frac{3087}{3250} = 0,9498 = 94,98\% \\
 4. \quad \text{PPV} & = \frac{2994}{2994 + 163} = \frac{2994}{3157} = 0,9484 = 94,84\% \\
 5. \quad \text{NPV} & = \frac{3087}{3087 + 256} = \frac{3087}{3343} = 0,9234 = 92,34\%
 \end{aligned}$$

Pengukuran ROC *curve* dengan menggunakan *Area Under Curve* (AUC) yang menghasilkan nilai AUC sebesar 0.968 seperti pada gambar 4.41



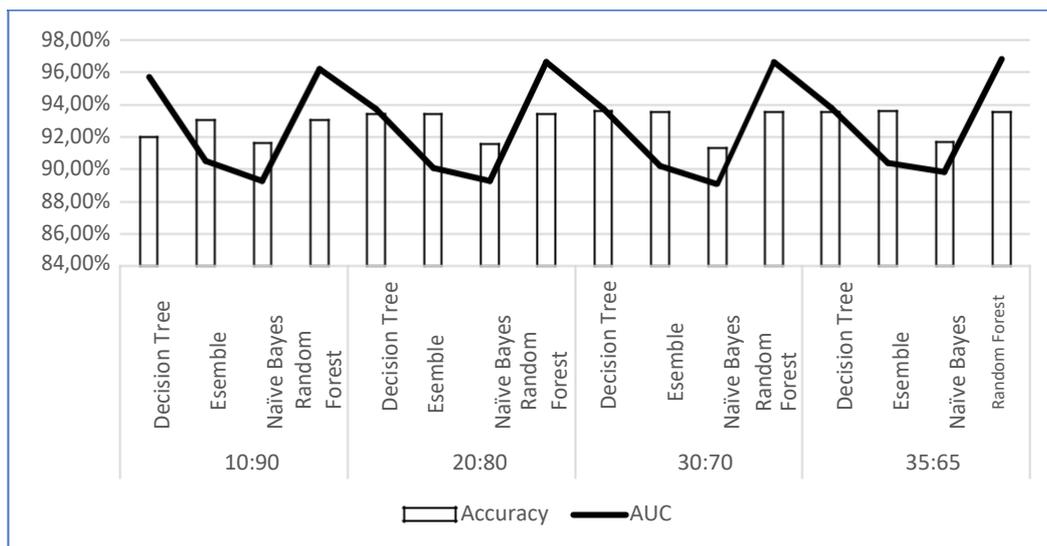
Gambar 4.41 Grafik Area Under Curve (AUC) Algoritma Ensemble, *data testing 35% dan data training 65%*

4.4 Perbandingan Keseluruhan Model

Dari tabel 4.8, dapat kita lihat hasil komparasi dari keempat algoritma yang digunakan dalam penelitian ini *Random Forest*, *Decision Tree*, *Naïve Bayes*, dan *Ensemble* dengan Fitur Vote, berdasarkan pembagian *data testing* : *data training* sebagai berikut; 10 : 90, 20 : 80, 30 : 70 dan 35 : 65 Pada perbandingan *data testing* 10% dan training 90%, nilai akurasi dari algoritma *Random Forest* sebesar 93,08% dan lebih tinggi dibandingkan dua algoritma lainnya, dengan nilai AUC-nya sebesar 0,962. Pada perbandingan *data testing* 20% dan data training 80%, nilai akurasi dari algoritma *Random Forest* sebesar 93,45% dengan nilai AUC-nya sebesar 0,966. Pada perbandingan terakhir dengan *data testing* 30% dan training 70%, algoritma *Decision Tree* dengan akurasi 93,60% dan dengan nilai AUC sebesar 0.937, yang terakhir adalah perbandingan 35% untuk *data testing* dan 65% untuk *data training* dengan hasil akurasi tertinggi pada algoritma *Ensemble* sebesar 93,60% dan AUC sebesar 0,904. Untuk rata-rata keseluruhan percobaan dapat dilihat pada tabel 4.8 dan gambar 4.42

Tabel 4.8 Rekapitulasi dataset percobaan *Rapidminer*

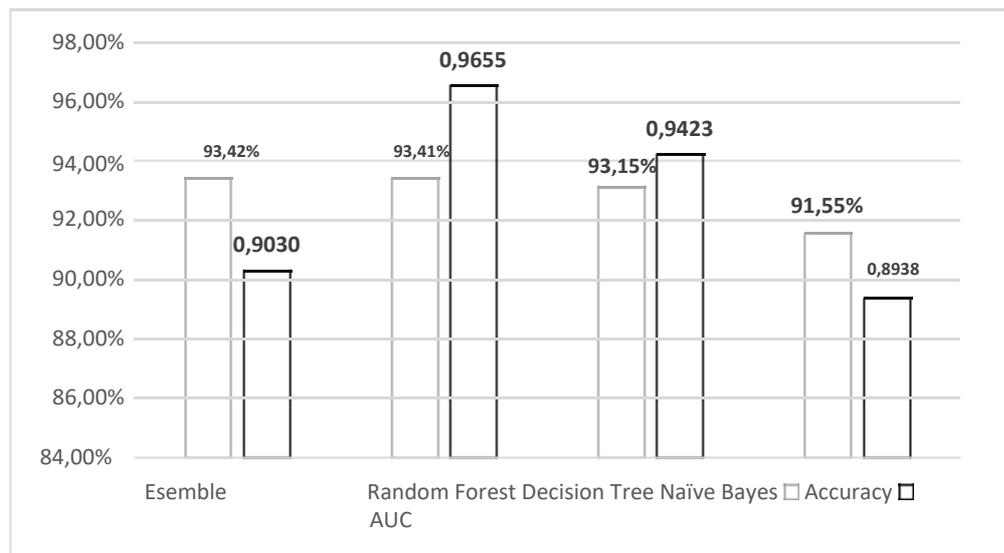
Algoritma	Testing	Training	Accuracy	AUC
<i>Decision Tree</i>	10%	90%	92,00%	0,957
<i>Ensemble</i>	10%	90%	93,06%	0,905
<i>Naïve Bayes</i>	10%	90%	91,64%	0,893
<i>Random Forest</i>	10%	90%	93,08%	0,962
<i>Decision Tree</i>	20%	80%	93,40%	0,937
<i>Ensemble</i>	20%	80%	93,44%	0,901
<i>Naïve Bayes</i>	20%	80%	91,60%	0,893
<i>Random Forest</i>	20%	80%	93,45%	0,966
<i>Decision Tree</i>	30%	70%	93,60%	0,937
<i>Ensemble</i>	30%	70%	93,57%	0,902
<i>Naïve Bayes</i>	30%	70%	91,30%	0,891
<i>Random Forest</i>	30%	70%	93,57%	0,966
<i>Decision Tree</i>	35%	65%	93,58%	0,938
<i>Ensemble</i>	35%	65%	93,60%	0,904
<i>Naïve Bayes</i>	35%	65%	91,66%	0,898
<i>Random Forest</i>	35%	65%	93,55%	0,968

Gambar 4.42 Grafik Perbandingan hasil percobaan *Rapidminer*

Sedangkan jika dirata-rata dari ke empat percobaan pada 4 algoritma, seperti pada table 4.9 dan gambar 4.43 berikut.

Tabel 4.9 Hasil Rata-rata dataset percobaan *Rapidminer*

Algorithm	Accuracy	AUC
<i>Ensembles</i>	93,42%	0,9030
<i>Random Forest</i>	93,41%	0,9655
<i>Decision Tree</i>	93,15%	0,9423
<i>Naïve Bayes</i>	91,55%	0,8938



Gambar 4.43 Grafik Perbandingan Rata-rata hasil percobaan *Rapidminer*

Pada tabel 4.9, dapat kita lihat bahwa rata-rata dari ke empat percobaan data didapat rata-rata akurasi dan rata-rata *Area Under Curve* (AUC) dari algoritma *Ensemble* dengan akurasi sebesar 93,42% dan AUC sebesar 0,9030. Untuk Algoritma *Random Forest* dengan akurasi 93,41 dan AUC sebesar 0,9655. Untuk Algoritma *Decision Tree* dengan Akurasi 93,15% dan AUC sebesar 0,9423. Sedangkan untuk Algoritma *Naïve Bayes* dengan akurasi 91,55% dan AUC sebesar 0,8938.

BAB V PENUTUP

5.1 Kesimpulan

Dari hasil komparasi algoritma *Random Forest*, *Decision Tree*, *Naïve Bayes*, dan *Ensemble*, dari percobaan dengan pembagian *data testing* : *data training* 10% : 90%, 20% : 80%, 30% : 70% dan 35% : 65%. Algoritma *Random Forest* unggul pada uji percobaan dengan komposisi *data testing* 10% : *data training* 90% dengan Akurasi 93,08% dan AUC sebesar 0,962, kemudian unggul juga pada *data testing* 20 % : *data training* 80% dengan akurasi 93,45 dan AUC 0,966. Sedangkan pada uji percobaan *data testing* 30% : *data training* 70%, unggul pada *Decision Tree* dengan Akurasi 93,60% dan AUC 0,937. Uji coba terakhir adalah pada *data testing* 35% : *data training* 65%, didapat akurasi 93,60% dan AUC 0,904 untuk Algoritma *Ensemble*.

