KLASIFIKASI TINGKAT KEPARAHAN BANJIR DENGAN PENDEKATAN *SUPPORT VECTOR MACHINE* DI JAWA TIMUR

SKRIPSI

Disusun Oleh:

NUR DIANA FAHMA SALSABILA

H96219058

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL
SURABAYA

2023

PERNYATAAN KEASLIAN

Saya yang bertanda tangan di bawah ini,

Nama	: Nur Diana Fahma Salsabila
NIM	: H96219058
Program Studi	: Sistem Informasi
Angkatan	: 2019

Menyatakan bahwa saya tidak melakukan plagiat dalam penulisan skripsi saya yang berjudul: "KLASIFIKASI TINGKAT KEPARAHAN BANJIR DENGAN PENDEKATAN SUPPORT VECTOR MACHINE DI JAWA TIMUR".
Apabila suatu saat nanti terbukti saya melakukan tindakan plagiat, maka saya bersedia menerima sanksi yang telah ditetapkan.

Demikian pernyataan keaslian ini saya buat dengan sebenar-benarnya.

Surabaya, 26 Juni 2023
Yang menyatakan,

Nur Diana Fahma Salsabila
NIM H96219058

LEMBAR PERSETUJUAN PEMBIMBING

Skripsi Oleh

NAMA : NUR DIANA FAHMA SALSABILA

NIM : H96219058

JUDUL : KLASIFIKASI TINGKAT KEPARAHAN BANJIR DENGAN PENDEKATAN *SUPPORT VECTOR MACHINE* DI JAWA TIMUR

.

Ini telah diperiksa dan disetujui untuk diujikan.

Surabaya, 26 Juni 2023

Dosen Pembimbing 1	Dosen Pembimbing 2
Dwi Rolliawati, M.T	Khalid, M. Kom
NIP. 197909272014032001	NIP. 197906092014031002

PENGESAHAN TIM PENGUJI SKRIPSI

Skripsi Nur Diana Fahma Salsabila ini telah dipertahankan
di depan tim penguji skripsi di Surabaya, 6 Juli 2023

Mengesahkan,
Dewan Penguji

Penguji 1

Dr. Eng. Anang Kunaefi, M.Kom.
NIP. 197911132014031001

Penguji 2

Subhan Nooriansyah, M.Kom.
NIP. 199012282020121010

Penguji 3

Dwi Rolliawati, M.T.
NIP. 197909272014032001

Penguji 4

Khalid, M.Kom.
NIP. 197906092014031002

Mengetahui,
Dekan Fakultas Sains dan Teknologi
UIN Sunan Ampel Surabaya

Dr. Saepul Hamdani, M.Pd
NIP. 196507312000031002

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
PERPUSTAKAAN
Jl. Jend. A. Yani 117 Surabaya 60237 Telp. 031-8431972 Fax.031-8413300
E-Mail: perpus@uinsby.ac.id

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama : NUR DIANA FAHMA SALSABILA
NIM : H96219058
Fakultas/Jurusan : Sains dan Teknologi / Sistem Informasi
E-mail address : h96219058@student.uinsby.ac.id

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :
■ Skripsi ☐ Tesis ☐ Desertasi ☐ Lain-lain (……)
yang berjudul : KLASIFIKASI TINGKAT KEPARAHAN BANJIR DENGAN PENDEKATAN SUPPORT VECTOR MACHINE DI JAWA TIMUR

beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Ekslusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara ***fulltext*** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya, 16 Juli 2023
Penulis

NUR DIANA FAHMA SALSABILA

ABSTRAK

KLASIFIKASI TINGKAT KEPARAHAN BANJIR DENGAN PENDEKATAN SUPPORT VECTOR MACHINE DI JAWA TIMUR

Oleh:

Nur Diana Fahma Salsabila

Berdasarkan laporan Badan Nasional Penanggulangan Bencana (BNBP) tahun 2022, wilayah Jawa Timur termasuk salah satu wilayah yang rawan akan bencana banjir sebanyak 92 kali dengan kategori tingkat keparahan yang dimulai dari rendah, sedang, dan tinggi. Telah banyak upaya pemerintah dalam penyebaran informasi banjir melalui media sosial seperti *Twitter*. Pemanfaatan data *Twitter* dalam bidang *text classification* dengan menggunakan *Support Vector Machine* (SVM) sudah banyak dilakukan untuk mengekstraksi dan mengklasifikasikan informasi penting, khususnya dalam domain kebencanaan. Penelitian ini bertujuan untuk membantu masyarakat dan pemerintah dalam memberikan informasi tingkat keparahan banjir di Jawa Timur dengan pendekatan SVM. Metode *extraction feature* yang digunakan adalah *tfidfvectorizer* dengan pembagian dataset 80% *training* dan 20% *testing*. Pembuatan model SVM dengan menggunakan *multiclass* OVR, kernel RBF, kemudian divalidasi dengan 5 *fold cross validation*. Model tersebut menghasilkan performa yang sangat baik dengan nilai *Accuracy* 89%, *F1-Score* 83%, AUC *class* rendah 0.998, AUC *class* sedang 0.993, AUC *class* tinggi 0.983, dan *Logloss* 0.1160. Berdasarkan hasil tersebut dapat disimpulkan, bahwa model yang dibuat memiliki kemampuan yang sangat baik dalam mengklasifikasikan tingkat keparahan banjir.

Kata Kunci: *Natural Language Processing, Tingkat Keparahan Banjir, Support Vector Machine, multiclass OVR, ROC AUC, Logloss.*

ABSTRACT

CLASSIFICATION OF FLOOD SEVERITY WITH THE SUPPORT VECTOR MACHINE APPROACH IN EAST JAVA

By:

Nur Diana Fahma Salsabila

Based on the 2022 National Disaster Management Agency (BNBP) report, the East Java region is one of the areas prone to flooding 92 times with severity categories starting from low, medium and high. There have been many government efforts in disseminating flood information through social media such as Twitter. The use of Twitter data in the field of text classification using Support Vector Machine (SVM) has been widely used to extract and classify important information, especially in the disaster domain. This study aims to assist the community and government in providing information on the severity of floods in East Java using the SVM approach. The feature extraction method used is tfidfvectorizer with 80% training and 20% testing dataset division. SVM model creation using OVR multiclass, RBF kernel, then validated with 5 fold cross validation. This model produces very good performance with an Accuracy value of 89%, F1-Score 83%, low class AUC 0.998, medium class AUC 0.993, high class AUC 0.983, and Logloss 0.1160. Based on these results it can be concluded that the model created has a very good ability to classify the severity of floods.

Keywords*: Natural Language Processing, Flood Severity, Support Vector Machine, multiclass OVR, ROC AUC, Logloss.*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Indonesia termasuk salah satu negara yang terletak di garis khatulistiwa dengan iklim tropis dan curah hujan yang tinggi, sehingga rawan bencana alam. Salah satu bencana alam yang sering terjadi beberapa akhir tahun ini, yaitu banjir. Berdasarkan laporan Badan Nasional Penanggulangan Bencana (BNBP) tahun 2022, wilayah Jawa Timur termasuk salah satu wilayah yang rawan akan bencana banjir dengan total kejadian sebanyak 92 kali. Banjir merupakan peristiwa alam yang terjadi akibat volume air tidak tersalur dan tidak tertampung dengan baik dalam penampung air seperti sungai dan danau (Utami & Tyas, 2021).

Banjir dapat dicegah dengan menjaga kebersihan lingkungan sekitar dan melakukan reboisasi. Telah banyak upaya pemerintah dalam mengatasi banjir seperti memberikan informasi melalui media sosial maupun media massa. Di era pesatnya pengguna media sosial di kalangan masyarakat, banyak Informasi yang dibagikan seperti *tweet* pada aplikasi *Twitter*. Pada salah satu penelitian yang membahas pemanfaatan data *Twitter* dalam penanggulangan bencana banjir dan longsor hasilnya menunjukkan, bahwa aplikasi *Twitter* dapat memberikan informasi lebih cepat terkait dengan kejadian bencana banjir dan longsor (Fariz et al., 2021).

