PENERAPAN METODE NAIVE BAYES DALAM ANALISIS SENTIMEN PADA DATA TWITTER (STUDI KASUS: HASIL DEBAT CALON PRESIDEN 2019)

SKRIPSI

Oleh
Tis Atul Aliah
NIM 152410101042

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS ILMU KOMPUTER
UNIVERSITAS JEMBER
2019

PENERAPAN METODE NAIVE BAYES DALAM ANALISIS SENTIMEN PADA DATA TWITTER (STUDI KASUS: HASIL DEBAT CALON PRESIDEN 2019)

SKRIPSI

Diajukan guna melengkapi tugas akhir dan memenuhi salah satu syarat untuk menyelesaikan Pendidikan Sarjana (S1) Program Studi Sistem Informasi Universitas Jember dan mencapai gelar Sarjana Komputer

Oleh
Tis Atul Aliah
NIM 152410101042

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS ILMU KOMPUTER
UNIVERSITAS JEMBER
2019

PERSEMBAHAN

Skripsi ini saya persembahkan untuk:

1. Allah SWT yang senantiasa memberikan rahmat dan hidayah-Nya untuk mempermudah dan melancarkan dalam pengerjaan skripsi;
2. Kedua orang tua saya, Bapak Masjen dan Ibu Maymunah yang selalu senatiasa mendoakan saya dalam berkuliah hingga menyelesaikan skripsi ini
3. Saudara-saudara saya;
4. Dosen Pembimbing saya, Achmad Maududie ST, M.Sc. dan Diksy Media Firmansyah S.Kom., yang telah meluangkan waktu, pikiran, dan perhatian dalam menyelesaikan penulisan skripsi;
5. Keluarga besar Selection yang selalu menemani dan membantu selama di perkuliahan;
6. Civitas Akademik Fakultas Ilmu Komputer atas pelayanan yang sangat baik selama di perkuliahan;
7. Almamater Fakultas Ilmu Komputer Universitas Jember.

MOTTO

“Tuhan tidak menuntut kita untuk sukses. Tuhan hanya menuntut kita berjuang tanpa henti.”

(Emha Ainun Nadjib)

PERNYATAAN

Saya yang bertanda tangan di bawah ini:

Nama : Tis Atul Aliah

NIM : 152410101042

menyatakan dengan sesungguhnya bahwa karya ilmiah yang berjudul "Penerapan Metode Naive Bayes dalam Analisis Sentimen pada Twitter (Studi Kasus: Hasil Debat Calon Presiden 2019)" adalah benar-benar hasil karya sendiri, kecuali jika dalam pengutipan substansi disebutkan sumbernya, belum pernah diajukan pada institusi mana pun, dan bukan karya jiplakan. Saya bertanggung jawab atas keabsahan dan kebenaran isinya sesuai dengan sikap ilmiah yang harus dijunjung tinggi.

Demikian pernyataan ini saya buat dengan sebenarnya, tanpa adanya tekanan dan paksaan dari pihak manapun serta bersedia mendapat sanksi akademik jika di kemudian hari pernyataan ini tidak benar.

Jember, 10 Desember 2019

Yang menyatakan,

Tis Atul Aliah

NIM 152410101042

SKRIPSI

PENERAPAN METODE NAIVE BAYES DALAM ANALISIS SENTIMEN PADA DATA TWITTER
(STUDI KASUS: HASIL DEBAT CALON PRESIDEN 2019)

Oleh :

Tis Atul Aliah

NIM 152410101042

Pembimbing :

Dosen Pembimbing Utama : Achmad Maududie ST, M.Sc.

Dosen Pembimbing Pendamping : Diksy Media Firmansyah S.Kom., M.Kom

PENGESAHAN PEMBIMBING

Skripsi berjudul "Penerapan Metode *Naïve Bayes* dalam Analisis Sentimen pada Data Twitter (Studi Kasus: Hasil Debat Calon Presiden 2019)", telah diuji dan disahkan pada:

Hari, tanggal : Selasa, 10 Desember 2019

Tempat : Fakultas Ilmu Komputer Universitas Jember

Disetujui oleh:

Pembimbing I	Pembimbing II
Achmad Maududie,ST., M.Sc.	Diksy Media Firmansyah, S.Kom.,M.Kom
NIP 197004221995121001	NIP 760016853

PENGESAHAN PENGUJI

Skripsi berjudul "Penerapan Metode *Naïve Bayes* dalam Analisis Sentimen pada Data Twitter (Studi Kasus: Hasil Debat Calon Presiden 2019)", telah diuji dan disahkan pada:

Hari, tanggal : Selasa, 10 Desember 2019

Tempat : Fakultas Ilmu Komputer Universitas Jember

Tim Penguji,

Penguji I	Penguji II
Nelly Oktavia A, S.Si., MT	Nova El Maidah, S.Si., M.Cs
NIP. 198410242009122008	NIP. 198411012015042001

Mengesahkan

Dekan Fakultas Ilmu Komputer,

Prof. Dr. Saiful Bukhori, ST., M. Kom
NIP. NIP. 196811131994121001

RINGKASAN

Penerapan Metode *Naïve Bayes* dalam Analisis Sentimen pada Data Twitter (Studi Kasus: Hasil Debat Calon Presiden 2019); Tis Atul Aliah, 152410101042, 2019; 66 Halaman, Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Jember.

Tahun 2019 Indonesia melaksanakan pemilihan umum presiden dan wakil presiden. Sebagai penyelenggara pemilihan umum presiden , KPU mengadakan debat calon presiden dan wakil presiden agar masyarakat mengetahui visi dan misinya. Kritik maupun saran. Salah satu cara untuk mengekspresikan sentimen melalui jejaring sosial, salah satunya adalah Twitter.

Penelitian ini bertujuan untuk menganalisis tingkat sentimen masyarakat pada Twitter terhadap hasil debat capres dan cawapres 2019. Metode yang digunakan dalam penelitian ini yaitu *Naive Bayes*, dan *output class* yang dituju adalah sentimen positif dan sentimen negatif.

Jumlah data pada penelitian ini yaitu 370 yang terbagi menjadi empat kelompok yaitu tweet yang mengarah pada Jokowi 85 data, yang mengarah ke Ma'ruf amin 115 data, yang mengarah ke Prabowo 85 data dan yang mengarah ke Sandiaga Uno 85. Dengan perbandingan data latih 80% dan data uji 20%, hasil pengujian menunjukkan pada sentimen tweet terhadap Jokowi, warganet memiliki sentimen positif terhadap Jokowi sebesar 8.33% dan sentimen negatif 91.67% dengan tingkat akurasi klasifikasi sebesar 70.588% . Pada sentimen tweet terhadap Ma'ruf Amin warganet memiliki sentimen positif terhadap Ma'ruf Amin sebesar 25% dan sentimen negatif 75% dengan tingkat akurasi klasifikasi sebesar 17.391% . Pada sentimen tweet terhadap Prabowo warganet memiliki sentimen positif terhadap Prabowo sebesar 75% dan sentimen negatif 25% dengan tingkat akurasi klasifikasi sebesar 35.294%. Pada sentimen tweet terhadap Sandiaga Uno warganet memiliki sentimen positif terhadap Sandiaga Uno sebesar 0% dan sentimen negatif 100% dengan tingkat akurasi klasifikasi sebesar 35.294%.

PRAKATA

Puji syukur kehadirat Allah SWT atas segala rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi dengan judul "Penerapan Metode *Naïve Bayes* dalam Analisis Sentimen pada Data Twitter (Studi Kasus: Hasil Debat Calon Presiden 2019)". Skripsi ini disusun untuk memenuhi salah satu syarat menyelesaikan pendidikan Strata Satu (S1) Fakultas Ilmu Komputer Universitas Jember.

Penyusunan skripsi ini tidak lepas dari bantuan berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih kepada:

1. Prof. Dr. Saiful Bukhori, ST.,M.Kom selaku Dekan Fakultas Ilmu Komputer Universitas Jember;
2. Achmad Maududie ST, M.Sc.selaku Dosen Pembimbing Utama dan Diksy Media Firmansyah S.Kom., selaku Dosen Pembimbing Pendamping yang telah meluangkan waktu, pikiran, dan perhatian dalam penulisan skripsi;
3. Diah Ayu Retnani Wulandari S.T.,M.Eng selaku Dosen Pembimbing Akademik (DPA), yang telah mendampingi penulis sebagai mahasiswa;
4. Nelly Oktavia A, S.Si., MT selaku Dosen Penguji I yang telah meluangkan waktu dan memberikan masukan demi sempurnanya skripsi ini;
5. Nova El Maidah, S.Si., M.Cs selaku Dosen Penguji II yang telah meluangkan waktu dan memberikan masukan demi sempurnanya skripsi ini;
6. Sahabat-sahabat seperjuangan semasa kuliah, Syarif, Dwiki, Irfan, Epaw, Ica, Ludfi, Rida, dan teman-teman Selection yang selalu membantu saya dalam penyelesaian tugas kuliah selama ini;
7. Seluruh Bapak dan Ibu dosen beserta staff karyawan di Fakultas Ilmu Komputer Universitas Jember;
8. Semua mahasiswa Fakultas Ilmu Komputer yang telah menjadi keluarga bagi penulis selama menempuh pendidikan S1;

Penulis menyadari bahwa laporan ini masih jauh dari sempurna, oleh sebab itu penulis mengharapkan adanya masukan yang bersifat membangun dari semua pihak. Penulis berharap skripsi ini dapat bermanfaat bagi semua pihak.

