ANALISIS SENTIMEN KEBENCANAAN DI INDONESIA MENGGUNAKAN *SUPPORT VECTOR MACHINE* BERBASIS *FASTTEXT* DENGAN PARAMETER OPTIMASI *FIREFLY*

SKRIPSI

FADILAH AMIRUL ADHEL

H051191064

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2023

ANALISIS SENTIMEN KEBENCANAAN DI INDONESIA MENGGUNAKAN *SUPPORT VECTOR MACHINE* BERBASIS *FASTTEXT* DENGAN PARAMETER OPTIMASI *FIREFLY*

SKRIPSI

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

FADILAH AMIRUL ADHEL
H051191064

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
AGUSTUS 2023

LEMBAR PERNYATAAN KEOTENTIKAN

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

ANALISIS SENTIMEN KEBENCANAAN DI INDONESIA MENGGUNAKAN *SUPPORT VECTOR MACHINE* BERBASIS *FASTTEXT* DENGAN PARAMETER OPTIMASI *FIREFLY*

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

Makassar, 01 Agustus 2023

FADILAH AMIRUL ADHEL
H051191064

ANALISIS SENTIMEN KEBENCANAAN DI INDONESIA MENGGUNAKAN *SUPPORT VECTOR MACHINE* BERBASIS *FASTTEXT* DENGAN PARAMETER OPTIMASI *FIREFLY*

Disetujui oleh:

Pembimbing Utama	Pembimbing Pertama
Sri Astuti Thamrin, S.Si., M.Stat., Ph.D.	**Siswanto, S.Si., M.Si.**
NIP. 197407131999032001	NIP. 199201072019031012

Ketua Program Studi

Dr. Anna Islamiyati, S.Si., M.Si.
NIP. 197708082005012002

Pada 01 Agustus 2023

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh:

Nama	: Fadilah Amirul Adhel
NIM	: H051191064
Program Studi	: Statistika
Judul skripsi	:Analisis Sentimen Kebencanaan Di Indonesia Menggunakan *Support Vector Machine* Berbasis *Fasttext* Dengan Parameter Optimasi *Firefly*

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Fakultas Matematikan 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 : Dra. Nasrah Sirajang, M.Si (.....)
4. Anggota : Sitti Sahriman, S.Si., M.Si. (.....)

Ditetapkan di : Makassar

Tanggal : 01 Agustus 2023

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Allah SWT atas limpahan rahmat dan karunia-Nya sehingga sehingga penyusunan skripsi ini dapat terselesaikan dengan baik. Skripsi dengan judul "Analisis Sentimen Kebencanaan Di Indonesia Menggunakan *Support Vector Machine* (SVM) Berbasis *Fasttext* Dengan Parameter Optimasi *Firefly*" ini merupakan salah satu rangkaian syarat akademik yang harus dipenuhi untuk memperoleh Gelar Sarjana Sains pada Program Studi Statistika, Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin.

Skripsi ini merupakan penelitian yang bertujuan untuk Memperoleh hasil pembobotan data teks menjadi vektor numerik menggunakan model *fasttext* dan Mendapatkan hasil klasifikasi sentimen kebencanaan Indonesia menggunakan SVM optimasi *firefly* dengan ekstraksi *fasttext* dan hasil evaluasi model menggunakan *accuracy*. Metode penelitian ini melibatkan algoritma *Support Vector Machine* dan optimasi *firefly* yang kemudian diharapkan dapat memberikan wawasan dan referensi kepada pembaca dalam klasifikasi data dengan menggunakan metode *machine learning*.

Penulis menyadari bahwa penyelesaian skripsi ini tidak terlepas dari bantuan dan dorongan yang diberikan oleh berbagai pihak yang secara konsisten memberikan bantuan baik secara moril maupun materil. Meskipun penulis memiliki keterbatasan dalam kemampuan dan pengetahuan, namun berkat bantuan dan dukungan tersebut, penulis berhasil menyelesaikan skripsi ini. Oleh karena itu, penulis ingin mengucapkan terima kasih yang sebesar-besarnya dan penghargaan yang tulus kepada semua pihak yang terlibat. Oleh karena itu, dengan penuh kesadaran dan kerendahan hati, pada kesempatan ini perkenankanlah penulis menyampaikan ucapan terima kasih dan penghargaan setinggi-tingginya kepada yang terhormat:

1. Terima kasih dan apresiasi yang setinggi-tingginya kepada diri sendiri yang telah berusaha dengan gigih dan tekun selama proses penyelesaian skripsi ini. Tidak mudah untuk melewati tantangan dan rintangan yang dihadapi selama proses penyusunan, namun penulis membuktikan keberhasilannya dengan terselesaikannya skripsi ini.

2. Terima kasih yang tak terhingga kepada kedua orang tua tercinta atas dukungan dan pengorbanan yang tidak bisa penulis ungkapkan dengan kata-kata terhadap proses penyelesaian skripsi ini.
3. Terima kasih kepada Prof. Dr. Ir. Jamaluddin Jompa, M.Sc., selaku Rektor Universitas Hasanuddin beserta seluruh staf jajarannya.
4. Terima kasih kepada Bapak Dr. Eng. Amiruddin, selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin beserta seluruh staf jajarannya.
5. Terima kasih kepada Ibu Dr. Anna Islamiyati, S.Si., M.Si., , selaku Ketua Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin beserta seluruh staf jajarannya.
6. Terima kasih kepada Ibu Sri Astuti Thamrin, S.Si., M.Stat., Ph.D sebagai pembimbing utama yang telah bersedia untuk meluangkan waktu untuk memberi masukan serta memberikan solusi atas semua permasalahan yang ada pada penelitian ini.
7. Terima kasih kepada Pak Siswanto, S.Si., M.Si., sebagai pembimbing pertama yang telah bersedia untuk meluangkan waktu untuk senantiasa menerima kendala penyusunan dan memberikan koreksi serta solusi yang terbaik.
8. Terima kasih kepada Ibu Dra. Nasrah Sirajang, M.Si. dan Sitti Sahriman, S.Si., M.Si. sebagai dosen penguji yang telah bersedia meluangkan waktunya untuk memberikan penilaian dan masukan terhadap skripsi ini.
9. Terima kasih kepada segenap jajaran dosen matakuliah dan staf Departemen Statistika yang telah banyak membantu, memberikan ilmu-ilmunya, serta berbagai kemudahan lainnya yang diberikan selama menempuh pendidikan sarjana di Departemen Statistika.
10. Terima kasih kepada teman-teman dekat di Istana Tercinta yang setia menemani penulis dari masa kuliah perdana sampai akhir perkuliahan. Terima kasih kanda-kanda: Arief Rahman Nur yang selalu mau direpotkan atas jasa transportasinya, Agus Hermawan, Muhammad Syamsul Bahri, Muhammad Ferdiansyah, Andi Muhammad Rajab, dan Sapriadi Rasyid.
11. Terima kasih kepada semua teman-teman di angkatan Statistika 2019 yang telah memberikan banyak ilmu dan pengalamannya.

12. Terima kasih yang setinggi-tingginya kepada seluruh pihak yang mungkin tidak sempat penulis sebutkan satu persatu. Terima kasih atas segala dukungan, partisipasi, dan apresiasinya yang diberikan kepada penulis.

Penulis juga menyadari bahwa skripsi ini masih jauh dari kata sempurna, namun ini hasil terbaik yang dapat diberikan oleh penulis pada penelitian ini. Oleh karena dengan segala kerendahan hati penulis mengucapkan permohonan maaf yang sebesar-besarnya. Akhir kata, semoga tulisan ini dapat memberikan manfaat untuk berbagai pihak.