Data *Twitter* memiliki peran penting dalam sebuah penelitian, seperti pada penelitian (Aziz et al., 2019) menjelaskan bahwa data *Twitter* dapat memberikan informasi secara cepat seperti ketika terjadinya bencana alam di suatu daerah, banyak masyarakat yang memberikan informasi melalui *Twitter*. Selain itu, data *Twitter* dapat dimanfaatkan dalam bidang *Natural Language Processing* (NLP) untuk mengekstraksi informasi penting. *Text classification* merupakan bagian dari NLP untuk mempersiapkan data latih yang dapat diterapkan pada satu dokumen atau lebih secara otomatis.

Pada penelitian (Mirończuk & Protasiewicz, 2018) menjelaskan, bahwa dengan pemanfaatan berbagai teknik dan dikombinasikan dengan teknik tersebut dalam sistem yang kompleks, *text classification* menjadi topik penelitian yang menonjol dan berkembang dengan baik. Selain itu, *text classification* dapat

diterapkan di berbagai domain seperti *text classification* lintas bahasa, analisis sentimen, *text streaming*, dsb. Dalam penerapannya, *text classification* dapat diselesaikan secara manual maupun otomatis. *Text classification* secara otomatis dilakukan dengan menggunakan pendekatan *machine learning* maupun *deep learning*.

Pada penelitian (Yovellia Londo et al., 2019) telah berhasil mengklasifikasikan artikel berita berbahasa Indonesia dengan menggunakan tiga algoritma *machine learning* yaitu, *Support Vector Machine* (SVM), *Multinomial Naïve Bayes* (MNB), dan *Decision Tree* (DT). Penelitian tersebut menghasilkan nilai akurasi tertinggi sebesar 93% pada model SVM, kemudian 86% pada model MNB dan 80% pada model DT. Dalam penelitian lain, penggunaan *automated text classification* untuk terjemahan bahasa arab dengan menggunakan metode *Naïve Bayes* (NB), DT, dan *K-Nearest Neighbors* (KNN) (Ababneh, 2019). Penelitian tersebut menghasilkan nilai akurasi tertinggi pada model NB sebesar 88%, KNN 84,65%, dan DT 82%.

Text classification tidak hanya diterapkan pada domain artikel berita dan terjemahan bahasa, tetapi dapat digunakan dalam domain kebencanaan. Pada penelitian (Sreenivasulu & Sridevi, 2020) yang menerapkan *text classification* untuk deteksi kejadian gempa bumi dengan menggunakan KNN, DT, *Neural Network,* SVM, dan *Random Forest* (RF). Penelitian tersebut menghasilkan nilai akurasi tertinggi pada model SVM sebesar 77%. Penelitian lanjutan diperlukan untuk memeriksa skalabilitas dan kompatibilitas SVM. Penelitian tersebut dapat digunakan untuk mendeteksi bencana lain dengan bahasa berbeda dan menggunakan data lebih banyak, agar menghasilkan akurasi yang lebih baik.

SVM merupakan salah satu algoritma *Supervised Learning* yang dapat digunakan untuk berbagai masalah klasifikasi seperti klasifikasi email spam atau *non spam.* Pada penelitian (Mosavi et al., 2018) menjelaskan, bahwa SVM merupakan metode yang sangat populer dan efisien dalam pemodelan kebencanaan khususnya banjir. Selain itu, pada penelitian (Kowsari et al., 2019) menjelaskan bahwa SVM dapat mengatasi masalah *overfitting* pada kinerja model terutama dalam kumpulan teks yang berdimensi tinggi. Selain itu, pada penelitian (Khalaf et al., 2018) yang telah berhasil memprediksi tingkat keparahan banjir

dengan menggunakan beberapa algoritma klasifikasi, salah satunya SVM *multiclass* yang menghasilkan nilai akurasi sebesar 87% pada kelas pertama, 72% pada kelas kedua, dan 77% pada kelas ketiga.

Berdasarkan uraian diatas, penelitian ini dengan mengajukan "KLASIFIKASI TINGKAT KEPARAHAN BANJIR DENGAN PENDEKATAN *SUPPORT VECTOR MACHINE* DI JAWA TIMUR" diharapkan dapat membantu masyarakat dan pemerintah dalam memberikan informasi terkait tingkat keparahan banjir yang terjadi di daerah Jawa Timur.