Jember, 10 Desember 2019

Penulis

DAFTAR ISI

HALAMAN JUDUL i
PERSEMBAHAN ii
MOTTO iii
PERNYATAAN iv
SKRIPSI v
PENGESAHAN PEMBIMBING vi
PENGESAHAN PENGUJI vii
RINGKASAN viii
PRAKATA ix
DAFTAR ISI xi
DAFTAR GAMBAR xiii
DAFTAR TABEL xiv

BAB I PENDAHULUAN 1
1.1. Latar Belakang 1
1.2. Rumusan Masalah 2
1.3. Tujuan 2
1.4. Manfaat 3
1.5. Batasan Masalah 3

BAB II TUNJAUAN PUSTAKA 4
2.1. Twitter 4
2.2. Sentimen Analisis 4
2.3. Data Mining 5
2.4. Pengelompokan Data Mining 7
2.5. Klasifikasi 9
2.6. Metode *Naive Bayes* 10
2.7. Confusion Matrix 13
2.8. Penelitian Terdahulu 14

BAB III METODOLOGI PENELITIAN 16
3.1. Jenis Penelitian 16
3.2. Tahapan Penelitian 16
3.3. Pengumpulan Data 18
3.4. Pembuatan Kamus Data 20
3.5. Preprocessing 21
3.6. Penyusunan Model 24
3.7. Pengujian Model 27
3.7.1. Confusion Matrix 27
3.7.2. Pengukuran Akurasi 27

BAB IV HASIL DAN PEMBAHASAN 28
4.1. Hasil Pengumpulan Data 28
4.2. Pembuatan Kamus Data 34
4.3. Preprocessing 44
4.4. Hasil Penyusunan Model 47
4.5. Pengujian Model 54
4.5.1. Confusion Matrix 54
4.5.2. Pengukuran Akurasi 54

BAB V PENUTUP 57
5.1. Kesimpulan 57
5.2. Saran 58

DAFTAR PUSTAKA

DAFTAR GAMBAR

Gambar 3.1. Tahapan Penelitian 17
Gambar 3.2 Alur pengumpulan data 18
Gambar 3.3 Gambaran proses *case folding* 22
Gambar 3.4 Gambaran *cleansing* 22
Gambar 3.5 Gambaran *tokenizing*. 23
Gambar 3.6 Gambaran *stopword* 23
Gambar 3.7 Gambaran *stemming* 24
Gambar 3.8 Alur perhitungan *Naive Bayes* 25
Gambar 4.1 Grafik sentimen masyarakat terhadap calon presiden dan wakil presiden 56

DAFTAR TABEL

Tabel 2.1 Tabel Confusion Matrix 14
Tabel 3.1 Contoh Penghapusan Retweet 19
Tabel 3.2 Partisi Data 20
Tabel 3.3 Teknik Pembentukan Kamus Positif dan Negatif 20
Tabel 3.4 Contoh Kamus Negasi 21
Tabel 3.5 Contoh Hasil Hitung Prior Probability 26
Tabel 3.6 Potongan hasil hitung *Naive Bayes* pada data latih 26
Tabel 4.1 Potongan Hasil Crawling 28
Tabel 4.2 Potongan Data Tweet Terhadap Jokowi 30
Tabel 4.3 Potongan Data Tweet Terhadap Ma'ruf Amin 31
Tabel 4.4 Potongan Data Tweet Terhadap Prabowo 32
Tabel 4.5 Potongan Data Tweet Terhadap Sandiaga Uno 33
Tabel 4.6 Potongan Data Tweet Kuesioner Arah Sentimen Pada Jokowi 34
Tabel 4.7 Potongan Kamus Positif pada Data Latih Tweet Jokowi 35
Tabel 4.8 Potongan Kamus Negatif pada Data Latih Tweet Jokowi 36
Tabel 4.9 Potongan Data Tweet Kuesioner Arah Sentimen Pada Ma'ruf 37
Amin 36
Tabel 4.10 Potongan Kamus Positif Pada Data Latih Tweet Ma'ruf Amin 38
Tabel 4.11 Potongan Kamus Negatif Pada Data Latih Tweet Ma'ruf Amin .. 38
Tabel 4.12 Potongan Data Tweet Kuesioner Arah Sentimen Pada Prabowo . 39
Tabel 4.13 Potongan Kamus Positif pada Data Latih Tweet Prabowo 40
Tabel 4.14 Potongan Kamus Negatif Pada Data Latih Tweet Prabowo 41
Tabel 4.15 Potongan Data Tweet Kuesioner Arah Sentimen Pada
Sandiaga Uno 42
Tabel 4.16 Potongan Kamus Positif Pada Data Latih Tweet Sandiaga Uno .. 43
Tabel 4.17 Potongan Kamus Negatif Pada Data Latih Tweet Sandiaga
Uno 43
Tabel 4.18 Hasil Penghimpunan Kamus Negasi Positif 44
Tabel 4.19 Hasil Penghimpunan Kamus Negasi Negatif 45

Tabel 4.20 Potongan data stopword 46
Tabel 4.21 Potongan data kata dasar 46
Tabel 4.22 Hasil prior probability tiap calon 47
Tabel 4.23 Hasil perhitungan *Naive Bayes* data uji Jokowi 49
Tabel 4.24 Hasil perhitungan *Naive Bayes* data uji Ma'ruf Amin 50
Tabel 4.25 Hasil perhitungan *Naive Bayes* data uji Prabowo 51
Tabel 4.26 Hasil perhitungan *Naive Bayes* data uji Sandiaga Uno 52
Tabel 4.27 Hasil perbandingan *output Naive Bayes* dan nilai aktual pada data uji Jokowi 55
Tabel 4.28 Hasil perbandingan *output Naive Bayes* dan nilai aktual pada data uji Ma'ruf Amin 56
Tabel 4.29 Hasil perbandingan *output Naive Bayes* dan nilai aktual pada data uji Prabowo 58
Tabel 4.30 Hasil perbandingan *output Naive Bayes* dan nilai aktual pada data uji Sandiaga Uno 60
Tabel 4.31 Hasil penghitungan *confusion matrix* pada data uji 62
Tabel 4.32 Hasil pengukuran akurasi dan sentimen pada data uji 63

BAB I PENDAHULUAN

1.1. Latar belakang

Tahun 2019 Indonesia melaksanakan pemilihan umum presiden dan wakil presiden. Sebagai penyelenggara pemilihan umum presiden , KPU mengadakan debat calon presiden dan wakil presiden agar masyarakat mengetahui visi dan misinya. Pemilihan umum tersebut diselenggarakan sedemokratis mungkin dengan melibatkan partisipasi masyarakat. Partisipasi ini antara lain berupa sarana penyampaian opini, kritik maupun saran. Salah satu cara untuk mengekspresikannya adalah melalui jejaring sosial, salah satunya adalah *Twitter*. Penggunaan *Twitter* di Indonesia saat ini berkembang sangat pesat. Kementerian Komunikasi dan Informatika (Kemenkominfo) mengungkapkan pengguna internet di Indonesia saat ini mencapai 63 juta orang. Tercatat Indonesia menduduki posisi ketiga sebagai Negara dengan pengguna *Twitter* terbanyak. Menurut CEO *Twitter* Dick Costlo, pada pertengahan tahun 2015 mencapai lima puluh juta (CNN Indonesia, 2016).

Tweet adalah teks status pengguna yang digunakan untuk memberikan informasi di *Twitter*. Pada umumnya *tweet* digunakan untuk mem-*posting* hal tentang diri pengguna dan berbagi informasi. Isi *tweet* juga dapat mengekspresikan perasaan pengguna, hal ini bersifat penilaian subjektif atau opini. Selain karakter *alphabet*, teks dalam *tweet* juga sering melibatkan karakter *emoticon* untuk lebih menekankan opini yang dikemukakan. Opini melalui *tweet* inilah yang dapat dimanfaatkan untuk melihat bagaimana sentimen yang dimunculkan salah satunya adalah mengenai opini seseorang terhadap hasil debat calon presiden 2019. Seperti yang ditulis akun @Rizalxiau pada *Twitter* “di forum resmi #DebatPilpres2019 Jokowi memberi data salah”.