Makassar, 01 Agustus 2023
Penulis

FADILAH AMIRUL ADHEL
NIM. H051191064

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIK

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

Nama	: Fadilah Amirul Adhel
NIM	: H051191064
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 Kebencanaan Di Indonesia Menggunakan *Support Vector Machine* Berbasis *Fasttext* Dengan Parameter Optimasi *Firefly*"

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, 01 Agustus 2023
Yang menyatakan,

FADILAH AMIRUL ADHEL
NIM. H051191064

ABSTRAK

Analisis sentimen adalah proses untuk melakukan analisis terhadap opini, sentimen, penilaian serta emosi dari pernyataan seseorang terhadap suatu domain atau juga merupakan proses untuk mengekstrak dan mengolah data berupa teks. SVM adalah sebuah teknik *Supervised Machine learning* yang berfungsi untuk menganalisa suatu data untuk klasifikasi. Penelitian ini bertujuan untuk Memperoleh *vector* numerik menggunakan *fasttext* yang kemudian diklasifikasi dengan *metode Support Vector Machine* (SVM). SVM tidak dapat memilih parameter yang sesuai sehingga penggunaan parameter menjadi tidak optimal, untuk memperoleh parameter yang optimal maka dilakukan optimasi *firefly* supaya mendapatkan hasil klasifikasi yang lebih baik. Data pada penelitian ini adalah *tweet* dengan kata kunci "Kebencanaan Indonesia" yang di *crawling* menggunakan aplikasi *twitter.* Hasil dari penelitian ini adalah mendapatkan 128 dimensi yang membentuk bobot setiap kata. Hal ini berarti setiap kata direpresentasikan dalam ruang vektor 128 dimensi dan nilai persentase evaluasi model klasifikasi SVM optimasi *firefly*: *accuracy* 89%. Sehingga, dapat disimpulkan bahwa metode klasifikasi SVM menunjukkan persentase performa klasifikasi yang cukup baik dibandingkan dengan model SVM tanpa optimasi dengan hanya mendapatkan *accuracy* sebesar 61%.

Kata Kunci: Analisis sentimen, *fasttext,* klasifikasi, optimasi *firefly*, *support vector machine*

ABSTRACT

Analysis sentiment is the process for do analysis to opinion, sentiment, judgment as well as emotion from statement somebody to a domain or is also a process for extract and processing data in the form of text. SVM is A technique *Supervised Machine learning* that works For analyze data for classification. This research aims to Obtain *vectors* numeric use *fasttext* which later classified with *Support Vector Machine* (SVM) *method* . SVM ca n't choose the appropriate parameters so that the use of parameters to be not optimal, for obtain the optimal parameters then done optimization *firefly* so get better classification results. The data in this study are *tweets* with the keyword " Indonesian Disaster " which is *crawled* use application *twitter*. Results of this research is to get 128 dimensions that make up weight every word. This is meaningful every word is represented in room vector of 128 dimensions and evaluation percentage values classification models SVM optimization *firefly*: *accuracy* 89%. So, got concluded that method classification SVM shows percentage performance sufficient classification Good compared to with the SVM model without optimization with only get *accuracy* by 61%.

Keywords: Analysis sentiment , *fasttext* , classification , optimization *firefly*, *support vector machine*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Letak geografis Indonesia yang terletak pada pertemuan tiga lempeng aktif, yaitu Indo-Australia, Eurasia, dan Pasifik mengakibatkan kondisi negara Indonesia memiliki tingkat kerawanan tinggi terhadap bencana geologis dan hidro-klimatologis. Dampak terjadinya bencana sangat bervariasi, mulai dari kerusakan, kerugian, hingga menimbulkan korban jiwa (Pahleviannur, 2019). Hal ini menggambarkan perlunya kesiapsiagaan terhadap bencana dengan memperhatikan faktor histori kejadian dimasa lalu sebagai antisipasi dalam penanggulangan bencana di Indonesia. Indonesia merupakan negara yang rawan akan bencana. Beberapa bencana alam yang sering terjadi di Indonesia mulai dari gempa, tsunami, banjir, tanah longsor, gunung meletus dan masih banyak lagi yang lainnya. Adanya kejadian ini mendorong pengguna sosial media atau lembaga penanggulangan kebencanaan untuk mengunggah informasi tentang kondisi bencana dari tempat terjadinya bencana (Giarsyani, 2020).

Kemajuan teknologi informasi dan komunikasi membuat informasi terkait bencana alam menjadi lebih cepat tersebar. Media memiliki peran penting dalam bencana alam. Melalui media informasi mengenai bencana alam dapat tersebar ke berbagai penjuru dunia. Informasi mengenai jenis bencana, informasi mengenai kapan terjadinya bencana, informasi mengenai lokasi bencana, dampak, dan kebutuhan korban bencana alam dapat terekam dan tersampaikan melalui pemberitaan (Sulthan *et al.*, 2021). Media sosial memberikan data yang dimanfaatkan sebagai sumber informasi untuk menanggapi terhadap kejadian bencana alam di suatu wilayah. Penelitian yang telah dilakukan oleh Wu, telah memberikan konsep baru dengan menggunakan media sosial sebagai *social network* (Wu & Li, 2016).

Salah satu sosial media yang dikenal oleh masyarakat adalah *twitter* yang digunakan masyarakat untuk membagikan hal yang dirasakan oleh penggunanya. Melalui postingan pada *twitter*, masyarakat dapat membagikan dan mendapatkan informasi terbaru terkait bencana yang terjadi. Dengan memanfaatkan data dari media sosial *twitter*, dapat dilakukan analisis pendapat dan opini masyarakat

terhadap kebencanaan di Indonesia melalui analisis sentimen dengan cara melakukan klasifikasi pendapat dan opini ke dalam 2 kelas yaitu negatif dan positif. Analisis sentimen adalah proses mengekstraksi, memahami dan mengolah data berupa teks yang tidak terstruktur secara otomatis guna mendapatkan informasi sentimen yang terdapat pada sebuah kalimat pendapat atau opini (Fitriana *et al*., 2021). Dalam penerapan analisis sentimen menggunakan metode *machine learning* terdapat beberapa metode yang sering digunakan (Arsi & Waluyo, 2021). Pada data teks diperlukan ekstraksi fitur untuk mengubah data teks menjadi data numerik. Terdapat beberapa jenis ekstraksi fitur *word embedding* seperti *word2vec*, *glove* dan *fasttext.* Ekstraksi fitur yang digunakan adalah *fasttext, fasttext* merupakan suatu model *word embedding* yang dikembangkan oleh *Facebook*. *Fasttext* ini adalah pengembangan dari model *word embedding word2Vec*. Salah satu keunggulan utama *fasttext* adalah kemampuannya untuk memperhitungkan informasi *subword* dalam kata. Dengan membagi kata menjadi unit *subword*, seperti suku kata atau karakter, *fasttext* dapat mewakili kata yang tidak ditemukan dalam kamus atau kata yang tidak umum dengan lebih baik. Ini berguna dalam bahasa dengan kata terkait, morfologi yang kompleks. *Fasttext* mampu mencapai kinerja yang sangat baik untuk representasi kata dan klasifikasi kalimat, khususnya dalam kasus kata langka dengan memanfaatkan informasi tingkat karakter (Rahman *et al*., 2021). Setelah proses fitur ekstraksi, maka akan dibuat model menggunakan *metode Support Vector Machine* (SVM).

