SKRIPSI

Analisis Sentimen Masyarakat Indonesia Terhadap Kebijakan Pemerintah Dalam Menangani Pandemi Covid-19 Menggunakan Klasifikasi *Random Forest* pada Media Sosial Twitter

Disusun dan diajukan oleh:

MUH. FATHIN ABDILLAH HALIK

D42116017

DEPARTEMEN TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS HASANUDDIN

MAKASSAR

2023

LEMBAR PENGESAHAN SKRIPSI

ANALISIS SENTIMEN MASYARAKAT INDONESIA TERHADAP KEBIJAKAN PEMERINTAH DALAM MENANGANI PANDEMI COVID-19 MENGGUNAKAN KLASIFIKASI RANDOM FOREST PADA MEDIA SOSIAL TWITTER

Disusun dan diajukan oleh

MUH FATHIN ABDILLAH HALIK

D42116017

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika Fakultas Teknik Universitas Hasanuddin pada tanggal 8 Maret 2023 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,

Dr. Amil Ahmad Ilham, S.T., MIT.
Nip. 197310101998021001

Pembimbing Pendamping,

Dr. Eng. Ir. Zulkifli Tahir, ST.,M.Sc.
Nip. 198404032010121004

Ketua Program Studi,

Prof. Dr. Ir. Indrabayu, S.T., M.T., M.Bus.Sys., IPM, ASEAN. Eng
Nip. 19750716 200212 1 004

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama : Muh. Fathin Abdillah Halik

NIM : D42116017

Departemen : Teknik Informatika

Jenjang : S1

Menyatakan dengan ini karya tulisan saya berjudul:

ANALISIS SENTIMEN MASYARAKAT INDONESIA TERHADAP KEBIJAKAN PEMERINTAH DALAM MENANGANI PANDEMI COVID-19 MENGGUNAKAN KLASIFIKASI *RANDOM FOREST* PADA MEDIA SOSIAL TWITTER

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilalihan tulisan orang lain bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Maros, 08 Maret 2023

Yang menyatakan,

Muh. Fathin Abdillah Halik

KATA PENGANTAR

Alhamdulillah, Puji dan Syukur kita panjatkan kepada Allah Subhanahu Wata'ala. Dzat yang hanya kepada-Nya memohon pertolongan. Alhamdulillah atas segala pertolongan, rahmat, dan kasih sayang-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "**Analisis Sentimen Masyarakat Indonesia Terhadap Kebijakan Pemerintah Dalam Menangani Pandemi Covid-19 Menggunakan Klasifikasi *Random Forest* pada Media Sosial Twitter**". Shalawat dan salam kepada Rasulullah Shallallahu Alaihi Wasallam yang senantiasa menjadi sumber inspirasi dan teladan terbaik untuk umat manusia. Penyusunan skripsi ini sebagai salah satu syarat untuk menyelesaikan Program Sarjana (S1) pada Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin.

Penulis menyadari bahwa penulisan ini tidak dapat terselesaikan tanpa dukungan dari berbagai pihak baik morel maupun materil. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih kepada semua pihak yang telah membantu dalam penyusunan skripsi ini terutama kepada:

1. Kedua orang tua, yang telah memberikan dukungan dan motivasi serta doa yang tiada henti-hentinya kepada penulis.
2. Pak Dr. Amil Ahmad Ilham, S.T., M.IT., selaku dosen pembimbing I dan Pak Dr-Eng. Zulkifli Tahir, S.T., M.Sc., selaku dosen pembimbing II, yang selalu menyediakan waktu, tenaga dan pikirannya yang luar biasa untuk mengarahkan penulis dalam penyusunan Tugas Akhir ini.

3. Segenap Dosen dan Staff Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin, yang telah banyak membantu penulis selama masa perkuliahan.
4. Saudara-saudari IGNITER 16, terutama kepada Ayu Lestari Ramadani S.T. yang selalu membersamai, menyemangati dan membantu penyelesaian skripsi ini serta mengisi hari-hari menjadi sangat menyenangkan.
5. Semua pihak atas dukungan dan bantuannya yang tidak dapat penulis tuliskan satu persatu.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna dikarenakan terbatasnya pengalaman dan pengetahuan yang dimiliki penulis. Oleh karena itu, penulis mengharapkan segala bentuk saran serta masukan bahkan kritik yang membangun dari berbagai pihak. Semoga skripsi ini dapat bermanfaat bagi para pembaca dan semua pihak.