Terkait dengan penyampaian opini hasil debat calon presiden 2019, polaritas positif atau negatif opini menggunakan teknik klasifikasi dokumen. Salah satu metode yang populer digunakan dalam pengklasifikasian dokumen adalah metode *Naïve Bayes Classifier*. Metode *Naïve Bayes Classifier* mempunyai kecepatan dan akurasi yang tinggi ketika diaplikasikan dalam basis data yang besar dan data yang

beragam. Hasil klasifikasi sentimen yang dihasilkan dapat menjadi komponen pendukung untuk masyarakat menentukan pilihan calon presiden. Masyarakat dapat mengetahui pendapat masyarakat umum tentang hasil debat. Sesuai dengan penelitian dari Buntoro (2017) tentang "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 di *Twitter*" mengatakan tingkat akurasi tertinggi 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%.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, maka dapat dibuat rumusan masalah sebagai berikut:

1. Bagaimana mengimplementasikan metode *Naive Bayes* untuk mengklasifikasikan sentimen data Twitter terhadap hasil debat calon presiden dan wakil presiden 2019?
2. Berapa tingkat akurasi dari hasil implementasi metode *Naive Bayes* dalam mengklasifikasikan sentimen data Twitter terhadap hasil debat calon presiden dan wakil presiden 2019?

1.3. Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Mengetahui hasil metode *Naive Bayes* dalam mengklasifikasikan sentimen data Twitter terhadap hasil debat calon presiden dan wakil presiden 2019.
2. Mengetahui tingkat akurasi dari hasil implementasi metode *Naive Bayes* dalam mengklasifikasikan sentimen data Twitter terhadap hasil debat calon presiden dan wakil presiden 2019.

1.4. Manfaat

Hasil penelitian ini diharapkan dapat memberi manfaat, antara lain:

1. Pengembangan metode data *mining* terhadap *social network analityc*.
2. Dengan mengetahui persentase sentimen masyarakat dapat dijadikan bahan pertimbangan sebagai kritik dan saran yang valid serta dapat digunakan untuk mengukur sentimen analisis yang lain seperti pemilihan legislatif dan kepala daerah.

1.5. Batasan Masalah

Agar dalam penelitian ini dapat mencapai sasaran dan tujuan yang diharapkan, maka permasalahan yang ada hanya dibatasi pada:

1. Klasifikasi sentimen masyarakat berdasarkan sentimen positif dan sentimen negatif.
2. Data *Twitter* yang dianalisis tidak mencakup data *retweet*, gambar atau foto, video dan *link*.
3. Rentang waktu *crawling* data yaitu 17 Januari 2019 sampai 13 April 2019.
4. Hanya menganalisis *tweet* dari Indonesia.
5. Pembagian data pada penelitian ini yaitu 80% data latih dan 20% data uji.
6. Metode klasifikasi yang digunakan adalah *Naive Bayes*.

BAB II TINJAUAN PUSTAKA

Bab pendahuluan merupakan bab pertama dari suatu penulisan yang berisi gambaran topik terkait isi yang akan disajikan. Bab ini berisi latar belakang, rumusan masalah, tujuan, manfaat, dan batasan masalah.

2.1. *Twitter*

Menurut Wikipedia, *Twitter* adalah layanan jejaring sosial dan mikroblog dalam jaringan (daring) yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis *text* hingga 140 karakter akan tetapi pada tanggal 07 November 2017 bertambah hingga 280 karakter yang dikenal dengan sebutan kicauan (*tweet*). Secara standar, *tweet* pengguna dapat terlihat oleh umum, namun pengguna dapat membatasi pengiriman kicauan hanya bagi pengikut mereka. Pengguna bisa "berkicau" melalui situs *Twitter*, aplikasi eksternal yang kompatibel (seperti untuk telepon pintar), ataupun melalui layanan pesan singkat (SMS) yang tersedia di negara-negara tertentu. Layanan-layanan tersebut bersifat gratis, kecuali layanan SMS, yang dikenakan biaya oleh penyedia layanan seluler. Pengguna bisa berlangganan kicauan pengguna lain dengan cara mengikuti (*follow*) pengguna yang bersangkutan, dan pengguna yang mengikuti tersebut akan menjadi pengikut (*followers*) bagi pengguna yang diikutinya. Pengguna dapat mengelompokkan kicauan menurut topik atau jenis dengan menggunakan tagar (*hashtag*), kata atau frasa yang diawali dengan tanda #. Sedangkan tanda @, yang diikuti dengan nama pengguna, digunakan untuk mengirim atau membalas kicauan pada pengguna lain.

2.2. Sentimen Analisis

Sentimen analisis adalah merupakan salah satu bidang dari ilmu komputer yang mempelajari komputasi linguistik, pengolahan bahasa alami, dan *text mining* yang bertujuan untuk menganalisa emosi, penilaian, sikap, pendapat, sentimen, evaluasi seseorang terhadap seorang pembicara atau penulis berkenaan dengan

suatu produk, layanan, organisasi, individu, topik publik, topik, acara, ataupun kegiatan tertentu (Liu, 2012). Proses utama dalam analisis sentimen yaitu mengelompokkan teks yang terdapat dalam sebuah kalimat atau dokumen kemudian menentukan pendapat yang dikemukakan tersebut apakah bersifat posistif, negatif, atau netral. Analisis sentimen dapat digunakan untuk mencari pendapat tentang produk, merek atau tokoh publik dan menentukan apakah mereka dilihat positif atau negatif (Saraswati, 2011). Analisis sentimen atau *opinion mining* adalah deteksi sikap-sikap terhadap objek atau orang (beineke *et al*. 2014). Analisis sentimen dapat digunakan untuk mendapatkan persentase sentimen positif dan sentimen negatif terhadap seseorang, perusahaan, institusi, produk atau pada sebuah kondisi tertentu. Nilai dari analisis sentimen bisa dipecah menjadi 3 yakni, sentimen positif, sentimen negatif dan sentimen netral atau diperdalam lagi sehingga dapat menemukan siapa atau kelompok yang menjadi sumber sentimen positif atau sentimen negatif. Tugas analisis sentimen yaitu mengelompokkan teks ke dalam kalimat atau dokumen kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen yang dianalisis apakah bersifat positif, negatif, atau netral (Dehhaf, 2010). Sentimen atau opini mengacu pada fokus topik tertentu, pernyataan pada satu topik mungkin akan berbeda makna dengan pernyataan yang sama pada *subject* yang berbeda. Alasan tersebut menyebabkan beberapa penelitian terutama pada *review* produk didahului dengan menentukan elemen dari sebuah produk yang sedang dibicarakan sebelum memulai proses *opinion mining*(Manalu, 2014).

Analisis sentimen bertujuan untuk melakukan penilaian terhadap emosi, sikap, pendapat, evaluasi yang disampaikan oleh seseorang pembicara atau penulis terhadap sebuah produk atau terhadap tokoh masyarakat. Alasan tersebut menyebabkan beberapa penelitian terutama pada review produk didahului dengan menentukan elemen dari sebuah produk yang sedang dibicarakan sebelum memulai proses *opinion mining*. Contoh kalimat yang mengandung sentimen positif yaitu “Jokowi sangat cerdas cocok menjadi Presiden”, contoh kalimat yang mengandung sentimen negatif yaitu “saya benci dengan Prabowo”, dan contoh

kalimat yang mengandung sentimen netral yaitu “sipapun presidennya saya selalu mendukung untuk kemajuan Indonesia”.

2.3. Data Mining

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar (Turban, dkk. 2005). Definisi umum dari data mining itu sendiri adalah proses pencarian pola-pola yang tersembunyi (*hidden patern*) berupa pengetahuan (*knowledge*) yang tidak diketahui sebelumnya dari suatu sekumpulan data yang mana data tersebut dapat berada di dalam database, data *werehouse*, atau media penyimpanan informasi yang lain. Hal penting yang terkait di dalam data mining adalah:

1. Data mining merupakan suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan diproses berupa data yang sangat besar.
3. Tujuan data mining adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat

Data mining dilakukan dengan *tool* khusus, yang mengeksekusi operasi data mining yang telah didefinisikan berdasarkan model analisis. Data mining merupakan proses analisis terhadap data dengan penekanan menemukan informasi yang tersembunyi pada sejumlah data besar yang disimpan ketika menjalankan bisnis perusahaan. Menurut Larose (2005), kemajuan luar biasa yang terus berlanjut dalam bidang data mining didorong oleh beberapa faktor antara lain:

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 peningkataan akses data melalui navigasi web dan internet.
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

Istilah data mining dan *knowledge discovery in databases* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lainnya. Salah satu tahapan dalam keseluruhan proses KDD adalah data mining. Proses KDD itu ada 5 tahapan yang dilakukan secara terurut, yaitu:

1. Data selection

   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.
2. Pre-processing / cleaning

   Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.
3. Transformation

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.
4. Data mining

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan

metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keselurahan.