Analisis sentimen biasanya melibatkan data berdimensi tinggi. Misalnya, ketika merubah teks menjadi vektor fitur melalui proses seperti *word embedding* atau tf-idf terdapat jumlah fitur yang sangat besar (yang sebanding dengan ukuran kosakata). SVM memiliki performa yang baik dalam mengolah data berdimensi tinggi ini. SVM adalah sebuah teknik *Supervised Machine learning* yang berfungsi untuk menganalisa suatu data untuk klasifikasi. SVM berjalan dengan cara mengkategorikan titik data yang berasal dari kumpulan data latih. Tujuan dari algoritma SVM adalah untuk mencari *hyperplane* pada ruang dimensi N (jumlah fitur) yang mengklasifikasikan secara jelas titik data (Awad & Khanna, 2015). Prinsip dasar SVM adalah *linear classifier*, dan selanjutnya dikembangkan agar

dapat bekerja pada *problem non-linear*. dengan memasukkan konsep *kernel trick* pada ruang kerja berdimensi tinggi. Perkembangan ini memberikan stimulus minat penelitian di bidang *pattern recognition* untuk investigasi potensi kemampuan SVM secara teoritis maupun dari segi aplikasi (Nafi & Aulia, 2022).

Ada dua kasus dalam melakukan klasifikasi, yaitu *linear* dan *non-linear*. Pada kasus *linear*, data terpisah secara sempurna sehingga pengklasifikasi hanya perlu mencari garis *linear* sebagai pemisah antar kelas. Akan tetapi, permasalahan klasifikasi yang sering dijumpai bersifat *non-linear*, yakni data tidak dapat dipisahkan secara *linear*. Untuk itu diperlukan metode yang dapat meningkatkan dimensi data agar pengklasifikasi bisa memisahkan data secara sempurna. Permasalahan klasifikasi *non-linear* dapat diatasi dengan menerapkan fungsi kernel pada SVM, untuk mentransformasikan data ke ruang vektor berdimensi lebih tinggi. Ada beberapa *kernel* yang dapat digunakan untuk proses penyeleksian parameter diantaranya *Radial Basis Function* (RBF), *Sigmoid*, dan *Polynomial*. Pada penelitian kali ini akan digunakan *kernel* RBF, karena *kernel* RBF dapat menangani pemisahan *linear* pada data input *non-linear* berdimensi tinggi seperti dalam klasifikasi teks (Noor & Islam, 2019). Serta juga bermanfaat mengurangi sulitnya membaca data numerik, karena nilai *kernel* berada diantara nol dan satu. selain itu *kernel* RBF juga menunjukkan *tradeoff* parameter C dalam algoritma SVM yang sangat mempengaruhi hasil dari klasifikasinya (Zhou et al., 2009). Kernel RBF dalam SVM akan mendapatkan performa klasifikasi dan akurasi yang lebih baik ketika melakukan pemilihan nilai parameter γ dan konstanta *soft margin* C yang tepat. Parameter *cost* atau biasa disebut sebagai C adalah parameter yang bekerja sebagai pengoptimalan SVM untuk menghindari kesalahan klasifikasi di setiap sampel dalam dataset *training* dan Parameter γ menentukan seberapa jauh pengaruh dari satu sampel dataset pelatihan.

SVM tidak dapat memilih parameter yang sesuai sehingga penggunaan parameter menjadi tidak optimal. Dengan penggunaan parameter yang sesuai diharapkan dapat meningkatkan *accuracy* SVM. *Firefly* merupakan metode *metaheuristik* dalam kelompok *Swarm Intelligence* (SI) yang mengadopsi perilaku sosial dan cara berkomunikasi sekelompok kunang-kunang melalui cahaya di bagian ekornya, optimasi merupakan proses mencari solusi optimal untuk masalah

tertentu yang menarik, dan proses pencarian ini dapat dilakukan dengan menggunakan beberapa agen yang pada dasarnya membentuk sistem agen yang berkembang. Terdapat banyak teknik optimasi diantaranya yaitu *Particle Swarm Optimization (PSO), Bat Algorithm (BA), Cuckoo Search (CS), Ant Colony Optimization (ACO)* dan *Flower* (Styawati et al., 2021).

Pardede & Pakpahan (2023) telah melakukan penelitian mengenai analisis sentimen penanganan covid-19 menggunakan metode *Long Short-Term Memory* pada media sosial *twitter* dengan menggunakan *fasttext*, kata divektorkan menggunakan *fasttext*, tujuannya untuk mengubah tipe data *string* menjadi vektor array, sehingga kata dapat diproses di dalam model *machine learning*. Performa akhir model diukur berdasarkan nilai *accuracy*. Styawati (2021) telah melakukan penelitian optimasi parameter SVM berbasis optimasi *firefly* pada data opini film, dengan mencari nilai parameter SVM yang optimal berdasarkan *accuracy*. Hasil penelitian menunjukkan bahwa optimasi *firefly* mampu mendapatkan kombinasi parameter SVM yang optimal berdasarkan *accuracy*, sehingga tidak diperlukan cara *trial and error* untuk mendapatkan nilai tersebut. Hal ini dibuktikan dengan hasil evaluasi menghasilkan *accuracy* tertinggi yaitu 87.84%. Pada penelitian lain yang dilakukan klasifikasi data opini film algoritma *Support Vector Machine-Firefly* (Rohim & Aminuallah, 2022). Hasil penelitian menunjukkan bahwa optimasi *firefly* dapat membantu SVM untuk mendapatkan kombinasi parameter yang sesuai berdasarkan *accuracy* dengan waktu eksekusi lebih singkat dan mendapat nilai *accuracy* sebanyak 87,15%.

Berdasarkan uraian yang telah dipaparkan, maka pada penelitian tugas akhir ini akan dilakukan analisis sentimen terkait kebencanaan Indonesia di *twitter* menggunakan *support vector machine* berbasis *fasttext* dengan parameter optimasi *firefly*.

1.2 Rumusan Masalah

Rumusan masalah yang dibahas pada penelitian ini adalah sebagai berikut:

1. Bagaimana hasil pembobotan data teks menjadi vektor numerik menggunakan model *fasttext*?

2. Bagaimana hasil klasifikasi sentimen kebencanaan Indonesia menggunakan SVM optimasi *firefly* dengan ekstraksi *fasttext* dan hasil evaluasi model menggunakan *accuracy*?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Dataset yang diperoleh dari hasil *crawling* menggunakan aplikasi *twitter* menggunakan API.
2. *Fitur Ekstraksi* yang digunakan adalah *word embedding* dengan metode *fasttext.*
3. Parameter yang dioptimasi adalah $C = 0{,}1$ dan $\gamma = 0{,}1$ dengan menggunakan optimasi *firefly*.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Memperoleh hasil pembobotan data teks menjadi vektor numerik menggunakan model *fasttext.*
2. Mendapatkan hasil klasifikasi sentimen kebencanaan Indonesia menggunakan SVM optimasi *firefly* dengan ekstraksi *fasttext* dan hasil evaluasi model menggunakan *accuracy*.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari hasil penelitian ini adalah sebagai berikut:

1. Memberikan informasi tentang pemodelan *machine learning* dengan menggunakan *support vector machine* dengan berbasis *fasttext.*
2. Memberikan informasi terkait pemodelan dari proses optimasi *firefly* setelah model *support vector machine* berbasis *fasttext.*
3. Menambah pemahaman pembaca tentang permasalahan *machine learning* dan penyelesaian permasalahannya dengan *firefly*.