Dengan Algoritma *Random Forest* dapat diprediksi hal yang mempengaruhi *user ID Twitter* atau *netizen* apakah *tweet* yang di *post* ke akun @aniesbawedan mayoritas mengarah ke *sentiment positive* atau *negative*. Hal-hal yang mempengaruhi *user ID* atau *netizen* mengarah ke *positive* atau *negative* seperti frekuensi *netizen* men-*tweet*, kemudian isi dari *tweet* tersebut bersifat *positive* atau *negative*, juga jenisnya apakah *tweet* atau *re-tweet*.

Pada *dataset* yang ada, akurasi dipengaruhi oleh pemilihan algoritma dan perbandingan *data testing* dengan *data training*. Hal ini dibuktikan dengan 4 percobaan yang dilakukan didapat hasil 3 dari 4 algoritma bisa unggul akurasinya. Hal lain yang juga mempengaruhi naiknya nilai akurasi dan AUC adalah *pre-processing* atas *dataset* yang *download*.

5.2 Saran

Untuk keperluan penelitian lebih lanjut mengenai komparasi metode klasifikasi algoritma, dapat dilakukan pengembangan untuk dapat menghasilkan model yang lebih baik lagi, diantaranya:

1. Untuk mendapatkan nilai akurasi yang lebih baik lagi, dapat Menggunakan algoritma peng-klasifikasi lain yang mungkin diluar *supervised learning* agar dapat dilakukan penelitian yang berbeda dari umumnya yang sudah ada.

2. Menggunakan tambahan fitur lainnya pada *Rapidminer*.
3. Variasi *pre-processing* di perdalam untuk mendapatkan akurasi yang lebih baik. Seperti menggunakan teori lain selain *Pareto* untuk membantu penelitian dalam mengolah *dataset*.
4. Membuat aplikasi sebagai pembuktian atas *dataset* dan algoritma yang digunakan.

DAFTAR PUSTAKA

- Al-Rubaiee, H., Qiu, R., & Li, D. (2016). Analysis of the relationship between Saudi twitter posts and the Saudi stock market. *2015 IEEE 7th International Conference on Intelligent Computing and Information Systems, ICICIS 2015, December*, 660–665. <https://doi.org/10.1109/IntelCIS.2015.7397193>
- Alfisahrin, S. N. N., & Mantoro, T. (2013). Data mining techniques for optimization of liver disease classification. *Proceedings - 2013 International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2013*. <https://doi.org/10.1109/ACSAT.2013.81>
- Alhamad, A., Azis, A. I. S., Santoso, B., & Taliki, S. (2019). *Prediksi Penyakit Jantung Menggunakan Metode-Metode Machine Learning Berbasis Ensemble – Weighted Vote*. 5(3), 352–360.
- Attenberg, J., & Ertekin, Ş. (2013). Class imbalance and active learning. In *Imbalanced Learning: Foundations, Algorithms, and Applications*. <https://doi.org/10.1002/9781118646106.ch6>
- Blatnik, A., Jarm, K., & Meža, M. (2014). Movie sentiment analysis based on public tweets. *Elektrotehniski Vestnik/Electrotechnical Review*, 81(4), 160–166.
- Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *Integer Journal Maret*, 1(1), 32–41. https://www.researchgate.net/profile/Ghulam_Buntoro/publication/316617194_Analisis_Sentimen_Calon_Gubernur_DKI_Jakarta_2017_Di_Twitter/links/5907eee44585152d2e9ff992/Analisis-Sentimen-Calon-Gubernur-DKI-Jakarta-2017-Di-Twitter.pdf
- Cureg, M. Q., De La Cruz, J. A. D., Solomon, J. C. A., Saharkhiz, A. T., Balan, A. K. D., & Samonte, M. J. C. (2019). Sentiment analysis on tweets with punctuations, emoticons, and negations. *ACM International Conference Proceeding Series, Part F1483(1)*, 266–270. <https://doi.org/10.1145/3322645.3322657>
- Da Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*. <https://doi.org/10.1016/j.dss.2014.07.003>
- Flux, A. W., & Pareto, V. (1897). Cours d'Economie Politique. *The Economic Journal*. <https://doi.org/10.2307/2956966>
- Jiawei Han, & Kamber, M. (2013). Data Mining: Concepts and Techniques Second Edition. In *Morgan Kaufmann*. <https://doi.org/10.1017/CBO9781107415324.004>
- Junianto, E., & Riana, D. (2017). Penerapan PSO Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan NBC. 4(1), 38–45.
- Kartiko, M., & Sfenrianto. (2019). Accuracy for Sentiment Analysis of Twitter Students on ELearning in Indonesia using Naive Bayes Algorithm Based on Particle Swarm Optimization. *Journal of Physics: Conference Series*, 1179(1). <https://doi.org/10.1088/1742-6596/1179/1/012027>
- Kothari, C. (2004). Research methodology: methods and techniques. In *New Age International*. <https://doi.org/http://196.29.172.66:8080/jspui/bitstream/123456789/2574/1/Research%20Methodology.pdf>

- Lee, K., Agrawal, A., & Choudhary, A. (2013). Real-Time disease surveillance using twitter data: Demonstration on flu and cancer. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Part F1288*, 1474–1477. <https://doi.org/10.1145/2487575.2487709>
- Mentari, N. D., Fauzi, M. A., & Muflikhah, L. (2018). Analisis Sentimen Kurikulum 2013 Pada Sosial Media Twitter Menggunakan Metode K-Nearest Neighbor dan Feature Selection Query Expansion Ranking. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, 2(8), 2739–2743.
- Mukminin, A., & Riana, D. (2017). Komparasi Algoritma C4.5, Naïve Bayes Dan Neural Network Untuk Klasifikasi Tanah. *Jurnal Informatika*.
- Nuraeni, N. (2017). Penentuan Kelayakan Kredit Dengan Algoritma Naïve Bayes Classifier : Studi Kasus Bank Mayapada Mitra Usaha Cabang PGC. *Jurnal Teknik Komputer AMIK BSI (JTK)*.
- Padmavathi, M., Suresh, R. M., & Mangadu, N. (2015). A Study of Fuzzy Based Block Selection Stratagem in Bittorrent Like P2P Network. *Australian Journal of Basic and Applied Sciences*, 9(January), 183–193.
- Paprotny, D., Andrzejewski, P., Terefenko, P., & Furmańczyk, K. (2014). Application of empirical wave run-up formulas to the polish baltic sea coast. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0105437>
- Pratama, B., Saputra, D. D., Novianti, D., Purnamasari, E. P., Kuntoro, A. Y., Hermanto, Gata, W., Wardhani, N. K., Sfenrianto, S., & Budilaksono, S. (2019). Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM and NB Methods. *Journal of Physics: Conference Series*, 1201(1). <https://doi.org/10.1088/1742-6596/1201/1/012038>
- Puyalnithi, T., V, M. V., & Singh, A. (2016). *Comparison of Performance of Various Data Classification Algorithms with Ensemble Methods Using RAPIDMINER*. 6(5), 1–6.
- Rachmat, A., & Lukito, Y. (2016). Implementasi Sistem Crowdsourced Labelling Berbasis Web dengan Metode Weighted Majority Voting. *Jurnal ULTIMA InfoSys*, 6(2), 76– 82. <https://doi.org/10.31937/si.v6i2.223>
- Ratul, A. R., & Engineering, F. (n.d.). *A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining*.
- Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (Google eBook). In *Complementary literature None*. <http://books.google.com/books?id=bDtLM8CODsQC&pgis=1>
- Akdon, dan Ridwan. (2013). *Rumus dan Data dalam Aplikasi Statistika*. Bandung: Alfabeta.
- Blaxter, L., Hughes, C., & Tight, M. (2010). *How to Research* (4th ed). Maidenhead: Open University Press.
- Breiman, L. (1996). *Bagging Predictors*. *Machine Learning*, 123-140.
- C.-M. Liaw, Yi-Ching, Leou Maw-Lin, “Fast exact k nearest neighbors search using anorthogonal search tree,” *Pattern Recognit.*, vol. 43, no. 6, pp. 2351–2358, Feb. 2010.

Dawson, C. W. (2009). *Projects in Computing and Information Systems a student's guide*. Harlow, UK: Addison-Wesley.

Gorunescu, F. (2011). *Data mining Concepts, Models and Techniques*. Verlag Berlin Heidelberg: Springer

Han, J., & Kamber, M. (2007). *Data mining Concepts and Technique*. Morgan Kaufmann publisher.

Larose, D. T. (2005). *Discovering Knowledge in Data*. New Jersey: John Willey & Sons, Inc.

Maimon, O., & Rokach, L. (2010). *Data mining and Knowledge Discovery Handbook*. London: Springer.