Makassar, Januari 2023

Penulis

ABSTRAK

Sejak adanya kasus pertama dengan dua orang dinyatakan positif di Indonesia, topik covid-19 selalu menjadi pembicaraan dalam berbagai media dan tentu saja media sosial. Joko Widodo sebagai Presiden Republik Indonesia pastinya menjadi pusat perhatian masyarakat terutama tentang kebijakan-kebijakan yang diterapkan dalam upaya penanganan covid-19 di Indonesia. Salah satu media sosial yang dapat digunakan dengan bebas oleh masyarakat untuk menuangkan pendapatnya adalah Twitter. Sebagai salah satu raksasa sosial media, rata-rata 500 juta *tweet* diunggah ke twitter setiap harinya. Berdasarkan *tweet* dari masyarakat tersebut, pemerintah dapat menganalisis sentimen masyarakat terhadap kebijakan pemerintah dalam menangani covid-19 sehingga kedepannya pemerintah dapat memberikan kebijakan yang efektif. Maka dari itu, penelitian ini melakukan percobaan analisa terhadap sentimen masyarakat untuk melihat kecenderungan sentimen masyarakat terhadap kebijakan pemerintah dalam menangani covid-19 pada media sosial twitter serta melihat pengaruh ekstraksi fitur terhadap akurasi analisis sentimen. Pada penelitian ini menggunakan 1500 data yang diberikan label dengan tiga kelas yaitu positif, negative, dan netral secara manual. Algoritma yang digunakan untuk pembuatan model analisis sentimen yaitu *Random Forest*. Untuk proses pengamatan ekstraksi fitur terdapat dua metode yang dibandingkan yaitu TF-IDF dan *Count Vectorizer*. Dari kelima percobaan yang telah dilakukan dapat dilihat bahwa nilai akurasi tertinggi didapatkan pada rasio 35% *data testing* dan 65% *data training* menggunakan ekstraksi fitur TF-IDF dengan *range* n-gram (1,1). Kemudian nilai ROC AUC *score* tertinggi didapatkan pada rasio 20% *data testing* dan 80% *data training* menggunakan ekstraksi fitur TF-IDF dengan *range* n-gram (1,2).

Kata Kunci : *Random Forest*, Twitter, Covid-19, TF-IDF, *Count Vectorizer*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang

Virus corona merupakan penyakit menular yang menyerang organ pernapasan manusia. Virus ini bermula dari Negara China, lebih tepatnya di Kota Wuhan, Provinsi Hubei. Pada awal bulan maret tahun 2020, virus ini pertama kali di deteksi masuk ke Indonesia (Firmansyah & Puspitasari, 2021). Sejak adanya kasus pertama dengan dua orang dinyatakan positif di Indonesia, topik covid-19 ini selalu menjadi pembicaraan dalam berbagai media berita dan tentu saja media sosial. Joko Widodo sebagai Presiden Republik Indonesia pastinya menjadi pusat perhatian masyarakat terutama tentang kebijakan-kebijakan yang diterapkan dalam upaya penanganan covid-19 di Indonesia.

Salah satu media sosial yang dapat digunakan dengan bebas oleh masyarakat untuk menuangkan pendapatnya adalah Twitter. Sebagai salah satu raksasa sosial media, rata-rata 500 juta *tweet* diunggah ke Twitter setiap harinya (*Safina*, 2020). Dalam beberapa tahun terakhir, Twitter memberikan banyak pengaruh dalam menghasilkan sumber informasi. Dalam *data mining,* banyak hal menarik yang dapat digali pada Twitter seperti bagaimana opini yang terdapat pada masyarakat tentang suatu kebijakan pemerintah (Hikmawan et al., 2020). Seperti halnya dengan kebijakan pemerintah Indonesia yang telah dikeluarkan dalam upaya penanganan rantai penyebaran

covid-19. Hal ini menjadi sangat penting karena dapat menjadi bahan pertimbangan oleh pemerintah dalam menanggapi sikap publik.

