ANALISIS SENTIMEN KEBIJAKAN MERDEKA BELAJAR KAMPUS MERDEKA MENGGUNAKAN MESIN VEKTOR PENDUKUNG DENGAN EKSTRAKSI FITUR WORD2VEC DAN PEMODELAN TOPIK *LATENT DIRICHLET ALLOCATION*

SKRIPSI

NURUL REZKI

H051181026

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

AGUSTUS 2022

ANALISIS SENTIMEN KEBIJAKAN MERDEKA BELAJAR KAMPUS MERDEKA MENGGUNAKAN MESIN VEKTOR PENDUKUNG DENGAN EKSTRAKSI FITUR WORD2VEC DAN PEMODELAN TOPIK *LATENT DIRICHLET ALLOCATION*

SKRIPSI

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

NURUL REZKI

H051181026

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

AGUSTUS 2022

LEMBAR PERNYATAAN KEOTENTIKAN

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

Analisis Sentimen Kebijakan Merdeka Belajar Kampus Merdeka Menggunakan Mesin Vektor Pendukung dengan Ekstraksi Fitur Word2Vec dan Pemodelan Topic *Latent Dirichlet Allocation*

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

Makassar, 16 Agustus 2022

Nurul Rezki

NIM H051181026

ANALISIS SENTIMEN KEBIJAKAN MERDEKA BELAJAR KAMPUS MERDEKA MENGGUNAKAN MESIN VEKTOR PENDUKUNG DENGAN EKSTRAKSI FITUR WORD2VEC DAN PEMODELAN TOPIK *LATENT DIRICHLET ALLOCATION*

Disetujui Oleh:

Pembimbing Utama,	Pembimbing Pertama,
Sri Astuti Thamrin, S.Si., M.Stat., Ph.D.	**Siswanto, S.Si., M.Si.**
NIP. 19740713 199903 2 001	**NIP. 19920107 201903 1 012**

Ketua Departemen Statistika

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

NIP. 19720117 199703 2 002

Pada 16 Agustus 2022

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh:

Nama	: Nurul Rezki
NIM	: H051181026
Program Studi	: Statistika
Judul Skripsi	: Analisis Sentimen Kebijakan Merdeka Belajar Kampus Merdeka Menggunakan Mesin Vektor Pendukung dengan Ekstraksi Fitur Word2vec dan Pemodelan Topik *Latent Dirichlet Allocation*

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

DEWAN PENGUJI

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 : Andi Kresna Jaya, S.Si., M.Si. (.....)

Ditetapkan di : Makassar

Tanggal : 16 Agustus 2022

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Segala puji hanya milik Allah *Subhanallahu Wa Ta'ala* atas segala limpahan rahmat dan hidayah-Nya yang telah diberikan kepada penulis sampai saat ini. Shalawat dan salam senantiasa tercurahkan kepada baginda Rasulullah *Shallallahu 'Alaihi Wa sallam*. *Alhamdulillaahirobbil'aalamiin*, berkat rahmat dan kemudahan yang diberikan oleh Allah *Subhanallahu Wa Ta'ala*, penulis dapat menyelesaikan skripsi yang berjudul "Analisis Sentimen Kebijakan Merdeka Belajar Kampus Merdeka Menggunakan Mesin Vektor Pendukung dengan Ekstraksi Fitur Word2vec dan Pemodelan Topik *Latent Dirichlet Allocation*" sebagai salah satu syarat memperoleh gelar sarjana pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

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 dalam bentuk 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 Sunre dan Ibunda tercinta Habibah yang telah membesarkan dan mendidik penulis dengan penuh kesabaran dan dengan limpahan cinta, kasih sayang dan doa kepada penulis, saudara-saudara penulis yaitu Sappe dan Masati serta kakak ipar Siti Aminah yang selalu membantu dan menjadi penyemangat untuk segera menyelesaikan masa studi penulis, serta keponakan tersayang penulis yaitu Muhammad Nur Fajri, Muhammad Afrijal, Muhammad Nurfadilah dan Ahmad Raihan Ashap yang selalu menghibur penulis.

Penghargaan yang tulus dan ucapan terima kasih dengan penuh keikhlasan juga penulis ucapkan kepada:

1. Bapak Prof. Dr. Ir. Jamaluddin Jompa, M.Sc., selaku Rektor Universitas Hasanuddin berserta 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, segenap Dosen Pengajar dan Staf 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 di tengah kesibukannya.
5. Bapak Siswanto, S.Si., M.Si. selaku pembimbing pertama penulis yang telah meluangkan waktunya di tengah kesibukan untuk memberikan arahan, pengetahuan, motivasi, bimbingan untuk penulis.
6. Ibu Dr. Anna Islamiyati, S.Si., M.Si. dan Bapak Andi Kresna Jaya, S.Si., M.Si. selaku tim penguji yang telah memberikan saran dan kritikan yang membangun dalam penyempurnaan penyusunan tugas akhir ini.
7. Sahabat tercinta penulis, A. Annisa Miftahul Sakinah, Alfiana Wahyuni, Fitra Damayanti, Naura Alfatiyya Arda dan Marsya Anggun Prisila yang telah menjadi sahabat terbaik yang senantiasa mendengarkan keluhan, memberikan dorongan, semangat dan motivasi dalam setiap keadaan sehingga penulis bisa mendapatkan lebih banyak inspirasi khususnya dalam menjalani pahit manisnya kehidupan perkuliahan.
8. Sahabat penulis sejak SMA, Rabihul Fauziah dan Zahra Nur Azizah, yang senantiasa menjadi rumah kedua bagi penulis di tengah kesibukan masing-masing.
9. Sahabat penulis, Yustika dan Sri Indriani Amil yang senantiasa memberikan dorongan serta senantiasa hadir untuk berbagi suka duka masa perkuliahan.
10. Sahabat “Ukhti”, Andi Sri Yulianti, Musdalifah dan Nurhidayah L yang senantiasa memberikan energi positif dan pembelajaran hidup yang berarti bagi penulis.
11. Komunitas Data Science Indonesia khususnya kepada Alfi Fauzia Hanifah, keluarga besar Data Science Indonesia Chapter Sulsel serta teman-teman alumni program MBKM Studi Independen PT. Microsoft Indonesia *learning track Data and Artificial Intelligence* terkhusus kelas DAI-006 yang

senantiasa memberikan dukungan dan inspirasi bagi penulis dalam menyelesaikan tugas akhir ini.

12. Teman-teman Statistika 2018, terkhusus kepada Adhiyaksa Prananda RS, Fadhil Al-Anshory, La Ade, Nehemia Millenium Payung, Taufiq Akbar, Viktor Liman, Hajratul Ashwad K dan Nur Anugrah Yusuf yang selalu membantu dan menjadi sosok guru bagi penulis, terima kasih atas kebersamaan, suka dan duka selama menjalani pendidikan di Departemen Statistika.
13. Keluarga besar INTEGRAL 2018, terkhusus kepada Akidah Amaliah, Juni Wahdaniyah, Abdul Jalil Saleh, Muh. Lutfi, Ahmad Ilham B dan Ardi S, 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 bersama.
14. Keluarga Mahasiswa FMIPA Unhas terkhusus anggota keluarga Himatika FMIPA Unhas dan Himastat FMIPA Unhas, terima kasih atas ilmu dan telah menjadi keluarga selama penulis kuliah di Universitas Hasanuddin.
15. 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 menyadari bahwa masih banyak kekurangan dalam skripsi ini, untuk itu dengan segala kerendahan hati penulis memohon maaf. Akhir kata, semoga tulisan ini memberikan manfaat untuk pembaca.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Makassar, 16 Agustus 2022

Nurul Rezki

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIK

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

Nama	: Nurul Rezki
NIM	: H051181026
Program Studi	: Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam
Jenis Karya	: Skripsi

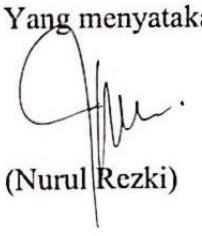
Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin **Hak Bebas Royalti Non-eksklusif** (***Non-exclusive Royalty- Free Right***) atas tugas akhir saya yang berjudul:

"Analisis Sentimen Kebijakan Merdeka Belajar Kampus Merdeka Menggunakan Mesin Vektor Pendukung dengan Ekstraksi Fitur Word2Vec dan Pemodelan Topik *Latent Dirichlet Allocation*"

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 16 Agustus 2022.