1.2 Perumusan Masalah

Sesuai dengan penjabaran permasalahan pada latar belakang, rumusan masalah pada penelitian ini yaitu:

1. Bagaimana Penerapan Algoritma *Support Vector Machine* Untuk Klasifikasi Tingkat Keparahan Banjir di Jawa Timur?
2. Bagaimana Tingkat Performa Algoritma *Support Vector Machine* Untuk Klasifikasi Tingkat Keparahan Banjir di Jawa Timur?

1.3 Batasan Masalah

Agar penelitian ini dapat berjalan sesuai dengan harapan, berikut batasan masalah penelitian ini yaitu:

1. *Dataset* diambil dari *tweet*s berbahasa Indonesia dengan kata kunci banjir pada bulan Oktober 2022 – Februari 2023 yang disesuaikan dengan data musim hujan pada BMKG Jawa Timur.
2. Filter lokasi di Jawa Timur hanya lingkup kabupaten/kota.
3. Jumlah *tweet* yang didapatkan maksimal 500.000.
4. Kata kunci tingkat keparahan yang digunakan terdiri dari label rendah, sedang, dan tinggi.

1.4 Tujuan Penelitian

Sesuai dengan perumusan masalah diatas, penelitian ini bertujuan untuk:

1. Untuk mengetahui penerapan Algoritma *Support Vector Machine* Untuk Klasifikasi Tingkat Keparahan Banjir di Jawa Timur.
2. Untuk mengetahui Tingkat Performa Algoritma *Support Vector Machine* Untuk Klasifikasi Tingkat Keparahan Banjir di Jawa Timur.

1.5 Manfaat Penelitian

Dengan adanya penelitian ini, terdapat beberapa manfaat yang diberikan, diantaranya:

1. Penelitian ini memberikan informasi tingkat keparahan dari kejadian banjir di Jawa timur dengan pendekatan *Support Vector Machine*.
2. Penelitian ini dapat menjadi pertimbangan bagi pemerintah dalam mengambil keputusan penanggulangan banjir.
3. Penelitian ini dapat menjadi referensi ilmiah bagi peneliti lain yang akan mengembangkan *text classification* khususnya dalam tingkat keparahan banjir.
4. Penelitian ini dapat dijadikan referensi ilmiah bagi penelitian lain yang akan mengatasi kejadian *non* banjir seperti banjir hadiah, banjir selamat, dsb. dengan menggunakan corpus yang telah dibuat pada penelitian ini.

BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Penelitian Terdahulu

Tinjauan penelitian terdahulu dilakukan agar dapat memberikan pemahaman dan menjadi dasar pengetahuan penelitian terkait *text classification* berbahasa Indonesia. Pada subbagian ini menjelaskan terkait dengan penelitian terdahulu yang memecahkan masalah sejenis dengan berbagai metode yang dijelaskan pada Tabel 2.1 dan Tabel 2.2.

Tabel 2. 1 Penelitian Terdahulu

No	Judul	Pembahasan & Hasil	GAP
1.	"*Comparative Study of Statistical Features to Detect the Target Event During Disaster*" (Sreenivasulu & Sridevi, 2020)	Penelitian ini bertujuan untuk deteksi gempa dengan menggunakan SVM dan menghasilkan akurasi tertinggi sebesar 77%.	Penambahan proses *preprocessing* seperti *handel slang words* dan *stemming* digunakan pada penelitian yang diajukan agar dapat menghasilkan akurasi yang lebih baik.
2.	"*A Study of Text Classification for Indonesian: New Article*" (Yovellia Londo et al., 2019)	Penelitian ini bertujuan untuk *text classification* artikel bahasa Indonesia menggunakan SVM, MNB, dan DT. Hasil akurasi terbaik 93% diperoleh model SVM.	*Feature extraction* pada penelitian yang diajukan menggunakan TF-IDF seperti pada penelitian ini.
3.	"*Application of Naïve Bayes, Decision Tree, and K-Nearest Neighbors for Automated Text Classification*" (Ababneh, 2019)	Penelitian ini bertujuan untuk menyelidiki kinerja KNN, DT, dan NB dalam *text classification* bahasa arab otomatis yang menghasilkan nilai akurasi tertinggi pada model NB sebesar 88%.	Untuk meningkatkan hasil evaluasi pada *text classification* penelitian ini, dibutuhkan penelitian lanjutan dengan menggunakan algoritma yang lain.