Percakapan yang diunggah di Twitter dapat diklasifikasikan berdasarkan sentimennya. Melalui analisis sentimen, data dapat dikelompokkan menjadi sentimen positif, negatif, dan netral (*Safina*, 2020). Analisis sentimen bertujuan untuk menganalisis, mengolah, mengekstraksi data tekstual yang berupa tanggapan terhadap suatu entitas terhadap suatu topik guna memperoleh informasi (Yulianita et al., 2020). Analisis sentimen dilakukan untuk melihat kecenderungan tanggapan terhadap sebuah masalah apakah cenderung bertanggapan negatif, positif, ataupun netral dengan menggunakan berbagai pendekatan. Analisis sentimen mengacu pada bidang yang sangat luas mulai dari peneraan Bahasa alami, komputasi linguistic dan *text mining* yang bertujuan untuk menganalisa sikap, evaluasi, sentimen, pendapat, penilaian serta emosi seseorang yang berkaitan dengan suatu produk, layanan, topik, individu, maupun kegiatan tertentu (Tokoh & Peserta, 2020). Sehingga *tweet* yang dilakukan oleh masyarakat merupakan sumber data yang valid untuk melakukan analisis sentimen.

Random Forest (RF) merupakan suatu algoritma yang digunakan pada klasifikasi data, umumnya dengan jumlah yang besar. Klasifikasi *random forest* dilakukan melalui penggabungan pohon (*tree*) dengan melakukan *training* pada sampel data yang dimiliki. *Random Forest* merupakan algoritma yang cocok digunakan untuk masalah klasifikasi pada *machine learning* dan *data mining* (Hanun et al., 2020). Maka dari itu, penelitian ini

mencoba melakukan analisa terhadap sentimen menggunakan algoritma klasifikasi *random forest* untuk melihat kecenderungan persepsi masyarakat terhadap kebijakan pemerintah dalam menangani covid-19 pada media sosial Twitter.

1.2 Rumusan Masalah

Berdasarkan latar belakang, maka rumusan masalah adalah sebagai berikut:

1. Bagaimana menggunakan algoritma *Machine Learning Random Forest* pada sentimen masyarakat terhadap kebijakan pemerintah dalam menangani covid-19 pada Twitter?
2. Bagaimana pengaruh ekstraksi fitur terhadap akurasi analisis sentimen?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah pada sub-bab sebelumnya maka dapat dibuat tujuan sebegai berikut:

1. Untuk menganalisis sentimen masyarakat terhadap kebijakan pemerintah dalam menangani covid-19 pada twitter menggunakan algoritma *Machine Learning Random Forest*.
2. Untuk mengetahui pengaruh ekstraksi fitur terhadap akurasi analisis sentiment.

1.4 Manfaat Penelitian

Dengan dilakukannya penelitian ini, manfaat yang diharapkan dari penelitan ini adalah:

1. Membantu pemerintah Indonesia dalam mengevaluasi kebijakan-kebijakan yang diberlakukan dalam menangani covid-19.
2. Mengetahui pengaruh feature extraction dalam keberhasilan sistem klasifikasi sentimen.

1.5 Batasan Masalah

Yang menjadi batasan masalah dari penelitian ini adalah asebagai berikut:

1. Data yang digunakan merupakan data dari *Twitter* berbahasa Indonesia.
2. Ekstraksi fitur yang digunakan yaitu TF-IDF dan *Count Vectorizer*
3. Proses *modelling* untuk uji akurasi sentimen menggunakan algoritma *Machine Learning Random Forest*.

1.6 Sistematika Penulisan

Untuk memberikan gambaran umum mengenai isi tulisan secara keseluruhan, berikut diuraikan sistematika penulisan dari laporan tugas akhir ini:

BAB I PENDAHULUAN

Bab ini berisi penjelasan tentang latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini berisi teori-teori terkait hal-hal yang mendasari dan yang berhubungan dengan penelitian yang dilakukan termasuk di dalamnya

membahas mengenai twitter, konsep dasar Natural Language Processing, serta metode-metode yang digunakan dalam penelitian.