Yang menyatakan



(Nurul Rezki)

ABSTRAK

Analisis sentimen merupakan analisis data teks yang mengklasifikasikan data ke dalam sentimen positif dan negatif. Analisis sentimen dapat dilanjutkan dengan pemodelan topik untuk merepresentasikan topik yang dibahas pada setiap kelas sentimen. Penelitian ini bertujuan untuk memperoleh hasil klasifikasi sentimen terkait kebijakan Merdeka Belajar Kampus Merdeka di Twitter serta pemodelan topik pada kelas sentimen positif dan negatif. Algoritma klasifikasi yang digunakan adalah mesin vektor pendukung dengan ekstraksi fitur Word2Vec, sedangkan metode pemodelan topik menggunakan *Latent Dirichlet Allocation*. Data pada penelitian ini adalah *tweet* dengan kata kunci "Kampus Merdeka" yang diunggah di Twitter. Diperoleh hasil klasifikasi sentimen terdiri dari 648 *tweet* bersentimen positif dan 931 *tweet* bersentimen negatif dengan akurasi model klasifikasi 89.87%, presisi 91.20%, *recall* 84.44% dan *F-Measure* 87.68%, sedangkan pemodelan topik *Latent Dirichlet Allocation* pada kelas sentimen positif dan negatif masing-masing menghasilkan 5 topik dengan nilai *coherence* 0.44 pada pemodelan topik kelas sentimen positif dan 0.38 pada pemodelan topik kelas sentimen negatif. Klasifikasi sentimen menggunakan mesin vektor pendukung dengan ekstraksi fitur Word2Vec pada penelitian ini menghasilkan model yang baik, namun pemodelan topik *Latent Dirichlet Allocation* menghasilkan topik-topik yang sulit diinterpretasikan karena memiliki nilai *coherence* yang rendah.

Kata Kunci: Analisis Sentimen, Mesin Vektor Pendukung, Word2Vec, *Latent Dirichlet Allocation*, Merdeka Belajar Kampus Merdeka.

ABSTRACT

Sentiment analysis is a data text analysis that classifies data into positive and negative sentiments. Sentiment analysis can be continued with topic modeling to represent the topics discussed in each sentiment class. This research is to get the results of the sentiment classification Merdeka Belajar Kampus Merdeka policy on Twitter and topic modeling results in the positive and negative sentiment class using support vector machine classification algorithm with Word2Vec feature extraction and Latent Dirichlet Allocation topic modeling. Data in this research are tweets with the keyword "Kampus Merdeka" uploaded on Twitter. Sentiment classification results are 648 positive sentiment tweets and 931 negative sentiment tweets with classification model accuracy 89.87%, precision 91.20%, recall 84.44% and F-Measure 87.68%, while the Latent Dirichlet Allocation topic modeling results each 5 topics in the positive and negative sentiment class with 0.44 coherence value on the positive sentiment class topic modeling and 0.38 on the negative sentiment class topic modeling. Sentiment classification with support vector machine and Word2Vec feature extraction resulted a good classification model, but Latent Dirichlet Allocation topic modeling produced topics that were difficult to understand because they had low coherence value.

***Keywords**: Sentiment Analysis, Support Vector Machine, Word2Vec, Latent Dirichlet Allocation, Merdeka Belajar Kampus Merdeka Policy.*

DAFTAR ISI

HALAMAN SAMPUL i
HALAMAN JUDUL i
HALAMAN PERNYATAAN KEOTENTIKAN ii
HALAMAN PERSETUJUAN PEMBIMBING iv
HALAMAN PENGESAHAN v
KATA PENGANTAR vi
PERSETUJUAN PUBLIKASI KARYA ILMIAH viii
ABSTRAK x
ABSTRACT xi
DAFTAR ISI xii
DAFTAR GAMBAR xiv
DAFTAR TABEL xv
DAFTAR LAMPIRAN xv
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Tujuan Penelitian 3
1.4 Batasan Masalah 3
BAB II TINJAUAN PUSTAKA 4
2.1 *Text Mining* 4
2.2 Analisis Sentimen 4
2.3 Praproses Data Teks 4
2.4 Ekstraksi Fitur Word2Vec 5
2.5 *K-fold Cross Validation* 9
2.6 Mesin Vektor Pendukung 9
2.6.1 Klasifikasi pada *Linearly Separable Data* 10
2.6.2 Klasifikasi pada *Non-Linearly Separable Data* 14
2.7 *Confusion Matrix* 20
2.8 *Latent Dirichlet Allocation* 22
2.9 Kebijakan Merdeka Belajar Kampus Merdeka 23

2.10 Twitter 24
BAB III METODOLOGI PENELITIAN 25
3.1 Sumber Data 25
3.2 Struktur Data 25
3.3 Tahapan Analisis 26
BAB IV HASIL DAN PEMBAHASAN 27
4.1 Deskripsi Data 27
4.2 Praproses Data Teks 29
4.3 Ekstraksi Fitur Word2Vec 36
4.4 Klasifikasi Mesin Vektor Pendukung 37
4.4.1 *Data Training* dan *Data Testing* 37
4.4.2 Klasifikasi Mesin Vektor Pendukung Kernel *Radial Basis Function* 38
4.4.3 Model Klasifikasi Mesin Vektor Pendukung 38
4.4.4 *Confusion Matrix* pada Mesin Vektor Pendukug 39
4.5 Pemodelan Topik *Latent Dirichlet Allocation* 42
4.5.1 Pemodelan Topik pada Sentimen Positif 42
4.5.2 Pemodelan Topik pada Sentimen Negatif 45
BAB V KESIMPULAN DAN SARAN 48
5.1 Kesimpulan 48
5.2 Saran 48
DAFTAR PUSTAKA 49
LAMPIRAN 54

DAFTAR GAMBAR

Gambar 2.1 Arsitektur CBOW dan *Skip-Gram* 6
Gambar 2.2 Model Mesin Vektor Pendukung Linear 10
Gambar 2.3 Proses Kerja *Probabilistic Graphical Model* 22
Gambar 4.1 *Bar Chart* Kelas Sentimen Data Kebijakan MBKM 28
Gambar 4.2 Nilai *Coherence* Pemodelan Topik Sentimen Positif 44
Gambar 4.3 Nilai *Coherence* Pemodelan Topik Sentimen Negatif..... 46

DAFTAR TABEL

Tabel 2.1 Ilustrasi Pembagian Data dengan *K-fold Cross Validation* 9
Tabel 2.2 *Confusion Matrix* 21
Tabel 3.1 Contoh Struktur Data Penelitian 25
Tabel 4.1 Hasil *Data Crawling* 27
Tabel 4.2 Struktur Data Sebelum Praproses 29
Tabel 4.3 Stuktur Data Sebelum dan Setelah Penghapusan URL 30
Tabel 4.4 Stuktur Data Sebelum dan Setelah Penghapusan *Username* 30
Tabel 4.5 Stuktur Data Sebelum dan Setelah Penghapusan *Hashtag* 31
Tabel 4.6 Stuktur Data Sebelum dan Setelah Penghapusan Angka dan Tanda Baca 32
Tabel 4.7 Struktur Data Sebelum dan Setelah Proses *Case Folding* 33
Tabel 4.8 Struktur Data Sebelum dan Setelah Proses *Spelling Normalization* 33
Tabel 4.9 Struktur Data Sebelum dan Setelah Proses *Stemming* 34
Tabel 4.10 Struktur Data Sebelum dan Setelah Proses *Stopword Removal* ... 35
Tabel 4.11 Struktur Data Sebelum dan Setelah Proses *Tokenizing* 35
Tabel 4.12 Struktur Data Sebelum dan Setelah Praproses Data Teks 36
Tabel 4.13 Vektor Kata Dasar dengan Ekstraksi Fitur Word2Vec 37
Tabel 4.14 Proporsi *Data Training* dan *Data Testing* 37
Tabel 4.15 *Confusion Matrix* Model Klasifikasi Mesin Vektor Pendukung .. 39
Tabel 4.16 Data FN *Tweet* Kebijakan MBKM 40
Tabel 4.17 Data FP Tweet Kebijakan MBKM 41
Tabel 4.18 Sentimen Positif Terprediksi 43
Tabel 4.19 Pemodelan Topik LDA pada Sentimen Positif Kebijakan MBKM 45
Tabel 4.20 Sentimen Negatif Terprediksi 45
Tabel 4.20 Pemodelan Topik LDA pada Sentimen Negatif Kebijakan MBKM 47