5. Interpretation / evalution

   Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagaian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.4. Pengelompokan Data Mining

Menurut Larose, Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

1. Deskripsi

   Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data. Sebagai contoh, petugas pengumpul suara mungkin tidak menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecendrungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecendrungan.

2. Estimasi

   Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model dibangun dengan record lengkap menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

3. Prediksi

   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa metode dan

teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

4. Klasifikasi

   Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

5. Pengklusteran

   Pengklusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record-record dalam kluster lain. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.

6. Asosiasi

   Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (market basket analysis).

2.5. Klasifikasi

Klasifikasi adalah sebuah proses untuk menemukan sebuah model yang menjelaskan dan membedakan konsep atau kelas data dengan tujuan memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui (Tan et al.,

2004). Dalam klasifikasi, diberikan sejumlah *record* yang dinamakan data latih, yang terdiri dari beberapa atribut yang dapat berupa kontinu ataupun kategoris, salah satu atribut menunjukkan kelas untuk *record*. Tujuan dari klasifikasi adalah untuk:

1. Menemukan model dari data latih yang membedakan *record* ke dalam kategori atau kelas yang sesuai, model tersebut kemudian digunakan untuk mengklasifikasikan *record* yang kelasnya belum diketahui sebelumnya pada *testing set*.
2. Mengambil keputusan dengan memprediksi suatu kasus, berdasarkan hasil klasifikasi yang diperoleh.

Untuk mendapatkan model, harus dilakukan analisis terhadap data latih, Sedangkan data uji digunakan untuk mengetahui tingkat akurasi dari model yang telah dihasilkan. Klasifikasi dapat digunakan untuk memprediksi nama atau nilai kelas dari suatu objek data. Proses klasifikasi data dapat dibedakan dalam dua tahap, yaitu:

1. Pembangunan Model

   Tiap-tiap *record* yang digunakan dalam pembangunan model dianalisis berdasarkan nilai-nilai atributnya dengan menggunakan suatu algoritma klasifikasi untuk mendapatkan model.

2. Klasifikasi

   Pengujian dilakukan untuk mengetahui tingkat akurasi dari model yang dihasilkan. Jika tingkat akurasi yang diperoleh sesuai dengan nilai yang ditentukan, maka model tersebut dapat digunakan untuk mengklasifikasikan *record* data baru yang belum pernah dilatihkan atau diujikan sebelumnya. Pembuatan model menguraikan sebuah set dari penentuan kelas-kelas sebagai:

   a. Setiap *record* diasumsikan sudah mempunyai kelas yang dikenal seperti ditentukan oleh label kelas atribut.

b. Kumpulan *record* yang digunakan untuk membuat model disebut data pelatihan.
c. Model direpresentasikan sebagai pola dalam penentuan klasifikasi.

Penggunaan model menguraikan pengklasifikasian data yang akan diuji ataupun objek yang belum diketahui. Adapun parameter keberhasilan dari model yang terdiri dari:

a. Label yang telah diketahui dari data latih dibandingkan dengan hasil klasifikasi dari model.
b. Nilai akurasi adalah persentase dari kumpulan data uji yang diklasifikasikan secara tepat oleh model.
c. Data uji tidak sama dengan data latih.
d. Jika sesuai, gunakan model untuk mengklasifikasi data *record* yang label kelasnya belum diketahui.

2.6. Metode *Naive Bayes*

Naïve Bayes merupakan metode probabilistik pengklasifikasian sederhana berdasarkan Teorema Bayes di mana pengklasifikasian dilakukan melalui *training set* sejumlah data secara efisien (Hadiyani, 2013). *Naïve Bayes* mengasumsikan bahwa nilai dari sebuah *input* atribut pada kelas yang diberikan tidak tergantung dengan nilai atribut yang lain (Pang-Ping, etc, 2006). Teorema Bayes sendiri dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema *Bayes*. Di mana persamaan teori *Bayes* tersebut adalah:

$$P(C|X) = \frac{p(X|C)p(c)}{p(x)} \quad \text{.....(1)}$$

Keterangan:

X : Data dengan kelas yang belum diketahui

C : Hipotesis data X merupakan suatu kelas spesifik

P(C|X): Probabilitas hipotesis C berdasar kondisi X (probabilitas posterior)

P(C) : Probabilitas hipotesis C (probabilitas prior)

P(X|C) : Probabilitas X berdasarkan kondisi pada hipotesis C

P(X) : Probabilitas X

Untuk menjelaskan Teorema *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, Teorema *Bayes* pada persamaan (1) disesuaikan menjadi persamaan (2):

$$P(C|X1 \dots Xn) = \frac{p(c)p(X1,\dots Xn \curlyvee C)}{p(X1,\dots Xn)} \quad \text{.....(2)}$$

Di mana Variabel C merepresentasikan kelas, sementara variabel X1 ... Xn merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi atau kriteria. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (Posterior) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, sering kali disebut *prior*), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (disebut juga *likelihood*), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (disebut juga *evidence*). Karena itu, rumus di atas dapat pula ditulis secara sederhana pada persamaan:

$$Posterior = \frac{Priorxclikelihood}{evidence} \quad \text{.....(3)}$$

Nilai *Evidence* selalu tetap untuk setiap kelas pada satu sampel. Nilai dari posterior tersebut nantinya akan dibandingkan dengan nilai nilai posterior kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. Penjabaran lebih lanjut rumus *Bayes* tersebut dilakukan dengan menjabarkan (C|X1, …, Xn) menggunakan aturan perkalian sebagai berikut:

P(C|X1, ..., Xn)= P(C) P(X1,…,Xn|C).....(5)

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor-faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Di sinilah digunakan asumsi independensi yang sangat tinggi (naïf), bahwa masing-masing kriteria (X1,X2,…,Xn) saling bebas (independen) satu sama lain.

Naïve Bayes Classifier adalah konsep probabilitas penentuan kelompok kelas dokumen. Algoritma klasifikasi ini dapat mengolah data dalam jumlah besar

dengan hasil akurasi yang tinggi. Algoritma *Naive Bayes Classifier* terdiri dari dua tahap. Tahap pertama adalah pelatihan terhadap himpunan dokumen contoh (data latih) dan tahap kedua adalah proses klasifikasi dokumen yang belum diketahui kategorinya (kelas).

Algoritma ini memanfaatkan teori probabilitas yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. Karena asumsi atribut tidak saling terkait, maka:

$$Vmap = \frac{argmax}{vj \in V} P(Vj) \pi P(Wk|Vj) \quad \text{.....(6)}$$

Setelah diperoleh perhitungan untuk masing-masing kategori, maka kategori yang dipilih adalah yang memiliki nilai Vmap terbesar. Nilai P(Vj) ditentukan pada saat pelatihan, yang nilainya berdasarkan persamaan:

$$P(Vj) = \frac{|\ docs\ j\ |}{|\ contoh\ |} \quad \text{.....(7)}$$

Keterangan :

P(Vj) : probabilitas setiap dokumen terhadap sekumpulan dokumen.

|docs j| : banyaknya dokumen yang memiliki kategori j dalam pelatihan.

|contoh|: banyaknya dokumen dalam contoh yang digunakan saat pelatihan

$$\text{Prior prob} = \frac{\sum sentimen}{\sum kamus} \quad \text{.....(8)}$$

Keterangan :

$\sum sentimen$ = total sentimen pada kamus positif atau negatif

$\sum kamus$ = total kamus positif dan negatif yang terbentuk

Untuk membobot tiap kata dengan pendekatan kamus yaitu titik pembobotan dengan nilai lebih terdapat pada kata yang telah terhimpun pada data latih, berikut persamaan yang digunakan pada pembobotan tiap kata.

$$\text{bobot} = \frac{1}{\sum kamus + \sum sentimen} \quad \text{.....(9)}$$

Dengan syarat pada sentimen data uji tidak memiliki kamus atau kata yang serupa pada kamus data latih sehingga nilai akan dikembalikan pada term frequency. Untuk pembobotan kata pada data uji yang memiliki kamus kata yang sama pada data latih akan membentuk persamaan seperti berikut.

$$\text{bobot} = \frac{1}{\sum kamus + \sum sentimen}\ x\ 2 \quad \text{.....(10)}$$

kondisi pada persamaan diatas akan membentuk bobot lebih terhadap kata pada data uji yang memiliki bentuk yang sama terhadap kamus data latih. Sehingga akan membentuk persamaan klasifikasi seperti berikut.

$$\text{class} = n1\ x\ n2\ x\ n3\ x\ n4 \ldots\ldots\ldots x\ prior\ prob \quad \text{.....(11)}$$

keterangan :

class : output, dalam kasus ini adalah sentimen yaitu positif dan negatif.

N : kata pada tiap tweet.

Prior prob : prior probabilitas pada output tiap class yang dihasilkan data latih.

2.7. *Confusion Matrix*

Data mining digunakan untuk mengukur kinerja model, ada beberapa cara untuk mengukur kinerja dari model yang dihasilkan salah satunya menggunakan *confusion matriks* (akurasi). *Confusion Matrix* adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep *data mining*. Presisi atau *confidence* adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. *Recall* atau *sensitivity* adalah proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar.