BAB III METODOLOGI PENELITIAN

Bab ini berisi tentang langkah-langkah apa saja yang dilakukan pada saat penelitian. Mulai dari studi literatur, persiapan kebutuhan perancangan sistem, perancangan dan pembuatan sistem.

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini berisi tentang hasil penelitian dan pembahasan terkait sistem yang telah dibuat.

BAB V PENUTUP

Pada bab ini berisi mengenai kesimpulan yang diperoleh dari hasil penelitian yang dilakukan serta saran-saran untuk pengembangan penelitian yang lebih lanjut.

BAB II

TINJAUAN PUSTAKA

2.1 Twitter

Twitter adalah layanan social media yang memungkinkan penggunanya untuk mengirim pesan yang biasa disebut dengan *tweet*. Pesan yang dikirimkan dapat berupa teks, gambar, video, dan audio. *Twitter* pertama kali didirikan pada tahun 2006 oleh Jack Dorsey .

Twitter menyediakan akses data kepada perusahaan, pengembang, dan penggunanya melalui API (*Application User Interface*). Salah satu API yang disediakan oleh *twitter* adalah *search* API. *Search* API memungkinkan perusahaan, pengembang, atau pengguna untuk memperoleh *tweet* yang telah diunggah ke twitter (Twitter, 2019).

2.2 Covid-19

Coronavirus Disease 2019 (COVID-19) adalah penyakit jenis baru yang belum pernah diidentifikasi sebelumnya pada manusia. Virus penyebab COVID-19 ini dinamakan Sars-CoV-2. Virus corona adalah zoonosis (ditularkan antara hewan dan manusia). Adapun, hewan yang menjadi sumber penularan COVID-19 ini masih belum diketahui. Berdasarkan bukti ilmiah, COVID-19 dapat menular dari manusia ke manusia melalui percikan batuk/bersin (droplet), Orang yang paling berisiko tertular penyakit ini adalah orang yang kontak erat dengan pasien COVID-19 termasuk yang merawat pasien COVID-19 (Putri, 2020). Tanda dan gejala umum infeksi covid-19 termasuk gejala gangguan pernapasan akut seperti demam, batuk, dan sesak

napas. Masa inkubasi rata-rata adalah 5 - 6 hari dengan masa inkubasi demam, batuk, dan sesak napas. Pada kasus yang parah, covid-19 dapat menyebabkan pneumonia, sindrom pernapasan akut, gagal ginjal, dan bahkan kematian (Putri, 2020).

2.3 *Natural Languange Processing*

Natural Language Processing (NLP) adalah area integral dari ilmu komputer antara pembelajaran mesin dan linguistik yang mana komputasi digunakan secara luas (Patel & Prajapati, 2018), yang ditujukan untuk membuat komputer mengerti pernyataan yang ditulis menggunakan bahasa manusia. Pengolahan bahasa alami timbul untuk meringankan pekerjaan *user* dan untuk memenuhi keinginan terhubung dengan komputer dalam bahasa murni. Karena semua pengguna mungkin tidak fasih dalam bahasa khusus mesin, NLP melayani pengguna yang tidak memiliki cukup waktu untuk mempelajari bahasa baru atau mendapatkan kesempurnaan di dalamnya (Khurana et al., 2017).

Natural Language Processing berdasarkan basisnya bisa dikelompokkan dalam dua komponen ialah *Natural Language Understanding* dan *Natural Language Generation* yang mengembangkan tugas guna memahami dan menghasilkan teks. NLP dapat dimanfaatkan untuk melakukan tugas-tugas seperti peringkasan otomatis, penerjemahan bahasa, analisis sentimen, pengenalan ucapan, dan segmentasi topik. Namun proses kerja NLP tidaklah mudah, terdapat beberapa kendala terkait bahasa alami dalam berkomunikasi dengan komputer. Ambiguitas adalah salah satu

kendala utama bahasa alami yang biasanya dihadapi di tingkat sintaksis yang memiliki sub-tugas sebagai leksikal dan morfologi yang berkaitan dengan studi kata dan pembentukan kata. Masing-masing level ini menyebabkan ambiguitas yang dapat diselesaikan dengan mengetahui kalimat secara utuh dan lengkap. Selain itu, terdapat kendala lain yaitu *non-standard word* yang terkait dengan variasi kata dalam suatu kalimat. Bentuk kalimat yang tidak terstruktur juga dapat menyulitkan dalam fase *preprocessing* teks, misalnya tidak menggunakan tata bahasa yang tepat, mengandung banyak singkatan, kesalahan pengetikan atau pengejaan, dan lain sebagainya. Karena itu, normalisasi dibutuhkan untuk membuat kalimat yang tidak terstruktur bisa dipahami oleh mesin komputer.