DAFTAR LAMPIRAN

Lampiran 1 Ketetapan Model Klasifikasi Mesin Vektor Pendukung..... 54
Lampiran 2 Ketetapan Klasifikasi Mesin Vektor Pendukung Kernel RBF .. 63
Lampiran 3 Ketetapan Parameter Pada Pemodelan Topik LDA Sentimen Positif 64
Lampiran 4 Ketetapan Parameter Pada Pemodelan Topik LDA Sentimen Negatif..... 65

BAB I
PENDAHULUAN

1.1 Latar Belakang

Sentimen adalah pendapat seseorang tentang perasaan, sikap atau pikiran yang bisa diungkapkan. Sentimen individu terhadap suatu peristiwa, merek, produk atau perusahaan tertentu dapat diperoleh dari laporan berita, ulasan pengguna, pembaruan media sosial atau situs *microblogging* (Mäntylä dkk., 2018). Data terkait sentimen yang diperoleh melalui situs-situs web atau media sosial pada umunya berbentuk teks dan tidak terstruktur.

Text mining merupakan proses ekstraksi pola dari sejumlah data tidak terstruktur dan akan diperoleh pola-pola data, *trend* serta ekstraksi pengetahuan yang potensial dari data teks (Turban, 2011). Salah satu tujuan penggunaan *text mining* adalah analisis sentimen yaitu proses memahami, mengestrak dan mengolah data tekstual untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini terhadap sebuah masalah atau objek. Informasi Sentimen yang diperoleh dapat berupa sentimen positif atau negatif. (Zhang dkk., 2018).

Salah satu pendekatan untuk melakukan analisis sentimen adalah *supervised learning* yang merupakan proses *machine learning* yang membutuhkan pengawasan melalui persiapan *data training* yang telah melalui proses pelabelan sehingga performa dari pendekatan *supervised learning* sangat bergantung pada *data training* dan cara data tersebut diproses. Beberapa algoritma yang umum digunakan untuk melakukan analisis sentimen dengan pendekatan *supervised learn*ing adalah mesin vektor pendukung, Naive Bayes dan *Maximum Entropy*. Mesin vektor pendukung adalah salah satu algoritma yang banyak digunakan pada penelitian-penelitian terkait analisis sentimen karena menghasilkan akurasi yang baik dibandingkan algoritma atau metode klasifikasi yang lain. Salah satunya adalah analisis sentimen terhadap ulasan sebuah maskapai penerbangan menggunakan metode mesin vektor pendukung dan Naive Bayes yang menghasilkan akurasi tertinggi sebesar 82.48% untuk metode mesin vektor pendukung (Rahat dkk., 2020). Penelitian terkait lainnya yaitu “Analisis Sentimen Media Sosial Universitas Amikom Yogyakarta Sebagai Sarana Penyebaran Informasi Menggunakan Algoritma Klasifikasi SVM” (Maulana dkk., 2018) dan

"Sentimen Analisis Publik Terhadap Kebijakan *Lockdown* Pemerintah Jakarta Menggunakan Algoritma SVM" (Isnain dkk., 2021).

Data sentimen berupa data teks yang dianalisis dengan analisis sentimen perlu diubah ke bentuk numerik agar dapat dibaca oleh komputer. Mengubah data teks menjadi data numerik dapat dilakukan dengan ekstraksi fitur (Pravina dkk., 2019). Terdapat beberapa jenis ektsraksi fitur seperti *Term Frequency* (TF), *Term Frequency Inverse Document Frequency* (TF-IDF), Word2Vec, *Glove* dan sebagainya (Rusli dkk., 2019). Pada penelitian ini, ekstraksi fitur yang digunakan adalah Word2Vec yang merepresentasikan sebuah kata pada bentuk vektor (Aldiansyah dkk., 2019). Pada penelitian yang dilakukan oleh Naufal dan Setiawan (2021) tentang analisis sentimen di Twitter terkait kebijakan publik, dibandingkan dengan TF-IDF, ekstraksi fitur Word2Vec memberikan peningkatan akurasi pada metode SVM.

Analisis sentimen dapat dilanjutkan dengan pemodelan topik yaitu proses merepresentasikan topik yang dibahas dalam dokumen teks (Febrianta dkk., 2021). *Latent Dirichlet Allocation* (LDA) merupakan metode pemodelan topik populer yang digunakan untuk meringkas, melakukan klasterisasi, menghubungkan ataupun memproses data yang menghasilkan daftar topik (Rachman dan Pramana, 2020). Beberapa penelitian terkait LDA diantaranya adalah "*Exploring Public Response to COVID-19 on Weibo with LDA Topic Modeling and Sentiment Analysis*" (Xie dkk., 2020) dan "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter" (Rachman dan Pramana, 2020).

Salah satu isu yang saat ini ramai dibicarakan di berbagai media sosial maupun situs-situs web adalah kebijakan Merdeka Belajar Kampus Merdeka (MBKM) oleh Kementerian Pendidikan dan Kebudayaan Republik Indonesia. Berdasarkan artikel yang dirilis pada 29 Januari 2020 oleh tirto.id, sejak awal diluncurkan, program ini telah banyak mendapat reaksi pro maupun kontra oleh masyarakat (Prabowo, 2020). Reaksi tersebut disampaikan melalui media sosial seperti Twitter, Instagram, Facebook dan sebagainya. Twitter merupakan salah satu media sosial yang banyak mendapatkan perhatian masyarakat Indonesia. Selain itu, Twitter menyediakan *Application Programming Interface* (API) bagi pengguna

untuk mengembangkan sebuah aplikasi dengan data yang bersumber dari Twitter (Blanchette, 2008).

Berdasarkan uraian yang telah dipaparkan, maka pada penelitian tugas akhir ini akan dilakukan analisis sentimen terkait kebijakan MBKM di Twitter dengan metode mesin vektor pendukung menggunakan ekstraksi fitur Word2Vec dan pemodelan topik LDA.

1.2 Rumusan Masalah

Berdasarkan latar belakang, maka rumusan masalah dari penelitian ini adalah sebagai berikut:

1. Bagaimana hasil klasifikasi sentimen kebijakan MBKM di Twitter menggunakan mesin vektor pendukung dengan ekstraksi fitur Word2Vec?
2. Bagaimana hasil pemodelan topik LDA pada sentimen positif dan negatif terkait kebijakan MBKM di Twitter?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan penelitian ini adalah sebagai berikut:

1. Memperoleh hasil klasifikasi sentimen kebijakan MBKM di Twitter menggunakan mesin vektor pendukung dengan ekstraksi fitur Word2Vec.
2. Memperoleh hasil pemodelan topik LDA pada sentimen positif dan negatif terkait kebijakan MBKM di Twitter.

