PERBANDINGAN KINERJA METODE MESIN VEKTOR PENDUKUNG TANPA DAN DENGAN FITUR BERBASIS LEKSIKON PADA ANALISIS SENTIMEN VAKSINASI COVID-19 DI INDONESIA

SKRIPSI

SRI MULYANI

H051171517

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

JANUARI 2022

PERBANDINGAN KINERJA METODE MESIN VEKTOR PENDUKUNG TANPA DAN DENGAN FITUR BERBASIS LEKSIKON PADA ANALISIS SENTIMEN VAKSINASI COVID-19 DI INDONESIA

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin

SRI MULYANI

H051171517

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

JANUARI 2022

LEMBAR PERNYATAAN KEOTENTIKAN

Saya yang bertanda tangan di bawah ini menyatakan dengan sungguh-sungguh bahwa Skripsi yang saya buat dengan judul

PERBANDINGAN KINERJA METODE MESIN VEKTOR PENDUKUNG TANPA DAN DENGAN FITUR BERBASIS LEKSIKON PADA ANALISIS SENTIMEN VAKSINASI COVID-19 DI INDONESIA

Adalah benar hasil karya saya sendiri, bukan hasil plagiat dan belum pernah dipublikasikan dalam bentuk apapun.

Makassar, 24 Januari 2022

SRI MULYANI

NIM. H051171517

iii

PERBANDINGAN KINERJA METODE MESIN VEKTOR PENDUKUNG TANPA DAN DENGAN FITUR BERBASIS LEKSIKON PADA ANALISIS SENTIMEN VAKSINASI COVID-19 DI INDONESIA

Disetujui Oleh:

Pembimbing Utama,	Pembimbing Pertama,
Sri Astuti Thamrin, S.Si., M.Stat., Ph.D.	**Siswanto, S.Si., M.Si.**
NIP. 19740713 199903 2001	**NIP. 19920107 201903 1 012**

Ketua Departemen Statistika

Dr. Nurtiti Sunusi, S.Si., M.Si.

NIP. 19720117 199703 2002

Pada Tanggal: 24 Januari 2022

iv

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh:

Nama : Sri Mulyani

NIM : H051171517

Program Studi : Statistika

Judul Skripsi : Perbandingan Kinerja Metode Mesin Vektor Pendukung Tanpa Fitur dan dengan Fitur Berbasis Leksikon Pada Analisis Sentimen Vaksinasi COVID-19 di Indonesia

Telah berhasil dipertahankan dihadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Sains program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

DEWAN PENGUJI

		Tanda Tangan
1. Ketua	: Sri Astuti Thamrin, S.Si., M.Stat., Ph.D	(.....)
2. Sekretaris	: Siswanto, S.Si., M.Si	(.....)
3. Anggota	: Dr. Anna Islamiyati, S.Si., M.Si	(.....)
4. Anggota	: Sitti Sahriman, S.Si., M.Si	(.....)

Ditetapkan di : Makassar

Tanggal : 24 Januari 2022

KATA PENGANTAR

Assalamu'alaikum Warohmatullahi Wabarokatuh.

Alhamdulillahi rabbil'alamin, Puji syukur kepada Allah *Subhanahu Wa Ta'ala* atas segala limpahan rahmat, nikmat, dan hidayah yang diberikan kepada penulis sehingga dapat menyelesaikan penulisan skripsi dengan judul "Perbandingan Kinerja Metode Mesin Vektor Pendukung Tanpa dan dengan Fitur Berbasis Leksikon pada Analisis Sentimen Vaksinasi COVID-19 di Indonesia" sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam.

Salam dan sholawat *InsyaAllah* senantiasa tercurah kepada Nabi Muhammad *Shallallahu'alaihi Wasallam*, sang kekasih tercinta yang telah memberikan petunjuk cinta dan kebenaran dalam kehidupan.

Dalam penyelesaian skripsi ini, penulis telah melewati perjuangan panjang dan pengorbanan yang tidak sedikit. Namun berkat rahmat dan izin-Nya serta dukungan dari berbagai pihak yang turut membantu baik moril maupun material sehingga akhirnya tugas akhir ini dapat terselesaikan. Oleh karena itu, penulis menyampaikan ucapan terima kasih yang setinggi-tingginya dan penghargaan yang tak terhingga kepada Ayahanda *M.Rizal Mustaka* dan Ibunda tercinta *Dahlia* yang telah membesarkan dan mendidik penulis dengan penuh kesabaran dan dengan limpahan cinta, kasih sayang, dan doa kepada penulis yang tak pernah habis, saudara-saudara penulis *A.Kiki Hardiyanti, Amd.Kep*. *M.Arif Budiman* dan kakak ipar *Moh. Adnin Firdaus, S.E* yang selalu membantu dan menjadi penyeman gat untuk segera menyelesaikan masa studi penulis, serta ponakan tersayang *Moh. Qayyum Ahsanul Barraq* yang selalu menghibur penulis.

Ucapan terima kasih dengan penuh keikhlasan juga penulis ucapkan kepada:

1. Ibu Prof. Dr. Dwia Aries Tina Pulubuhu, MA, selaku Rektor Universitas Hasanuddin beserta seluruh jajarannya.
2. Bapak Dr. Eng. Amiruddin, selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin beserta seluruh jajarannya.
3. Ibu Dr. Nurtiti Sunusi, S.Si., M.Si., selaku Ketua Departemen Statistika yang telah seperti orang tua sendiri. Segenap dosen pengajar dan staf Departemen

Statistika yang telah membekali ilmu dan kemudahan kepada penulis dalam berbagai hal selama menjadi mahasiswa di Departemen Statistika.

4. Ibu Sri Astuti Thamrin, S.Si.,M.Stat.,Ph.D. selaku Pembimbing Utama sekaligus penasehat akademik penulis yang telah ikhlas meluangkan waktu dan pemikirannya untuk memberikan arahan, pengetahuan, motivasi dan bimbingan ditengah kesibukan beliau.
5. Bapak Siswanto, S.Si.,M.Si. selaku Pembimbing Pertama penulis yang telah meluangkan waktunya ditengah kesibukan untuk memberikan arahan, pengetahuan, motivasi, bimbingan serta menjadi tempat berkeluh kesah untuk penulis.
6. Ibu Siti Sahriman, S.Si.,M.Si. dan Ibu Dr. Anna Islamiyanti., S.Si.,M.Si. selaku tim penguji yang telah memberikan saran dan kritikan yang membangun dalam penyempurnaan penyusunan tugas akhir ini.
7. Seluruh guru saat TK, SD, SMP, hingga SMA serta Para Dosen atas ilmunya, perhatian dan pengalaman yang telah diberikan sejak menempuh pendidikan.
8. Agung Hasan, S.Si. selaku teman dekat penulis yang bertemu dipenghujung perkuliahan yang menjadi motivasi diri, selalu memberikan, membantu, tempat berkeluh kesah dan selalu menjadi penyemangat dalam mengerjakan skripsi ini.
9. Sahabat tercinta penulis selama perkuliahan, Triana Rahayu, Surefti Bato' Sau', Nurul Hidayah Magfira, Rahma Fitriani Maradi Ibrahim, Mirna, Salsabilah Ramadhani, Novilia Jao, Indry Angelin dan Gabriele Jesica Tulandi yang telah menjadi sahabat terbaik sejak awal perkuliahan dan senantiasa mendengarkan curhatan, memberikan dorongan, semangat, dan motivasi dalam setiap keadaan sehingga penulis bisa mendapatkan lebih banyak pelajaran hidup.
10. Sahabat HIMAKOS, Kanda Muh. Dadang Kurniawan, dan Muh. Ilyas yang selalu memberikan semangat kepada penulis.
11. Saudara-saudariku Fitri Ainun Malahayati, Putri Ainun Mutia Datau, Ayu Feratiwy, Firdayanti, Fira Fadhilah Firman , Ainun Halia. Dj, Ilham Askari dan Keluarga Besar PROTON SMAELI yang menjadi rumah kedua sejak bangku

SMA yang sampai saat ini selalu ada dan masih setia mendengarkan keluh kesah penulis.