Tabel 2.1 Tabel *Confusion Matrix*

Aktual	*Classifield* as	
	+	-
+	TP (*True positives)*	FN (*False negative*)
-	FP (*False positives*)	TN (*True negative*)

Perhitungan akurasi dengan tabel *Confusion Matrix* adalah sebagai berikut:

Akurasi = $\frac{TP+TN}{TP+FP+TN+FN}$ .....(12)

Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. Presisi dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. Rumus presisi adalah:

$Precison = \frac{TP}{TP+FP}$ .....(13)

Presisi dapat diberi nilai dalam bentuk angka dengan menggunakan perhitungan persentase (1-100%) atau dengan menggunakan bilangan antara 0-1. Sistem rekomendasi akan dianggap baik jika nilai presisi tinggi.

2.8. Penelitian Terdahulu

Penelitian terdahulu digunakan sebagai bahan pertimbangan dalam penelitian yang sedang dilakukan. Berikut penelitian yang di gunakan:

1. Berdasarkan penelitian yang dilakukan oleh Parveen dan Pandey dalam judulnya "*Sentiment Analysis on Twitter Data-set using Naïve Bayes Algorithm*", Pada analisis ini, penelitian dilakukan terhadap opini, *feedback* dam *review* pada film melalui *tweet* masyarakat. Dengan mengklasifikasikan menjadi positif, netral dan negatif.
2. Berdasarkan penelitian yang dilakukan oleh Falahah dan Nur (2015) dalam judulnya "Pengembangan Aplikasi Sentimen Analisis Menggunakan Metode *Naïve Bayes* (Studi Kasus Sentimen Analisis dari Media *Twitter*)", Penelitian ini mengklasifikasikan opini publik *Twitter* terkait layanan pemerintah terhadap masyarakat, berdasarkan sentimen posistif, negatif atau netral. Metode *Naïve Bayes Classifier* dapat diterapkan sebagai metode untuk melakukan klasifikasi sentimen analisis.
3. Berdasarkan penelitian yang dilakukan oleh Lestari, Perdana, dan Fauzi (2017) dalam judulnya "Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen *Twitter* Berbahasa Indonesia Menggunakan *Näive Bayes* dan Pembobotan Emoji", Penelitian ini mengklasifikasikan Opini Pilkada DKI 2017 berdasarkan sentimen positif dan negatif dengan pembobotan

emoji. Dari hasil penelitian disimpulkan bahwa pembobotan non-tekstual sangat mempengaruhi hasil dari klasifikasi sentimen.

2.9. Penelitian yang diajukan

Berdasarkan penelitian-penelitian terdahulu sebagai referensi serta landasan untuk menyusun penelitian ini, penulis menyusun penelitian yang beda dari penelitian-penelitian terdahulu. Penyusun melakukan pendekatan melalui kamus, penulis melakukan pembobotan menggunakan *term frequency* (TF) dan klasifikasi menggunakan *Naive Bayes* serta pembagian data menjadi empat menurut calon presiden dan wakil presiden. Hal ini akan berpengaruh pada pemodelan yang nantinya akan menghasilkan empat model.

BAB III METODOLOGI PENELITIAN

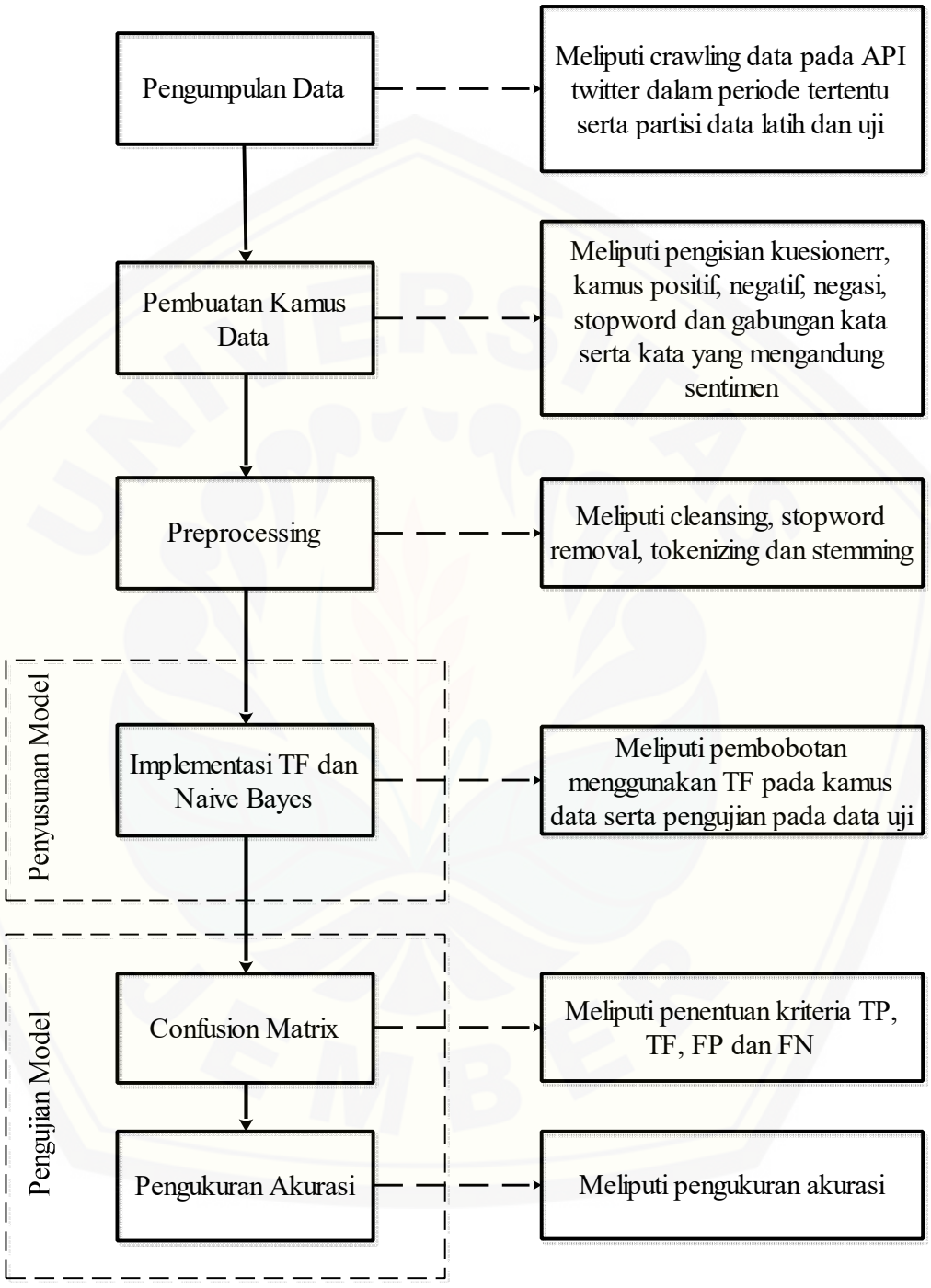
Bab ini menjelaskan tentang jenis penelitian, tempat dan waktu penelitian, serta tahap penelitian yang digunakan dalam menganalisa sentiment pada Twitter terhadap hasil debat calon presiden menggunakan metode *Naive Bayes* .

3.1 Jenis Penelitian

Penelitian ini adalah fokus terhadap menganalisis sentimen masyarakat terhadap sebuah topik, dimana sentimen masyarakat yang akan diteliti adalah opini masyarakat pada linimasa Twitter. Jenis penelitian yang dilakukan merupakan penelitian kuantitatif, mengingat kebutuhan akan banyaknya data dikarenakan penelitian ini menggunakan pendekatan pada kamus yang dipakai masyrakat dalam mengutarakan sentimennya. Penelitian kuantitatif dilakukan pada tahap penghitungan dan pemrosesan data berupa angka, hasil *tweet* berupa kalimat akan melalui berbagai macam proses dan pembobotan sehingga setiap kata akan mempunyai nilai. Perhitungan tersebut dilakukan menggunakan term *frequency* serta pengklasifikasian menggunakan metode Naïve Bayes.