2.4 Analisis Sentimen

Analisis Sentimen adalah bidang ilmu yang mempelajari tentang opini public terhadap suatu topik, baik berupa produk, individu, masalah, dan sebagainya. Analisis sentiment merupakan salah satu aplikasi dari *Natural Language Processing*, yang digunakan untuk mengidentifikasi dan mengklasifikasi pendapat dari suatu sumber. Secara garis besar, analisis sentimen bertujuan untuk menentukan sikap dari penulis terhadap suatu topik atau polaritas kontekstual keseluruhan dokumen. Sikap dari penulis dapat berupa penilaian, keadaan emosi penulis saat menulis, atau emosi yang ingin disampaikan penulis kepada pembaca.

Tujuan analisis sentimen adalah untuk mengidentifikasi dan mengklasifikasi apakah suatu dokumen atau kalimat mengungkapkan

pendapat dan apakah pendapat tersebut bersifat positif, negatif, atau netral. Dengan kata lain, tujuan analisis sentimen yaitu mengidentifikasi apakah suatu teks mengekspresikan pendapat atau tidak, dan menentukan orientasi teks yang mengekspresikan pendapat.

2.5 *Random Forest*

Random Forest adalah metode klasifikasi terbaru berbasis komputasi yang diperkenalkan oleh Breiman tahun 2001. Random Forest juga merupakan pengembangan dari metode CART sehingga mempunyai beberapa kelebihan yang tidak dimiliki metode CART.

Random forest adalah suatu metode klasifikasi yang terdiri dari gabungan pohon klasifikasi (CART) yang saling independen. Prediksi klasifikasi diperoleh melalui proses *voting* (jumlah terbanyak) dari pohon-pohon klasifikasi yang terbentuk. *Random forests* merupakan pengembangan dari metode *ensemble* yang pertama kali dikembangkan oleh Leo Breiman (2001) dan digunakan untuk meningkatkan ketepatan klasifikasi. Bila dalam proses *bagging* digunakan *resampling bootstrap* untuk membangkitkan pohon klasifikasi dengan banyak versi yang kemudian mengkombinasikannya untuk memperoleh prediksi akhir, maka dalam *random forests* proses pengacakan untuk membentuk pohon klasifikasi tidak hanya dilakukan untuk data sampel saja melainkan juga pada pengambilan variabel prediktor. Sehingga, proses ini akan menghasilkan kumpulan pohon klasifikasi dengan ukuran dan bentuk yang berbeda-beda. Hasil yang diharapkan adalah suatu kumpulan pohon klasifikasi yang memiliki korelasi

kecil antar pohon. Korelasi yang kecil akan menurunkan hasil kesalahan prediksi *Random Forests* (Fachruddin, 2015).

Peneliti menggunakan algoritma klasifikasi *random forest* karena berdasarkan beberapa penelitian terkait yang membandingkan beberapa algoritma klasifikasi lain dengan algoritma *random forest* mendapatkan hasil bahwa nilai algoritma *random forest* lebih baik. Penelitian terkait dapat dilihat sebagai berikut.

Perbandingan Algoritma Random Forest, Naïve Bayes, dan Support Vector Machine Pada Analisis Sentimen Twitter Mengenai Opini Masyarakat Terhadap Penghapusan Tenaga Honorer, (Akhmad Miftahusalam, 2022). Dengan hasil akurasi pada *random forest* = 62%, SVM = 58%, dan naïve bayes = 61%.

Perbandingan Metode Klasifikasi Random Forest dan SVM pada Analisis Sentimen, (M. R. Adrian dkk, 2021). Dengan hasil akurasi pada *random forest* = 57% dan SVM = 55%.