12. Sahabat terbaik sejak SMP, Suciati Faisal yang selalu ada dan berbagi suka duka meskipun terkadang jarang bertemu karena kesibukan masing-masing.
13. Teman-teman KKN Pinrang 02 Gelombang 104 dan teman terbaik Faishal Saini dan Azhardi Hamzah, terima kasih untuk hiburan, dukungan, doa serta menjadikan penulis menemukan teman sedaerah di penghujung perkuliahan.
14. Teman-teman Statistika 2017, terima kasih atas kebersamaan, suka, dan duka selama menjalani pendidikan di Departemen Statistika.
15. Keluarga besar DISKRIT 2017, terima kasih telah memberikan pelajaran yang berharga dan arti kebersamaan selama ini kepada penulis. Pengalaman yang berharga telah penulis dapatkan dari teman-teman selama berproses.
16. Keluarga Besar HIMASTAT FMIPA UNHAS dan HIMATIKA FMIPA UNHAS yang telah memberikan banyak kesan serta membantu dalam mengukir kisah yang tidak akan terlupakan dan ilmu yang mungkin tidak bisa didapatkan di proses perkuliahan.
17. Keluarga Mahasiswa FMIPA Unhas terkhusus anggota keluarga Himatika FMIPA Unhas dan Himastat FMIPA Unhas, terima kasih atas ilmu yang mungkin tidak bisa didapatkan di proses perkuliahan dan telah menjadi keluarga selama penulis kuliah di Universitas Hasanuddin.
18. Kepada seluruh pihak yang tidak dapat penulis sebutkan satu persatu, terima kasih setinggi-tingginya untuk segala dukungan dan partisipasi yang diberikan kepada penulis semoga bernilai ibadah di sisi Allah *Subhanahu Wa Ta'ala.*

Penulis berharap skripsi ini dapat memberikan tambahan pengetahuan baru bagi para pembelajar statistika. Penulis menyadari bahwa dalam penulisan tugas akhir ini masih banyak terdapat kekurangan. Oleh karena itu, dengan segala kerendahan hati penulis memohon maaf. Akhir kata, semoga dapat bermanfaat bagi pihak-pihak yang berkepentingan. *Aamiin Yaa Rabbal Alamin.*

Makassar, 24 Januari 2022

Sri Mulyani

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai civitas akademik Universitas Hasanuddin, saya yang bertandatangan di bawah ini:

Nama	: Sri Mulyani
NIM	: H051171517
Program Studi	: Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam
Jenis Karya	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin Hak Bebas Royalti Non eksklusif (*Non-exclusive Royalty-Free Right*) atas tugas akhir saya yang berjudul:

"Perbandingan Kinerja Metode Mesin Vektor Pendukung dan Fitur Berbasis Leksikon pada Analisis Sentimen Vaksinasi COVID-19 di Indonesia"

beserta perangkat yang ada (jika diperlukan). Terkait dengan hal di atas, maka pihak universitas berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di Makassar pada tanggal 24 Januari 2022

Yang menyatakan

Sri Mulyani

ABSTRAK

Twitter merupakan salah satu media sosial yang sering digunakan untuk mendapatkan berita secara cepat dan singkat. Setelah merebaknya wabah virus COVID-19 dan kebijakan pemerintah untuk melakukan vaksinasi COVID-19 di Indonesia, opini masyarakat semakin banyak yang dituangkan melalui *tweet*. Hal ini menyebabkan topik vaksinasi COVID-19 menarik untuk dilakukan analisis sentiment. Analisis sentimen adalah suatu teknik mengekstrak data teks menggunakan metode klasifikasi untuk mendapatkan informasi tentang sentimen bernilai positif maupun negatif. Oleh karena itu dalam penelitian ini dilakukan klasifikasi opini masyarakat terhadap vaksinasi COVID-19 menggunakan metode mesin vektor pendukung tanpa fitur dan dengan fitur berbasis leksikon. Data hasil pelabelan manual yang digunakan dalam penelitian sebanyak 2981 data *tweet*. Hasil ketetapan klasifikasi menggunakan mesin vektor pendukung pada data vaksinasi COVID-19 di Indonesia diperoleh akurasi, g-mean dan AUC sebesar 83%, 50% dan 61.35%, sedangkan dengan fitur berbasis leksikon pada data vaksinasi COVID-19 di Indonesia diperoleh akurasi, g-mean dan AUC sebesar 90%, 86.63% dan 87%. Dari hasil pengujian diperoleh bahwa mesin vektor pendukung dengan fitur berbasis leksikon memiliki hasil klasifikasi yang lebih baik dibandingkan dengan mesin vektor pendukung tanpa fitur.

Kata Kunci: Analisis Sentimen, Fitur Berbasis Leksikon, Mesin Vektor Pendukung, Vaksinasi COVID-19

ABSTRACT

Twitter is one of the social media that is often used to get news quickly and briefly. Following the breakout of the COVID-19 virus and the vaccination campaign against it adopted by the government in Indonesia, public opinion has been increasingly vocal through tweets about this activity. As a result, sentiment analysis on the issue of COVID-19 immunization is a suitable fit. Sentiment analysis is a technique of extracting text data using a classification method to obtain information about positive or negative sentiment. Therefore, in this study, a classification of public opinion on COVID-19 vaccination was carried out using the support vector machine without and with lexicon-based features method. The data from manual labeling used in the study was 2981 tweet data. The results of the classification determination using a support vector machine on the COVID-19 vaccination data in Indonesia obtained accuracy, g-mean, and AUC of 83%, 50%, and 61.35%, while with lexicon-based features on the COVID-19 vaccination data in Indonesia obtained accuracy, g -mean and AUC of 90%, 86.63%, and 87%. Based on the result, it is found that the support vector machine with lexicon based features has a better classification result compared to using support vector machine without features.

Keywords: Sentiment Analysis, Lexicon-Based Features, Support Vector Machine, COVID-19 Vaccination

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Semakin berkembangnya teknologi internet pertumbuhan jumlah data digital juga semakin besar. Media sosial adalah salah satu contoh penghasil data internet terbesar (Dhawan, 2014). Salah satu media sosial yang sering digunakan yaitu *twitter,* sebuah media baru berjenis *microblogging* yang memberikan kemudahan untuk mendapatkan berita secara cepat dan singkat. Pengguna *twitter* dapat mengemukakan pendapatnya terhadap suatu produk atau mengomentari suatu masalah melalui *tweet* (Kurniawan, 2017). *Snscrape* merupakan suatu *library* pada *python* yang dapat memesan dan mengatur informasi pada *twitter* dan juga memudahkan pengguna untuk mengambil data *tweet* tanpa batasan secara *real time*.

Salah satu topik yang sekarang menarik dibahas pada laman *twitter* adalah vaksinasi COVID-19 akibat peristiwa pandemi COVID-19 yang merebak di seluruh dunia pada akhir tahun 2019. Berdasarkan situs covid19.go.id grafik laju penyebaran COVID-19 di Indonesia sejak virus itu dinyatakan masuk pada Maret 2020 belum menunjukkan garis landai sehingga pemerintah mengadakan beberapa kebijakan, salah satunya vaksinasi yang menjadi solusi penting untuk menekan angka penyebarannya dengan asumsi semakin banyak orang yang mempunyai kekebalan maka sebaran virus dapat ditekan. Namun vaksinasi ini masih banyak menuai pro dan kontra masyarakat karena banyaknya informasi yang beredar di media sosial.

Cara mengetahui tanggapan masyarakat mengenai vaksinasi COVID-19 pada aplikasi *twitter* dapat menggunakan *text mining* dengan menerapkan analisis sentimen. *Text mining* merupakan proses dari sejumlah data tidak terstruktur yang nantinya akan diperoleh pengetahuan yang potensial dari data teks (Turban, 2011). Salah satu tujuan penggunaan *text mining* adalah analisis sentimen. Analisis sentimen atau *opinion mining* merupakan metode yang digunakan untuk menganalisa pendapat untuk melihat kecenderungan seseorang menilai suatu objek yang bersentimen positif atau negatif (Liu, 2012). Terdapat banyak metode klasifikasi dalam menyelesaikan analisis sentimen, diantaranya: K-*Nearest*

Neighbors (KNN), *decision tree*, *naïve bayes*, mesin vektor pendukung (*support vector machine*), dan lain-lain.

Salah satu metode klasifikasi yang sering digunakan adalah metode mesin vektor pendukung. Konsep dasar mesin vektor pendukung adalah *linear classifier*, dan selanjutnya dikembangkan agar dapat bekerja pada *problem* non-linear dengan memasukkan konsep *kernel trick* pada ruang berdimensi tinggi (Nugroho, dkk. 2003) dimana mesin vektor pendukung adalah metode yang mempelajari area yang memisahkan antar kategori dalam sebuah observasi (Williams, 2011). Pada penelitian sebelumnya, yang termuat pada penelitian Vidya, dkk. (2015) tentang Analisis Sentimen pada Tingkat Reputasi Merk Penyedia Layanan Telekomunikasi Seluler, membuktikan bahwa metode mesin vektor pendukung memiliki akurasi yang lebih tinggi dibandingkan *naïve bayes* dan *decision tree* yaitu 82.40%.