BAB II
TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Analisis sentimen adalah proses untuk melakukan analisis terhadap opini, sentimen, penilaian serta emosi dari pernyataan seseorang terhadap suatu domain atau juga merupakan proses untuk mengekstrak dan mengolah data berupa teks secara otomatis untuk mendapatkan suatu gambaran mengenai kecenderungan opini terhadap sebuah objek, apakah cenderung beropini positif atau negatif (Parlika et al., 2020). Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen atau pendapat. Polaritas mempunyai arti apakah teks yang ada dalam dokumen, kalimat, atau pendapat memiliki aspek positif atau negatif (Nugroho et al., 2015). Proses analisis sentimen bertujuan untuk menghitung total skor sentimen dan mengklasifikasikan opini. Proses ini menggunakan kamus kata positif dan negatif untuk dicocokkan ke setiap kata dalam kalimat uji sehingga dapat ditetapkan skor sentimen untuk setiap kata dalam sebuah kalimat. Setelah mendapatkan skor sentimen untuk setiap kata, kemudian dihitung secara keseluruhan nilai sentimen dari kalimat tersebut (Sasmito Aribowo, 2018). Informasi yang menyebabkan pertambahan data yang kebanyakan berupa data teks dapat dijadikan sumber yang sangat potensial untuk digali lebih dalam. Salah satu contohnya adalah data teks yang diambil dari *twitter* (Ratnawati, 2018).

2.2 *Text Preprocessing*

Text preprocessing adalah proses yang dilakukan sebelum data dapat diolah untuk dilakukan seleksi atau *filter* untuk mendapatkan kata yang penting namun tetap mempertahankan karakter dari subjek teks tersebut. Adapun selain mendapatkan kata penting pada subjek teks, *text preprocessing* efisien untuk mengurangi fitur yang berdimensi tinggi dan mengurangi *noise* sehingga akan mengefisiensi waktu pemrosesan data dan meningkatkan *accuracy* (Chen et al., 2020). Tahapan ini adalah tahapan yang berfungsi untuk membersihkan teks sebelum diolah lebih lanjut. Data teks mentah yang tidak terstruktur memiliki cukup banyak *noise* seperti tanda baca, angka, imbuhan, karakter khusus, *slang word* dan lain sebagainya. Dalam tahapan ini, data teks tersebut dibersihkan

sehingga tersisa bentuk dasarnya saja untuk keperluan analisis teks lebih lanjut (Priyanto & Ma'arif, 2018).Berikut tahapan *text preprocessing*:

1. *Case Folding*

Case folding merupakan proses dalam *text preprocessing* yang dilakukan untuk menyeragamkan karakter pada data. Proses *case folding* adalah proses mengubah seluruh huruf menjadi huruf kecil. Pada proses ini karakter-karakter A-Z yang terdapat pada data diubah kedalam karakter a-z (Shiri, 2004).

2. *Filtering*

Filtering merupakan langkah untuk menghilangkan karakter-karakter ilegal pada dokumen seperti tanda baca, simbol, angka, html, dan *mention*. Proses dalam menghilangkan karakter-karakter ilegal dapat disebut *filtering*. Contoh karakter ilegal yang dihilangkan antara lain %, &, >, (, {,], 1-9, @uluwatu, http://tripadvisor.com.

3. *Stemming*

Stemming merupakan proses memetakan variasi kata ke bentuk dasar. Proses *stemming* dilakukan dengan menghapus imbuhan, baik awalan maupun akhiran dari suatu kata untuk mendapatkan kata dasarnya. *Stemming* yang umum digunakan pada teks berbahasa Indonesia menggunakan *library stemmer sastrawi* yang dikembangkan berdasarkan algoritma Nazief-Adriani.

4. *Stopword Removal*

Stopword removal merupakan tahap pengambilan kata-kata penting dan membuang kata yang dianggap tidak penting. Cara untuk membuang kata yang tidak penting disebut *stopword removal*. *Stopword removal* bertujuan untuk menghilangkan kata yang sering muncul namun tidak memiliki kontribusi dalam proses analisis data. *Stopword removal* berusaha memperkecil dimensi data dan mempercepat waktu komputasi (Symeonidis et al., 2018). Contoh kata yang tidak penting di Bahasa Indonesia seperti kata "dan", " yang", "di", "ke".

5. *Tokenization*

Tokenization adalah tahap memecah kalimat menjadi beberapa bagian dinamakan token. Sebuah token dianggap sebagai suatu bentuk kata, frasa, atau suatu elemen yang berarti. Tahapan ini juga menghilangkan karakter tertentu seperti tanda baca dan mengubah semua token ke bentuk huruf kecil. *Tokenization*

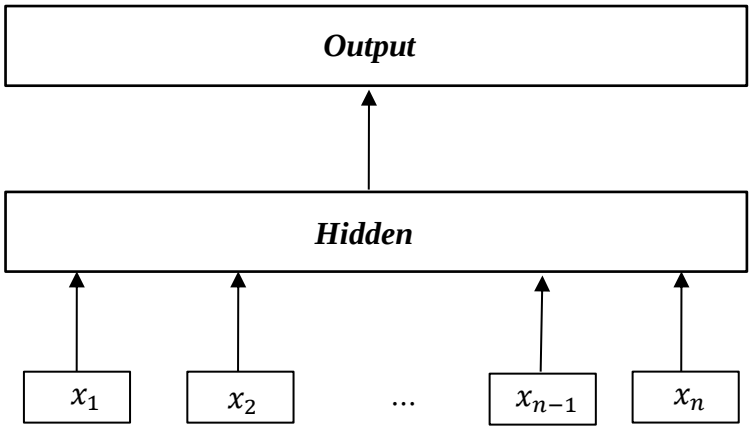
memiliki kemampuan untuk memecah dokumen menjadi kata, frasa, simbol atau elemen lain yang memiliki makna. Pada proses *tokenization*, data teks ulasan dipecah menjadi token yang terdiri dari satu kata yang bermakna dan disimpan dalam *array* kata.

2.3 *Word Embedding*

Word embedding adalah istilah yang digunakan untuk teknik mengubah sebuah kata menjadi sebuah vektor atau *array* yang terdiri dari kumpulan angka. Ketika membuat *model machine learning* yang menerima *input* sebuah teks, tentu *machine learning* tidak bisa langsung menerima teks mentah yang kita miliki, kata tersebut harus diubah dulu menjadi angka dengan acuan sebuah kamus kata. Biasanya jika tidak menggunakan *word embedding*, setiap kata akan diubah menjadi angka dalam bentuk *integer* sesuai dengan posisi angka tersebut dalam kamus, misalkan kata "sembuh" diubah menjadi angka "4" dan kata "meninggal" diubah menjadi angka "7". Angka tersebut diubah kembali menjadi sebuah vektor (*array* 1 dimensi) yang memiliki panjang sebanyak kata yang dimiliki oleh kamus. *Array* tersebut hanya akan bernilai 1 atau 0 (disebut *one hot encoding*). Nilai 1 diposisikan pada indeks yang merupakan nomor kata tersebut sedangkan elemen lainnya bernilai 0 (A. Rahman et al., 2021).