Perbandingan Algoritma Naïve Bayes, K-Nearest Neighbors dan Random Forest untuk Klasifikasi Sentimen Terhadap BPJS Kesehatan pada Media Twitter, (Tamrizal A.M dkk, 2022). Dengan hasil dari tiga kali percobaan yang dilakukan algoritma *random forest* menunjukkan tingkat akurasi terbaik pada percobaan kedua sebesar 68% dengan *data training* 75% dari *dataset* dan percobaan ketiga sebesar 87% dengan *data training* 90% dari *dataset*.

2.6 Pembobotan Kata TF-IDF

Pembobotan kata (*term weighting*) adalah proses pemberian bobot *term* pada dokumen. Pembobotan dasar dilakukan dengan menghitung frekuensi kemunculan *term* dalam dokumen. Frekuensi kemunculan (*term frequency*) merupakan gambaran sejauh mana *term* tersebut mewakili isi dokumen. Semakin besar kemunculan suatu *term* dalam dokumen akan memberikan nilai kesesuian yang semakin besar. Pembobotan dapat mencerminkan betapa pentingnya sebuah kata dalam dokumen. Pembobotan ini nantinya digunakan oleh algoritma *machine learning* untuk klasifikasi dokumen .

1. *Term Frequency–Inverse Document Frequency* (TF-IDF) adalah bobot *term* terhadap dokumen. TF-IDF dirumuskan sebagai berikut.

$$w_{ij} = tf_{ij} \times idf_j \qquad (1)$$

2. *Term Frequency* (TF) adalah jumlah kemunculan *term* dalam dokumen. Pada *Term Frequency* (TF), terdapat beberapa jenis formula yang dapat digunakan (Febrianti, 2020)
    a. TF biner (*binary* TF): hanya memperhatikan jumlah kemunculan suatu *term* ada atau tidak dalam dokumen, jika ada diberi nilai satu (1), jika tidak ada diberi nilai nol (0).
    b. TF murni (*raw* TF): nilai TF dihitung berdasarkan jumlah kemunculan suatu *term* di dokumen, jika *term* tersebut muncul sebanyak tiga kali dalam dokumen maka TF bernilai tiga.

c. TF logaritmik: perhitungan TF logaritmik digunakan untuk menghindari dominasi dokumen yang mengandung sedikit *term* dalam *query*, namun mempunyai frekuensi yang tinggi.

$$tf_{ij} = 1 + \log(tf) \qquad (2)$$

d. TF normalisasi: perhitungan TF normalisasi menggunakan perbandingan antara frekuensi *term* dengan nilai maksimum dari keseluruhan atau kumpulan frekuensi *term* yang ada pada suatu dokumen.

$$tf_{ij} = 0{,}5 + 0{,}5\ x \frac{tf}{\max tf} \qquad (3)$$

3. *Inverse Document Frequency* (IDF) berfungsi mengurangi bobot suatu *term* jika kemunculannya banyak tersebar diseluruh dokumen. Perhitungan IDF diperlukan, karena jika hanya menggunakan TF saja dikhawatirkan akan muncul kata umum yang akan dominan, yang seharusnya kata tersebut dihilangkan. IDF dirumuskan sebagai berikut.

$$idf_j = \log\left(\frac{D}{df_j}\right) \qquad (4)$$

Di mana,

D = Jumlah semua dokumen yang ada dalam data

df_j = Jumlah dokumen yang mengandung *term*

2.7 *Count Vectorizer*

Count Vectorizer digunakan untuk menghitung tingkat kepentingan setiap kata pada seluruh dokumen dan menghasilkan bobot dari setiap kata yang telah dihitung tingkat kepentingannya. *Count vectorizer* berfungsi untuk

menghitung frekuensi kata dalam dokumen. *Count vectorizer* dapat mengubah fitur teks menjadi sebuah representasi vector (Munawar, 2019)

Contoh dari *count vectorizer* adalah sebagai berikut. Terdapat dua kalimat data teks.