Mayoritas analisis sentimen yang telah dilakukan hanya melihat dari polaritas kalimat tanpa memperhatikan nilai kata yang ada pada sebuah *tweet.* Fitur berbasis leksikon adalah fitur atau kata yang telah diberi bobot berdasarkan kamus atau *lexical* untuk mengolah informasi sentimen berbahasa Indonesia. Jumlah bobot kata-kata bersentimen baik dan buruk dapat mengarahkan suatu teks pada suatu kelas sentimen. Data yang diperoleh dari *twitter* dilakukan praproses dan pelabelan dengan melakukan sentimen *score*, lalu dijumlahkan nilainya untuk menentukan nilais sentimennya. Selanjutnya dilakukan pemodelan untuk memperoleh nilai klasifikasi terbaik.

Berdasarkan hal tersebut, sehingga penelitian yang dilakukan adalah analisis sentimen pada opini masyarakat berupa sentimen yang didapatkan dari media sosial *twitter* menggunakan metode mesin vektor pendukung tanpa fitur dan dengan fitur berbasis leksikon untuk mengetahui pro dan kontra masyarakat mengenai vaksinasi COVID-19 di Indonesia, serta membandingkan hasil klasifikasi terbaik dari pengujian sistem yang dihasilkan dari masing-masing metode sehingga dapat diketahui mana yang lebih unggul dari kedua metode tersebut.

1.2 Rumusan Masalah

Adapun rumusan masalah dalam penelitian ini sebagai berikut:

1. Bagaimana hasil klasifikasi sentimen analisis pada opini masyarakat terhadap vaksinasi COVID-19 di Indonesia dengan metode mesin vektor pendukung tanpa fitur dan dengan fitur berbasis leksikon ?
2. Bagaimana perbandingan kinerja metode mesin vektor pendukung tanpa fitur dan dengan fitur berbasis leksikon dalam klasifikasi opini masyarakat tentang vaksinasi COVID-19 di Indonesia?

1.3 Batasan Masalah

Untuk mendekati sasaran yang diharapkan, maka perlu diadakan pembatasan permasalahan dalam penelitian, adapun batasan masalah yaitu:

1. *Tweet* yang diambil dan dianalisis hanya *tweet* pada *twitter* yang berbahasa Indonesia menggunakan *library snscrape* pada *python* pada tanggal 1 Januari 2021 sampai 9 April 2021.
2. Sentimen *twitter* hanyalah sentimen yang berhubungan dengan vaksinasi COVID-19 di Indonesia.
3. Sentimen positif berisi tentang dukungan dan partisipasi masyarakat mengikuti vaksinasi COVID-19. Sedang sentimen negatif berisi tentang penolakan dan ketidakpercayaan terhadap vaksinasi COVID-19.
4. Metode yang digunakan adalah mesin vektor pendukung dengan kernel *radial basis function* (RBF) dan fitur berbasis leksikon.

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini sebagai berikut:

1. Memperoleh hasil klasifikasi sentimen analisis tentang opini masyarakat terhadap vaksinasi COVID-19 di Indonesia dengan metode mesin vektor pendukung tanpa fitur dan dengan fitur berbasis leksikon.
2. Membandingkan kinerja klasifikasi analisis metode mesin vektor pendukung tanpa fitur dan dengan fitur berbasis leksikon.

BAB II
TINJAUAN PUSTAKA

2.1 Virus Corona

Virus corona merupakan keluarga besar virus yang menyebabkan penyakit pada manusia dan hewan. Pada manusia biasanya menyebabkan penyakit infeksi saluran pernapasan, mulai flu biasa hingga penyakit yang serius seperti *Middle East Respiratory Syndrome* (MERS) dan Sindrom Pernafasan Akut Berat / *Severe Acute Respiratory Syndrome* (SARS). Virus corona adalah virus jenis baru yang ditemukan pada manusia sejak kejadian luar biasa muncul di Wuhan, Cina pada Desember 2019, kemudian diberi nama *Severe Acute Respiratory Syndrome Coronavirus 2* (SARS-COV2), dan penyakit *Coronavirus Disease-2019* (COVID-19) (Stoppneumonia.id, 2020). Grafik laju penyebaran COVID-19 di Indonesia sejak virus itu dinyatakan masuk pada Maret 2020 belum menunjukkan garis landai sehingga selain kebijakan Pembatasan Sosial Berskala Besar (PSBB), menerapkan protokol kesehatan dan 5 M, vaksinasi juga menjadi solusi penting untuk menekan angka penyebarannya. Vaksinasi COVID-19 merupakan salah satu upaya pemerintah Indonesia dalam menangani masalah COVID-19. Vaksinasi COVID-19 bertujuan untuk menciptakan kekebalan kelompok (*herd immunity*) agar masyarakat menjadi lebih produktif dalam menjalankan aktivitas kesehariannya (Kemenkes, 2021).

2.2 *Twitter*

Sosial media adalah sebuah media untuk bersosialisasi satu sama lain dan dilakukan secara *online* yang memungkinkan manusia untuk saling berinteraksi tanpa dibatasi ruang dan waktu (Rustian, 2012). Salah satu sosial media adalah *twitter*. *Twitter* merupakan sosial media yang dimiliki dan dioperasikan oleh *Twitter Inc,*. sebuah media baru berjenis *microblog* yang digunakan untuk menyebarkan pesan secara singkat dan padat dengan 140 karakter kepada pembacanya di seluruh dunia. Pesan itu disebut dengan *tweet*, yang bisa digunakan sebagai sarana penyebar informasi kepada semua orang baik yang dikenal maupun tidak. *Tweet* bisa dilihat secara publik, namun pengirim dapat membatasi pengiriman pesan ke daftar teman-

teman mereka saja. Pengguna dapat melihat *tweet* pengguna lain yang dikenal dengan sebutan pengikut (*follower*).

2.3 *Text Mining*

Text mining adalah lintas disiplin ilmu yang mengacu pada pencarian informasi, *data mining*, *machine learning*, statistik, dan komputasi linguistik. *Text mining* juga dikenal sebagai data mining teks. *Text mining* dapat didefinisikan sebagai suatu proses menggali informasi dari seorang *user* berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang merupakan komponen-komponen dalam *data mining* (Vidya, dkk. 2015).

Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Sumber data yang digunakan dalam *text mining* adalah sekumpulan teks. Proses kerja dari *text mining* adalah pola yang digunakan diambil dari sekumpulan bahasa alami yang tidak terstruktur. Tahap-tahap *text mining* secara umum adalah praproses teks dan *feature selection* (Kurniawan, 2017).

2.4 Sentimen Analisis

Sentimen analisis atau *opinion mining* merupakan bidang ilmu yang menganalisa pendapat, evaluasi, penilaian, sikap dan emosi publik terhadap entitas seperti produk, jasa, organisasi, individu, masalah dan peristiwa (Liu, 2012). Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen yang tergolong dalam sentimen positif atau negatif.

Besarnya pengaruh dan manfaat dari analisis sentimen menyebabkan penelitian dan aplikasi berbasis analisis sentimen berkembang pesat bahkan di Amerika terdapat sekitar 20-30 perusahaan yang memfokuskan pada layanan analisis sentimen. Sentimen analisis juga dapat menyatakan perasaan emosional sedih, gembira atau marah. Dalam sentimen analisis juga dapat mencari pendapat tentang produk-produk, *merk* dan lain-lain yang tergolong sentimen positif atau negatif pada *web.*

Ekspresi atau sentimen mengacu pada fokus topik tertentu, pernyataan pada satu topik mungkin akan berbeda makna dengan pernyataan yang sama pada subjek

yang berbeda. Oleh karena itu pada beberapa penelitian, pekerjaan didahului dengan menentukan elemen dari sebuah produk yang sedang dibicarakan sebelum memulai proses *opinion mining*.

2.5 *Twitter Crawling*

Untuk mendapatkan sekumpulan data *tweet* pada *twitter* diperlukan sebuah program untuk memperoleh informasi. *Snscrape* merupakan suatu *library* pada *python*. Dirilis pada 8 Juli 2020, *snscrape* adalah *library* yang digunakan untuk mengakses informasi untuk media sosial, salah satunya *twitter*. *Snscrape* dapat memperoleh informasi seperti profil pengguna, tagar, serta *tweet* yang dapat ditemukan tanpa menggunakan *API Twitter*. *Snscrape* memudahkan pengguna untuk mengambil data *tweet* tanpa batasan secara real time.