DAFTAR RIWAYAT HIDUP

I. Biodata Mahasiswa

NIM : 14002185
Nama Lengkap : Antonius Yadi Kuntoro
Tempat & Tanggal Lahir : Jakarta, 02 Oktober 2072
Jenis Kelamin : Laki-Laki
Agama : Katolik
Alamat lengkap : Jl. H.Taiman No. 18, RT. 005 RW. 010,
Kelurahan Gedong, Kecamatan Pasar Rebo
Jakarta Timur 13760
No HP : 0881 123 0210

II. Pendidikan

1. SDN 01 Manggarai, Jakarta Selatan, lulus tahun 1985
2. SMPN 03 Manggarai Jakarta Selatan, lulus tahun 1988
3. SMA ST. Antonius Jakarta Timur, lulus tahun 1992
4. Diploma Tiga (DIII) Universitas Pembangunan Nasional “Veteran” Jakarta, Fakultas Ilmu Komputer, Manajemen Informatika, lulus tahun 1996
5. Strata Satu (S1) Universitas Pembangunan Nasional “Veteran” Jakarta, Fakultas Ilmu Komputer, Sistem Informasi, lulus tahun 2007
6. Strata Dua (S2) Universitas Persada Indonesia Y.A.I Fakultas Ekonomi, Magister Ekonomi, lulus tahun 2015

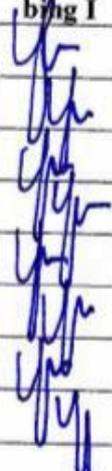


Jakarta, 23 January 2020

Antonius Yadi Kuntoro

	LEMBAR KONSULTASI BIMBINGAN TESIS
	SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER NUSA MANDIRI

NIM : 14002185
 Nama Lengkap : Antonius Yadi Kuntoro
 Dosen Pembimbing : Dr. Yan Rianto, M.Eng
 Judul Tesis : "Prediksi Tweet Netizen Menggunakan Algoritma
 Random Forest, Decision Tree, Ensemble, dan Naïve
 Bayes (Studi Kasus Gubernur DKI)".

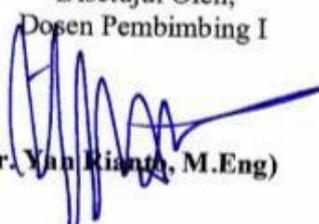
No.	Tanggal Bimbingan	Pokok Bahasan	Paraf Dosen Pembimbing I
1	12 Oktober 2019	Pengajuan Judul	
2	20 Oktober 2019	Acc Judul & Pengajuan Bab 1	
3	27 Oktober 2019	Acc Bab 1 & Pengajuan Bab 2	
4	03 November 2019	Acc Bab 2 & Pengajuan Bab 3	
5	17 November 2019	Acc Bab 3 & Pengajuan Bab 4	
6	15 Desember 2019	Revisi Bab 4	
7	31 Desember 2019	Acc Bab 4 & Pengajuan Bab 5	
8	02 Januari 2020	Acc Bab 5 & Acc Keseluruhan	

Catatan untuk dosen pembimbing

Bimbingan Tesis

- Dimulai pada tanggal : 12 Oktober 2019
- Diakhiri pada tanggal : 02 Januari 2020
- Jumlah pertemuan bimbingan : 8 (Delapan) Kali Bimbingan

Disetujui Oleh,
Dosen Pembimbing I


(Dr. Yan Rianto, M.Eng)

Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Source	Geo-Location - Latitude	Geo-Location - Longitude	Retweet-Count	Id	Text
2019-09-28 04:31:17	FaisalRT	214669644	smaya96	416381744	in	Twitter for Android				117769724215879270	@smaya96 @mas piyuu @aniesbaswedan Astaghfirullah.. pak anis dan para pemimpin daerah lain.. tolong bantu evakuasi.. jgn nunggu mrk mati?? @aniesbaswedan @sandiuno @PEPESOfficial
2019-09-28 04:34:35	????????????? of 常識。	489850577	Imad_kun	2245061161	in	Twitter Web App				117769807084003737	@Imad_kun Itulah kenapa saya makin sangsi si Jae bisa turun. Perlu kelompokan Jakarta agar mampu memaksa si Jae turun. Tapi gubernur DKI itu mas @aniesbaswedan. Saya paham benar kapasitas beliau sebagai gubernur. Disguising in bless. ??
2019-09-28 04:37:53	Dkizamannow	963011589149356032	Coldice10911773	968953864798969857	in	Twitter for Android				117769890058661478	@Coldice10911773 @Dennysiregar7 @pmdkijakarta @aniesbaswedan Buka dong bang, penasarannya neh. Sebenarnya batunya ada apa nggak seh
2019-09-28 04:38:56	raizy fath	324919350	dektampu	241909291	in	Twitter for Android				117769916737364787	@dektampu @joe_Amstroeng @KPK_RI Tak bisakah bung @aniesbaswedan periksa periksa kembali data pengadaan bus rongsokan cina ini
2019-09-28 04:39:02	BODO.	849377349435092992	detektive88	395984870	in	Twitter for Android			1,0	117769919351255040	@detektive88 @jokowi @aniesbaswedan Dimohon dengan amat sangat segera ke RSJ secepatnya. Semoga lekas sembuh!
2019-09-28 04:39:13	Eddoludwig	104773605892897587	detektive88	395984870	in	Twitter for Android				117769923915916492	@detektive88 @jokowi @aniesbaswedan Hahaha..matamu picek ccookk????????
2019-09-28 04:39:51	Eddoludwig	104773605892897587	ellykellyk	103529827229689036	in	Twitter for Android				117769939665110630	@ellykellyk @aniesbaswedan @Sofy_Beeeeeee Banyak orang gila baru jadi di sosmed????
2019-09-28 04:39:57	Qu Rosa	111628310292865433	Hilmi28	55067181	in	Twitter for Android				117769942305018675	@Hilmi28 @aniesbaswedan Awalnya radikal, fpi, hti dianggap kue yg paling laris untuk di jajakan. Tp saat masyarakat mulai cerdas, kue ini dirubah jadi kambing hitam... Penyebab semua rasa tidak enak. La makar Allah lain, si kambing ternyata tetaplah putih...
2019-09-28 04:40:14	cobiramasenju	100157404895699353	dias6969	892847922701770752	in	Twitter for Android				117769949221416550	@dias6969 @aniesbaswedan @fadlizon Ini kejadian papua 2 taun lalu bro
2019-09-28 04:41:21	Beneathbluesky	228222862	bos_sir	715696111	in	Twitter for Android				117769977401686835	@bos_sir @aniesbaswedan @chozin_ID @firmanyursak @musniumar Sabtu barokah. Semoga Aksi Mujahid 212 Selamatkan NKRI hari ini berjalan aman dan damai

Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Source	Geo-Location - Latitude	Geo-Location - Longitude	Retweet-Count	Id	Text
2019-09-28 04:44:06	shantiedta	283391380	JustinTrudeau	14260960	in	Twitter for iPhone				117770046689908736	@JustinTrudeau Boleh dicontoh pak @aniesbaswedan Sepertinya kita belum perlu art installation. Tapi perlu banyak pohon di sini. Belum perlu bongkar2 trotoar juga, pak.
2019-09-28 04:44:21	KADAL GURUN KADRUN?????? ??	2473830139	detektive88	395984870	in	Twitter for Android				117770053029178982	@detektive88 @jokowi @aniesbaswedan Baru 5 orang yg bersalaman sdh mau ganti jokowi... Terlepas dari itu dalam hal pencitraan memang top markotop si anies ini
2019-09-28 04:48:18	Siti Khumaidi_1974	921423146019905536	faizalassegaf	2259000780	in	Twitter for Android				117770152526057062	@faizalassegaf @bima_kencana19 @NajwaShihab @aniesbaswedan kok iso ngunu gemez z jadinya
2019-09-28 04:51:23	Madun Borneo-DKI	11493334507771468	Milleni81817560	110453474893900595	in	Twitter for Android				117770229789729587	@Milleni81817560 @detektive88 @jokowi @aniesbaswedan Yang merasakan keberhasilan presiden saat ini dibandingkan presiden yg lalu adalah rakyat diluar jakarta can pulau jawa, kalau soal pelaku korup tentu saja membenci pak @jokowi karena selama belio memimpin byk para pjabat & pemegang kekuasaan lainnya yg kehilangan lahan basah
2019-09-28 04:51:46	Hertik Nila Riyanti	988197633075236865	TofaGarisLurus	113610134795720704	in	Twitter for Android			1,0	117770239669830041	@TofaGarisLurus @aniesbaswedan @PMI_DKI_Jakarta Selama ini Cebong selalu melaporkan kampret dalam ketidaksukaannya. Tapi kampret tidak melaporkan apapun yang dilakukan oleh Cebong karena merasa tidak ditanggapi. Seharusnya kampret tidak cengeng dan tidak menyerah untuk menegakkan keadilan. Laporkan Cebong ditanggapi atau
2019-09-28 04:52:16	BUZZER BAYARAN	983323657572270080	detektive88	395984870	in	Twitter for Android				117770252359701299	@detektive88 @jokowi @aniesbaswedan Si carmuk..
2019-09-28 04:52:45	Habibi Lana	748323497127010304	detektive88	395984870	in	Twitter Web App				117770264394418176	@detektive88 @jokowi @aniesbaswedan Biar semua kali dikasih jaring
2019-09-28 04:53:51	Cak_Moendez	107690580752494182	dadesamsulrais	778049729770860545	in	Twitter for Android				117770292094860492	@dadesamsulrais @Hilmi28 @aniesbaswedan Ati2 nanti diblokir ??????
2019-09-28 04:54:10	K4K4_88	117750654968767283	Hilmi28	65067181	in	Twitter for Android				117770300103047987	@Hilmi28 @aniesbaswedan BOSNYA GOBLOK ANAK BUAH YANG BIKIN ATURAN GOBLOK JADINYA BEGO SEMUA
2019-09-28 04:58:35	iwanSdr	1853201329	detektive88	395984870	in	Twitter for Android				117770411299910861	@detektive88 @mas_piyuuu @jokowi @aniesbaswedan Sumpah saya terharu.. Gubernur rasa presiden

Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Source	Geo-Location - Latitude	Geo-Location - Longitude	Retweet-Count	Id	Text
2019-09-28 04:58:53	Tantearine	109538973532376268	Iswaldi_A	233498937	in	Twitter Web App				117770418899153305	@Iswaldi_A @detektive88 @jokowi @aniesbaswedan Yg ono katanya pleciden ,nyebut kota kendari letaknya dimana aja SALAH,ini bukan gk becus lgi jdi pleciden wajib mundur
2019-09-28 05:00:33	Tri	825102357880926208	BatakModeren	715734015895072768	in	Twitter for Android				117770460427374182	@BatakModeren Yup. Bisa jadi Hanya bertujuan skn mnjthkn @aniesbaswedan dan mefitnah sendana. Merusak rekon
2019-09-28 05:03:28	putra bangsa822	102949064803699916	ellykelyk	103529827229689036	in	Twitter for Android				117770533955626188	@ellykelyk @aniesbaswedan @Sofy_Beeyeee Gkgkg lucu klo oling ini jd pleciden
2019-09-28 05:04:33	2beAssertive-One	113833927447284531	2beAssertiveOne	113833927447284531	in	Twitter for Android				117770561316326195	@krisgesit @prabowo @sandiuno @aniesbaswedan @bachtiarnasir @Gus_Sholah @msaid_didu @hanafrais @TofaGarisLurus @Dahnilanzar @Gamal_Albinsaid @ustadtengkuzul @ZaitunRasmin @TeladanRasul @GNPF_MUI @TausiyahHadits Assalaamu'alaikum wr wb Sdhkah kita Subuh berjama'ah, sempurnakan sholat kmrn & hari ini di Masjid/Musholla? 28092019 Hanya 'Alash Sholaah Hanya 'Alal Falaah Ayo kita Sholat Ayo #KitaSambutKEMENANGAN Bangun..Jgn tertidur!! Berjama'ah..Jgn sendirian!! https://t.co/Vfi5FzSti
2019-09-28 05:06:37	Agus Wahyudhi	2813234707	CintaNKRI08	819930164092174336	in	Twitter for Android				117770613160018329	@CintaNKRI08 @Arief324787 @aniesbaswedan Adem bersih aja msh dibilang radikal, saking bencinya dgn musuh politik.
2019-09-28 05:08:28	#NegeriParaCusong	135743417	interisti2001	514876304	in	Twitter for Android				117770659717637324	@interisti2001 @detektive88 @mas piyuuu @jokowi @aniesbaswedan Enggak, itu kejadian beneran. Aktor itu tuuuhh yg suka selfie ngga jelas, yg cengar cengir disaat anak bangsa meregang nyawa
2019-09-28 05:11:46	Choki Juliano HRP	220496545	MudasirRomini	799108167170949121	in	Twitter for Android				117770743016094105	@MudasirRomini @inimaliki @Dennysiregar7 @aniesbaswedan @jokowi Lah si tolol bin dungu
2019-09-28 05:13:11	Erik Fadhil	832473756475260928	IpungLombok	86233947	in	Twitter for Android				117770778441608806	@IpungLombok @InfoBerkarya @aniesbaswedan @DivHumas_Polri Percuma nggak mungkin diproses, sudah buta tuli.
2019-09-28 05:13:13	Aquilla	114673156864110592	Mrmrshall_rama	421603051	in	Twitter for Android				117770779478860185	@Mrmrshall_rama @msaid_didu @aniesbaswedan Masya ALLAH,,, cobalah pak jokowi bisa seperti beliau pak Anies pemimpin yg di idamkan rakyatnya,, #THENEXTPRESIDENT

Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Source	Geo-Location - Latitude	Geo-Location - Longitude	Retweet-Count	Id	Text
2019-09-28 05:17:22	Mutiara Hitam Dari Timur ????	132743506	detektive88	895984870	in	Twitter for Android				-117770883676273050	@detektive88 @jokowi @aniesbaswedan lagi ada drama apaan sih ??
2019-09-28 05:17:34	Aira Afni Amalia	110966926763969740	AnangWa7943915	115633587999401164	in	Twitter for Android				-117770888702730649	@AnangWa79439150 @Hilmi28 @aniesbaswedan Siapa yg difitnah? Siapa yg ngefitnah? Apa isi fitnahnya?
2019-09-28 05:19:50	RT-TAN??(B L D P ???????????????)	897708050	edelweis2901	765760543185600513	in	Twitter for iPhone			1,0	117770945908000768	@edelweis2901 @aniesbaswedan Serahkan ke abud hehehe kiamat bujur Indonesia
2019-09-28 05:19:54	Alex	111848097714620416	MT_Reborn	113289530180979507	in	Twitter for Android				-117770947580688793	@MT_Reborn @CebongInsaf2 @aniesbaswedan @jokowi Protap jend. Merah sby dkk pancing kemarahan umat islam -demo besar yg melibatkan umat islam pasti banyak korban -stigma dibuat umat islam perusuh = arab perusuh = anis mendukung perusuh -komandan kosgama demokrat didongkrak utk menggantikan anis..
2019-09-28 05:21:42	????WONG??? BENG????	100854484171259904	MARSUDL	725675453768310784	in	Twitter for Android				-117770992801579008	@MARSUDI @detektive88 @jokowi @aniesbaswedan Bukannya udah gendut ya...?! ????
2019-09-28 05:23:03	Paja leseak	233498937	tantearine	109538973532376268	in	Twitter for Android				-117771026931882803	@tantearine @detektive88 @jokowi @aniesbaswedan Setiap orang pasti pernah salah, apalagi hanya salah sebut letak kota, sedang pin atm sendiri aja kita pernah salah, yg ono juga pernah salah sebut haiti di afrika, kesalahan dlm hal2 kecil itu manusiawi kok, yg ga boleh itu salah mengambil keputusan yg menyangkut khalayak banyak
2019-09-28 05:23:31	Singgasana	115452395501053132	CintaNKRI08	819930164092174336	in	Twitter for Android				-117771038612183040	@CintaNKRI08 @aniesbaswedan Aamiin ya Robb..
2019-09-28 05:24:07	Suyamto TL	62769025	insinyurgot	942610234643644417	in	Twitter for iPhone			1,0	117771053516644352	@insinyurgot @FerdinandHaeen2 @aniesbaswedan Mosok ferceb @FerdinandHaeen2 primata? Bukannya amfibi?
2019-09-28 05:24:11	Marcelino	887127198167031808	JKFC23456789	874602937737138178	in	Twitter for Android				-117771055338228121	@JKFC23456789 @Gerindra @DivHumas_Polri @hnurwahid @aniesbaswedan @hmskaban @BareskrimPolri @SuaraAnies @ZAEffendy @P3rnj3l4j4h @andre_rosiade @RizmaWidiono Ebong Dari dulu Sampe Sekarang gk berubah Masih 200
2019-09-28 05:24:15	Maldini Milano	108300958008959795	Gladislagiwoy	117239552346402816	in	Twitter for Android				-117771057046169600	@Gladislagiwoy @AyunindyaRizkya @aniesbaswedan Semua loe nyinyir anies ya?????bawaanya loe pada dengki semua sih???????semoga loe pda dampuni dosanya...

Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Source	Geo-Location - Latitude	Geo-Location - Longitude	Retweet-Count	Id	Text
2019-09-28 05:27:24	bang ipoel	4210959973	Hilmi28	65067181	in	Twitter for Android				- 117771136544151142	@Hilmi28 @aniesbaswedan Orangnya ya itu aja
2019-09-28 05:28:21	@acun	102310414771165184	detektive88	395984870	in	Twitter for Android				- 117771160343890329	@detektive88 @jokowi @aniesbaswedan anies apa kemampuan nya?
2019-09-28 05:28:52	haoesfir	2611093016	lpg_01	115908650316464538	in	Twitter for Android				- 117771173050952908	@lpg_01 @AhmadZa89201150 @detektive88 @mas_piyuuu @jokowi @aniesbaswedan yg goblog yg mecat, biar ga keliatan gobloknya
2019-09-28 05:30:15	Mazaya	8223891542	AhmadZa8920115	106434831858565120	in	Twitter for iPhone				- 117771208208447897	@AhmadZa89201150 @detektive88 @mas_piyuuu @jokowi @aniesbaswedan Biarlah gua goblok tpi waras timbang lo kaga waras wkwkwkw
2019-09-28 05:30:49	azzana	117286574602982195	fahiraidrisc	68304724	in	Twitter for iPhone				- 117771222438460211	@fahiraidrisc @Dahnilar @DKIJakarta @DivHumas_Polri @TMCPOldMetro @aniesbaswedan Yakini nggak kalau yg masuk ke ambulans itu bukan suruhan dari yg nuduh ? Trik spt itu kan trik basi
2019-09-28 05:30:54	Rinto	870650416106586112	MT_Reborn	113289530180979507	in	Twitter for Android				- 117771224222740070	@MT_Reborn @aniesbaswedan @jokowi Ni orang ga lebih dari abu janda
2019-09-28 05:31:26	Azis Ramlan	110054475456196608	ariyan_ana	983840696	in	Twitter for Android				- 117771238022819020	@ariyan_ana Ko sedih yah ???????? smoga sehat dan slalu ada dalam lindungan allah SWT pa @aniesbaswedan
2019-09-28 05:32:43	Yeni Nur	113116471379707904	Mrmrshall_rama	421603051	in	Twitter for Android				- 117771269951540428	@Mrmrshall_rama @aniesbaswedan Ini REZIM MALUNYA RIBUAN KALI ..TIAP MEMFITNAH SELALU ALKAH BUKA BOROKNYA
2019-09-28 05:33:57	Kimi	117612158596866048	mas_piyuuu	2885191782	in	Twitter for Android				- 117771301077468774	@mas_piyuuu @jokowi @aniesbaswedan anda terlibat kerusuhan ya ? Bencong kalau ngga berani lapor polisi ... dan kita anggap benar anda provokator
2019-09-28 05:33:58	Yeni Nur	113116471379707904	Willy_Vaizard	921190333	in	Twitter for Android				- 117771301753590579	@Willy_Vaizard @Mrmrshall_rama @aniesbaswedan @SutariUu111 @Aa_ti_Bdg @sunendar_agus @baee_wee Sarua ma akoh jawabarat pimpinannya selebgram bapak sosialita Jawa barat
2019-09-28 05:39:19	fokus bela nkri	744042211369713664	Mrmrshall_rama	421603051	in	Twitter for iPad				- 117771436370620006	@Mrmrshall_rama @aniesbaswedan Santun, ramah, intelek, berkuwalitas..