1.4 Batasan Masalah

Batasan masalah penelitian ini adalah menggunakan ekstraksi fitur Word2Vec dengan arsitektur *Skip-Gram* dan algoritma *negative sampling.*

BAB II
TINJAUAN PUSTAKA

2.1 *Text Mining*

Text mining adalah proses analisis terhadap data teks yang bersumber dari dokumen. Konsep *text mining* digunakan untuk melakukan klasifikasi dokumen tekstual agar dokumen-dokumen tersebut dapat diklasifikasikan sesuai dengan topik yang diinginkan. Dengan menggunakan *text mining*, kategori dokumen teks dapat diketahui melalui kata-kata yang terkandung di dalamnya (Herwijayanti dkk., 2018).

2.2 Analisis Sentimen

Analisis sentimen adalah studi komputasi opini, sentimen, emosi, penilaian dan sikap individu terhadap entitas seperti produk, layanan, organisasi, individu lain, masalah, peristiwa, topik serta atribut yang berkaitan (Zhang dkk., 2018). Tugas dasar analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen lalu menentukan pendapat yang dikemukakan dalam kalimat atau dokumen tersebut bersifat positif atau negatif (Kurniawan, 2017).

2.3 Praproses Data Teks

Praproses merupakan proses mempersiapkan data yang belum terstruktur menjadi data yang terstruktur sehingga dapat digunakan untuk proses berikutnya (Ipmawati dkk., 2017). Berikut tahapan-tahapan dari praproses data teks:

1. *Cleaning*, yaitu membersihkan data dari *noise* seperti *hashtag*, *username*, URL, dan tanda baca (Pertiwi, 2019).
2. *Case folding*, yaitu proses penyamaan *case* dalam sebuah dokumen. Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu peran *case folding* dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (dalam hal ini huruf kecil atau *lowercase*).
3. *Spelling normalization*, merupakan proses perbaikan atau subtitusi kata-kata yang salah eja atau disingkat dalam bentuk tertentu. Subtitusi kata dilakukan

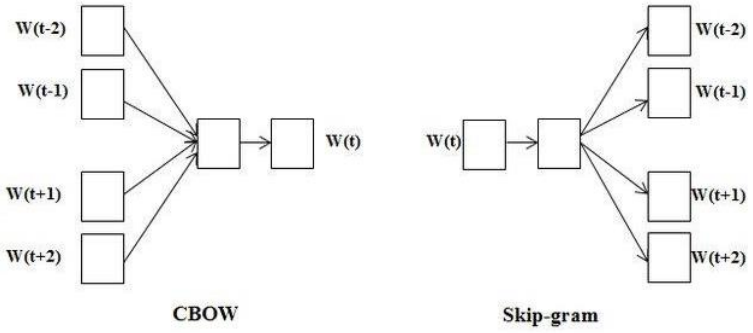
untuk menghindari jumlah perhitungan dimensi kata yang melebar (Khotimah, 2019).

4. *Stopword removal*, yaitu menghapus kata-kata yang sangat umum. Kata yang termasuk dalam *stopword* adalah yang, dan, di, itu, dengan, untuk, dari, dalam, akan, pada, ini, juga, saya, serta, adalah, bahwa, lain, kamu, dan masih banyak lagi (Jumeilah, 2017).
5. *Tokenizing*, yaitu proses memecah kalimat menjadi kata-kata. Proses tokenisasi mengandalkan karakter spasi pada dokumen teks untuk melakukan pemisahan. Hasil dari proses ini adalah kumpulan kata (Khotimah, 2019).
6. *Stemming* adalah tahapan mengubah kata menjadi kata dasar menurut kaidah Bahasa Indonesia yang benar dengan menghilangkan imbuhan, awalan serta akhiran kata (Yulita, 2021).

2.4 Ekstraksi Fitur Word2Vec

Ekstraksi fitur adalah proses mengubah kata dari suatu teks menjadi nilai numerik yang dapat dibaca oleh komputer (Pravina dkk., 2019). Word2Vec merupakan salah satu jenis ekstraksi fitur yang mengubah kata menjadi vektor. Kata-kata yang memiliki arti yang cukup mirip akan memiliki *output* yang berdekatan satu sama lain sehingga hubungan semantik antar kata dapat dipelajari (Triyantono dkk., 2021). Word2Vec merupakan model *neural network* dengan kata teks sebagai *input* dan ruang vektor sebagai *output*.

Terdapat dua jenis model arsitektur Word2Vec, yaitu *Skip-Gram* dan *Continuous Bag of Words* (CBOW) yang ditunjukkan pada Gambar 2.1. Arsitektur model *Skip-Gram* memprediksi kata konteks berdasarkan kata tengah, sedangkan model CBOW memprediksi kata tengah berdasarkan kata konteks (Rahman dkk., 2021). Kecepatan *training* CBOW lebih cepat dibandingkan dengan *Skip-Gram*, tetapi *Skip-Gram* lebih akurat dibandingkan dengan CBOW (Chang dkk., 2017).



Gambar 2.1 Arsitektur CBOW dan *Skip-Gram*

Berdasarkan Gambar 2.1, *input* dari model *Skip-Gram* adalah kata tengah dan *output* berupa kata konteks. Tujuannya adalah memaksimalkan peluang memprediksi kata konteks berdasarkan kata tengah dengan mengoptimalkan matriks pembobot. Secara matematis dapat dituliskan sebagai berikut:

$$\arg\max_{\boldsymbol{\theta}} \ln P(w_c \mid w_t; \boldsymbol{\theta})$$

dengan $\boldsymbol{\theta}$ sebagai matriks pembobot yang dioptimalkan, w_c adalah kumpulan kata konteks dan w_t adalah kata tengah. Matriks pembobot $\boldsymbol{\theta}$ diinisiasi secara acak dan akan dioptimalkan pada model *neural network*. Secara matematis, optimasi matriks pembobot $\boldsymbol{\theta}$ dituliskan sebagai berikut:

$$\boldsymbol{\theta}^{new} = \boldsymbol{\theta}^{old} - \eta \nabla J(\theta)$$

dengan $\boldsymbol{\theta}^{new}$ adalah matriks pembobot baru yang telah dioptimasi, $\boldsymbol{\theta}^{old}$ adalah matriks pembobot yang dioptimasi, η adalah *learning rate*, $\nabla J(\boldsymbol{\theta})$ gradien matriks pembobot dan $J(\boldsymbol{\theta})$ adalah fungsi tujuan.

Algoritma *gradien derivate* digunakan untuk mendapatkan turunan fungsi tujuan sehubungan dengan matriks pembobot $\boldsymbol{\theta}$. Secara umum, fungsi tujuan dari model *Skip-Gram* diperoleh dengan algoritma *softmax* yang secara matematis ditunjukkan pada persamaan berikut:

$$J(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-C\le j\le C, j\neq 0} \ln P(w_c|w_t;\boldsymbol{\theta})$$

dengan:

$J(\theta)$: Fungsi Tujuan

T, t : Indeks kata unik pada *corpus*

w_t : Kata tengah

w_c : Kata konteks

C : Ukuran *window*

Fungsi tujuan dengan algoritma *softmax* dinilai tidak efektif secara komputasi karena ukuran indeks T sangat besar. Untuk mengatasi hal tersebut, fungsi tujuan dapat diperoleh dengan algortima *negative sampling* menggunakan fungsi sigmoid. Fungsi sigmoid adalah fungsi yang digunakan untuk regresi logistik biner dengan persamaan umum sebagai berikut:

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{2.1}$$

dengan σ(x) adalah fungsi sigmoid dari x.