Crawling adalah proses pengambilan sejumlah besar halaman web dengan cepat ke dalam suatu tempat penyimpanan lokal dan mengindeksnya berdasar sejumlah kata kunci. *Crawling* data di *twitter* dapat menggunakan dua sistem pencarian, *by user* dan *by keyword*. Pencarian menggunakan *by keyword* yaitu pencarian menggunakan penggalan kata maupun *hashtag*. Sedangkan pencarian dengan *by user* yaitu pencarian berdasarkan nama akun user *twitter* (Sembodo, dkk. 2016).

2.6 Praproses Teks

Praproses teks merupakan tahapan awal dalam pengolahan teks yang digunakan untuk pengubahan bentuk dokumen menjadi data yang terstruktur sesuai kebutuhannya agar dapat diolah lebih lanjut dalam proses *text mining*. Tahapan praproses teks dalam klasifikasi bertujuan untuk meningkatkan akurasi klasifikasi data (Kurniawan, 2017). Berikut tahapan dalam praproses teks (Kurniawan, 2017):

1. *Case folding* adalah proses penyamaan *case* dalam sebuah dokumen. Ini dilakukan untuk mempermudah pencarian. Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu peran *case folding* dibutuhkan dalam mengonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (biasanya huruf kecil).

2. *Data Cleansing* adalah proses untuk membersihkan data atau karakter yang tidak baku yang dapat mengganggu pengolahan data.
3. *Stemming* adalah proses untuk mendapatkan kata dasar dengan cara menghilangkan awalan, akhiran, sisipan dan *confixes* (kombinasi awalan dan akhiran).
4. *Stopword* didefinisikan sebagai sebuah kata yang sangat sering muncul dalam suatu dokumen teks yang kurang memberikan arti penting terhadap isi dokumen. Kata depan dan konjungsi merupakan kandidat besar dari daftar *stopword* yang harus dihilangkan. Untuk dokumen berbahasa Indonesia, contoh dari kata-kata penghubung adalah "yang", "di", "dan", "itu", "dengan". Pada *stopword* ini juga peneliti mengabaikan emoji, sehingga peneliti berfokus pada penelitian dengan fitur *text* karena emoji dapat mengganggu proses analisis sentimen.
5. *Tokenizing* adalah proses untuk membagi teks yang berasal dari kalimat atau paragraf menjadi bagian-bagian tertentu. Sebagai contoh, tokenisasi dari kalimat "azizah senang sekali keliling papua" menghasilkan lima token, yakni: "azizah", "senang", "sekali", "keliling", "papua". Biasanya, yang menjadi acuan pemisah antar token adalah spasi dan tanda baca. Tokenisasi seringkali dipakai dalam ilmu linguistik dan hasil tokenisasi berguna untuk analisis teks lebih lanjut.

2.7 *Term Frequency Inverse Document Frequency*

Pembobotan *Term Frequency-Inverse Document Frequency* atau TF-IDF adalah suatu proses untuk melakukan transformasi data dari data tekstual ke dalam data numerik untuk dilakukan pembobotan pada tiap kata atau fitur. TF-IDF ini adalah sebuah ukuran statistik yang digunakan untuk mengevaluasi seberapa penting sebuah kata di dalam sebuah dokumen. TF adalah frekuensi kemunculan kata pada di tiap dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam tiap dokumen tersebut. DF adalah frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. IDF adalah invers dari nilai DF. Hasil dari pembobotan kata menggunakan TF-IDF ini adalah hasil perkalian dari TF dikalikan dengan IDF. Bobot kata semakin besar jika sering

muncul dalam suatu dokumen dan semakin kecil jika muncul dalam banyak dokumen (Jeremy, dkk. 2019).

Adapun rumus dari pembobotan kata TF-IDF adalah:

$$W_{t,d} = tf_{t,d} \text{ x } idf_t = tf_{t,d} \text{ x } log \frac{N}{df_t} \quad (2.1)$$

dengan:

$W_{t,d}$: Bobot TF-IDF

$tf_{t,d}$: Jumlah frekuensi kata

idf_t : Jumlah *inverse* frekuensi dokumen tiap kata

df_t : Jumlah frekuensi dokumen tiap kata

N : Jumlah total dokumen

2.8 *K-Fold Cross Validation*

K-fold cross validation adalah salah satu metode yang digunakan untuk mempartisi data menjadi data latih dan data uji. Metode ini banyak digunakan peneliti karena dapat mengurangi bias yang terjadi dalam pengambilan sampel *K*. *K-fold cross validation* secara berulang-ulang membagi data menjadi data latih dan data uji, untuk setiap data mendapat kesempatan menjadi data uji (Gokgoz & Subasi, 2015). *K* merupakan besar angka partisi data yang digunakan untuk pembagian data latih dan data uji. Menurut Max Kuhn & Kjell Johnson (2013) tidak ada aturan formal dalam penentuan nilai K, namun nilai K yang bagus digunakan adalah 5-10. Tabel 2.1 merupakan ilustrasi pembagian data menggunakan *k-fold cross validation*.

Tabel 2.1 Ilustrasi Pembagian Data

Percobaan 1	Uji	Latih	Latih	Latih	Latih
Percobaan 2	Latih	Uji	Latih	Latih	Latih
Percobaan 3	Latih	Latih	Uji	Latih	Latih
Percobaan 4	Latih	Latih	Latih	Uji	Latih
Percobaan 5	Latih	Latih	Latih	Latih	Uji

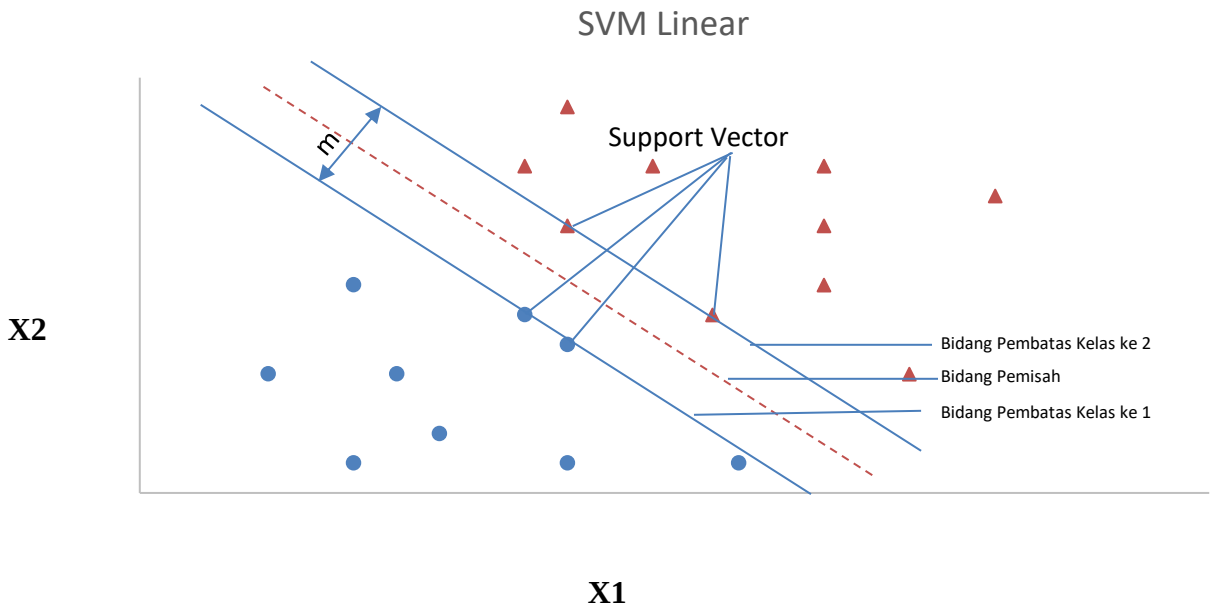
2.9 Mesin Vektor Pendukung

Mesin vektor pendukung atau *Support Vector Machine* (SVM) adalah metode yang mempelajari area yang memisahkan antar kategori dalam sebuah observasi

(Williams, 2011). Formulasi optimasi mesin vektor pendukung untuk masalah klasifikasi dibedakan menjadi dua kelas yaitu klasifikasi *linear* dan klasifikasi *non-linear*.