1. Kue itu berwarna putih
2. Pisang itu berwarna kuning

Dari data teks tersebut (atau Bernama korpus) makan dapat disusun sebuah *vocabulary* yang terdiri dari 6 kata yaitu “berwarana”, “itu”, “kue”, “kuning”, “pisang”, dan “putih”. Kemudian menjadikan data untuk merepresentasi vector 6 dimensi (masing-masing untuk setiap kata). Tiap elemen dari vector menunjukkan jumlah fitur kata yang ada pada data dengan hasil sebagai berikut (Hakim, 2020).

1. Kue itu berwarna putih : [1,1,1,0,0,1]
2. Pisang itu berwarna kuning : [1,1,0,1,1,0]

2.8 *N-gram*

N-gram adalah model probabilistik yang dapat digunakan untuk membangkitkan karakter kata serta memprediksi kata berikutnya dalam urutan kata tertentu. Dalam pembangkitan karakter, *N-gram* terdiri dari *substring* sepanjang n karakter dari sebuah *string*. *N-gram* digunakan untuk mengambil potongan-potongan karakter huruf sejumlah n dari sebuah kata yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen. Semakin besar nilai n dari sebuah kata maka berbanding terbalik dengan

jumlah frekuensi keluar yang didapat, yaitu semakin kecil atau lebih jarang keluar. Penggunaan model *bigram* dan *trigram* untuk model Bahasa masih memungkinkan, karena hasil dari jumlah frekuensi keluar pada suku *N-gram*-nya masih cukup besar dan datanya masih valid apabila diproses lebih lanjut (Fitria K., 2019).

Pendekatan berbasis *N-gram* digunakan pada identifikasi Bahasa juga didasarkan bahasa manusia selalu memiliki beberapa kata yang lebih sering terjadi daripada yang lain, bahwa ada selalu set kata-kat yang mendominasi Sebagian besar dari kata-kata lainnya pada suatu bahasa-bahasa dalam hal frekuensi penggunaan. Karena itu *N-gram* dari suatu bahasa bersifat unik disebabkan frekuensi penggunaan huruf, atau pasangan huruf baik itu vocal atau konsonan dari suatu bahasa yang umumnya berbeda dengan bahasa yang lain. Seperti misalnya pada teks bahasa Indonesia vocal "a" akan merupakan vocal yang frekuensi munculnya paling tinggi, sementara untuk bahasa inggris vocal "e" merupakan vocal yang frekuensinya paling tinggi (Airlangga et al., 2010).

2.9 *Confussion Matrix*

Confusion Matrix merupakan metode evaluasi yang dapat digunakan untuk menghitung kinerja atau tingkat kebenaran dari proses klasifikasi. *Confusion Matrix* adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan nilai aktual. Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada *Confusion Matrix* yaitu *True Positive* (TP), *True Negative*

(TN), *False Positive* (FP), dan *False Negative* (FN). Tabel *Confusion Matrix* dapat dilihat pada Tabel 2.1.

Table 2.1 *Confusion Matrix*

		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FN
	Negatif	FP	TN

Keterangan:

a) TP (*True Positive*) merupakan banyaknya data yang kelas aktualnya adalah kelas postif dengan kelas prediksinya merupakan kelas positif.
b) FN (*False Negative*) merupakan banyaknya data yang kelas aktualnya adalah kelas positif dengan kelas prediksinya merupakan kelas negatif.
c) FP (*False Positive*) merupakan banyaknya data yang kelas aktualnya adalah kelas negatif dengan kelas prediksinya merupakan kelas positif.
d) TN (*True Negative*) merupakan banyaknya data yang kelas aktualnya adalah kelas negatif dengan kelas prediksinya merupakan kelas negatif.

Untuk evaluasi pada sistem klasifikasi dengan *multiclass*, *confusion matrix* juga bisa digunakan, namun dengan beberapa ketentuan tambahan.

Tabel 2.2 *Multiclass Confusion Matrix*

		Prediksi		
		Positif	Negatif	Netral
Akurasi	Positif	TPos	FPosNeg	FPosNet
	Negatif	FNegPos	TNeg	FNegNet
	Netral	FNetPos	FNetNeg	TNet

Tidak jauh berbeda dengan klasifikasi biner, *multiclass confusion matrix* juga memiliki elemen TP (*True Positive*), FN (*False Negative*), FP (*False Positive*), dan TN (*True Negative*). Berikut adalah ketentuan dalam menetapkan nilai elemen tersebut:

a) TP (*True Positive*) merupakan banyaknya data yang kelas aktualnya sama dengan kelas prediksinya.
b) FN (*False Negative*) merupakan total dari seluruh baris yang ditunjuk kecuali TP yang dicari.
c) FP (*False Positive*) merupakan total dari seluruh kolom yang ditunjuk kecuali TP yang dicari.
d) TN (*True Negative*) merupakan total dari seluruh kolom dan baris selain yang ditunjuk.