Seluruh kata dalam *corpus* akan dikategorikan menjadi dua kelas, yaitu positif dan negatif. Kata konteks akan dikategorikan sebagai kata positif dan kata lainnya akan dikategorikan sebagai kata negatif. Dengan algoritma *negative sampling*, untuk setiap kata konteks dari sebuah kata tengah, akan ditarik sebanyak K kata negatif sehingga akan lebih efektif secara komputasi karena hanya sebanyak $(K + 1)$ untuk setiap kata positif pada matriks pembobot $\boldsymbol{\theta}$ yang akan dioptimasi. Model *Skip-Gram* dengan algoritma *negative sampling* akan memaksimalkan peluang kata positif dan meminimalkan peluang kata negatif untuk setiap kata tengah. Secara matematis dituliskan sebagai berikut:

$$\begin{aligned}
&\arg\max_{\boldsymbol{\theta}} \ln\left(P\left(D = 1 \mid w, c_{pos}; \boldsymbol{\theta}\right) \prod_{c_{neg}} P\left(D = 0 \mid w, c_{neg}; \boldsymbol{\theta}\right)\right) \\
&\arg\max_{\boldsymbol{\theta}} \ln\left(P\left(D = 1 \mid w, c_{pos}; \boldsymbol{\theta}\right) \prod_{c_{neg}} 1 - P\left(D = 1 \mid w, c_{neg}; \boldsymbol{\theta}\right)\right) \\
&\arg\max_{\boldsymbol{\theta}} \ln P\left(D = 1 \mid w, c_{pos}; \boldsymbol{\theta}\right) + \prod_{c_{neg}} \ln\left(1 - P\left(D = 1 \mid w, c_{neg}; \boldsymbol{\theta}\right)\right) \\
&\arg\max_{\boldsymbol{\theta}} \ln P\left(D = 1 \mid w, c_{pos}; \boldsymbol{\theta}\right) + \sum_{c_{neg}} \ln\left(1 - P\left(D = 1 \mid w, c_{neg}; \boldsymbol{\theta}\right)\right)
\end{aligned} \tag{2.2}$$

dengan K merupakan indeks maksimum kata negatif. Menggunakan fungsi sigmoid pada Persamaan 2.1, maka Persamaan 2.2 dapat dituliskan sebagai berikut:

$$\begin{aligned}
&\arg\max_{\boldsymbol{\theta}} \ln\left(\frac{1}{1+\exp\left(-\boldsymbol{c}_{pos}(\boldsymbol{w})\right)}\right) + \sum_{c_{neg}} \ln\left(1 - \frac{1}{1+\exp\left(-\boldsymbol{c}_{neg}(\boldsymbol{w})\right)}\right) \\
&\arg\max_{\boldsymbol{\theta}} \ln\left(\frac{1}{1+\exp\left(-\boldsymbol{c}_{pos}(\boldsymbol{w})\right)}\right) + \sum_{c_{neg}} \ln\left(\frac{1}{1+\exp\left(\boldsymbol{c}_{neg}(\boldsymbol{w})\right)}\right) \\
&\arg\max_{\boldsymbol{\theta}} \ln \sigma\left(\boldsymbol{c}_{pos}(\boldsymbol{w})\right) + \sum_{c_{neg}} \ln \sigma\left(-\boldsymbol{c}_{neg}(\boldsymbol{w})\right)
\end{aligned}$$

Sehingga fungsi tujuan untuk algoritma *negative sampling* dengan fungsi sigmoid adalah sebagai berikut:

$$J(\boldsymbol{\theta}; w, c_{pos}) = -\ln \sigma\left(\boldsymbol{c}_{pos}(\boldsymbol{w})\right) - \sum_{c_{neg}} \ln \sigma\left(-\boldsymbol{c}_{neg}(\boldsymbol{w})\right)$$

dengan $\boldsymbol{w}$ adalah vektor kata input yang ekuivalen dengan $\boldsymbol{h}$ yang merupakan *projection layer* pada model *neural network* dengan ukuran dimensi N yang diinisasi. Optimasi matriks pembobot $\boldsymbol{\theta}$ pada model *Skip-Gram* dengan algoritma *negative sampling* adalah sebagai berikut:

$$\boldsymbol{\theta}^{new} = \boldsymbol{\theta}^{old} - \eta \frac{\partial \mathrm{J}}{\partial \boldsymbol{\theta}}$$

Penurunan fungsi sigmoid adalah Nilai $\frac{\partial \sigma}{\partial x} = \sigma(x)\big(1 - \sigma(x)\big)$ sehingga nilai $\frac{\partial \mathrm{J}}{\partial \boldsymbol{\theta}}$ adalah sebagai berikut:

$$\frac{\partial \mathrm{J}}{\partial \boldsymbol{\theta}} = \left(\sigma\left(\boldsymbol{c}_{pos}(\boldsymbol{h})\right) - 1\right) \frac{\partial\left(\boldsymbol{c}_{pos}(\boldsymbol{h})\right)}{\partial \boldsymbol{\theta}} + \sum_{c_{neg}} 1 - \sigma\left(-\boldsymbol{c}_{neg}(\boldsymbol{h})\right) \frac{\partial\left(-\boldsymbol{c}_{neg}(\boldsymbol{h})\right)}{\partial \boldsymbol{\theta}}$$

Karena $\boldsymbol{\theta}$ terdiri dari $\boldsymbol{W}_{input}$ berukuran $V \times N$ dan $\boldsymbol{W}_{output}$ berukuran $N \times V$ dengan V merupakan ukuran dimensi korpus yang digunakan, maka fungsi tujuan perlu dibedakan menjadi $\frac{\partial \mathrm{J}}{\partial W_{input}}$ dan $\frac{\partial \mathrm{J}}{\partial W_{output}}$ yang diuraikan sebagai berikut (Goldberg dan Levy, (2014):

Kondisi untuk $\boldsymbol{W}_{input}$

Turunan fungsi tujuan pada $\boldsymbol{W}_{input}$ dapat dituliskan sebagai berikut:

$$\frac{\partial \mathrm{J}}{\partial \boldsymbol{h}} = \left(\sigma\left(\boldsymbol{c}_{pos}(\boldsymbol{h})\right) - 1\right) \boldsymbol{c}_{pos} + \sum_{c_{neg}} 1 - \sigma\left(-\boldsymbol{c}_{neg}(\boldsymbol{h})\right) \boldsymbol{c}_{neg}$$

Maka optimasi nilai vektor kata input secara matematis ditunjukkan sebagai berikut:

$$\boldsymbol{w}^{(new)} = \boldsymbol{w}^{(old)} - \eta \left(\left(\sigma\left(\boldsymbol{c}_{pos}(\boldsymbol{h})\right) - 1\right) \boldsymbol{c}_{pos} + \sum_{c_{neg}} \left(1 - \sigma\left(-\boldsymbol{c}_{neg}(\boldsymbol{h})\right)\right) \boldsymbol{c}_{neg}\right)$$

Kondisi untuk $\boldsymbol{W}_{output}$

Sebanyak $K + 1$ vektor kata pada $\boldsymbol{W}_{output}$ akan diperbaharui yang terdiri dari K kata negatif dan 1 kata positif yaitu kata konteks. Turunan fungsi tujuan untuk kata positif dan kata negatif pada $\boldsymbol{W}_{output}$ adalah sebagai berikut:

$$\frac{\partial \mathrm{J}}{\partial \boldsymbol{c}_{pos}} = \left(\sigma\left(\boldsymbol{c}_{pos}(\boldsymbol{h})\right) - \mathbf{1}\right)\boldsymbol{h}$$

$$\frac{\partial \mathrm{J}}{\partial \boldsymbol{c}_{neg}} = 1 - \sigma\left(-\boldsymbol{c}_{neg}(\boldsymbol{h})\right)\boldsymbol{h}$$

Sehingga optimasi vektor kata pada $\boldsymbol{W}_{output}$ secara matematis dituliskan sebagai berikut:

$$\boldsymbol{c}_{pos}^{(new)} = \boldsymbol{c}_{pos}^{(old)} - \eta\left(\sigma\left(\boldsymbol{c}_{pos}(\boldsymbol{h})\right) - \mathbf{1}\right)\boldsymbol{h}$$

$$\boldsymbol{c}_{neg}^{(new)} = \boldsymbol{c}_{neg}^{(old)} - \eta\left(1 - \sigma\left(\boldsymbol{c}_{neg}(\boldsymbol{h})\right)\boldsymbol{h}\right)$$

2.5 *K-fold Cross Validation*

K-fold cross validation adalah salah satu metode yang digunakan untuk mempartisi data menjadi *data training* dan *data testing*. *K-fold cross validation* secara berulang-ulang membagi data menjadi *data training* dan *data testing* sehingga setiap data memperoleh kesempatan menjadi *data testing* (Alkaff dkk., 2021). *K* merupakan besar angka partisi data yang digunakan untuk *data training* dan *data testing*. Menurut Max Kuhn & Kjell Johnson (2013) tidak ada aturan formal dalam penentuan nilai *K*, namun nilai *K* yang bagus digunakan adalah 5 hingga 10. Tabel 2.1 menunjukkan ilustrasi pembagian data menggunakan *K-fold cross validation*.