2.9.1 SVM pada *Linearly Separable Data*

SVM pada *linearly separable data* adalah penerapan metode SVM pada data yang dapat dipisahkan secara linier. Misalkan $x_i = \{x_1, x_2, \dots, x_n\}$ adalah titik data dan $y_i \in \{-1, +1\}$ adalah label kategori atau kelas data untuk *dataset*. Penggambaran *linearly separable data* dapat dilihat pada gambar berikut:



Gambar 2.1 Model SVM Linear

Gambar 2.1 kedua kelas data dapat dipisahkan oleh sepasang bidang pembatas yang sejajar (linier). Data yang berada pada bidang pembatas disebut dengan *support vector*. Persamaan *hyperplane* dapat ditulis sebagai berikut:

$$f(x) = (\boldsymbol{w^T x}) + b$$

$$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_N x_N + b \qquad (2.2)$$

Dengan $\boldsymbol{w}$ merupakan vektor normal terhadap *hyperplane* yang akan menentukan orientasi dari *hyperplane*. b merupakan suatu konstanta skalar yang menentukan lokasi fungsi *hyperplane* terhadap titik asal (Sari, 2017). Oleh karena itu *hyperplane* pemisah pada kasus ini dapat dimodelkan sebagai berikut:

$$f(x) = (\boldsymbol{w^T x}) + b$$
$$= \begin{cases} 1, & f(x) > 0 \\ -1, & f(x) < 0 \end{cases}$$

Pada kasus ini akan dipisahkan 2 *hyperplane* yang sejajar dengan *hyperplane* pertama akan membatasi kelas 1 sedangkan *hyperplane* kedua akan membatasi kelas 2, sehingga dapat dibentuk pertidaksamaan model matematika:

$$\boldsymbol{w^T x_i} + b \geq 1, \forall_i \in Kelas\ Positif \quad (2.3)$$

$$\boldsymbol{w^T x_i} + b \leq -1, \forall_i \in Kelas\ Negatif \quad (2.4)$$

Berdasarkan persamaan (2.3) dan (2.4) maka dapat diperoleh nilai margin (jarak) di antara 2 *hyperplane* dengan menggunakan prinsip jarak antara 2 garis sejajar pada persamaan berikut:

$$margin\ (m) = \frac{|(b-1)-(b+1)|}{\sqrt{\boldsymbol{w^T w}}}$$
$$= \frac{|-2|}{\|\boldsymbol{w}\|}$$
$$= \frac{2}{\|\boldsymbol{w}\|}$$

Untuk mendapatkan dua *hyperplane* pemisah dengan jarak sejauh mungkin, maka nilai dari *hyperplane* optimal dapat diperoleh melalui suatu solusi optimasi.
fungsi tujuan:

$$max\ \frac{2}{||\boldsymbol{w}||}$$

dengan kendala:

$$\boldsymbol{w^T x_i} + b \geq 1\ , \forall_i \epsilon\ Kelas\ Positif$$
$$\boldsymbol{w^T x_i} + b \leq -1\ , \forall_i \epsilon\ Kelas\ Negatif \quad (2.5)$$

atau dapat diringkas
fungsi tujuan:

$$max\ \frac{2}{||\boldsymbol{w}||}$$

dengan kendala:

$$y_i(\boldsymbol{w^T x_i} + b) \geq 1\ , i = 1,2,3, \dots, N \quad (2.6)$$

Menurut Sari, 2017 memaksimalkan margin $\frac{1}{||\boldsymbol{w}||}$ sama dengan meminimumkan $\frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2$, maka persamaan 2.6 dapat ditulis:

fungsi tujuan:

$$min\frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2$$

dengan kendala:

$$y_i(\boldsymbol{w^T x_i} + b) \geq 1, i = 1,2,3, \dots, N \tag{2.7}$$

Nilai *hyperplane* optimal yang memisahkan data dengan margin maksimum dapat diperoleh dengan menyelesaikan kendala optimasi kuadratik. Penyelesaian optimal dari kasus optimasi ini memiliki ekspansi subset pada pola latih yang berada dekat dengan *hyperplane.* Subset dari pola ini disebut *Support Vector* (SV).

Permasalahan optimasi kasus data yang dapat dipisahkan secara linear pada 2.7 dapat dituliskan dalam bentuk:

fungsi tujuan:

$$\min_{w,b}(\boldsymbol{w}, b) = \frac{1}{2}\boldsymbol{w^T w}$$

dengan kendala:

$$y_i(\boldsymbol{w^T x_i} + b) - 1 \geq 0\ , i = 1,2,3, \dots\ , N \tag{2.8}$$

Untuk mendapatkan solusi dari permasalahan optimasi bentuk primal pada Persamaan 2.8 maka akan digunakan metode *Lagrange* dengan kendala pertidaksamaan atau dikenal sebagai *Karush Kuhn Tucker* (KKT). Berdasarkan persamaan 2.8 maka diperoleh persamaan *lagrange* untuk kasus ini menjadi:

$$\begin{aligned} L_p(\boldsymbol{w}, b, \alpha) &= (\boldsymbol{w}, b) + \textstyle\sum_{i=1}^{n} \alpha_i\, [\, 1 - \, y_i(\, \boldsymbol{w^T x_i} + b)] \\ &= (\boldsymbol{w}, b) - \sum_{i=1}^{n} \alpha_i\, [\, y_i(\, \boldsymbol{w^T x_i} + b) - 1] \\ &= \left(\frac{1}{2}\boldsymbol{w^T w}\right) - \sum_{i=1}^{n} \alpha_i\, [\, y_i\big(\, \boldsymbol{w^T x_i} + b\big) - 1 \end{aligned} \tag{2.9}$$

Peubah α_i merupakan *Lagrange Multiplier* atau pengali *Lagrange*. Nilai dari *Lagrange Multiplier* ini adalah $\alpha_i > 0$. Fungsi Persamaan 2.9 akan diminimalkan terhadap $\boldsymbol{w}$ dan $\boldsymbol{b}$ serta dimaksimalkan terhadap peubah α (Sari, 2017). Turunan pertama dari fungsi pada Persamaan 2.9 adalah sebagai berikut:

Kondisi 1

$$\frac{\partial L_p}{\partial b} = 0$$

$$\frac{\partial L_p}{\partial b} = \frac{\partial L_p(\boldsymbol{w}, b, \alpha)}{\partial b}$$

$$= \frac{\partial\{(\boldsymbol{w}, b) - \sum_{i=1}^{n} \alpha_i \,[\, y_i(\, \boldsymbol{w^T x_i} + b) - 1]\}}{\partial b}$$

$$= \frac{\partial \left\{\frac{1}{2}\boldsymbol{w^T w} - \sum_{i=1}^{n} \alpha_i \,[\, y_i(\, \boldsymbol{w^T x_i} + b) - 1]\right\}}{\partial b}$$

$$= 0 - 0 - \sum_{i=1}^{n} \alpha_i \, y_i + 0$$

diperoleh:

$$\sum_{i=1}^{n} \alpha_i \, y_i = 0 \tag{2.10}$$

Kondisi 2

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = 0$$

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \frac{\partial L_p(\boldsymbol{w}, b, \alpha)}{\partial b}$$

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \frac{\partial\{(\boldsymbol{w}, b) - \sum_{i=1}^{n} \alpha_i \,[\, y_i(\, \boldsymbol{w^T x_i} + b) - 1]\}}{\partial \boldsymbol{w}}$$

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \frac{\partial \left\{\frac{1}{2}\boldsymbol{w^T w} - \sum_{i=1}^{n} \alpha_i \,[\, y_i(\, \boldsymbol{w^T x_i} + b) - 1]\right\}}{\partial \boldsymbol{w}}$$

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \frac{\partial \left\{\frac{1}{2}\boldsymbol{w^T w} - \sum_{i=1}^{n} \alpha_i \;\, y_i \, \boldsymbol{w^T x_i} - \boldsymbol{b} \sum_{i=1}^{n} \alpha_i \;\, y_i + \sum_{i=1}^{n} \alpha_i\right\}}{\partial \boldsymbol{w}}$$

$$= \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i \;\, y_i \, \boldsymbol{x_i} - 0 + 0$$

diperoleh:

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \;\, y_i \, \boldsymbol{x_i} \tag{2.11}$$

Tampak bahwa fungsi tujuan (2.8) mengandung fungsi kuadrat pada peubah $\boldsymbol{w}$, sehingga hal tersebut akan mengakibatkan cukup sulitnya penyelesaian secara komputasi dan akan memakan waktu yang panjang. Dengan demikian, maka

permasalahan tersebut akan lebih mudah dan lebih efisien jika diselesaikan dalam bentuk dual (Rashif, 2007).