3.2 Tahapan Penelitian

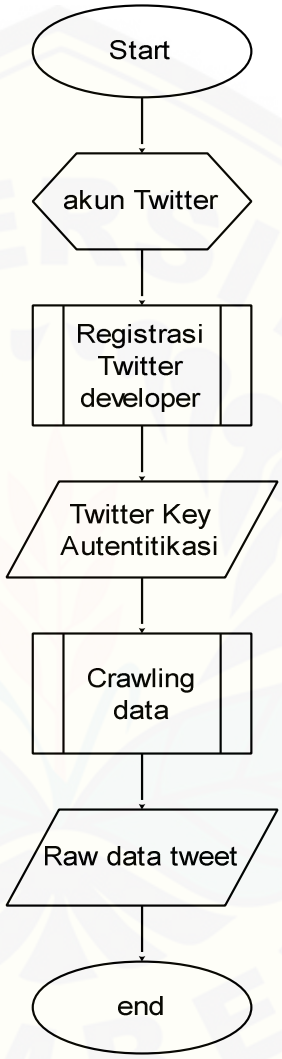
Tahapan penelitian untuk mengetahui klasifikasi sentimen masyarakat pada Twitter terhadap hasil debat calon presiden 2019 dilakukan dalam beberapa tahapan. Gambar proses dari tahapan penelitian dapat dilihat pada Gambar 3.1



Gambar 3.1. Tahapan Penelitian

3.3. Pengumpulan Data

Pengumpulan data didapat dari hasil *crawling* opini masyarakat terhadap debat Pilpres dengan kata kunci #DebatPilpres2019 pada API Twitter. Berikut alur flowchart pengambilan data.



Gambar 3.2 Alur pengumpulan data

Berdasarkan gambar 3.2 dapat dijelaskan bahwa langkah awal untuk mendapatkan data *tweet* adalah dengan melakukan registrasi pada laman Twitter developer yaitu https://apps.Twitter.com/. Twitter akan melakukan verifikasi tujuan dan data kita dan memberikan key autentikasi untuk melakukan *crawling* data *tweet*. Peneliti juga membuat program pada Rstudio yang berbasis bahasa R untuk pengambilan data *tweet*. Menggunakan key autentikasi yang diberikan

Twitter dan topik #DebatPilpres2019 peneliti melakukan *crawling* data. Data hasil *crawling* masih dalam berupa raw data berbentuk csv. Selanjutnya data hasil akan dipecah sesuai kebutuhan penelitian, dalam kasus ini berdasar tujuan sentimen, data latih dan uji.

a. Partisi data berdasar calon

Hasil *crawling* data dari Rstudio yang berupa raw data berbentuk csv selanjutnya atribut yang tidak dibutuhkan akan dihapus. Pada tahap ini atribut yang dihapus adalah *favorited*, *favoriteCount*, *replyToSN*, *created*, *truncated*, *replyToSID*, *id*, *replyToUID*, *statusSource*, *retweetCount*, *isRetweet*, *retweeted*, *longitude* dan *latitude*. Pembersihan *noise* selanjutnya adalah Re*tweet*, Re*tweet* dihapus karena bersifat sama dengan *tweet* asli serta tidak ada penambahan dan perbedaan kamus data. Proses pembersihan atau penghapusan Re*tweet* dilakukan secara manual pada excel dengan cara mencari kata kunci "RT" lalu menghapus semuanya. Contoh penghapusan data Re*tweet* dijelaskan pada tabel 3.1.

Tabel 3.1 Contoh Penghapusan Re*tweet*

No	*Tweet* asli	Re*tweet*
1	Pak @prabowo memang Ok, karna sudah jelas bahwa beliaulah yg membela umat Islam sejak dulu	RT @hartantowelas: Pak @prabowo memang Ok, karna sudah jelas bahwa beliaulah yg membela umat Islam sejak dulu

Tweet dengan awalan "RT" yang artinya adalah Re*tweet* akan dihapus keseluruhannya. Dari total data *crawling* selanjutnya data *tweet* akan dipisahkan berdasarkan *tweet* yang digunakan dan tidak digunakan. Dalam penelitian ini data *retweet* akan dihapus (tidak dipakai) karena bersifat sama seperti *tweet* lain atau memiliki perasaan yang sama tanpa membuat *tweet* baru. Data hasil penghapusan *retweet* akan dikelompokkan berdasarkan arah tujuan *tweet* ke dalam 4 calon presiden wakil presiden yaitu Jokowi, Ma'ruf Amin, Prabowo dan Sandiaga Uno.

b. Penyusunan data latih dan uji

Setelah *tweet* terkelompokkan ke setiap calon, selanjutnya akan disusun menjadi data latih dan data uji. Dalam penelitian ini persentase data latih adalah 80% dari total data dan data uji 20% dari total data.

Tabel 3.2 Partisi Data

Total Data							
Jokowi		Ma'ruf Amin		Prabowo		Sandiaga Uno	
Latih 80%	Uji 20%	Latih 80%	Uji 20%	Latih 80%	Uji 20%	Latih 80%	Uji 20%

3.4. Pembuatan Kamus Data

Pembuatan kamus data terdiri atas kamus positif, negatif, *stopword* dan negasi.

a. Kamus positif dan negatif

Hal pertama yang dilakukan dalam pembentukan kamus positif dan negatif adalah menentukan nilai aktual dari sentimen tiap *tweet*. Tahap ini membutuhkan pengisian kuesioner oleh ahli bahasa. Peneliti menggunakan ahli bahasa dari Dosen Sastra Indonesia Universitas Jember yaitu Dra. A. Erna Rochiyati S. M.Hum. Dengan rujukan data *tweet* yang memiliki arah sentimen, selanjutnya peneliti dapat menganalisis dan menghimpun kata yang dianggap memiliki arah sentimen pada tiap *tweet* dan menjadikannya kamus data. Untuk contoh penerapan dapat dilihat pada Tabel 3.2.

Tabel 3.3 Teknik Pembentukan Kamus Positif dan Negatif

No.	ScreenName	Text	Aktual
1	Maidilsyahputr8	Saya **setuju** dengan pak sandi menghapus program UN @sandiuno#debatcapres2019 #SandiMenangDebat #prabowomenang #IndonesiaMenang	Positif
2	Jaka19_official	Addeuh... @sandiuno jangan bawa hal lain. **Ndak Visioner**. #2019NgayalPresiden	Negatif

Berdasarkan Tabel 3.3 dapat dijelaskan bahwa nilai aktual adalah hasil dari kuesioner oleh ahli bahasa yang dijadikan nilai asli atau aktual. Teknik pembentukan kamus dengan menganalisis satu persatu tiap kata yang merujuk pada hasil sentimen. Pada contoh tabel 3.2, pada dokumen pertama memiliki arah

sentimen positif, hasil analisis menunjukkan kata “setuju” merujuk pada sentimen positif, maka kata “setuju akan dijadikan kamus positif”. Sedangkan pada dokumen nomor dua, gabungan kata “Ndak Visioner” memiliki arti negasi bahwa *tweet* tersebut mengarah pada sentimen negatif. Selanjutnya kata “Ndak Visioner dijadikan kamus negatif.

b. *Stopword*

Stopword adalah kamus data yang berisi kata- kata yang keberadaannya dianggap tidak berpengaruh pada sentimen *tweet*. Dalam penelitian ini penulis memakai dua cara untuk penghimpunan kamus *stopword*. Pertama penulis menggunakan pihak ketiga yang telah menyediakan kamus *stopword* yaitu Sastrawi yang merujuk pada KBBI, artinya kamus dari pihak ketiga ini memiliki tatanan bahasa baku. Kedua, penulis menganalisis tiap *tweet*. Dengan mencari kata-kata yang disingkat oleh peng-*tweet*.

c. Negasi

Pembentukan kamus negasi akan digabung dengan pembentukan kamus yang diambil dari dua kata atau lebih yang memiliki satu arti.

Tabel 3.4 Contoh Kamus Negasi

No.	Kamus	Hasil konversi
1	gak jelas	gakjelas
2	ndak visioner	ndakvisioner
3	ga akan menang	gaakanmenang

Berdasarkan Tabel 3.4 dapat dijelaskan bahwa teknik pembentukan negasi adalah dengan membuat persamaan yaitu, $ingkaran + kata\ bermakna\ positif = kata\ bermakna\ negatif$ dan jika $ingkaran + kata\ bermakna\ negatif = kata\ bermakna\ positif$, hal itu terjadi pada tabel No. satu dan 2. Sedangkan tabel No. 3 adalah gabungan dua kata atau lebih yang memiliki satu arti dan keberadaannya tidak dapat dipisah, jika dipisah maka sentimen tidak akan didapat.