#	Proses	Remark	Char Text	Isi Tweet	To-User ?	Jenis Tweet	Freq Re-Tweet	Active Re-Tweet	Freq Tweet	Active Tweet	Active Tweet/Re-Tweet	Freq All	Active Freq All	Freq Retweet-Count	Active Retweet-Count	From	To	Active To	User Active/Passive	Status Label User	# Proses	# Label	Label
20	Yes		69		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	2	Active	Passive	Positive	7	11	Positive
21	Yes		104		1	Tweet	0	Passive	2	Active	Active Tweet	2	Active	0	Passive	1	3	Active	Active	Negative	5	9	Negativ
23	Yes		17		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	4	Active	Passive	Positive	8	12	Positive
26	Yes		139		1	Tweet	0	Passive	2	Active	Active Tweet	2	Active	1	Passive	1	5	Active	Active	Negative	6	11	Negativ
27	Yes		160		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	1	Passive	1	3	Active	Passive	Negative	7	12	Negativ
28	Yes		9		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	3	Active	Passive	Negative	8	13	Negativ
29	Yes		28		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	3	Active	Passive	Positive	9	14	Positive
31	Yes		16		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	3	Active	Passive	Negative	9	14	Negativ
32	Yes		62		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	2	Passive	Passive	Negative	10	15	Negativ
35	Yes		29		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	4	Active	Passive	Positive	10	16	Positive

#	Proses	Remark	Char Text	Isi Tweet	To-User ?	Jenis Tweet	Freq Re-Tweet	Active Re-Tweet	Freq Tweet	Active Tweet	Active Tweet/Re-Tweet	Freq All	Active Freq All	Freq Retweet-Count	Active Retweet-Count	From	To	Active To	User Active/Passive	Status Label User	# Proses	# Label	Label
36	Yes		109		1	Tweet	1	Passive	3	Active	Active Tweet	4	Active	4	Passive	1	4	Active	Active	Negative	11	18	Negativ
38	Yes		67		1	Tweet	2	Active	1	Passive	Active Re-Tweet	3	Active	1510	Active	1	2	Active	Active	Positive	11	17	Positive
40	Yes		36		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	3	Active	Passive	Negative	12	21	Negativ
41	Yes		184		1	Tweet	0	Passive	2	Active	Active Tweet	2	Active	0	Passive	1	16	Active	Active	Positive	12	18	Positive
43	Yes		65		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	3	Active	Passive	Positive	13	19	Positive
45	Yes		113		1	Tweet	0	Passive	2	Active	Active Tweet	4	Active	2	Passive	1	5	Active	Active	Positive	14	21	Positive
46	Yes		22		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	5	Active	Passive	Negative	13	22	Negativ
48	Yes		41		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	4	Active	Passive	Positive	15	22	Positive
49	Yes		72		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	3	Active	Passive	Positive	16	23	Positive

#	Proses	Remark	Char Text	Isi Tweet	To-User ?	Jenis Tweet	Freq Re-Tweet	Active Re-Tweet	Freq Tweet	Active Tweet	Active Tweet/Re-Tweet	Freq All	Active Freq All	Freq Retweet-Count	Active Retweet-Count	From	To	Active To	User Active/Passive	Status Label User	# Proses	# Label	Label
52	Yes		24		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	3	Active	Passive	Negative	14	23	Negativ
53	Yes		27		1	Tweet	0	Passive	3	Active	Active Tweet	3	Active	44	Passive	1	3	Active	Active	Positive	17	25	Positive
56	Yes		43		1	Tweet	0	Passive	2	Active	Active Tweet	2	Active	1	Passive	1	2	Active	Active	Negative	15	26	Negativ
57	Yes		178		1	Tweet	0	Passive	8	Active	Active Tweet	15	Active	1	Passive	1	4	Active	Active	Negative	16	27	Negativ
58	Yes		17		1	Tweet	0	Passive	8	Active	Active Tweet	8	Active	0	Passive	1	4	Active	Active	Negative	17	28	Negativ
59	Yes		120		1	Tweet	0	Passive	2	Active	Active Tweet	2	Active	0	Passive	1	4	Active	Active	Positive	18	26	Positive
60	Yes		14		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	2	Active	Passive	Positive	19	27	Positive
61	Yes		33		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	1	Passive	1	4	Active	Passive	Negative	18	29	Negativ
62	Yes		19		1	Tweet	1	Passive	1	Passive	Active Tweet	3	Active	531	Active	1	12	Active	Active	Positive	20	28	Positive
63	Yes		87		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	3	Active	Passive	Positive	21	29	Positive

#	Proses	Remark	Char Text	Isi Tweet	To-User ?	Jenis Tweet	Freq Re-Tweet	Active Re-Tweet	Freq Tweet	Active Tweet	Active Tweet/Re-Tweet	Freq All	Active Freq All	Freq Retweet-Count	Active Retweet-Count	From	To	Active To	User Active/Passive	Status Label User	# Proses	# Label	Label
66	Yes		20		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	2	Passive	Passive	Negative	19	31	Negativ
67	Yes		19		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	3	Active	Passive	Negative	20	32	Negativ
68	Yes		38		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	6	Active	Passive	Negative	21	33	Negativ
69	Yes		55		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	5	Active	Passive	Negative	22	34	Negativ
71	Yes		94		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	6	Active	Passive	Negative	23	35	Negativ
72	Yes		32		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	3	Active	Passive	Negative	24	36	Negativ
73	Yes		65		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	2	Passive	Passive	Positive	22	32	Positive
75	Yes		64		1	Tweet	0	Passive	3	Active	Active Tweet	3	Active	1	Passive	1	2	Active	Active	Positive	23	34	Positive
76	Yes		95		1	Tweet	1	Passive	1	Passive	Active Tweet	2	Active	162	Passive	1	3	Active	Passive	Negative	25	37	Negativ
77	Yes		67		1	Tweet	0	Passive	3	Active	Active Tweet	3	Active	1	Passive	1	7	Active	Active	Positive	24	35	Positive
82	Yes		33		1	Tweet	0	Passive	1	Passive	Active Tweet	1	Passive	0	Passive	1	2	Passive	Passive	Positive	25	39	Positive