Pada Tabel 2.1 menjelaskan terkait dengan tinjauan pustaka implementasi *text classification* dengan menggunakan berbagai metode. Penelitian tersebut digunakan sebagai wawasan dalam memahami penerapan *text classification* secara umum dan digunakan sebagai dasar penelitian ini.

Tabel 2. 2 Penelitian Terdahulu Lanjutan

No.	Topik	Pembahasan & Hasil	GAP
1.	"*Flood Prediction Using Machine Learning Models: Literature Review*"(Mosavi et al., 2018)	SVM merupakan model popular dan akurat yang digunakan untuk pemodelan banjir dengan hasil yang menjanjikan dan kinerja yang lebih baik.	Alasan penggunaan SVM pada penelitian yang diajukan mengacu pada penelitian ini.
2.	"Pemanfaatan Metode *Multiclass* – SVM pada Model Klasifikasi Pesan Bencana Banjir di *Twitter*" (Delimayanti et al., 2021)	Dalam penelitian ini menggunakan teknik *one versus all* (OVA) pada *multiclass* – SVM dengan klasifikasi 3 kelas berbeda yang menghasilkan nilai akurasi sebesar 87.03%.	Proses penerapan *multiclass* - SVM penelitian yang diajukan mengacu pada penelitian ini.
3.	"*A Data Science Methodology Based on Machine Learning Algorithms for Flood Severity Prediction*" (Khalaf et al., 2018)	Pada penelitian ini telah berhasil melakukan klasifikasi dengan RFC, SVM, dan LEVNN kemudian dilakukan prediksi dengan LNN dan ROM. Hasil akurasi tertinggi didapatkan RFC kemudian SVM, LEVNN, ROM, dan LNN.	Data yang digunakan pada penelitian ini menggunakan *dataset* yang dikumpulkan dari situs web badan lingkungan, sedangkan pada penelitian yang diajukan menggunakan data teks (*Twitter*).
4.	"*Text Classification Algorithms: A Survey*"(Kowsari et al., 2019)	Penelitian ini bertujuan untuk memberikan gambaran secara umum terkait *text classification* berdasarkan survei.	Proses *text classification* pada penelitian yang diajukan mengacu pada penelitian ini.

Pada Tabel 2.2 menjelaskan tentang tinjauan pustaka terkait dengan *text classification.* Penelitian tersebut membantu dalam memahami proses secara umum ke khusus, teknik, dan beberapa algoritma yang dapat digunakan dalam *text classification*.

2.2 Teori Dasar

2.2.1 *Machine Learning*

Machine Learning merupakan bagian dari *artificial intelligence* yang melalui proses pembelajaran dapat memahami pola pada data masukan dengan label yang beragam (Bickel et al., 2023). *Machine learning* merupakan komponen penting dalam perkembangan ilmu di bidang data yang mencakup beberapa kombinasi ilmu pengetahuan seperti matematika, statistika, data *mining*, dsb.