Pada penelitian ini, metode *word embedding* yang digunakan ialah *fasttext*. *Fasttext* adalah metode *word embedding* yang merupakan pengembangan dari *word2vec*. Metode ini mempelajari representasi kata dengan mempertimbangkan informasi *subword*. Setiap kata direpresentasikan sebagai sekumpulan karakter *n-gram*. Dengan demikian, dapat membantu menangkap arti kata yang lebih pendek dan memungkinkan *embedding* untuk memahami *sufiks* dan *prefiks* dari kata. Representasi vektor dikaitkan dengan setiap karakter *n-gram*, sedangkan kata direpresentasikan sebagai jumlah dari representasi vektor tersebut. Setelah kata direpresentasikan dengan karakter *n-gram*, *model skip-gram* dilatih untuk mempelajari *embedding* vektor dari kata. Pada umumnya, model yang mempelajari representasi kata ke dalam vektor dengan mengabaikan morfologi kata, setiap kata yang memiliki vektor berbeda. Hal ini menjadi keterbatasan untuk merepresentasikan kata dari bahasa dengan kosakata yang besar dan memiliki banyak kata langka. *Fasttext* memiliki kinerja yang baik, dapat melatih

model pada dataset yang besar dengan cepat dan dapat memberikan representasi kata yang tidak muncul dalam data latih. Jika kata tidak muncul selama pelatihan model, kata tersebut dapat dipecah menjadi *n-gram* untuk mendapatkan *embedding* vektornya (Bojanowski et al., 2017). *Fasttext* menginisiasikan suatu vektor kata dengan menjumlahkan nilai dari *n-gram* kata dan nilai dari penyisipan kata itu sendiri (Pardede & Pakpahan, 2023). Selanjutnya, token kata tersebut menjadi input di dalam jaringan saraf *skip-gram* untuk menghasilkan nilai probabilitas konteks. Arsitektur jaringan *skip-gram* dapat dilihat pada gambar 2.1.



Gambar 2.1 Ilustrasi *Fasttext*

Fasttext menginisiasikan suatu vektor kata dengan menjumlahkan nilai dari *n-gram* kata dan nilai dari penyisipan kata itu sendiri.

$$v(w) = z(w) + \sum_{i=1}^{n} x_i \tag{2.1}$$

keterangan:

$v(w)$: representasi kata w

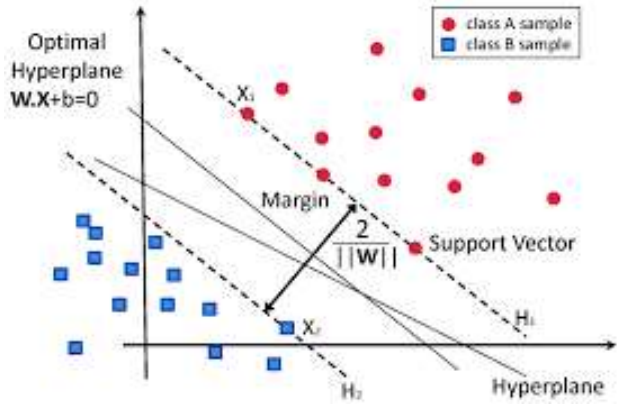
$z(w)$: vektor fitur untuk kata w

x_i : vektor fitur untuk kata $n - gram, i = 1, 2, \cdots, n$

Penggunaan *fasttext* pada penelitian ini dilakukan untuk mengubah kata menjadi sebuah vektor. Keunggulan *fasttext* adalah bisa menangani kata yang tidak pernah dijumpai sebelumnya karena prosesnya adalah memanfaatkan *sub-word* dari setiap kata (Sani & Sarwani, 2022).

2.4 *Support Vector Machine*

Support Vector Machine (SVM) merupakan metode yang dikembangkan oleh Boser, Guyon, dan Vapnik yang dipresentasikan pertama kali pada tahun 1992 di acara *annual workshop on computational learning theory*. Konsep khusus SVM adalah meminimalkan kesalahan klasifikasi empiris dan memaksimalkan *margin geometric*. Oleh karena itu, SVM disebut *maximum margin classifiers*. SVM adalah metode *machine learning* yang bekerja dengan prinsip *Structural Risk Minimization* (SRM). SVM diasumsikan bahwa apabila semakin besar *margin* atau jarak antara *hyperplane*, maka semakin baik pula generalisasi kesalahan pengklasifikasiannya (Sunori *et al*., 2021).



Gambar 2.2 Ilustrasi *Support Vector Machine*

Terkait kasus klasifikasi pada ruang berdimensi tinggi, untuk mencari *hyperplane* yang dapat memaksimalkan jarak (*boundary*) seluruh kelas data (Husada & Paramita, 2021). Tujuan SVM adalah menemukan *hyperplane* terbaik yang memisahkan dua buah kelas atau kelompok pada *input space*. *Hyperplane* adalah garis batas pemisah data antar kelas. SVM bekerja dengan memaksimalkan *margin* yang merupakan jarak pemisah antara kedua kelas data tersebut. *Hyperplane* terbaik adalah *hyperplane* yang terletak di tengah antara dua set objek dari dua kelas. *Hyperplane* pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimal *margin*. *Margin* adalah jarak antara *hyperplane* terbaik dengan *pattern* terdekat dari masing-masing kelas. *Pattern* yang paling dekat disebut sebagai *support vector*. Pada dasarnya, *training* yang dilakukan oleh SVM tidak menggunakan

keseluruhan data *training*, namun hanya data *training* yang terpilih untuk digunakan.

2.4.1 *Linear Support Vector Machine*

Misalkan data yang ada pada himpunan data *training* dinotasikan $x_i \in \Re^q$ sebagai sedangkan label masing-masing kelas dinyatakan sebagai $y_i \in \{-1, +1\}$ *model linear* secara umum yang dipakai dalam metode SVM untuk menghasilkan *hyperplane* (Fitriyah et al., 2020) yaitu:

$$y = f(x) = w^T x_i + b, i = 1,2 \dots, n \tag{2.2}$$

keterangan:

$y \in \{-1, +1\}$: label kelas dari himpunan data

w : $[w_1, w_2, \dots \dots . w_q]$ merupakan vektor bobot yang tegak lurus terhadap *hyperplane* berupa vektor kolom berukuran q merupakan banyaknya variabel bebas

x_i : $[w_{i1}, w_{i2}, \dots \dots . w_{iq}]$ merupakan vektor kolom berukuran q x 1

b : *error* atau bias merupakan skalar

n : banyak data

Kedua kelas +1 dan -1 diasumsikan dapat terpisah secara sempurna oleh *hyperplane* yang berdimensi q atau data yang bertempat di *hyperplane*, sehingga dapat dinotasikan sebagai berikut:

$$w^T x_i + b = 0 \tag{2.3}$$

Pattern x_i yang termasuk kelas positif yaitu +1 dapat dirumuskan sebagai berikut:

$$[(w^T . x_i) + b] \geq +1 \tag{2.4}$$

Pattern $x_i$yang termasuk kelas negatif yaitu -1 dapat dirumuskan sebagai berikut:

$$[(w^T . x_i) + b] \leq -1 \tag{2.5}$$

Ada permasalahan SVM, *margin* terbesar dapat ditemukan dengan memaksimalkan *margin* atau jarak antara dua set objek dari kelas yang berbeda. Nilai *margin* antara bidang pembatas dapat dirumuskan sebagai berikut:

$$margin\ (m) = \frac{|(b-1)-(b+1)|}{\sqrt{\boldsymbol{w^T w}}} \tag{2.6}$$

$$= \frac{|-2|}{\|\boldsymbol{w}\|}$$

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

Memaksimalkan nilai *margin* sama dengan meminimumkan nilai. Persamaan untuk mencari *hyperplane* terbaik dengan nilai margin terbesar pada permasalahan *linear* di dalam *primal space* dapat dilihat pada Persamaan (2.7) dengan memperhatikan *constraint* pada Persamaan (2.8). Meminimumkan nilai digunakan metode *Quadratic Programming* (QP) yaitu:

$$\min w\, \tau(w) = \frac{1}{2} \| w \|^2 \tag{2.7}$$

dengan:

$$y_i(w^t x_i > 1 = 1{,}2 \ldots\ldots, n \tag{2.8}$$

Persoalan tersebut akan menjadi lebih mudah diselesaikan apabila diubah ke dalam formula *lagrangian* yaitu menggunakan teknik komputasi *lagrange multiplier*. Solusi untuk optimalisasi hal tersebut adalah sebagai berikut:

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

Nilai a_i adalah *lagrange multiplier* yang berkorespondensi dengan nilai x_i adalah nol atau positif. Untuk menyelesaikan masalah tersebut, hal pertama yang dilakukan adalah meminimalkan L terhadap w dan memaksimalkan L terhadap b. Setelah memodifikasi persamaannya, optimasi pada Persamaan (2.9) dapat direpresentasikan dalam a_i seperti Persamaan (2.10).

$$\max_a \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i\, \alpha_j y_i y_j x_i . x_j \tag{2.10}$$

dengan kendala:

$$\sum_{j=1}^{n} \alpha_i y_i = 0, \forall_i \text{ dengan } \alpha_i > 0 \tag{2.11}$$

Solusi dari Persamaan (2.9) menghasilkan banyak α_i dengan nilai nol. Data yang berkorespondensi dengan α_i yang tidak nol merupakan *support vector*, yaitu data yang memiliki jarak terdekat dengan *hyperplane*. Setelah nilai α_i ditemukan, maka kelas dari data yang baru dapat ditentukan berdasarkan nilai fungsi batas keputusan berikut:

$$f(x_d) = \sum_{i=1}^{n} \alpha_i y_i (x_i x_d) + b \tag{2.12}$$

dengan n adalah jumlah *support vector* dan x_d adalah vektor data baru yang akan dicari kelasnya. Data akan digolongkan sebagai kelas +1 jika nilai $f(x_d) > 0$ ataupun sebagai kelas -1 jika nilai $(x_d) < 0$.

2.4.2 *Non Linear Support Vector Machine*

Untuk kasus data yang tidak terpisah secara *linear* diasumsikan bahwa kelas pada *input space* tidak dapat dipisahkan secara sempurna. Hal tersebut menyebabkan Persamaan (2.7) tidak dapat terpenuhi. Oleh karena itu, untuk mengklasifikasi dua kelas yang terpisah secara *non linear* dan lebih tahan terhadap pencilan maka dapat menggunakan teknik *soft margin* yaitu dengan mengubah masalah *Quadratic Programming* (QP) dan batasan diatas dengan ξ_i menambahkan *slack variable* yaitu $\xi_i > 0$ sehingga menjadi persamaan berikut:

$$\min \tau(w, \xi) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \tag{2.13}$$

dengan kendala

$$y_i(w^T . x_i) + b \geq 1\xi_i, \forall , i \ldots \ldots n \tag{2.14}$$

Parameter C berfungsi untuk mengontrol optimasi antara *margin* dan kesalahan klasifikasi ξ . Semakin besar nilai C maka semakin besar pula penalti terhadap kesalahan (*error*) klasifikasi. Persamaan L_p berdasarkan fungsi objektif pada Persamaan (2.15) dapat dirumuskan sebagai berikut:

$$L_p = \left(\frac{1}{2} \| w \|^2 - C \sum_{i=1}^{n} \xi_i \right) - \sum_{i=1}^{n} \alpha_i \left[y_i \left(\boldsymbol{w^T x_i} + b \right) - 1 + \xi_i \right] - \sum_{i=1}^{n} \beta_i (\xi_i) \tag{2.15}$$

dengan kendala

$$\alpha_i \geq 0 \ , \mu_i \geq 0.$$

dengan μ_i merupakan *koefisien* pengali *lagrange* untuk ξ_i. Turunan L_p terhadap w, b, dan ξ_i adalah sebagai berikut:

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^{n} a_i y_i x_i$$

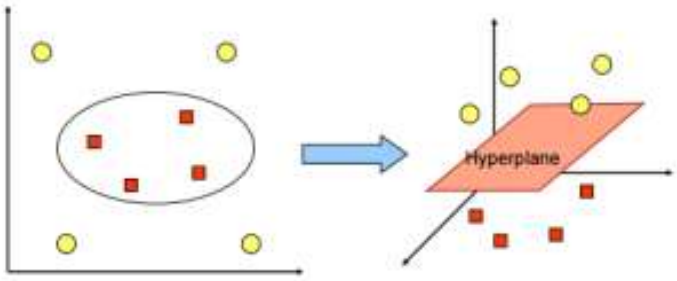
$$\frac{\partial L_p}{\partial b} = \sum_{i=1}^{n} a_i y_i$$

$$\frac{\partial L_p}{\partial \xi_i} = C - a_i - \mu_i = 0$$

dengan mensubstitusi turunan L_p terhadap w, b, dan ξ_i, maka didapatkan persamaan max sebagai berikut:

$$\max_a \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j y_i y_j \boldsymbol{x_i} . \boldsymbol{x_j} \tag{2.16}$$

Jika data tidak dapat dipisahkan secara *linear*, data terlebih dahulu ditransformasi ke dalam ruang dimensi yang lebih tinggi. Gambar 2.3 sebagai berikut:



Gambar 2.3 Ilustrasi *Kernel* Trick

Tidak semua data bersifat *linear*, sehingga sulit dicari bidang pemisah secara *linear*. Permasalahan ini dapat diselesaikan dengan mentransformasikan data ke dalam dimensi ruang fitur (*feature space*) yang lebih tinggi, sehingga dapat dipisahkan secara *linear* pada *feature space* yang baru. Caranya, data dipetakan dengan menggunakan fungsi pemetaan (Φ). Misal, terdapat $x = (x_1 , x_2)$, pada Gambar 2.3 diperlihatkan bahwa data pada kelas kuning dan data pada kelas merah yang berada pada *input space* berdimensi dua tidak dapat dipisahkan secara *linear*. Kemudian fungsi yang memetakan setiap data pada *input space* ke ruang vektor baru yang berdimensi lebih tinggi yaitu berdimensi tiga. Kedua kelas

dapat dipisahkan secara *linear* oleh sebuah *hyperplane*. Pada ruang vektor yang baru, SVM mencari *hyperplane* yang memisahkan kedua kelas secara *linear*.