Berdasarkan Tabel 2.3 dapat dijabarkan nilai TP, FN, FP, dan TN untuk masing-masing kelas sebagai berikut:

1. Kelas Positif

Rincian nilai TP, FN, FP, TN pada kelas positif sebagaimana ditunjukkan pada Tabel 2.4 adalah sebagai berikut:

- TP terletak pada sel (1,1) berwarna hijau
- FN terletak pada sel (1,2) dan (1,3) berwarna kuning
- FP terletak pada sel (2,1) dan (3,1) berwarna biru
- TN terletak pada sel (2,2), (2,3), (3,2) dan (3,3) berwarna merah

Table 2.3 Rincian TP, FN, FP, dan TN pada kelas positif

		Prediksi		
		Positif	Negatif	Netral
Akurasi	Positif	TP (1,1)	FN (1,2)	FN (1,3)
	Negatif	FP (2,1)	TN (2,2)	TN (2,3)
	Netral	FP (3,1)	TN (3,2)	TN (3,3)

2. Kelas Negatif

Rincian nilai TP, FN, FP, TN pada kelas positif sebagaimana ditunjukkan pada Tabel 2.4 adalah sebagai berikut:

- TP terletak pada sel (2,2) berwarna hijau
- FN terletak pada sel (2,1) dan (2,3) berwarna kuning
- FP terletak pada sel (1,2) dan (3,2) berwarna biru
- TN terletak pada sel (1,1), (1,3), (3,1) dan (3,3) berwarna merah

Tabel 2.4. Rincian TP, FN, FP, dan TN pada kelas negatif

		Prediksi		
		Positif	Negatif	Netral
Akurasi	Positif	TN (1,1)	FP (1,2)	TN (1,3)
	Negatif	FN (2,1)	TP (2,2)	FN (2,3)
	Netral	TN (3,1)	FP (3,2)	TN (3,3)

3. Kelas Netral

Rincian nilai TP, FN, FP, TN pada kelas positif sebagaimana ditunjukkan pada Tabel 2.4 adalah sebagai berikut:

- TP terletak pada sel (3,3) berwarna hijau
- FN terletak pada sel (3,1) dan (3,2) berwarna kuning
- FP terletak pada sel (1,3) dan (2,3) berwarna biru
- TN terletak pada sel (1,1), (1,2), (2,1) dan (2,2) berwarna merah

Tabel 2.5 Rincian nilai TP, FN, FP, dan TN pada kelas netral

		Prediksi		
		Positif	Negatif	Netral
Akurasi	Positif	TN (1,1)	TN (1,2)	FP (1,3)
	Negatif	TN (2,1)	TN (2,2)	FP (2,3)

	Netral	FN (3,1)	FN (3,2)	TP (3,3)

Dengan dasar tabel *Confusion Matrix* kemudian dapat dilakukan penghitungan nilai akurasi. Metriks tersebut sangat bermanfaat untuk mengukur performa dari *classifier* atau algoritma yang digunakan untuk melakukan prediksi.

Akurasi merupakan metode pengujian berdasarkan tingkat kedekatan antara nilai prediksi dengan nilai aktual. Dengan mengetahui jumlah data yang diklasifikasikan secara benar maka dapat diketahui akurasi hasil prediksi. Persamaan akurasi *confusion matrix* 2x2 ditunjukkan pada persamaan berikut.

$$Akurasi = \frac{TP+TN}{TP+FP} \times 100\% \qquad (5)$$

Sedangkan persamaan akurasi confusion matrix 3x3 dapat dilihat pada persamaan berikut.

$$Akurasi = \frac{Total\ TP}{Jumlah\ data} \times 100\% \qquad (6)$$