@Anotation Removal	Transformation: Remove URL	Tokenization: Regexp	Indonesian Stemming	Transformation: Not (Negative)	Indonesian Stop word removal
astaghfirullah.. pak anis dan para pemimpin daerah lain.. tolong bantu evakuasi.. jgn nunggu mrk mati??	astaghfirullah.. pak anis dan para pemimpin daerah lain.. tolong bantu evakuasi.. jgn nunggu mrk mati??	astaghfirullah pak anis dan para pempindaerahlintolongbantu evakuasi jgn nunggu mrk mati	astaghfirullah pak anis dan para pimpin daerah lain tolong bantu evakuasi jgn nunggu mrk mati	astaghfirullah pak anis dan para pimpin daerah lain tolong bantu evakuasi jgn nunggu mrk mati	astaghfirullah pak anis dan para pimpin daerah lain tolong bantu evakuasi jgn nunggu mrk mati
itulahkenapasayamakin sangsisi jae bisa turun. perlu kelumpuhan jakarta agar mampu memaksa si jae turun. tapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur. disguising in bless. ??	itulahkenapasayamakin sangsisi jae bisa turun. perlu kelumpuhan jakarta agar mampu memaksa si jae turun. tapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur. disguising in bless. ??	itulah kenapa saya makin sangsi si jae bisa turun perlu kelumpuhan jakarta agar mampu memaksa si jae turun tapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur disguising in bless	itulah kenapa saya makin sangsi si jae bisa turun perlu lumpuh jakarta agar mampu paksasi jaeturuntapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur disguising in bless	itulah kenapa saya makin sangsi si jae bisa turun perlu lumpuh jakarta agar mampu paksasi jaeturuntapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur disguising in bless	itulah kenapa saya makin sangsi si jae bisa turun perlu lumpuh jakarta agar mampu paksasi jaeturuntapi gubernur dki itu mas saya paham benar kapasitas beliau sebagai gubernur disguising in bless
buka dong bang, penasaran neh. sebenarnya batunya ada apa nggak seh	buka dong bang, penasaran neh. sebenarnya batunya ada apa nggak seh	buka dong bang penasaran neh sebenarnya batunya ada apa nggak seh	buka dong bang penasaran neh benar batu ada apa nggak seh	buka dong bang penasaran neh benar batu ada apa nggak seh	buka dong bang penasaran neh benar batu ada apa nggak seh
tak bisakah bung periksa periksa kembali data pengadaan bus rongsok cina ini	tak bisakah bung periksa periksa kembali data pengadaan bus rongsok cina ini	tak bisakah bung periksa periksa kembali data pengadaan bus rongsok cina ini	tak bisakah bung periksa periksa kembali data gada bus rongsok cina ini	tak bisakah bung periksa periksa kembali data gada bus rongsok cina ini	tak bisakah bung periksa periksa kembali data gada bus rongsok cina ini
dimohon dengan amat sangat segera ke rsj secepatnya. semoga lekas sembuh!	dimohon dengan amat sangat segera ke rsj secepatnya. semoga lekas sembuh!	dimohon dengan amat sangat segera ke rsj secepatnya semoga lekas sembuh	mohon dengan amat sangat segera ke rsj cepat semoga lekas sembuh	mohon dengan amat sangat segera ke rsj cepat semoga lekas sembuh	mohon rsj cepat semoga lekas sembuh
hahaha..matamu picek ccookk????????	hahaha..matamu picek ccookk????????	hahahamatamu picek ccookk	hahahamatamu picek ccookk	hahahamatamu picek ccookk	hahahamatamu picek ccookk
banyakoranggilabarujadidi sosmed????	banyakoranggilabarujadidi sosmed????	banyakoranggilabarujadidi sosmed	banyakoranggilabarujadidi sosmed	banyakoranggilabarujadidi sosmed	banyakoranggilabarujadidi sosmed
awalnya radikal, fpi, hti dianggap kue yg paling laris untuk di jajakan. tp saat masyarakat mulai cerdas, kue ini dirubah jadi kambing hitam... penyebab semua rasa tidak enak. la makar allah lain, si kambing ternyata tetaplah putih...	awalnya radikal, fpi, hti dianggap kue yg paling laris untuk di jajakan. tp saat masyarakat mulai cerdas, kue ini dirubah jadi kambing hitam... penyebab semua rasa tidak enak. la makar allah lain, si kambing ternyata tetaplah putih...	awalnya radikal fpi hti dianggap kue yg paling laris untuk di jajakan tp saat masyarakat mulai cerdas kue ini dirubah jadi kambing hitam penyebab semua rasa tidak enak la makar allah lain si kambing ternyata tetaplah putih	awal radikal fpi hti anggap kue yg paling laris untuk jaja tp saat masyarakat mulai cerdas kue ini rubah jadi kambing hitam penyebab semua rasa tidak enak la makar allah lain si kambing nyata tetap putih	awal radikal fpi hti anggap kue yg paling laris untuk jaja tp saat masyarakat mulai cerdas kue ini rubah jadi kambing hitam penyebab semua rasa tidak enak la makar allah lain si kambing nyata tetap putih	radikal fpi htianggap kue laris jaja masyarakat cerdas kue rubah kambing hitam penyebab tidak_enak la makar allah si kambing nyata putih
ini kejadian papua 2 taun lalu bro	ini kejadian papua 2 taun lalu bro	ini kejadian papua taun lalu bro	ini jadi papua taun lalu bro	ini jadi papua taun lalu bro	ini jadi papua taun lalu bro
sabtu barokah. semoga aksi mujahid 212 selamatkan nkri hari ini berjalan aman dan damai	sabtu barokah. semoga aksi mujahid 212 selamatkan nkri hari ini berjalan aman dan damai	sabtu barokah semoga aksi mujahid selamatkan nkri hari ini berjalan aman dan damai	sabtu barokah semoga aksi mujahid selamat nkri hari ini jalan aman dan damai	sabtu barokah semoga aksi mujahid selamat nkri hari ini jalan aman dan damai	sabtu barokah semoga aksi mujahid selamat nkri hari ini jalan aman dan damai

@Anotation Removal	Transformation: Remove URL	Tokenization: Regexp	Indonesian Stemming	Transformation: Not (Negative)	Indonesian Stop word removal
boleh dicontoh pak sepertinya kita belum perlu art installation. tapi perlu banyak pohon di sini. belum perlu bongkar2 trotoar juga, pak.	boleh dicontoh pak sepertinya kita belum perlu art installation. tapi perlu banyak pohon di sini. belum perlu bongkar2 trotoar juga, pak.	boleh dicontoh pak sepertinya kita belum perlu art installation tapi perlu banyak pohon di sini belum perlu bongkar trotoar juga pak	boleh contoh pak seperti kita belum perlu art installation tapi perlu banyak pohon di sini belum perlu bongkar trotoar juga pak	boleh contoh pak seperti kita belum_perlu art installation tapi perlu banyak pohon di sini belum_perlu bongkar trotoar juga pak	contoh belum_perlu art installation pohon belum_perlu bongkar trotoar pak
baru 5 orang yg bersalaman sdh mau ganti jokowi terlepas dari itu dalam hal pencitraan memang top markotop si anies ini	baru 5 orang yg bersalaman sdh mau ganti jokowi terlepas dari itu dalam hal pencitraan memang top markotop si anies ini	baru orang yg bersalaman sdh mau ganti jokowi terlepas dari itu dalam hal pencitraan memang top markotop si anies ini	baru orang yg salam sdh mau ganti jokowi lepas dari itu dalam hal citra memang top markotop si anies ini	baru orang yg salam sdh mau ganti jokowi lepas dari itu dalam hal citra memang top markotop si anies ini	baru orang yg salam sdh mau ganti jokowi lepas dari itu dalam hal citra memang top markotop si anies ini
kok iso ngunu gemezzz jadinya	kok iso ngunu gemezzz jadinya	kok iso ngunu gemezzz jadinya	kok iso ngunu gemezzz jadi	kok iso ngunu gemezzz jadi	iso ngunu gemezzz
yang merasakan keberhasilan presiden saat ini dibandingkan presiden yg lalu adalah rakyat diluar jakarta can pulau jawa, kalau soal pelaku korup tentu saja membenci pak karena selama belio memimpin byk para pjabat & pemegang kekuasaan lainnya yg kehilangan lahan basah	yang merasakan keberhasilan presiden saat ini dibandingkan presiden yg lalu adalah rakyat diluar jakarta can pulau jawa, kalau soal pelaku korup tentu saja membenci pak karena selama belio memimpin byk para pjabat & pemegang kekuasaan lainnya yg kehilangan lahan basah	yang merasakan keberhasilan presiden saat ini dibandingkan presiden yg lalu adalah rakyat diluar jakarta can pulau jawa kalau soal pelaku korup tentu saja membenci pak karena selama belio memimpin byk para pjabat pemegang kekuasaan lainnya yg kehilangan lahan basah	yang rasa keberhasilan presiden saat ini banding presiden yg lalu adalah rakyat luar jakarta can pulau jawa kalau soal laku korup tentu saja benci pak karena selama belio pimpinbykparapjabatgangkuasa lain yg hilang lahan basah	yang rasa keberhasilan presiden saat ini banding presiden yg lalu adalah rakyat luar jakarta can pulau jawa kalau soal laku korup tentu saja benci pak karena selama belio pimpinbykparapjabatgangkuasa lain yg hilang lahan basah	keberhasilan presiden banding presiden rakyat jakarta can pulau jawa laku korup benci belio pimpin byk pjabat gang kuasa hilang lahan basah
selama ini cebong selalu melaporkan kampret dalam ketidaksukaannya. tapi kampret tidak melaporkan apapun yang dilakukan oleh cebong karena merasa tidak ditanggapi. seharusnya kampret tidak cengeng dan tidak menyerah untuk menegakkan keadilan. laporkan cebong ditanggapi atau	selama ini cebong selalu melaporkan kampret dalam ketidaksukaannya. tapi kampret tidak melaporkan apapun yang dilakukan oleh cebong karena merasa tidak ditanggapi. seharusnya kampret tidak cengeng dan tidak menyerah untuk menegakkan keadilan. laporkan cebong ditanggapi atau	selama ini cebong selalu melaporkan kampret dalam ketidaksukaannya tapi kampret tidak melaporkan apapun yang dilakukan oleh cebong karena merasa tidak ditanggapi seharusnya kampret tidak cengeng dan tidak menyerah untuk menegakkan keadilan laporkan cebong ditanggapi atau	selama ini cebong selalu lapor kampret dalam ketidaksukaannya tapi kampret tidak lapor apapun yang laku oleh cebong karena rasa tidak tanggap harus kampret tidak cengeng dan tidak serah untuk tegak adil lapor cebong ditanggapi atau	selama ini cebong selalu lapor kampret dalam ketidaksukaannya tapi kampret tidak_lapor apapun yang laku oleh cebong karena rasa tidak_tanggap harus kampret tidak_cengeng dan tidak_serah untuk tegak adil lapor cebong ditanggapi atau	cebong lapor kampret ketidaksukaannya kampret tidak_lapor apapun laku cebong tidak_tanggap kampret tidak_cengeng tidak_serah tegak adil lapor cebong ditanggapi
si carmuk..	si carmuk..	si carmuk	si carmuk	si carmuk	si carmuk
biar semua kali dikasih jaring	biar semua kali dikasih jaring	biar semua kali dikasih jaring	biar semua kali kasih jaring	biar semua kali kasih jaring	biar semua kali kasih jaring
ati2 nanti diblokir ??????	ati2 nanti diblokir ??????	ati nanti diblokir	ati nanti blokir	ati nanti blokir	ati nanti blokir
bosnya goblok anak buah yang bikin aturan goblok jadinya bego semua	bosnya goblok anak buah yang bikin aturan goblok jadinya bego semua	bosnya goblok anak buah yang bikin aturan goblok jadinya bego semua	bosnya goblok anak buah yang bikin atur goblok jadi bego semua	bosnya goblok anak buah yang bikin atur goblok jadi bego semua	bosnya goblok anak buah yang bikin atur goblok jadi bego semua
sumpah saya terharu.. gubernur rasa presiden	sumpah saya terharu.. gubernur rasa presiden	sumpah saya terharu gubernur rasa presiden	sumpah saya haru gubernur rasa presiden	sumpah saya haru gubernur rasa presiden	sumpah haru gubernur presiden