Pada dasarnya, proses pembelajaran pada SVM dalam menemukan *support vector* hanya bergantung pada *dot product* dari data pada ruang fitur, yaitu $\Phi(x_i).\Phi(x_j)$ dan sulit dipahami maka perhitungan *dot product* dapat digantikan dengan fungsi *kernel* yang mendefinisikan secara implisit fungsi transformasi Φ. Formulasi *kernel trick* adalah sebagai berikut (Cortes & Vapnik, 1995):

$$\mathrm{K}\left(x_i, x_j\right) = \Phi(x_i).\Phi(x_j) \tag{2.17}$$

keterangan:

$\mathrm{K}\left(x_i, x_j\right)$: fungsi *Kernel*

$\Phi(x_i).\Phi(x_j)$: *dot product* dari data pada ruang fitur

$\mathrm{K}\left(x_i, x_j\right)$ merupakan sebuah *kernel* yang digunakan untuk mentransformasikan data *input* menjadi ruang fitur, sehingga data dapat dipisahkan secara *linear*, dan kemudian mencari *hyperplane* yang optimal. Fungsi *kernel* memungkinkan penanganan kasus yang *non linear* dalam ruang *input*, sehingga data dapat dipisahkan secara *linear* dalam ruang fitur. Selanjutnya, hal ini memungkinkan perhitungan *decision boundary* (batas keputusan) secara efisien. Penggunaan fungsi *kernel* dalam persamaan mengakibatkan modifikasi pada persamaan *Lagrange* sebagai berikut:

$$\max_a \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j y_i y_j K(\boldsymbol{x_i}.\boldsymbol{x_j}) \tag{2.18}$$

dengan kendala:

$$0 \le a_i \le C, untuk\ setiap\ i = 1, \dots\dots\dots m$$

untuk menghitung nilai w dan b, maka dapat menggunakan persamaan dari *soft margin* SVM dengan mengubah nilai x menjadi $\Phi(x)$, sehingga diperoleh persamaan sebagai berikut:

$$b = \frac{1}{s}\sum_{j=1}^{s}\left(y_k - w \cdot \Phi(x_k)\right) \tag{2.19}$$

dengan

$$w = \sum_{i=1}^{n} \alpha_i y_i \Phi_{\mathrm{i}}(x_i) \tag{2.20}$$

Keterangan:

$\boldsymbol{x}_k$: data *train* yang termasuk *support vector* dan $\boldsymbol{x}_k = [x_{k1}, x_{k2}, \dots, x_{kq}]$ dan merupakan *vector* kolom $q \times 1$

y_k : label kelas dari himpunan data

α_k : α yang termasuk *support vector*

s : jumlah *support vector*

n : jumlah data *train*

kelas dari suatu data *test* dapat ditentukan dengan persamaan sebagai berikut:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{2.21}$$

dengan $f(x) \geq 0$ merupakan kelas positif dan $f(x) < 0$ merupakan kelas negatif. Beberapa fungsi *kernel* yang dapat digunakan pada umumnya yaitu:

1. *Polynomial*

Polynomial kernel merupakan kernel dengan dua parameter c yang merepresentasikan nilai konstan, dan d yang merepresentasikan nilai derajat dari kernel. *Polynomial* kernel memiliki persamaan:

$$\mathrm{K}\left(x_i, x_j\right) = x_i^T x \tag{2.22}$$

2. *Radial Basis Function* (RBF)

Kernel RBF atau juga disebut kernel *Gaussian* adalah konsep kernel yang paling banyak digunakan untuk memecahkan masalah klasifikasi data yang tidak dapat dipisahkan secara *linear*. Kernel ini dikenal memiliki performa yang baik dengan parameter tertentu, dan hasil dari pelatihan memiliki nilai *error* yang kecil dibandingkan dengan kernel lainnya. Rumus persamaan untuk fungsi kernel RBF adalah:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|^2\right) \tag{2.23}$$

$\boldsymbol{x}_i$ dan $\boldsymbol{x}_j$ mewakili dua sampel input, $\left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|^2$menunjukkan jarak *Euclidean* kuadrat di antara keduanya, dan γ adalah parameter yang menentukan pengaruh setiap sampel pelatihan. Parameter gamma mengontrol lebar kernel dan memengaruhi fleksibilitas batas keputusan. Nilai γ yang lebih kecil mengarah ke kernel yang lebih luas, memungkinkan lebih banyak sampel untuk berkontribusi pada batas keputusan. Sebaliknya, nilai γ yang lebih besar menghasilkan kernel

yang lebih sempit, menyebabkan batas keputusan lebih dipengaruhi oleh sampel yang lebih dekat.

3. *Sigmoid*

Konsep kernel *sigmoid* merupakan pengembangan dari Jaringan Saraf Tiruan (JST). Pada dasarnya, konsep kernel sigmoid berkaitan dengan penggunaan fungsi *sigmoid* sebagai fungsi kernel dalam algoritma pemetaan kernel, terutama dalam konteks pembelajaran mesin, khususnya dalam metode SVM dan kernel trick. *Sigmoid* kernel memiliki persamaan:

$$\mathrm{K}\left(x_i, x_j\right) = tahn(\gamma x_i^T + r) \tag{2.24}$$

K adalah fungsi kernel sigmoid, x_i dan x_j adalah dua vektor data dalam ruang fitur, γ adalah parameter yang mengontrol seberapa curam kurva tangen hiperbolik ($tanh$), dan r adalah parameter bias.

2.5 Optimasi *Firefly*

Optimasi *firefly* merupakan optimasi *metaheuristic* berbasis populasi yang tergolong ke dalam kelompok *Nature Inspired Algorithm* (NIA). Optimasi *firefly* ini pertama kali dikembangkan (Yang, 2008). Optimasi *firefly* mensimulasikan kinerja sosial kunang-kunang di alam untuk memecahkan masalah pengoptimalan. Keuntungan utama optimasi *firefly* adalah Efektif pada Masalah *non linear*, optimasi *firefly* telah terbukti efektif dalam menangani masalah optimasi *non linier* yang kompleks. Hal ini menjadikannya pilihan yang baik untuk berbagai masalah optimasi. Hal ini disebabkan oleh penurunan intensitas cahaya dengan jarak, menyebabkan daya tarik di antara kunang-kunang bisa menjadi global atau lokal, tergantung pada koefisien penyerapan. Optimasi *firefly* memiliki banyak kemiripan dengan optimasi lain yang didasarkan pada kecerdasan kawanan, seperti *Particle Swarm Optimization* (PSO), *Artificial Bee Colony optimization* (ABC), *Bacterial Foraging Algorithm* (BFA). Optimasi *firefly* ini sangat efisien dan lebih unggul dibandingkan optimasi konvensional lainnya, seperti optimasi genetika yang digunakan untuk memecahkan banyak masalah optimasi (Yang, 2010). Berikut beberapa aturan optimasi *firefly*:

1. Kunang-kunang bersifat *unisex*, sehingga satu kunang-kunang dapat tertarik dengan kunang-kunang lain tanpa melihat jenis kelamin.

2. Ketertarikan antar kunang kunang akan sebanding dengan tingkat kecerahan kunang-kunang tersebut. Dengan ketentuan bahwa semakin jauh jarak antar kunang-kunang, maka tingkat kecerahan kunang-kunang akan menurun atau menghilang. Jadi untuk setiap dua kunang-kunang yang berkedipan, kunang-kunang yang kurang terang (redup) akan mendekati kunang-kunang yang lebih terang. Jika dari kedua kunang-kunang tidak ada yang lebih terang maka kunang-kunang akan bergerak secara acak.
3. Kecerahan pada kunang-kunang akan ditentukan oleh fungsi tujuan dari masalah yang diberikan. Berikut ini beberapa istilah yang digunakan dalam optimasi *firefly* menurut (Yang, 2010).