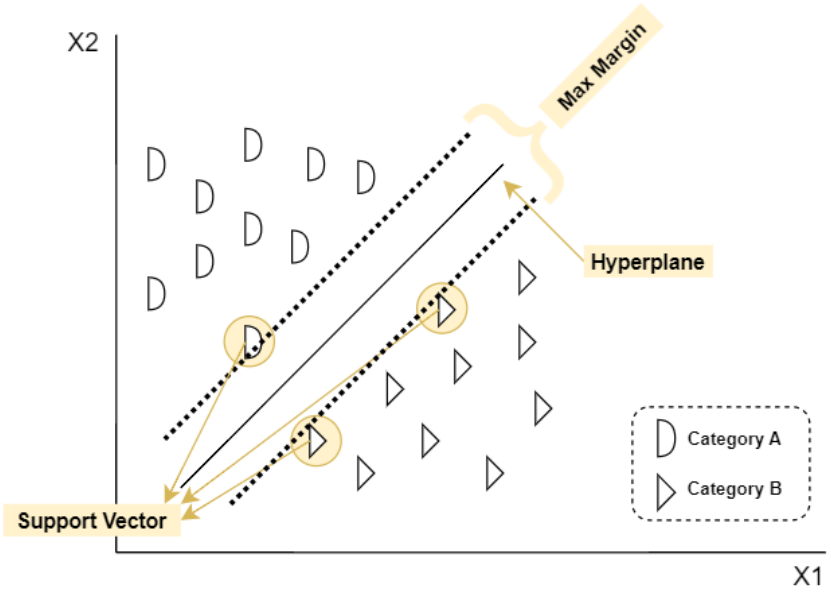
Penggunaan *machine learning* dapat ditemui dalam kegiatan sehari-hari seperti, penggunaan asisten *virtual* dan pengenalan suara dengan *google assistant, alexa,* serta *sirri*. Terdapat beberapa penelitian *machine learning* yang digunakan baik dalam *text classification* maupun bukan *text classification* diantaranya, (Delimayanti et al., 2021) untuk klasifikasi pesan bencana banjir di *Twitter*, (Yovellia Londo et al., 2019) untuk *text classification* artikel bahasa Indonesia, (Sreenivasulu & Sridevi, 2020) untuk deteksi gempa, dan (Mintarya et al., 2023) untuk memprediksi pasar saham.

Berdasarkan tipe pembelajaran, *machine learning* dibedakan menjadi tiga yaitu:

1. *Supervised Learning*

Supervised learning mencerminkan kemampuan suatu algoritma untuk menggeneralisasi pengetahuan dari data yang tersedia dengan target atau kasus berlabel (Berry, Mohamed and Yap, 2020). Algoritma yang termasuk dalam *supervised learning* adalah *estimation, forecasting,* dan *classification*. Terdapat beberapa metode yang termasuk *supervised learning* yaitu, *naïve bayes* (NB), *linear regression* (LR), *logistic regression* (LogR), *support vector machine* (SVM), dan *random forest* (RF).

Support Vector Machine (SVM) termasuk dalam metode machine learning yang menerapkan prinsip *Structural Risk Minimization* (SRM) guna menemukan *hyperplane* terbaik dengan menggunakan fungsi linier maupun *non* linear yang berdimensi tinggi. SVM menurut (Fesseha et al., 2020) merupakan model pengklasifikasian terawasi yang paling kuat dan populer. Model tersebut dibangun secara linier maupun *non* linier dengan melalui *hyperplane* dalam ruang berdimensi N yang digunakan untuk memisahkan data sesuai kelas dan menemukan bidang dengan margin maksimum. Implementasi SVM secara *non* linear dilakukan dengan menggunakan konsep kernel. Hasil evaluasi model SVM sangat bergantung pada pemilihan fungsi kernel dan parameter yang digunakan (Parapat et al., 2018).



Gambar 2. 1 Ilustrasi SVM

Sesuai ilustrasi SVM pada Gambar 2.1, terdapat beberapa istilah penting diantaranya:

a) *Hyperplane*, sebuah garis pembantu mengklasifikasi titik data dengan jumlah dimensi yang bergantung pada jumlah fitur.
b) *Support vector*, titik data terluar dengan posisi hampir *overlap* dengan kelas lain dan paling dekat dengan posisi *hyperplane*. Objek ini digunakan untuk menemukan *hyperplane* yang paling optimal.
c) *Margin*, jarak antara support vector dengan *hyperplane*.
d) Pola, anggota dari dua kelas yang dipisahkan dengan *hyperplane* dengan nilai 1 dan kelas lain dengan nilai -1.