@Anotation Removal	Transformation: Remove URL	Tokenization: Regexp	Indonesian Stemming	Transformation: Not (Negative)	Indonesian Stop word removal
yg ono katanya pleciden ,nyebut kota kendari letaknya dimana aja salah,ini bukan gk becus lgi jdi pleciden wajib mundur	yg ono katanya pleciden ,nyebut kota kendari letaknya dimana aja salah,ini bukan gk becus lgi jdi pleciden wajib mundur	yg ono katanya pleciden nyebut kota kendari letaknya dimana aja salah,ini bukan gk becus lgi jdi pleciden wajib mundur	yg ono kata pleciden nyebut kota kendari letak mana aja salahini bukangkbecuslgi jdiplecidenwajib mundur	yg ono kata pleciden nyebut kota kendari letak mana aja salahini bukan_gk becus lgi jdi pleciden wajib mundur	ygono kata pleciden nyebut kota kendari letak mana aja salahini bukan_gk becus lgi jdi pleciden wajib mundur
yup. bisa jadi hanya bertujuan akn mnjthkn dan mefitnah cendana. merusak rekon	yup. bisa jadi hanya bertujuan akn mnjthkn dan mefitnah cendana. merusak rekon	yup bisa jadi hanya bertujuan akn mnjthkn dan mefitnah cendana merusak rekon	yup bisa jadi hanya tuju akn mnjthkn dan fitnah cendana rusak rekon	yup bisa jadi hanya tuju akn mnjthkn dan fitnah cendana rusak rekon	yup bisa jadi hanya tuju akn mnjthkn dan fitnah cendana rusak rekon
gkgkglucu klo olng ini jd pleciden	gkgkglucu klo olng ini jd pleciden	gkgkglucu klo olng ini jd pleciden	gkgkglucu klo olng ini jd pleciden	gkgkglucu klo olng ini jd pleciden	gkgkglucu klo olng ini jd pleciden
assalaamu'alaikum wrwb sdhkah kita subuh berjama'ah, sempurnakan sholat kmrn& hari ini di masjid/musholla? 28092019 - hayya 'alash sholaah hayya 'alal falaah ayo kita sholat ayo #kitasambutkemenangan bangun..jgn tertidur!! berjama'ah..jgn sendirian!! https://t.co/vfis5zsti	assalaamu'alaikum wrwb sdhkah kita subuh berjama'ah, sempurnakan sholat kmrn& hari ini di masjid/musholla? 28092019 - hayya 'alash sholaah hayya 'alal falaah ayo kita sholat ayo #kitasambutkemenangan bangun..jgn tertidur!! berjama'ah..jgn sendirian!!	assalaamualaikum wr wb sdhkah kita subuh berjamaah sempurnakan sholat kmrn hari ini di masjidmusholla hayya alash sholaah hayyaalalfalaahayokita sholatayo kitasambutkemenangan bangunjgn tertidur berjamaahjgn sendirian	assalaamualaikum wr wb sdhkah kita subuh berjamaah sempurna sholat kmrn hari ini di masjidmusholla hayya alash sholaah hayyaalalfalaahayokita sholatayo kitasambutkemenangan bangunjgn tidur berjamaahjgn sendiri	assalaamualaikum wr wb sdhkah kita subuh berjamaah sempurna sholat kmrn hari ini di masjidmusholla hayya alash sholaah hayyaalalfalaahayokita sholatayo kitasambutkemenangan bangunjgn tidur berjamaahjgn sendiri	assalaamualaikum wr wb sdhkah subuh berjamaah sempurna sholat kmrn masjidmusholla hayya alash sholaah hayya alal falaah ayo sholat ayo kitasambutkemenangan bangunjgn tidur berjamaahjgn
adem bersih aja msh dibilang radikal, saking bencinya dgn musuh politik.	adem bersih aja msh dibilang radikal, saking bencinya dgn musuh politik.	adem bersih aja msh dibilang radikal saking bencinya dgn musuh politik	adem bersih aja msh bilang radikal saking benci dgn musuh politik	adem bersih aja msh bilang radikal saking benci dgn musuh politik	adem bersih aja msh bilang radikal saking benci dgn musuh politik
enggak, itu kejadian beneran. aktor itu tuuuuh yg suka selfie ngga jelas, yg cengar cengir disaat anak bangsa meregang nyawa	enggak, itu kejadian beneran. aktor itu tuuuuh yg suka selfie ngga jelas, yg cengar cengir disaat anak bangsa meregang nyawa	enggak itu kejadian beneran aktor itu tuuuuh yg suka selfie ngga jelas yg cengar cengir disaat anak bangsa meregang nyawa	enggak itu jadi beneran aktor itu tuuuuh yg suka selfie ngga jelas yg cengar cengir saat anak bangsa regang nyawa	enggak itu jadi beneran aktor itu tuuuuh yg suka selfie ngga jelas yg cengar cengir saat anak bangsa regang nyawa	enggak itu jadi beneran aktor itu tuuuuh yg suka selfie ngga jelas yg cengar cengir saat anak bangsa regang nyawa
lah si tolol bin dungu	lah si tolol bin dungu	lah si tolol bin dungu	lah si tolol bin dungu	lah si tolol bin dungu	lah si tolol bin dungu
percuma nggak mungkin diproses, sudah buta tuli.	percuma nggak mungkin diproses, sudah buta tuli.	percuma nggak mungkin diproses sudah buta tuli	cuma nggak mungkin proses sudah buta tuli	cuma nggak mungkin proses sudah buta tuli	cuma nggak mungkin proses sudah buta tuli
masya allah,,, cobalah pak jokowi bisa seperti beliau pak anies pemimpin yg di idamkan rakyatnya,, #thenextpresident	masya allah,,, cobalah pak jokowi bisa seperti beliau pak anies pemimpin yg di idamkan rakyatnya,, #thenextpresident	masya allah cobalah pak jokowi bisa seperti beliau pak anies pemimpin yg di idamkan rakyatnya thenextpresident	masya allah coba pak jokowi bisa sepertibeliau pakanies pimpinyg di idam rakyatthenextpresident	masya allah coba pak jokowi bisa sepertibeliau pakanies pimpinyg di idam rakyatthenextpresident	masya allah coba jokowi beliau anies pimpin idam rakyat thenextpresident