Menurut teorema dualitas Keckman (2005) jika problem primal memiliki solusi optimal, maka problem dual juga akan mempunyai solusi optimal yang nilainya sama. Untuk memperoleh bentuk dual dari Persamaan (2.8), maka akan disubtitusikan Persamaan (2.11) ke dalam Persamaan (2.9) sebagai berikut:

$$\begin{aligned} L_p(\alpha) &= (\boldsymbol{w}, b) + \sum_{i=1}^{n} \alpha_i \,[\, 1 - \, y_i(\, \boldsymbol{w^T x_i} + b)] \\ &= \tfrac{1}{2}\boldsymbol{w^T w} - \, \textstyle\sum_{i=1}^{n} \alpha_i \,[\, y_i(\, \boldsymbol{w^T x_i} + b) - 1 \\ &= \frac{1}{2}\boldsymbol{w^T w} - \sum_{i=1}^{n} \alpha_i \, y_i \, \boldsymbol{w^T x_i} - b \sum_{i=1}^{n} \alpha_i \, y_i + \sum_{i=1}^{n} \alpha_i \\ &= \frac{1}{2}\boldsymbol{w^T w} - \sum_{i=1}^{n} \alpha_i \, y_i \, \boldsymbol{w^T x_i} + \sum_{i=1}^{n} \alpha_i \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \, \alpha_j y_i y_j \boldsymbol{x}_i^T \, \boldsymbol{x}_j - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \, \alpha_j y_i y_j \boldsymbol{x}_i^T \, \boldsymbol{x}_j + \sum_{i=1}^{n} \alpha_i \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \, \alpha_j y_i y_j \boldsymbol{x}_i^T \, \boldsymbol{x}_j \end{aligned} \tag{2.12}$$

Menurut teori dualitas meminimumkan $L_p(w)$ sama dengan memaksimumkan $L_p(\alpha)$, sehingga masalah pencarian *hyperplane* yang optimal pada kasus *linear separable* dapat dirumuskan:

fungi tujuan:

$$f(x_d) = \sum_{i=1}^{n} \alpha_i y_i (x_i x_d) + b$$

dengan kendala:

$$\sum_{j=1}^{n} \alpha_i y_i = 0, \forall_i \text{ dengan } \alpha_i > 0$$

$$y_i = \begin{cases} 1, & i > 0 \\ -1, & i < 0 \end{cases} \tag{2.13}$$

Dengan demikian , nilai α_i dapat ditemukan dengan menyelesaikan kasus optimasi pada 2.13 dan nilai $\boldsymbol{w}$ akan didapatkan dengan mensubstitusikan α_i pada

persamaan 2.11. Selanjutnya fungsi *hyperplane* yang optimal pada kasus *linear separable* dapat terbentuk persaaan berikut:

$$f(x_d) = \sum_{i=1}^{n} \alpha_i y_i (x_i x_d) + b \tag{2.14}$$

dengan x_d merupakan data yang akan diklasifikasikan, α_i merupakan solusi optimasi dari masalah optimasi (2.13) dan $\boldsymbol{b}$ dicari dengan formula:

$$b = -\frac{1}{2}\,(wx^{+} + wx^{-}) \tag{2.15}$$

Dengan demikian, kelas berdasarkan *hyperplane* optimal pada kasus dataset yang *linear separable* terbentuk sebagai berikut:

$$\begin{aligned} f(x) &= (\boldsymbol{w^T x}) + b \\ &= \begin{cases} 1 & f(x_d) > 0 \\ -1, & f(x_d) < 0 \end{cases} \end{aligned} \tag{2.16}$$

2.9.2 SVM pada *Non Linearly Separable Data* dengan Metode Kernel

Klasifikasi data yang tidak dapat dipisahkan secara linier memerlukan modifikasi pada formula SVM agar dapat menemukan solusinya. Dalam praktiknya, jarang ditemui data latih yang dapat dipisahkan secara linear. Modifikasi formula tersebut dilakukan pada kedua bidang *hyperplane.* Agar lebih fleksibel maka kedua *hyperplane* diberi tambahan peubah *slack* (ξ_i) dengan $\xi_i \geq 0, i = 1,2,\dots,n$, sehingga diperoleh suatu modifikasi *hyperplane* baru pada persamaan berikut:

$$\begin{aligned} y_i\Big((\boldsymbol{w^T x_i}) + b\Big) + \xi_i &\geq 1 \\ y_i\Big((\boldsymbol{w^T x_i}) + b\Big) &\geq 1 - \xi_i \end{aligned} \tag{2.17}$$

atau dapat dituliskan:

$$\boldsymbol{w^T x_i} + b \geq 1 - \xi_i, \forall \in Kelas\ Positif \tag{2.18}$$

$$\boldsymbol{w^T x_i} + b \leq -1 + \xi_i, \forall \in Kelas\ Negatif \tag{2.19}$$

Pencarian *hyperplane* optimal dengan tambahan peubah *slack* ini sering disebut sebagai *soft margin hyperplane*. Selain itu, pada modifikasi formula untuk data yang tidak dapat dipisahkan secara linier juga memerlukan tambahan parameter *penalty C* (Sari. P. D, 2017). Sehingga formula untuk mendapatkan *hyperplane* optimal menjadi:

fungsi tujuan:

$$min\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right)$$

dengan kendala:

$$y_i\Big((\boldsymbol{w^T x}_i) + b\Big) \geq 1 - \xi_i\,, i = 1,2,3\ldots, n \tag{2.20}$$

Dengan C merupakan parameter yang menetukan besarnya penalti akibat keselahan dalam klasifikasi data dan nilai C ini ditentukan oleh pengguna. Semakin tinggi nilai C, maka kemungkinan terjadinya kesalahan dalam penentuan solusi akan semakin kecil. Sebaliknya, jika nilai C semakin rendah maka semakin tinggi proporsi kesalahan yang terjadi pada penentuan solusi.

Dari Persamaan 2.20, akan dimaksimalkan nilai margin antara 2 kelas dengan meminimalkan $\|\boldsymbol{w}\|^2$. Dalam formula ini, akan dicoba meminimalkan kesalahan yang dinyatakan oleh peubah ξ_i. Meminimalkan $\|\boldsymbol{w}\|^2$ ekuivalen dengan memaksimalkan $\frac{1}{\|\boldsymbol{w}\|}$. Penggunaan peubah *slack* bertujuan untuk mengatasi kasus ketidaklayakan (*infeasibility*) dari kendala tersebut. Untuk meminimalkan nilai dari peubah *slack*, akan diberikan pinalty dengan menggunakan harga dari parameter C (Sari, 2017).

Dengan menggunakan teori optimasi, maka didapatkan:

fungsi tujuan:

$$\min_{w,b,\xi}(\boldsymbol{w}, b, \xi) = \tfrac{1}{2}\,\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i$$

dengan kendala:

$$y_i(\boldsymbol{w^T x_i} + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,\ldots, n \tag{2.21}$$

atau dapat dituliskan:

fungsi tujuan:

$$\min_{w,b,\xi}(\boldsymbol{w}, b, \xi) = \tfrac{1}{2}\,\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i$$

dengan kendala:

$$y_i(\boldsymbol{w^T x_i} + b) - 1 + \xi_i \geq 0, \xi_i \geq 0, i = 1,2,\ldots, n \tag{2.22}$$

Untuk mendapatkan solusi dari permasalahan optimasi bentuk primal pada Persamaan 2.9, maka seperti pada kasus *linear separable* akan digunakan *Karush*

Khun Tucker (KKT). Namun pada kasus *non-linear separable* ini akan memiliki dua pengali *Lagrange* untuk kasus ini menjadi:

$$L_p(\boldsymbol{w}, b, \xi, \alpha, \beta) = (\boldsymbol{w}, b, \xi) + \sum_{i=1}^{n} \alpha_i [1 - \xi_i - y_i(\boldsymbol{w^T x_i} + b)] + \sum_{i=1}^{n} \beta_i(-\xi_i)$$

$$= (\boldsymbol{w}, b, \xi) - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w^T x_i} + b) - 1 + \xi_i] - \sum_{i=1}^{n} \beta_i(-\xi_i)$$

$$= \left(\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i\right) - \sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w^T x_i} + b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i)$$

$$= \frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w^T x_i} + b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i) \quad (2.23)$$

Peubah α_i dan β_i merupakan *Lagrange Multiplier* dengan $\alpha_i \geq 0$ dan $\beta_i \geq 0$. Peubah α_i dan β_i dapat disebut sebagai peubah non-negatif. Fungsi Persamaan 2.23 akan diminimalkan terhadap peubah $\boldsymbol{w}, b$ dan ξ serta harus dimaksimalkan terhadap peubah α dan β (Sari. P. D, 2017). Berikut turunan pertama dari fungsi Persamaan 2.23 yaitu:

Kondisi 1

$$\frac{\partial L_p}{\partial b} = 0$$

$$\frac{\partial L_p}{\partial b} = \frac{\partial L_p(\boldsymbol{w}, b, \xi, \alpha, \beta)}{\partial b}$$

$$= \frac{\partial\{(\boldsymbol{w},b,\xi) - \sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w^T x_i}+b)-1+\xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i)\}}{\partial b}$$

$$= \frac{\partial\left\{\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w^T x_i} + b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i)\right\}}{\partial b}$$

$$= \frac{\partial\left\{\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w^T x_i} + b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i)\right\}}{\partial b}$$

$$= \frac{\partial\left\{\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{w^T x_i} - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\beta_i(\xi_i)\right\}}{\partial b}$$

$$= 0 + 0 - 0 - \sum_{i=1}^{n}\alpha_i y_i + 0 - 0 - 0$$

diperoleh:

$$0 = \sum_{i=1}^{n} \alpha_i y_i \tag{2.24}$$

Kondisi 2

$$\frac{\partial L_p}{\partial w} = 0$$

$$\frac{\partial L_p}{\partial w} = \frac{\partial L_p\,(\boldsymbol{w},b,\xi,\alpha,\beta)}{\partial w}$$

$$= \frac{\partial\{(\boldsymbol{w},b,\xi) - \sum_{i=1}^{n}\alpha_i\,[y_i(\boldsymbol{w^T x_i}+b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i)\}}{\partial w}$$

$$= \frac{\partial\left\{\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\boldsymbol{w^T x_i}+b\right)-1+\xi_i\right] - \sum_{i=1}^{n}\beta_i(\xi_i)\right\}}{\partial w}$$

$$= \frac{\partial\left\{\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{w^T x_i} - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\beta_i\,(\xi_i)\right\}}{\partial b}$$

$$= \boldsymbol{w} + 0 - \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x_i} - 0 + 0 - 0 - 0$$

$$= \boldsymbol{w} - \sum_{\boldsymbol{i=n}}^{\boldsymbol{n}}\alpha_i y_i \boldsymbol{x_i}$$

diperoleh:

$$\boldsymbol{w} = \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x_i} \tag{2.25}$$

Kondisi 3

$$\frac{\partial L_p}{\partial \xi} = 0$$

$$\frac{\partial L_p}{\partial \xi} = \frac{\partial L_p\,(\boldsymbol{w},b,\xi,\alpha,\beta)}{\partial \xi}$$

$$= \frac{\partial\left\{(\boldsymbol{w},b,\xi) - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\boldsymbol{w^T x_i}+b\right)-1+\xi_i\right] - \sum_{i=1}^{n}\beta_i(\xi_i)\right\}}{\partial \xi}$$

$$= \frac{\partial\left\{\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w^T x_i}+b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i)\right\}}{\partial \xi}$$

$$= \frac{\partial\left\{\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{w^T x_i} - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\beta_i(\xi_i)\right\}}{\partial \xi}$$

$$= 0 + \sum_{i=1}^{n} C - 0 - 0 + 0 - \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \beta_1$$

$$= \sum_{i=1}^{n} C - \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \beta_i$$

diperoleh:

$$\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} C - \sum_{i=1}^{n} \beta_i$$

$$C = \alpha_i + \beta_i \tag{2.26}$$

Untuk memperoleh bentuk dual, maka akan disubtitusikan Persamaan (2.24), (2.25) dan (2.26) ke dalam Persamaan (2.23)

$$L_p(\alpha) = \frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w^T x_i} + b) - 1 + \xi_i] - \sum_{i=1}^{n} \beta_i(\xi_i)$$

$$= \frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{w^T x_i} - b\sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i \xi_i - \sum_{i=1}^{n} \beta_i(\xi_i)$$

$$= \frac{1}{2}\boldsymbol{w^T w} + \sum_{i=1}^{n} (C - \alpha_i)\xi_i - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{w^T x_i} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} (C - \alpha_i)\xi_i$$

$$= \tfrac{1}{2}\boldsymbol{w^T w} - \textstyle\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{w^T x_i} + \sum_{i=1}^{n} \alpha_i$$

$$= \tfrac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T x_j} - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T x_j} + \sum_{i=1}^{n} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T x_j} \tag{2.27}$$

Menurut teori dualitas meminimumkan $L_p(\boldsymbol{w})$ sama dengan memaksimumkan $L_p(\alpha)$, sehingga masalah pencarian *hyperplane* yang optimal pada kasus non-*linear separable* dapat dirumuskan dengan:

fungsi tujuan:

$$\max_{\alpha} L_p(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T x_j}$$

dengan kendala:

$\sum_{i=1}^{n} \alpha_i y_i = 0$ dan $\beta_i \xi_i = (C - \alpha_i)\xi_i = 0\ \forall_i$ dengan $\alpha_i > 0$

Selanjutnya fungsi *hyperplane* yang optimal pada kasus SVM *non-linear* hampir sama dengan kasus *linear separable* yaitu sebagai berikut:

$$f(x_d) = \sum_{i \in S} \alpha_i y_i \boldsymbol{x_i^T x} + b \tag{2.28}$$

dengan S merupakan data yang akan diklasifikasikan, α_i merupakan solusi optimal dari masalah optimasi (2.5) dan $\boldsymbol{b}$ dicari dengan formula dimana S adalah himpunan indeks *support vector*. Karena α_i tidak nol untuk *support vector* maka penjumlahan di (2.28) ditambahkan hanya untuk *support vector*. Untuk α_i tak berhingga, maka:

$$b = y_i - \boldsymbol{w^T x_i} \tag{2.29}$$

sudah memuaskan. Untuk memastikan ketepatan perhitungan, kita menghitung rata-rata bias (b) yang dihitung untuk *support vector* tak berhingga sebagai berikut

$$b = \frac{1}{|U|} \sum_{i \epsilon U} (y_i - \boldsymbol{w^T x_i}) \tag{2.30}$$

dengan U adalah himpunan indeks *support vector* tak berhingga. Dengan demikian, kelas berdasarkan *hyperplane* optimal pada kasus dataset yang *non-linear separable* terbentuk sebagai berikut:

$$\begin{aligned} f(x) &= \sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b + b \\ &= \begin{cases} 1, & f(x_d) > 0 \\ -1, & f(x_d) < 0 \end{cases} \end{aligned} \tag{2.31}$$

Selanjutnya pada data pelatihan untuk kasus *nonseparable*, klasifikasi yang diperoleh mungkin tidak memiliki kemampuan generalisasi yang tinggi meskipun *Hyperplane* ditentukan secara optimal. Sehingga diatasi dengan cara *input space* dipetakan ke dalam *dot-product space* berdimensi tinggi yang disebut *feature space*.

Fungsi kernel merupakan suatu fungsi yang memetakan data ke ruang dimensi yang lebih tinggi dengan harapan data akan memiliki struktur yang lebih baik sehingga lebih mudah dipisahkan. Menggunakan vektor fungsi *non-linear* sebagai $\phi(x_i): R^n \rightarrow R^{n_k}$ memetakan ruang input ke ruang dimensi yang lebih tinggi. Dalam fungsi *non linier* pengklasifikasian diperoleh dengan:

$$f(x) = sign\,[\boldsymbol{w^T}\,\phi(x_i) + b] \tag{2.32}$$

dengan menggunakan teori optimasi, maka didapatkan:

fungsi tujuan:

$$\min_{w,b,\xi}(\boldsymbol{w}, b, \xi) = \frac{1}{2}\,\boldsymbol{w^T w} + \frac{1}{2} C \sum_{i=1}^{n} \xi_i^2 \tag{2.33}$$

dengan kendala:

$$y_i(\boldsymbol{w}^T\,\phi(x_i) + b) = 1 - \xi_i^2$$

$$\xi_i \geq 0, i = 1,2, \ldots, n \tag{2.34}$$

Untuk menurunkan permasalahan (2.26), (2.27) dan (2.28) digunakan *Lagrange Multiplier* dimana nilai $\alpha_i \geq 0$. Optimal poin akan ada di dalam *saddle point* dari *Lagrange function* menjadi:

$$L_p = (\boldsymbol{w}, b, \xi, \alpha, \beta) = (\boldsymbol{w}, b, \xi) + \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w^T}\,\phi(x_i) + b) - 1 + \xi_i] \tag{2.35}$$

Kemudian dilakukan turunan pertama yaitu sebagai berikut:

Kondisi 1

$$\frac{\partial L_p}{\partial b} = 0$$

$$= \frac{\partial L_p(\boldsymbol{w}, b, \xi, \alpha, \beta)}{\partial b}$$

$$= \frac{\partial\{(\boldsymbol{w},b,\xi) - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w^T}\,\phi(x_i)+b) - 1 + \xi_i]\}}{\partial b}$$

$$= \frac{\partial \left\{\frac{1}{2}\boldsymbol{w^T w} + \frac{1}{2} C \sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w^T}\,\phi(x_i) + b) - 1 + \xi_i]\right\}}{\partial b}$$

$$= \frac{\partial \left\{\frac{1}{2}\boldsymbol{w^T w} + \frac{1}{2} C \sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{w^T}\,\phi(x_i) - b \sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i \xi_i\right\}}{\partial b}$$

$$= 0 + 0 - 0 - \sum_{i=1}^{n} \alpha_i\, y_i + 0 - 0$$

diperoleh:

$$\sum_{i=1}^{n} \alpha_i\, y_i = 0 \tag{2.36}$$

Kondisi 2

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = 0$$

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \frac{\partial L_p(\boldsymbol{w}, b, \xi, \alpha, \beta)}{\partial \boldsymbol{w}}$$

$$= \frac{\partial\{(\boldsymbol{w}, b, \xi) - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w^T}\, \phi(x_i) + b) - 1 + \xi_i]\}}{\partial \boldsymbol{w}}$$

$$= \frac{\partial \left\{\frac{1}{2}\boldsymbol{w^T w} + \frac{1}{2} C \sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w^T}\, \phi(x_i) + b) - 1 + \xi_i]\right\}}{\partial \boldsymbol{w}}$$

$$= \frac{\partial \left\{\frac{1}{2}\boldsymbol{w^T w} + \frac{1}{2} C \sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{w^T}\, \phi(x_i) - b \sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i \xi_i\right\}}{\partial \boldsymbol{w}}$$

$$= \boldsymbol{w} + 0 - \sum_{i=1}^{n} \alpha_i\, y_i x_i - 0 + 0 - 0 - 0$$

$$= \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i\, y_i \phi(x_i)$$

diperoleh:

$$\boldsymbol{w} - \sum_{i=1}^{n} \alpha_i\, y_i x_i = 0$$

$$\sum_{i=1}^{n} \alpha_i\, y_i \phi(x_i) = \boldsymbol{w} \tag{2.37}$$

Dengan demikian, kelas berdasarkan *hyperplane* optimal pada kasus dataset yang *non-linear separable* terbentuk sebagai berikut:

$$f(x_d) = \sum_{i=1}^{n} \alpha_i y_i K\,(x_i, x_d) + b \tag{2.38}$$

dan b diperoleh:

$$b = y_i - \frac{1}{2}\sum_{i=1}^{n} \alpha_i y_i K(x_i x) \tag{2.39}$$

K merupakan salah satu fungsi Kernel yang akan digunakan. Fungsi Kernel digunakan untuk memetakan data non linier menjadi linier. Menurut Hsu, dkk (2010) berikut ini adalah beberapa fungsi kernel yang umum digunakan yaitu:

1. *Polynomial*
   $\mathrm{K}\,(x_i, x_j) = x_i^T x$
2. RBF
   $$\mathrm{K}(x_i, x_j) = exp\left(-\frac{(x_i - x_j)^T (x_i - x_j)}{2\,\gamma^2}\right) \tag{2.40}$$
3. *Sigmoid*
   $\mathrm{K}\,(x_i, x_j) = tahn(\gamma x_i^T + r)$

2.10 Fitur berbasis leksikon

Leksikon adalah kumpulan kata sentimen yang telah diketahui dan terhimpun (Mehta, 2017). Fitur berbasis leksikon adalah fitur atau kata yang telah diberi bobot berdasarkan kamus atau *lexical*. Pemberian bobot dilakukan untuk setiap kata yang termasuk sentimen positif atau termasuk sentimen negatif. Tujuan penggunaan fitur berbasis leksikon yaitu untuk menentukan orientasi sentimen suatu kata (Kurniawan, dkk. 2019). Untuk proses pembobotan pada fitur ini, dibutuhkan kamus atau *lexical* yang berisi kata-kata yang mengandung sentimen yang disebut dengan *sentiment dictionaries* (Buntoro, 2014). Agar bobot fitur berbasis leksikon seimbang dengan bobot TF-IDF, maka fitur jumlah kata positif dan kata negatif perlu dinormalisasikan dengan metode Min-max.

2.11 Normalisasi Min-Max

Min-Max normalization merupakan metode normalisasi dengan melakukan transformasi linier terhadap data asli sehingga menghasilkan keseimbangan nilai perbandingan antar data saat sebelum dan sesudah proses (Junaedi, dkk. 2011). Normalisasi digunakan untuk mengurangi kesalahan pada proses *data mining* (Wirawan dan Eksistyanto, 2015). Metode normalisasi ini merupakan metode yang paling sederhana dengan melakukan transformasi linier terhadap data asli dan memiliki kelebihan yaitu terdapat keseimbangan nilai perbandingan antara nilai data sebelum dinormalisasi dengan nilai data yang telah dinormalisasi. Persamaan normalisasi Min-max:

$$x_i' = \frac{x_i - Xmin}{Xmax - Xmin}(BA - BB) + BB \tag{2.41}$$

dengan:

x_i' : Nilai data baru hasil normalisasi min-max.

x_i : Nilai data yang akan di normalisasi.

Xmin : Nilai minimum dari data.

Xmax : Nilai maksimum dari data.

BB : Nilai batas minimum yang diberikan.

BA : Nilai batas maksimum yang diberikan.

2.12 *Confusion Matrix*

Confusion matrix adalah tabel yang digunakan untuk menganalisis keakuratan sebuah metode klasifikasi untuk memprediksi kelas suatu data. Pada penelitian ini, data diklasifikasikan menjadi positif dan negatif atau disebut klasifikasi biner. Tabel 2.3 menunjukkan tabel *confusion matrix*.

Tabel 2. 2 *Confusion matrix*

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	TP (*True Positive*)	FN *(False Negative)*
Negatif	FP *(False Positive)*	TN (*True Negative*)

Pengukuran yang sering digunakan untuk menghitung ketepatan klasifikasi adalah akurasi, *specificity*, dan *sensitivity/recall*. Akurasi merupakan persentase dokumen yang teridentifikasi secara tepat dari total dokumen dalam proses klasifikasi. Akurasi digunakan untuk menghitung ketepatan klasifikasi sebuah dokumen yang mempunyai data yang *balanced* pada tiap kategorinya. Berikut merupakan rumus dalam menghitung akurasi :

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \text{ x } 100\% \tag{2.42}$$

Specificity / *precision* adalah nilai keserasian antara kelas data yang dihasilkan sistem dengan kelas sebenarnya. Perhitungan *precision* dapat dilihat pada persamaan dibawah *:*

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \text{ x } 100\% \tag{2.43}$$

Sensitivity / *recall* adalah nilai dari ketepatan banyaknya data yang berhasil dihasilkan sistem berdasarkan kelas sebenarnya. Perhitungan *recall* dapat dilihat pada persamaan dibawah :

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \text{ x } 100\% \tag{2.44}$$

F-measure adalah ukuran hubungan timbal balik antara *precision* dan *recall. F-measure* juga disebut sebagai alternatif dengan menggabungkan *precision* dan *recall*

menjadi satu ukuran evaluasi. Perhitungan *f-measure* dapat dilihat pada persamaan dibawah:

$$F = \frac{2 \text{ x } Specificity \text{ x } Sensitivity}{Specificity + Sensitivity} \quad (2.45)$$

Sedangkan untuk data *imbalanced*, pengukuran ketetapan klasifikasi yang digunakan adalah *g-mean. G-Mean* atau *geometric mean* merupakan rata-rata geometrik nilai *recall* dari data yang memiliki dua kategori (Sun, dkk. 2006). Dalam mengukur nilai performansi klasifikasi, *G-Mean* memiliki kelebihan yaitu nilai klasifikasi yang dihasilkan *robust*. Berikut merupakan rumus untuk mendapatkan nilai *G-Mean*. Selain *G-mean* juga digunakan nilai Area Under Curve (AUC). AUC merupakan indikator performansi kurva ROC (*Receiver Operating Characteristic*) yang dapat meringkas kinerja sebuah *classifier* menjadi satu nilai (Bekkar, Djemaa, & Alitouch, 2013).

$$G - mean = \sqrt{Sensitivity \text{ x } Specificity} \quad (2.46)$$

$$AUC = \frac{1}{2}(Sensitivity + Specificity) \quad (2.47)$$