Tabel 2.1 Ilustrasi Pembagian Data dengan *K-fold Cross Validation*

Percobaan 1	*Test*	*Train*	*Train*	*Train*	*Train*
Percobaan 2	*Train*	*Test*	*Train*	*Train*	*Train*
Percobaan 3	*Train*	*Train*	*Test*	*Train*	*Train*
Percobaan 4	*Train*	*Train*	*Train*	*Test*	*Train*
Percobaan 5	*Train*	*Train*	*Train*	*Train*	*Test*

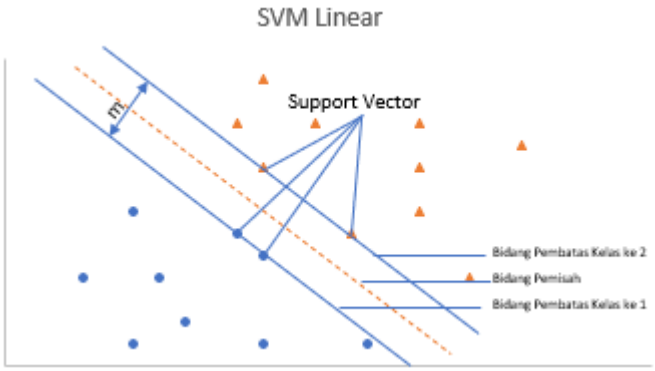
2.6 Mesin Vektor Pendukung

Mesin vektor pendukung merupakan salah satu metode klasifikasi pada *machine learning* dengan pendekatan *supervised learning* yang memprediksi kelas berdasarkan model atau pola dari hasil proses *training*. Klasifikasi dilakukan dengan mencari *hyperplane* atau garis pembatas (*decision boundary*) yang memisahkan satu kelas dengan kelas yang lain. *Hyperplane* yang optimal adalah

hyperplane yang memberikan margin optimal antara suatu sampel *data training* yang paling dekat dengan *hyperplane*. Kumpulan sampel *data training* disebut dengan *support vector* (Lutfiyanto, 2021). Mesin vektor pendukung digunakan untuk menyelesaikan masalah klasifikasi linier dan non-linier (Sari, 2017).

2.6.1 Klasifikasi pada *Linearly Separable Data*

Mesin vektor pendukung pada *linearly separable data* adalah penerapan metode mesin vektor pendukung 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 2.2 berikut.



Gambar 2.2 Model Mesin Vektor Pendukung Linear

Pada 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* sebagai berikut:

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

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

dengan **w** merupakan vektor normal terhadap *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:

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

Pada kasus ini akan dipisahkan dua *hyperplane* yang sejajar dengan *hyperplane* pertama akan membatasi kelas pertama sedangkan *hyperplane* kedua

akan membatasi kelas kedua, 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 dapat diperoleh nilai margin diantara dua *hyperplane* dengan menggunakan prinsip jarak antara dua 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, nilai dari *hyperplane* optimal dapat diperoleh melalui suatu solusi optimasi.

fungsi tujuan:

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

dengan kendala:

$$\boldsymbol{w^T x_i} + b \geq 1\ , \forall_i \epsilon\ Kelas\ Positif$$

$$\boldsymbol{w^T x_i} + b \leq -1\ , \forall_i \epsilon\ Kelas\ Negatif$$

atau dapat diringkas sebagai berikut:

fungsi tujuan:

$$max \frac{2}{||\boldsymbol{w}||} \quad (2.5)$$

dengan kendala:

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

Menurut Sari, 2017 memaksimalkan margin $\frac{1}{||\boldsymbol{w}||}$ sama dengan meminimumkan $\frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2$, maka Persamaan 2.5 dapat ditulis sebagai berikut:

fungsi tujuan:

$$min \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 \quad (2.6)$$

dengan kendala:

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

Nilai *hyperplane* optimal yang memisahkan data dengan margin maksimum dapat diperoleh dengan menyelesaikan kendala optimasi kuadratik. Permasalahan optimasi kasus data yang dapat dipisahkan secara linear pada Persamaan 2.6 dapat dituliskan dalam bentuk sebagai berikut:

fungsi tujuan:

$$\min_{w,b}(\boldsymbol{w}, b) = \frac{1}{2}\boldsymbol{w^T w} \tag{2.7}$$

dengan kendala:

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

Untuk mendapatkan solusi dari permasalahan optimasi bentuk primal pada Persamaan 2.7 maka akan digunakan metode *Lagrange* dengan kendala pertidaksamaan atau dikenal sebagai *Karush Kuhn Tucker* (KKT). Berdasarkan Persamaan 2.7 maka diperoleh persamaan *lagrange* untuk kasus ini menjadi:

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

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

$$L_p(\boldsymbol{w}, b, \alpha) = \left(\frac{1}{2}\boldsymbol{w^T w}\right) - \sum_{i=1}^{n} \alpha_i\, [y_i(\, \boldsymbol{w^T x_i} + b) - 1] \tag{2.8}$$

Peubah α_i merupakan *Lagrange Multiplier* atau pengali *Lagrange*. Nilai dari *Lagrange Multiplier* ini adalah $\alpha_i > 0$. Fungsi Persamaan 2.8 akan diminimalkan terhadap $\boldsymbol{w}$ dan $\boldsymbol{b}$ serta dimaksimalkan terhadap peubah α (Sari, 2017). Turunan pertama dari fungsi pada Persamaan 2.8 adalah sebagai berikut:

Kondisi 1

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

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

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

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

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

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

Kondisi 2

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

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

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

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

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

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

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \; y_i \; \boldsymbol{x_i} \qquad (2.9)$$

Tampak bahwa fungsi tujuan pada Persamaan 2.6 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 2.7, maka akan disubtitusikan Persamaan 2.9 ke dalam Persamaan 2.8 sebagai berikut:

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

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

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

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

$$= \frac{1}{2}\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j y_i y_j \boldsymbol{x}_i^T \, \boldsymbol{x}_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j y_i y_j \boldsymbol{x}_i^T \, \boldsymbol{x}_j + \sum_{i=1}^{n} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j y_i y_j \boldsymbol{x}_i^T \, \boldsymbol{x}_j$$

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:

$$\max_{\alpha} L_p(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \quad (2.10)$$

dengan kendala:

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

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

dengan demikian, nilai α_i dapat ditemukan dengan menyelesaikan kasus optimasi pada Persamaan 2.10 dan nilai $\boldsymbol{w}$ akan diperoleh dengan mensubstitusikan α_i pada persamaan 2.9. Selanjutnya fungsi *hyperplane* yang optimal pada kasus *linear separable* dapat terbentuk persaaan berikut:

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

dengan x_d merupakan data yang akan diklasifikasikan, α_i merupakan solusi optimasi dari masalah optimasi pada Persamaan 2.10 dan *b* diperoleh dengan cara berikut:

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

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

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

2.6.2 Klasifikasi pada *Non-Linearly Separable Data*

Klasifikasi data yang tidak dapat dipisahkan secara linier memerlukan modifikasi pada formula mesin vektor pendukung agar dapat menemukan solusinya. Dalam praktiknya, jarang ditemui *data training* 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:

$$y_i\Big((\boldsymbol{w^T x_i}) + b\Big) + \xi_i \geq 1$$

$$y_i\Big((\boldsymbol{w^T x_i}) + b\Big) \geq 1 - \xi_i$$

atau dapat dituliskan:

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

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

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, 2017). Sehingga formula untuk mendapatkan *hyperplane* optimal menjadi:

fungsi tujuan:

$$min\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right) \quad (2.11)$$

dengan kendala:

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

dengan C merupakan parameter yang menetukan besarnya penalti akibat kesalahan dalam klasifikasi data yang 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.