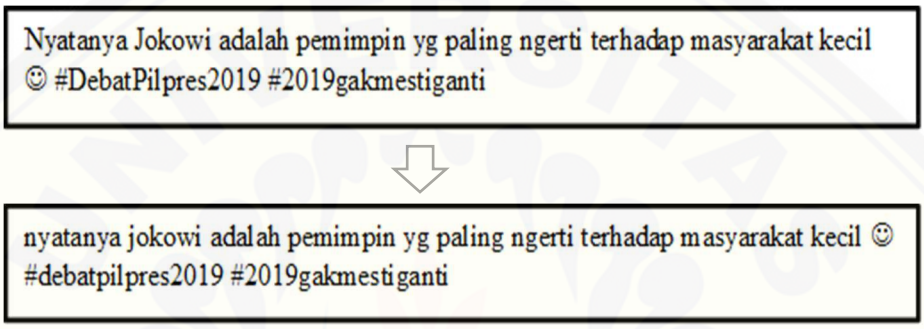
3.5. *Preprocessing*

Preprocessing bertujuan untuk mendapatkan data siap diproses pada pemodelan. Pada proses ini data *tweet*s yang digunakan untuk data training dan

data testing dibersihkan dari *noise* seperti link, "RT", "@", *stopword*, simbol, angka, conversi gambar dan video dan hashtag. Proses *preprocessing* terdiri dari berbagai tahapan yaitu case folding, cleansing, *stopword*, convert emoticon, convert negation, tokenizer dan *stemming*. Berikut adalah penjelasan dari masing-masing tahapan:

a. *Case Folding*

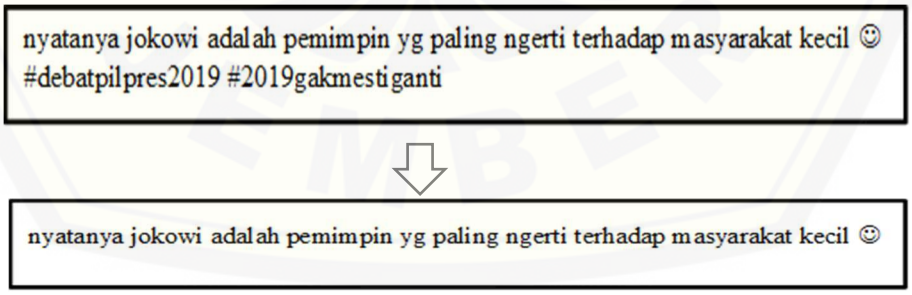
Case Folding ialah proses merubah huruf kapital (*uppercase*) menjadi huruf kecil (l*owercase*). Hal ini dilakukan agar semua huruf menjadi seragam. Berikut ini adalah contoh *case folding*.



Gambar 3.3 Gambaran proses *case folding*

b. *Cleansing*

Tweet yang berhubungan dengan Pilpres memiliki berbagai komponen atau karakterisitik *tweet* yaitu "#", *link* dan RT, angka dan simbol. Komponen-komponen tersebut tidak memiliki pengaruh apapun terhadap sentimen, maka akan dibuang. Berikut ini adalah contoh *cleansing*.



Gambar 3.4 Gambaran *cleansing*

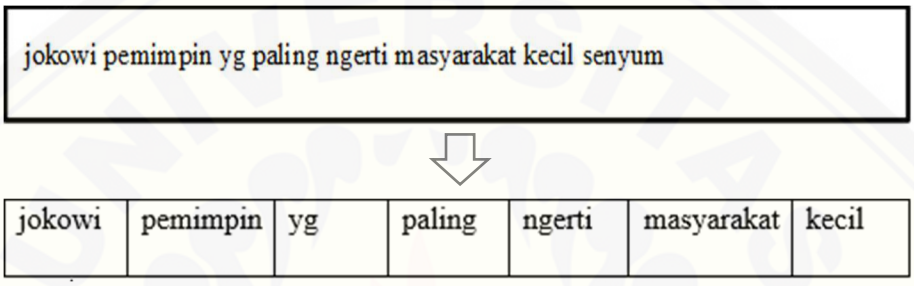
c. *Convert Negation*

Convert Negation merupakan proses konversi kata-kata negasi yang terdapat pada suatu *tweet*, karena kata negasi mempunyai pengaruh dalam merubah nilai

sentimen pada suatu *tweet*. Jika terdapat kata negasi pada suatu *tweet* maka kata tersebut akan disatukan dengan kata setelahnya. Contoh kata-kata negasi tersebut diantaranya "bukan", "bkn", "tidak", "enggak", "g", "ga", "jangan", "nggak", "tak", "tdk", dan "gak".

d. *Tokenizing*

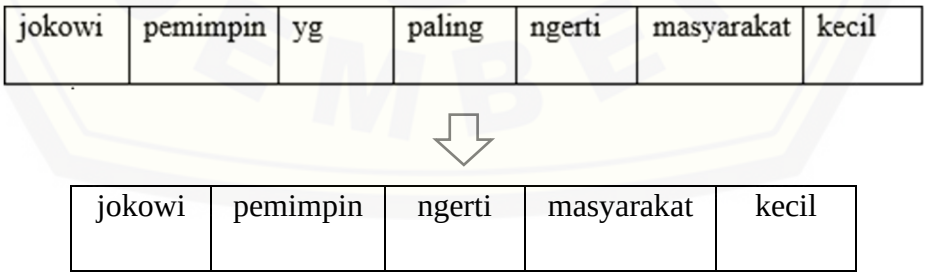
Proses *tokenizing* setiap kata pada *tweet* dipisahkan, pada proses ini tahap yang dilakukan adalah memisahkan setiap kata yang dipisahkan oleh spasi. Hal ini dilakukan agar tahap *preprocessing* selanjutnya dapat berjalan. Berikut ini adalah contoh *tokenizing.*



Gambar 3.5 Gambaran *tokenizing*.

e. *Stopword*

Data *tweet* pada tahap ini masih mengandung kata yang dianggap tidak dapat memberikan pengaruh dalam menentukan suatu kategori sentimen. Kata-kata tersebut dimasukkan ke dalam daftar *stopword* yang biasa berupa kata ganti orang, kata ganti penghubung, paranomial petunjuk dan lain sebagainya. Jika pada *cleansing* masih terdapat kata yang tercantum pada daftar *stopword* maka kata tersebut dihilangkan. Berikut ini adalah contoh *stopword removal.*



Gambar 3.6 Gambaran *stopword*

f. *Stemming*

Stemming adalah proses mengubah kata berimbuhan ke bentuk asalnya (kata dasar). Algoritma yang digunakan untuk proses *stemming* berbahasa Indonesia

adalah Nazief dan Andriani. Proses *stemming* juga termasuk salah satu proses dalam *Information Retreval*. Berikut ini adalah contoh *stemming*.

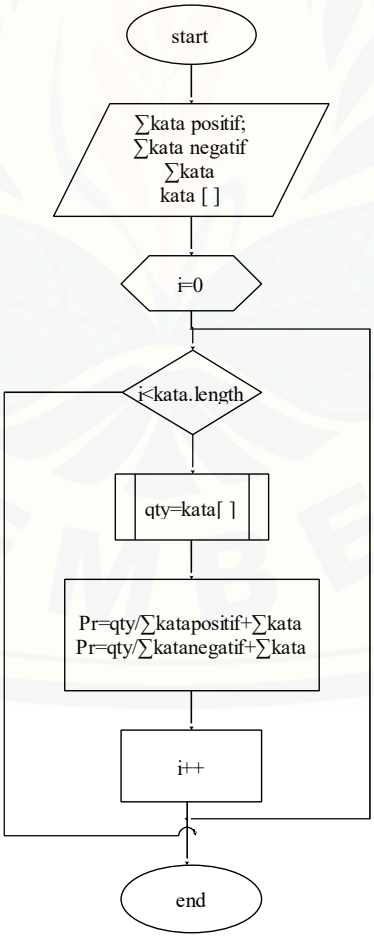
jokowi	pemimpin	ngerti	masyarakat	kecil

jokowi	pimpin	ngerti	masyarakat	kecil

Gambar 3.7 Gambaran *stemming*

3.6. Penyusunan Model

Tahap selanjutnya adalah tahap penyusunan model, dari hasil *preprocessing* akan dibobot menggunakan algoritma TF kemudian diklasifikasikan ke dalam dua class; positif dan negatif menggunakan metode Naïve Bayes Classifier. Model yang disusun berdasarkan 4 sub dataset yaitu Jokowi, Ma'ruf Amin, Prabowo dan Sandiaga Uno. Berikut alur flowchart perhitungan Naive Bayes.



Gambar 3.8 Alur perhitungan Naive Bayes

Langkah-langkah dalam implementasi algoritma TF dan metode *Naive Bayes* Classifier adalah sebagai berikut :

a. Menghitung *prior probability*

Prior probability adalah menghitung probabilitas dari total kamus data. Berikut tergambar pada Tabel 3.5.