Berdasarkan pada optimasi *firefly*, fungsi objektif didapatkan dari persamaan berikut:

$$\max f(x), x = (x_1, \dots \dots x_d)^T \tag{2.25}$$

2.5.1 Intensitas Cahaya

Dua masalah penting yang terdapat pada optimasi *firefly* yaitu variasi intensitas cahaya dan perumusan *attractiveness*. Kecerahan pada kunang-kunang akan ditentukan oleh fungsi tujuan dan *attractiveness* sebanding dengan kecerahan, dengan demikian untuk setiap dua kunang-kunang yang berkedip, kunang-kunang dengan cahaya yang kurang terang akan bergerak ke arah kunang-kunang yang cahaya lebih cerah. Intensitas cahaya pada kunang-kunang dipengaruhi oleh fungsi tujuan. Tingkat intensitas cahaya untuk masalah meminimumkan sebuah kunang-kunang x dapat dilihat sebagai berikut:

$$I(x) = \frac{1}{f(x)} \tag{2.26}$$

Nilai $I(x)$ merupakan tingkat intensitas cahaya pada kunang-kunang x yang berbanding terbalik terhadap solusi fungsi tujuan permasalahan yang akan dicari $f(x)$. *Attractiveness* β bernilai relatif, karena intensitas cahaya harus dilihat dan dinilai oleh kunang-kunang lain. Dengan demikian, hasil penilaian akan berbeda tergantung dari jarak antara kunang-kunang yang satu dengan yang lainnya. Selain itu, intensitas cahaya akan menurun dari sumbernya dikarenakan terserap oleh media, misalnya udara. Sehingga dapat ditentukan *attractiveness* (β) dengan jarak r sebagai berikut:

$$\beta = \beta_0 e^{-yr^2} \tag{2.27}$$

keterangan:

$\beta(r)$: fungsi daya tarik

β_0 : koefisien ketertarikan pada posisi awal

y : koefisien penyerapan cahaya

r : jarak antar kunang-kunang

2.5.2 *Distance*

Distance atau jarak antara dua kunang-kunang i dan j pada posisi x_i , dan x_j, masing-masing adalah jarak *euclidean* yang dirumuskan sebagai berikut:

$$r_{ij} = \| x_i - x_j \| = \sqrt{\sum_{k=1}^{n} \left(x_i^k - x_j^k\right)^2} \tag{2.28}$$

Dengan selisih dari kordinat kunang-kunang i terhadap kunang-kunang j merupakan jarak antara kedua kunang-kunang tersebut.

2.5.3 *Movement*

Movement adalah pergerakan yang dilakukan oleh *firefly* karena ketertarikan terhadap *firefly* lain yang intensitas cahaya lebih terang. Dengan adanya *movement*, maka posisi *firefly* atau solusi dari *firefly* tersebut akan berubah sesuai rumus berikut :

$$x_i' = x_i + \beta_0 e^{-y r_{ij}^2}\left(x_j - x_i\right) + a\left(rand - \frac{1}{2}\right) \tag{2.29}$$

keterangan:

x_i' : posisi kunang-kunang i yang baru

x_i : posisi kunang-kunang i yang sekarang

β_0 : koefisien ketertarikan pada posisi awal

y : koefisien penyerapan cahaya

x_j : kunang-kunang j

x_i : kunang-kunang i

a : [0,1]

$rand$: bilangan acak pada selang [0,1]

dengan suku pertama merupakan posisi lama dari *firefly*, suku kedua terjadi karena ketertarikan, suku ketiga adalah pergerakan *random firefly* dengan α

adalah *koefisien parameter random* dan $rand$ adalah bilangan *real random* pada interval [0,1]. Pada sebagian besar implementasi optimasi *firefly* menggunakan $\beta_0 = 1, \alpha \in [0,1]\ dan\ \gamma \in [0, \infty]$ (Yang, 2010).

2.6 *Confusion Matrix*

Evaluasi bertujuan untuk menilai hasil uji coba sistem yang dibuat apakah telah sesuai antara hasil sistem analisis sentimen dengan hasil sebenarnya (Kurniawan et al., 2019). *Confusion matrix* adalah alat evaluasi visual yang digunakan dalam *machine learning*. Kolom *confusion matrix* mewakili hasil kelas prediksi dan baris mewakili hasil kelas sebenarnya sehingga dapat menghitung semua kemungkinan kasus masalah klasifikasi (Chen et al., 2020). Dalam *confusion matrix* terdapat beberapa jenis yaitu *precision* dan *recall*. *Accuracy* adalah rasio prediksi benar terhadap keseluruhan data. *Precision* adalah perbandingan nilai prediksi benar positif dibandingkan dengan total hasil dengan prediksi positif. *Recall* adalah perbandingan nilai prediksi benar positif dengan seluruh data yang benar positif.

Tabel 2.1 *Confusion Matrix*

		Predicted	
		Positive (P) +	*Negative* (N) -
Actual	*Positive* +	*True Positive* (TP)	*False Positive* (FN)
	Negative (N) -	False Negative (FN)	*True Negative* (TP)

1. TP adalah *true positive* yang didapatkan dari jumlah data positif yang diprediksi benar.
2. TN adalah *true negative* yang didapatkan dari jumlah data negatif yang diprediksi benar.
3. FP adalah *false positive* yang didapatkan dari jumlah data negatif namun diprediksi sebagai data positif.

4. FN adalah *false negative* yang didapatkan dari jumlah data positif namun diprediksi sebagai data negatif.

Rumus *confusion matrix* untuk menghitung *accuracy*, *precision* dan *recall* seperti berikut:

1. *Accuracy* menggambarkan seberapa akurat model dapat mengklasifikasikan dengan benar. *Accuracy* merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Dengan kata lain, *accuracy* merupakan tingkat kedekatan nilai prediksi dengan nilai aktual (sebenarnya). Nilai *accuracy* dapat diperoleh sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.30}$$

2. *Precision* menggambarkan tingkat keakuratan antara data yang diminta dengan hasil prediksi yang diberikan oleh model. *Precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Dari semua kelas positif yang telah diprediksi dengan benar, berapa banyak data yang benar-benar positif. Nilai *precision* dapat diperoleh sebagai berikut:

$$Precision = \frac{TP}{FP + TP} \tag{2.31}$$

3. *Recall* menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. *Recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Nilai *recall* dapat diperoleh sebagai berikut:

$$Recall = \frac{TP}{TP + FN} \tag{2.32}$$

2.7 Kebencanaan Indonesia

Letak geografis Indonesia yang terletak pada pertemuan tiga lempeng aktif, yaitu Indo-Australia, Eurasia, dan Pasifik mengakibatkan kondisi negara Indonesia memiliki tingkat kerawanan tinggi terhadap bencana *geologis* dan *hidro-klimatologis*. Dampak terjadinya bencana sangat bervariasi, mulai dari kerusakan, kerugian, hingga menimbulkan korban jiwa (Pahleviannur, 2019). Hal ini menggambarkan perlunya kesiapsiagaan terhadap bencana dengan memperhatikan faktor histori kejadian dimasa lalu sebagai antisipasi dalam

penanggulangan bencana di Indonesia. Melihat sangat pentingnya histori bencana dan penanggulangannya yang terjadi di Indonesia. Terjadinya bencana alam sering dikaitkan dengan isu perubahan iklim di bumi dengan isu populernya adalah pemanasan global dan kerusakan lingkungan yang mendorong peningkatan terjadinya bencana alam. Isu tersebut memang kian gencar dibahas dan dikaji seiring dengan pembuktian-pembuktian fenomena bencana alam yang terjadi, termasuk di Indonesia. Apalagi jika melihat data Badan Nasional Penanggulangan Bencana (BNPB) tentang meningkatnya kejadian bencana alam dan korban jiwa seperti yang diuraikan di atas, seharusnya menjadi perhatian pemerintah dan masyarakat Indonesia. Perhatiannya tidak sebatas tanggap pada isu pemanasan global yang mendunia, tapi justru kesiapsiagaan menghadapi bencana alam yang terjadi sebagai akibat dari pemanasan global dan kerusakan lingkungan (Tahminden & Krismanto, 2019).