Berdasarkan Persamaan 2.11, 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. Penggunaan peubah *slack* bertujuan untuk mengatasi kasus ketidaklayakan (*infeasibility*) dari kendala tersebut. Untuk meminimalkan nilai dari peubah *slack* maka digunakan nilai dari parameter C (Sari, 2017).

Dengan menggunakan teori optimasi, maka didapatkan:

fungsi tujuan:

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

dengan kendala:

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

atau dapat dituliskan:

fungsi tujuan:

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

dengan kendala:

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

Untuk mendapatkan solusi dari permasalahan optimasi bentuk primal pada Persamaan 2.22, maka seperti pada kasus *linear separable* akan digunakan *KTT*. Namun pada kasus *non-linear separable* ini akan memiliki dua pengali *Lagrange* untuk kasus ini menjadi:

$$\begin{aligned} L_p(\boldsymbol{w}, b, \xi, \alpha, \beta) &= (\boldsymbol{w}, b, \xi) + \sum_{i=1}^{n}\alpha_i\left[1 - \xi_i - y_i\left(\boldsymbol{w^T x_i} + b\right)\right]\sum_{i=1}^{n}\beta_i(-\xi_i) \\ &= (\boldsymbol{w}, b, \xi) - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\boldsymbol{w^T x_i} + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n}\beta_i(-\xi_i) \\ &= \left(\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i\right) - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\boldsymbol{w^T x_i} + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n}\beta_i(\xi_i) \\ &= \frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w^T x_i} + b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i) \end{aligned} \quad (2.12)$$

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 2.9 akan diminimalkan terhadap peubah $\boldsymbol{w}, b$ dan ξ serta harus dimaksimalkan terhadap peubah α dan β (Sari, 2017). Berikut turunan pertama dari fungsi Persamaan 2.12 yaitu:

Kondisi 1

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

$$\begin{aligned} \frac{\partial L_p}{\partial b} &= \frac{\partial L_p(\boldsymbol{w},b,\xi,\alpha,\beta)}{\partial b} \\ &= \frac{\partial\{(\boldsymbol{w},b,\xi) - \sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w^T x_i}+b)-1+\xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i)\}}{\partial b} \\ &= \frac{\partial\left\{\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w^T x_i}+b)-1+\xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i)\right\}}{\partial b} \\ &= \frac{\partial\left\{\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w^T x_i}+b)-1+\xi_i] - \sum_{i=1}^{n}\beta_i(\xi_i)\right\}}{\partial b} \\ &= \frac{\partial\left\{\frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{w^T x_i} - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\beta_i(\xi_i)\right\}}{\partial b} \\ &= 0 + 0 - 0 - \sum_{i=1}^{n}\alpha_i y_i + 0 - 0 - 0 \\ 0 &= \sum_{i=1}^{n}\alpha_i y_i \end{aligned} \quad (2.13)$$

Kondisi 2

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

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

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

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

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

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

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

$$\boldsymbol{w} = \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x_i} \qquad (2.14)$$

Kondisi 3

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

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

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

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

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

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

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

$$C = \alpha_i + \beta_i \qquad (2.15)$$

Untuk memperoleh bentuk dual, maka akan disubtitusikan Persamaan 2.13, 2.14 dan 2.15 ke dalam Persamaan 2.13.

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

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

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

$$\sum_{i=1}^{n}(\ C - \ \alpha_i)\xi_i$$

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

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

$$= \ \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{x_i^T x_j} \qquad (2.16)$$

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:

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

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

$$f(x_d) = \ \sum_{i\in S}\alpha_i y_i \boldsymbol{x_i^T x} + b \qquad (2.17)$$

dengan S merupakan data yang akan diklasifikasikan, α_i merupakan solusi optimal dari masalah optimasi (2.5) dan $\boldsymbol{b}$ dicari dengan formula. S adalah himpunan indeks *support vector*. Karena α_i tidak nol untuk *support vector* maka penjumlahan di (2.17) ditambahkan hanya untuk *support vector*. Untuk α_i tak berhingga, maka:

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

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

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

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:

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

Pada data pelatihan untuk kasus *non-separable*, 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:

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

dengan menggunakan teori optimasi, maka didapatkan:

fungsi tujuan:

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

dengan kendala:

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

$$\xi_i \geq 0, i = 1,2, \ldots, n$$

Untuk menurunkan permasalahan (2.15), (2.16) dan (2.17) digunakan *Lagrange Multiplier* dengan nilai $\alpha_i \geq 0$. Optimal poin akan ada di dalam *saddle point* dari *Lagrange function* menjadi:

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

Kemudian dilakukan turunan pertama yaitu sebagai berikut:

Kondisi 1

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

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

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

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

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

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

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

Kondisi 2

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

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

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

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

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

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

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

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

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

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

$$f(x_d) = \sum_{i=1}^{n}\alpha_i y_i K\,(x_i, x_d) + b \qquad (2.18)$$

dan b diperoleh:

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

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

$$\mathrm{K}(x_i, x_j) = exp\left(-\frac{(x_i - x_j)^T(x_i - x_j)}{2\,\gamma^2}\right) \qquad (2.19)$$

3. *Sigmoid*

$$\mathrm{K}\,(x_i, x_j) = tahn(\gamma x_i^T + r)$$

2.7 *Confusion Matrix*

Confusion matrix adalah metode pengukuran kinerja yang menunjukkan ringkasan hasil prediksi pada masalah klasifikasi yang menunjukkan empat

kombinasi nilai prediksi dan aktual yang digunakan untuk menghitung kinerja sebuah model. *Confusion matrix* ditunjukkan pada Tabel 2.2.

Tabel 2.2. *Confusion Matrix*

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

Berikut adalah penjelasan dari masing-masing nilai tersebut:

1. TP yaitu kondisi ketika prediksi (*predicted values*) bermakna positif dan hasilnya (*actual values*) benar.
2. TN yaitu kondisi ketika prediksi (*predicted values*) bermakna negatif dan hasilnya (*actual values*) benar.
3. FP yaitu kondisi ketika prediksi (*predicted values*) bermakna positif dan hasilnya (*actual values*) salah.
4. FN yaitu kondisi ketika prediksi (*predicted values*) bermakna negatif dan hasilnya (*actual values*) salah.

Berdasarkan nilai-nilai tersebut, dapat dihitung performa sebagai berikut (Dellia dan Tjahyanto, 2017):

1. *Accuracy*, merupakan jumlah yang diprediksi benar.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.20}$$

2. *Precision*, merupakan jumlah prediksi yang benar dari semua kelas positif yang diprediksi.

$$\text{Presisi} = \frac{TP}{TP + FP} \tag{2.21}$$

3. *Recall*, merupakan jumlah prediksi yang benar dari semua kelas positif.