Tabel 3.5 Contoh Hasil Hitung *Prior Probability*

	Jumlah kata	Jumlah kata posistif	Jumlah kata negatif	Probabilitas positif	Probabilitas negatif
Kata keseluruhan	26	22	4	0,846153846	0,181818182

Berdasarkan Tabel 3.5 dapat dijelaskan jumlah kata adalah total kamus data positif + kamus data negatif sedangkan untuk menghitung probabilitasnya adalah $\frac{\sum kamus\ sentimen}{\sum kamus\ data}$.

b. Menghitung bobot dan klasifikasi *Naive Bayes*

Tabel 3.6 Potongan hasil hitung *Naive Bayes* pada data latih

	Ngakunya pro rakyat halah				total	Klasifikasi	aktual
	ngakunya	pro	Rakyat	halah			
	1	1	1	1			
Positif	$\frac{1}{22+26}$	$\frac{1}{22+26}$	$\frac{1}{22+26}$	$\frac{1}{22+26}$	0,000000012375	Negatif	negatif
Negatif	$\frac{1}{4+26}$	$\frac{1}{4+26}$	$\frac{1}{4+26}$	$\frac{1}{4+26}$ x 2	0,000000811043		

Berdasarkan Tabel 3.6 dapat dijelaskan bahwa, setiap dokumen yang dipecah dengan *tokenizing* diberi nilai sesuai kemunculan kata dengan

menggunakan TF dengan persamaan $\frac{n}{\sum kamus\ sentimen + \sum total\ kamus}$, hal ini berlaku untuk kata yang tidak ada pada kamus sentimen. Jika kata pada *tweet* sama seperti kata pada kamus sentimen maka persamaan pembobotan menjadi $\frac{n}{\sum kamus\ sentimen + \sum total\ kamus}$ x 2 lalu diproses menggunakan *Naive Bayes*, yaitu dengan mengalikan setiap kata yang telah terbobot dengan *prior probability*. Hasil akhir dari pengkalian akan membentuk dua hasil yaitu total positif dan negatif. Jika total positif lebih besar maka hasil klasifikasi positive dan sebaliknya.

3.7. Pengujian Model

Pengujian model yang dihasilkan menggunakan confusion matrix untuk menentukan kriteria dan pengukuran tingkat akurasi.

3.7.1. *Confusion matrix*

Hasil dari klasifikasi menggunakan Naïve Bayes classifier selanjutnya akan diukur tingkat akurasi dan presisi. Hasil klasifikasi ini dibandingkan dengan hasil klasifikasi manual dari kuesioner. Berdasarkan tabel *confusion matrix*, perhitungan di atas (Tabel 3.6) masuk dalam kriteria *True negative*.

Merujuk pada tabel *confusion matrix* pada bab 2 dapat dijabarkan untuk menentukan kriteria adalah sebagai berikut

a. Jika nilai aktual positive dan klasifikasi positive maka kriteria bernilai *true positive* (TP)
b. Jika nilai aktual positive dan klasifikasi negative maka kriteria bernilai *false negative* (FN).
c. Jika nilai aktual negative dan klasifikasi negative maka kriteria bernilai *true negative* (FN).
d. Jika nilai aktual negative dan klasifikasi positive maka kriteria bernilai *false positive* (FP).

3.7.2. Pengukuran Akurasi

Peneliti menggunakan pengukuran tingkat akurasi pada tiap calon presiden dan wakil presiden untuk mengetahui hasil kualitas model klasifikasi yang dihasilkan, dengan merujuk pada persamaan :

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

Persamaan di atas digunakan untuk mengukur tingkat akurasi semua calon.

BAB V PENUTUP

Bab penutup merupakan bab yang berisi kesimpulan dan saran dari penelitian. Kesimpulan merupakan inti dari suatu penelitian yang telah dilakukan, mulai dari awal hingga akhir. Kemudian dari kesimpulan tersebut dapat diperoleh saran atau masukan untuk penelitian selanjutnya.

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat ditarik beberapa kesimpulan pada penelitian sentimen analisis *twitter* terhadap debat capres-cawapres ini yang terklasifikasi menjadi 4 kontestan yaitu 2 capres dan 2 cawapres, memiliki data latih 370 dengan yaitu:

1. Dari implementasi metode Naive Bayes dalam mengklasifikasikan data Twitter terhadap hasil debat calon presiden dan wakil presiden 2019 menghasilkan 4 model yaitu:
    a. Pada data Tweet sentimen terhadap jokowi menghasilkan model prior probability positif sebesar 0.696969697 dan prior probability negatif sebesar 0.303030303.
    b. Pada data Tweet sentimen terhadap Ma'ruf Amin menghasilkan model prior probability positif sebesar 0.890243902 dan prior probability negatif sebesar 0.109756098.
    c. Pada data Tweet sentimen terhadap Prabowo menghasilkan model prior probability positif sebesar 0.4625 dan prior probability negatif sebesar 0.5375.
    d. Pada data Tweet sentimen terhadap Sandiaga Uno menghasilkan model prior probability positif sebesar 0.585714286 dan prior probability negatif sebesar 0.414285714.

2. Dari hasil pengukuran tingkat akurasi metode Naive Bayes dalam mengklasifikasikan sentimen data Twitter terhadap calon presiden dan wakil presiden 2019 dengan pembagian data 20% uji dan 80% latih yaitu:
    a. Pada sentimen tweet terhadap Jokowi dengan data latih 68 tweet dan data uji 17 tweet, warganet memiliki sentimen positif terhadap Jokowi sebesar 8.33% dan sentimen negatif 91.67% dengan tingkat akurasi klasifikasi sebesar 70.588% .
    b. Pada sentimen tweet terhadap Ma'ruf Amin dengan data latih 92 tweet dan data uji 23 tweet, warganet memiliki sentimen positif terhadap Ma'ruf Amin sebesar 25% dan sentimen negatif 75% dengan tingkat akurasi klasifikasi sebesar 17.391% .
    c. Pada sentimen tweet terhadap Prabowo dengan data latih 68 tweet dan data uji 17 tweet, warganet memiliki sentimen positif terhadap Prabowo sebesar 75% dan sentimen negatif 25% dengan tingkat akurasi klasifikasi sebesar 35.294%.
    d. Pada sentimen tweet terhadap Sandiaga Uno dengan data latih 68 tweet dan data uji 17 tweet, warganet memiliki sentimen positif terhadap Sandiaga Uno sebesar 0% dan sentimen negatif 100% dengan tingkat akurasi klasifikasi sebesar 35.294%.

5.2 Saran

Penulis menyadari bahwa penelitian ini jauh dari kesempurnaan, maka dari itu penulis membuka diri untuk dilakukannya pengembangan pada penelitian ini. beberapa aspek dalam penelitian ini yang dapat dikembangkan antara lain:

1. Pengembang dapat menambahkan data baru agar menghasilkan daftar kamus yang lebih kaya.
2. Pengembang dapat menambahkan metode skenario uji agar mendapatkan titik lipat data model yang terbaik.
3. Pengembang dapat menggunakan metode ini untuk topik politik lain.
4. Pengembang dapat menggunakan metode klasifikasi yang lain untuk membandingkan dan menemukan metode paling tepat.

DAFTAR PUSTAKA

Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. INTEGER: Journal of Information Technology, 2(1).

C. Aggarwal, C. 2015. *Data Classification: Algorithms and Applications. Berilustrasi peyunt.* New York: CRC Press.

Dehhaf (2010) *Sentiment Analysis, Available at*: http://customerthink.com/sentiment_analysis_hard_but_worth_it/ (diakses pada 12 Oktober 2019)

Grossman, D., dan Ophir, F. 1998. *Information Retrieval: Algorithm and Heuristics*. Kluwer Academic Publisher.

Hadiyani. Eka Pratiwi. 2013. "Rancang Bangun Sistem Pendukung Keputusan Untuk Pemilihan Anggota Terbaik AIESEC Surabaya Dengan Menggunakan Metode Naive Bayes". Fakultas Sains dan Teknologi. Universitas Airlangga.

Huma Parveen, Prof. Shikha Panddey. 2016. *Sentiment Analysist on TwitterDat-set using NaiveBayes Algorithm*. Rungta Collegeof Engineering and Technology India.

Jurdi, Fajlurrahman. 2018. *Pengantar Hukum Pemilihan Umum*. Jakarta: KENCANA.

Larose, Daniel T. 2005. *Discovering Knowledge in Data : An Introduction to Data Mining*. John Willey & Sons, Inc.

Liu, B. 2012. *Sentiment analysis and opinion mining*. Synthesis lectures on human language technologies, 5(1), 1-167.

Manalu B. 2014. Analisis Sentimen Pada Twitter Menggunakan Text Mining. Prodi Teknologi Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi. Universitas Sumatera Utara, Medan.

Natalius, Samuel. 2010. Metoda Naïve Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen. Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung. Bandung.

P. Beineke, T. Hastie and S. Vaithyanathan, "*The sentimental factor: Improving review classification via human-provided information", Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 263, 2003.

Robertson, Stephen. 2005. *Understanding Inverse Document Frequency: On theoretical arguments for IDF*. Journal of Documentation, Vol. 60, pp. 502–520

Saraswati, N. W. S. 2011. *Text mining dengan metode naive bayes classifier dan support vector machines untuk sentiment analysis*. Universitas Udayana Denpasar.

Tan P., Steinbach M., dan Kumar V. 2006. *Introduction to Data Mining. Pearson Education*. Boston.

Turban, E. 2005. Decision Support Systems and Intelligent Systems Edisi Bahasa Indonesia Jilid 1. Andi: Yogyakarta.

Wikipedia. *Twitter*. (*online*), (https://id.wikipedia.org/wiki, diakses 7 Februari 2019).