@Anotation Removal	Transformation: Remove URL	Tokenization: Regexp	Indonesian Stemming	Transformation: Not (Negative)	Indonesian Stop word removal
lagi ada drama apaan sih ??	lagi ada drama apaan sih ??	lagi ada drama apaan sih	lagi ada drama apaan sih	lagi ada drama apaan sih	lagi ada drama apaan sih
siapa yg difitnah? siapa yg ngefitnah? apa isi fitnahnya?	siapa yg difitnah? siapa yg ngefitnah? apa isi fitnahnya?	siapa yg difitnah siapa yg ngefitnah apa isi fitnahnya	siapa yg fitnah siapa yg ngefitnah apa isi fitnah	siapa yg fitnah siapa yg ngefitnah apa isi fitnah	fitnah ngefitnah isi fitnah
serahkan ke abud hehehe kiamat bujur indonesia	serahkan ke abud hehehe kiamat bujur indonesia	serahkan ke abud hehehe kiamat bujur indonesia	serah ke abud hehehe kiamat bujur indonesia	serah ke abud hehehe kiamat bujur indonesia	serah ke abud hehehe kiamat bujur indonesia
protap jend. merah sby dkk - pancing kemarahan umat islam - demo besar yg melibatkan umat islam pasti banyak korban - stigma dibuat umat islam perusuh = arab perusuh = anis mendukung perusuh -komandan kosgama demokrat didongkrak utk menggantikan anis..	protap jend. merah sby dkk - pancing kemarahan umat islam - demo besar yg melibatkan umat islam pasti banyak korban - stigma dibuat umat islam perusuh = arab perusuh = anis mendukung perusuh -komandan kosgama demokrat didongkrak utk menggantikan anis..	protap jend merah sby dkk pancing kemarahan umat islam demo besar yg melibatkan umat islam pasti banyak korban stigma dibuat umat islam perusuh arab perusuh anis mendukung perusuh komandan kosgama demokrat didongkrak utk menggantikan anis	protap jend merah sby dkk pancing marah umat islam demo besar yg libat umat islam pasti banyak korban stigma buat umat islam rusuh arab rusuh anis dukung rusuh komandan kosgama demokrat dongkrak utk ganti anis	protap jend merah sby dkk pancing marah umat islam demo besar yg libat umat islam pasti banyak korban stigma buat umat islam rusuh arab rusuh anis dukung rusuh komandan kosgama demokrat dongkrak utk ganti anis	protap jend merah sby dkk pancing marah umat islam demo libat umat islam korban stigma umat islam rusuh arab rusuh anis dukung rusuh komandan kosgama demokrat dongkrak ganti anis
bukannyaudah gendut ya...?!????	bukannyaudah gendutya...?!????	bukannya udah gendut ya	bukan udah gendut ya	bukan_udah gendut ya	bukan_udah gendut
setiap orang pasti pernah salah, apalagi hanya salah sebut letak kota, sedang pin atm sendiri aja kita pernah salah, yg ono juga pernah salah sebut haiti di afrika, kesalahan dlm hal2 kecil itu manusiawi kok, yg ga boleh itu salah mengambil keputusan yg menyangkut khalayak banyak	setiap orang pasti pernah salah, apalagi hanya salah sebut letak kota, sedang pin atm sendiri aja kita pernah salah, yg ono juga pernah salah sebut haiti di afrika, kesalahan dlm hal2 kecil itu manusiawi kok, yg ga boleh itu salah mengambil keputusan yg menyangkut khalayak banyak	setiap orang pasti pernah salah apalagi hanya salah sebut letak kota sedang pin atm sendiri aja kita pernah salah yg ono juga pernah salah sebut haiti di afrika kesalahan dlm hal kecil itu manusiawi kok yg ga boleh itu salah mengambil keputusan yg menyangkut khalayak banyak	setiap orang pasti pernah salah apalagi hanya salah sebut letak kota sedang pin atm sendiri aja kita pernah salah yg ono juga pernah salah sebut haiti di afrika salah dlm hal kecil itu manusiawi kok yg ga boleh itu salah ambil putus yg sangkut khalayakbanyak	setiap orang pasti pernah salah apalagi hanya salah sebut letak kota sedang pin atm sendiri aja kita pernah salah yg ono juga pernah salah sebut haiti di afrika salah dlm hal kecil itu manusiawi kok yg ga boleh itu salah ambil putus yg sangkut khalayakbanyak	orang salah salah letak kota pin atm salah ono salah haiti afrika salah dlm manusiawi salah ambil putus sangkut khalayak
aamiin ya robb..	aamiin ya robb..	aamiin ya robb	aamiin ya robb	aamiin ya robb	aamiin ya robb
mosok ferceb primata? bukannya amfibi?	mosok ferceb primata? bukannya amfibi?	mosok ferceb primata bukannya amfibi	mosok ferceb primata bukan amfibi	mosok ferceb primata bukan_amfibi	mosok ferceb primata bukan_amfibi
ebong dari dulu sampe sekarang gk berubah masih 200	ebong dari dulu sampe sekarang gk berubah masih 200	ebong dari dulu sampe sekarang gk berubah masih	ebong dari dulu sampe sekarang gk ubah masih	ebong dari dulu sampe sekarang gk ubah masih	ebong sampe gk ubah
semua loe nyinyir anies ya????bawaanya loe pada dengki semua sih?????semoga loe pda dampuni dosanya...	semua loe nyinyir anies ya????bawaanya loe pada dengki semua sih?????semoga loe pda dampuni dosanya...	semua loe nyinyir anies yabawaanya loe padadengki semuasihssemoga loe pda dampuni dosanya	semua loe nyinyir anies yabawaanya loe padadengki semuasihssemoga loe pda dampuni dosa	semua loe nyinyir anies yabawaanya loe padadengki semuasihssemoga loe pda dampuni dosa	semua loe nyinyir anies yabawaanya loe padadengkisemuasihssemoga loe pda dampuni dosa

@Anotation Removal	Transformation: Remove URL	Tokenization: Regexp	Indonesian Stemming	Transformation: Not (Negative)	Indonesian Stop word removal
orangnya ya itu itu aja	orangnya ya itu itu aja	orangnya ya itu itu aja	orang ya itu itu aja	orang ya itu itu aja	orang ya itu itu aja
anies apa kemampuan nya?	anies apa kemampuan nya?	anies apa kemampuan nya	anies apa mampu nya	anies apa mampu nya	anies apa mampu nya
yg goblog yg mecat, biar ga kelihatan gobloknya	yg goblog yg mecat, biar ga kelihatan gobloknya	yg goblog yg mecat biar ga kelihatan gobloknya	yggoblog yg mecat biarga liat goblok	yggoblog yg mecat biargaliat goblok	yggoblog yg mecat biarga liat goblok
biarlah gua goblok tpi waras timbang lo kaga waras wkwkww	biarlah gua goblok tpi waras timbang lo kaga waras wkwkww	biarlah gua goblok tpi waras timbang lo kaga waras wkwkww	biar gua goblok tpi waras timbang lo kaga waras wkwkww	biar gua goblok tpi waras timbang lo kaga waras wkwkww	biar gua goblok tpi waras timbang lo kaga waras wkwkww
yakin nggak kalau yg masuk ke ambulani itu bukan suruhan dari yg nuduh ? trik spt itu kan trik basi	yakin nggak kalau yg masuk ke ambulani itu bukan suruhan dari yg nuduh ? trik spt itu kan trik basi	yakin nggak kalau yg masuk ke ambulani itu bukan suruhan dari yg nuduh trik spt itu kan trik basi	yakin nggak kalau yg masuk ke ambulani itu bukan suruh dari yg nuduh trik spt itu kan trik basi	yakin nggak kalau yg masuk ke ambulani itu bukan_suruh dari yg nuduh trik spt itu kan trik basi	yakin nggak kalau yg masuk ke ambulani itu bukan_suruh dari yg nuduh trik spt itu kan trik basi
ni orang ga lebih dari abu janda	ni orang ga lebih dari abu janda	ni orang ga lebih dari abu janda	ni orang ga lebih dari abu janda	ni orang ga lebih dari abu janda	ni orang ga lebih dari abu janda
ko sedih yah ???????? smoga sehat dan slalu ada dalam lindungan alloh swt pa	ko sedih yah ???????? smoga sehat dan slalu ada dalam lindungan alloh swt pa	ko sedih yah smoga sehat dan slalu ada dalam lindungan alloh swt pa	ko sedih yah smoga sehat dan slalu ada dalam lindung alloh swt pa	ko sedih yah smoga sehat dan slalu ada dalam lindung alloh swt pa	ko sedih yah smoga sehat dan slalu ada dalam lindung alloh swt pa
ini rezim malunya ribuan kali ..tiap memfitnah selalu alkah buka borok nya	ini rezim malunya ribuan kali ..tiap memfitnah selalu alkah buka borok nya	ini rezim malunya ribuan kali tiap memfitnah selalu alkah buka borok nya	ini rezim malu ribu kali tiap fitnah selalu alkah buka borok nya	ini rezim malu ribu kali tiap fitnah selalu alkah buka borok nya	ini rezim malu ribu kali tiap fitnah selalu alkah buka borok nya
anda terlibat kerusuhan ya ? bencong kalau ngga berani lapor polisidan kita anggap benar anda provokator	anda terlibat kerusuhan ya ? bencong kalau ngga berani lapor polisi dan kita anggap benar anda provokator	anda terlibat kerusuhan ya bencong kalau ngga berani lapor polisidan kita anggap benar anda provokator	anda libat rusuh ya bencong kalau ngga berani lapor polisidan kita anggap benar anda provokator	anda libat rusuh ya bencong kalau ngga berani lapor polisidan kita anggap benar anda provokator	anda libat rusuh ya bencong kalau ngga berani lapor polisidan kita anggap benar anda provokator
sarua ma akoh jawabarat pimpinannya selebgram bapak sosialita jawa barat	sarua ma akoh jawabarat pimpinannya selebgram bapak sosialita jawa barat	sarua ma akoh jawabarat pimpinannya selebgram bapak sosialita jawa barat	sarua ma akoh jawabarat pimpin selebgram bapak sosialita jawa barat	sarua ma akoh jawabarat pimpin selebgram bapak sosialita jawa barat	sarua ma akoh jawabarat pimpin selebgram bapak sosialita jawa barat
santun, ramah, intelek, berkuwalitas..	santun, ramah, intelek, berkuwalitas..	santun ramah intelek berkuwalitas	santun ramah intelek berkuwalitas	santun ramah intelek berkuwalitas	santun ramah intelek berkuwalitas