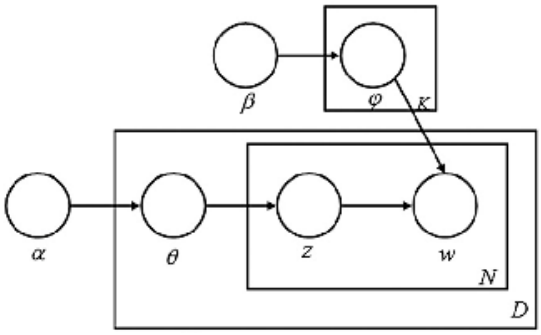
$$Recall = \frac{TP}{TP + FN} \tag{2.22}$$

4. *F-measure* untuk mengukur *recall* dan *precision* secara bersamaan.

$$F - Measure = \frac{2 \times Recall \times Presicion}{Recall + Presicion} \tag{2.23}$$

2.8 *Latent Dirichlet Allocation*

Ide dasar yang diusulkan metode LDA adalah setiap dokumen direpresentasikan sebagai campuran acak atas topik, setiap topik memiliki karakter yang ditentukan berdasarkan distribusi kata-kata yang terdapat di dalamnya. LDA membutuhkan pendefinisian jumlah topik yang akan dihasilkan oleh model (Rachman dan Pramana, 2020). Cara kerja LDA digambarkan menggunakan grafik yang disebut *probabilistic graphical model* yang ditunjukkan pada Gambar 2.3.



Gambar 2.3 Proses Kerja *Probabilistic Graphical Model*

Pada Gambar 2.2, α dan β merupakan parameter distribusi topik dari dokumen terlebih dahulu atau biasa disebut parameter *prior dirichlet* dan parameter distribusi kata dari sebuah topik. Nilai keduanya merupakan bilangan riil positif yang dapat dituliskan $0 \leq \alpha, \beta \leq 1$. Semakin tinggi nilai α menunjukkan bahwa setiap dokumen mengandung sebagian besar topik dan sebaliknya menunjukkan dokumen memiliki kemungkinan diwakili oleh beberapa topik. Sementara semakin tinggi nilai β menunjukkan suatu topik mengandung campuran sebagian besar kata-kata, sebaliknya suatu topik hanya mengandung campuran dari beberapa kata. Distribusi topik dari dokumen (α) mengakibatkan adanya nilai θ sebagai kumpulan campuran topik yang berbentuk matriks probabilitas topik terhadap dokumen seperti pada matriks berikut:

$$A = \begin{bmatrix} \theta_{11} & \cdots & \theta_{1K} \\ \vdots & \ddots & \vdots \\ \theta_{D1} & \cdots & \theta_{DK} \end{bmatrix}$$

θ_{DK} menujukkan probabilitas topik ke$-K$ pada dokumen ke$-D$. Dari kumpulan campuran topik (θ), dapat dipisahkan masing-masing topik (z) dari campuran topik tersebut. Sehingga diperoleh sebuah matriks baru yang berisikan nilai probabilitas kata terhadap topik untuk masing-masing dokumen sebagai berikut:

$$B = \begin{bmatrix} z_{11} & \cdots & z_{1K} \\ \vdots & \ddots & \vdots \\ z_{N1} & \cdots & z_{KN} \end{bmatrix}$$

Nilai z_{KN} menunjukkan topik ke$-K$ pada kata ke$-N$. Probabilitas topik yang diperoleh (z) dan distribusi kata pada topik (β) menghasilkan probabilitas kata-kata yang muncul sebagai hasil akhir pembentukan model (w), sehingga hasil dari model satu dokumen ini akan memunculkan kata-kata dari kelompok yang terbentuk, kata-kata ini dapat membantu dalam pendefinisian kategori setiap kelompok. Sehingga total probabilitas berdasarkan grafik model LDA secara matematis dapat dituliskan sebagai berikut:

$P(w, z, \theta|\alpha, \beta) = \prod_{d=1}^{D} P(\theta_j|\alpha) \prod_{k=1}^{K} P(\phi_k|\beta) \prod_{n=1}^{N} P(z_{dn}|\theta_j) P(w_{dn}|\phi, z_{dn})$

dengan d dan D menunjukkan indeks dokumen, k dan K menunjukkan indeks topik, n dan N menujukkan indeks kata dalam korpus serta ϕ merupakan korpus untuk kumpulan dokumen yang terjadi akibat adanya β seperti pada Gambar 2.2 (Rangkuti, 2020).

Evaluasi topik yang dihasilkan melalui LDA dilakukan dengan melihat nilai *coherence* topik yaitu seberapa mudah topik tersebut diinterpretasikan menggunakan *Point-wise Mutual Information* (PMI) *Topic Coherence* sebagai berikut:

$$PMI(k) = \sum_{j=2}^{N} \sum_{i=2}^{j-1} log \left(\frac{P(w_i, w_j)}{P(w_i) P(w_j)} \right)$$

dengan N adalah jumlah kata teratas pada topik k, $P(w_i, w_j)$ adalah peluang munculnya kata ke-i dan kata ke-j secara bersama-sama pada dokumen serta $P(w_i)$ dan $P(w_j)$ adalah peluang munculnya kata ke-i dan kata ke-j dalam dokumen (Yao dkk., 2017).

2.9 Kebijakan Merdeka Belajar Kampus Merdeka

Merdeka Belajar Kampus Merdeka (MBKM) merupakan salah satu kebijakan Menteri Pendidikan dan Kebudayaan Nadiem Makarim. Ada dua konsep yang esensial yaitu "Merdeka Belajar" dan "Kampus Merdeka". Konsep merdeka belajar bermakna adanya kemerdekaan berpikir. Kampus merdeka merupakan upaya untuk melepaskan belenggu untuk bisa bergerak lebih mudah. Tujuan kebijakan MBKM adalah meningkatkan kompetensi lulusan, baik *soft skills* maupun *hard skills* agar

lebih siap dan relevan dengan kebutuhan zaman. Bentuk kegiatan pembelajaran mengacu pada Permendikbud No. 3 Tahun 2020 Pasal 15 Ayat 1 bahwa dilakukan pada delapan bentuk program yang meliputi pertukaran pelajar, magang/praktik kerja, mengajar di instansi pendidikan, proyek di desa, penelitian/riset, kegiatan kewirausahaan, studi/proyek independen dan proyek kemanusiaan (Fuadi dan Aswita 2021).

2.10 Twitter

Twitter didirikan pada Maret 2006 oleh Jack Dorsey dan situs jejaring sosialnya diluncurkan pada Juli 2006. Sejak diluncurkan, Twitter telah menjadi salah satu dari 10 situs yang paling sering dikunjungi di internet (Zuhdi dkk., 2019). Twitter menyediakan *Application Programming Interface* (API) yang memungkinkan pengguna untuk mengakses dan mendapatkan informasi mengenai *tweet*, profil pengguna, data pengikut dan lainnya. Hal tersebut menjadikan Twitter sebagai *microblog* yang banyak diminati perusahaan, organisasi, maupun individu dalam mendapatkan opini publik mengenai suatu topik tertentu (Kurniawan, 2017).

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan pada penelitian ini merupakan data primer yang diperoleh dari Twitter berupa *tweet* berbahasa Indonesia dengan kata kunci "Kampus Merdeka" yang diunggah pada tanggal 20 Januari 2020 hingga 31 Maret 2022. Data sebanyak 10000 *tweet* yang digunakan berbentuk teks.

3.2 Struktur Data

Struktur data yang digunakan dalam penelitian ini setelah praproses pada data teks *tweet* terdiri dari variabel prediktor yaitu kata dasar setiap *tweet* dan variabel respon yaitu klasifikasi sentimen *tweet* (positif dan negatif). Data klasifikasi pada metode mesin vektor pendukung dibagi menjadi *data training* dan *data testing* dengan perbandingan 80:20 menggunakan 10-*fold cross validation*. Tabel 3.1 menunjukkan contoh struktur data penelitian sebelum praproses.

Tabel 3.1 Contoh Struktur Data Penelitian

No	*Tweet*	Sentimen
1	@Batutuo terus berjuang wujudkan kampus merdeka ...disain program2nya agar dihasilkan SDM yg berkwalitas di Era 4.0	Positif
2	Program Kampus Merdeka ini keren banget sih, ngasih harapan buat perubahan sistem pendidikan Indonesia	Positif
3	Ngiri saya sama program kampus merdeka	Positif
⋮	⋮	⋮
1579	labelnya kampus merdeka, tapi sesungguhnya sivitasnya terjajahnya sama berlapis-lapis regulasi	Negatif