Pada dasarnya implementasi SVM dilakukan secara linear, namun seiring dengan perkembangannya SVM dapat digunakan pada data *non* linear dengan menggunakan bantuan kernel. Terdapat empat jenis kernel SVM, yaitu:

a) Linear Kernel

Teknik linear SVM sangat efisien untuk aplikasi data berdimensi tinggi seperti klasifikasi dokumen, disambiguasi kata. Kompleksitas komputasi dalam *training* SVM linear jauh lebih cepat daripada SVM *non* linear (Chauhan et al., 2019).

$$F(x) = w^T x + b \quad (1)$$

Dimana w^T adalah vektor bobot dan b adalah *deviation* yang bertujuan untuk menentukan klasifikasi dengan mencari nilai *deviation* terkecil yang paling maksimum.

b) *Radial Basis Function* (RBF) Kernel

$$K(x,y) = \exp(-\frac{||x-y||^2}{2\sigma^2}) \quad (2)$$

c) *Polynomial* Kernel

$$K(x,y) = (<x, y> + 1)^d \quad (3)$$

d) *Sigmoid* Kernel

$$K(x, x') = \tanh(k1(x,x') + k2)) \quad (4)$$

Dengan nilai tanh,

$$\text{Tanh } x = \frac{e^x - e^{-x}}{e^x - e^{-x}} \quad (5)$$

Dalam penelitian (Kalcheva et al., 2020) menghasilkan nilai akurasi tertinggi pada kernel linear sebesar 84,80% dalam mengklasifikasikan *review* film. Penelitian tersebut menyebutkan bahwasannya sigmoid kernel merupakan pilihan yang tidak tepat dalam *text classification*.

Secara dasar, algoritma SVM digunakan untuk klasifikasi ke dalam dua kelas (biner). Seiring dengan perkembangannya, SVM dapat digunakan untuk klasifikasi lebih dari dua kelas dengan menggunakan *multiclass* SVM. Terdapat tiga pendekatan umum yang digunakan dalam *multiclass* SVM (Zhang, 2020) yaitu:

a) *One Versus One* (OVO), Teknik klasifikasi SVM dengan cara membandingkan satu kelas dengan satu kelas lainnya. Pendekatan ini dilakukan pada kasus yang memiliki dua kelas label.
b) *One Versus All* (OVA), Teknik klasifikasi OVA SVM ini disebut dengan "*one versus the rest*" yang dilakukan dengan membandingkan sebuah kelas dengan semua kelas. Teknik ini menggunakan kernel linear secara *default*. Pendekatan ini digunakan pada kasus yang memiliki kelas lebih dari dua.
c) *Directed Acyclic Graph* (DAG), Teknik klasifikasi pada DAG SVM ini bersifat hierarki dengan sistem gugur yang artinya kelas yang sudah kalah, maka tidak diikutsertakan lagi.

2. *Unsupervised Learning*

Proses *unsupervised learning* mengacu pada pengelompokan data dalam *cluster* dengan menggunakan metode atau algoritma untuk data yang belum dikategorikan (Berry, Mohamed and Yap, 2020). Salah satu algoritma yang termasuk dalam *unsupervised learning* adalah *clustering*. Terdapat empat metode yang digunakan yaitu *Fuzzy C-Means, K-Means, K-Medoids, dan Self-Organizing Map*.

3. Semi-*Supervised Learning*

Semi supervised learning dapat diilustrasikan ketika seorang pelajar dapat secara bertahap meningkatkan kinerjanya dengan belajar dari sampel yang tidak berlabel tanpa interaksi eksternal (Zhou, 2021). Dengan kata lain, pembelajaran *semi supervised* mengambil asumsi dunia terbuka, yaitu model yang akan dipelajari dapat diterapkan pada sampel yang tidak teramati. Dalam proses pembelajarannya, *semi supervised learning* menggunakan data berlabel dan tidak berlabel secara bersamaan. Penggunaan *semi supervised learning* ini dapat digunakan pada *speech recognition*, dan *web content classification*.

Dalam pembuatan model dengan menggunakan *machine learning*, namun tidak memungkinkan untuk terjadinya permasalahan. Terdapat dua macam permasalah dalam model *machine learning* yang sering dijumpai, yaitu:

a) *Overfitting*

Model dikatakan *overfitting* apabila model menghasilkan nilai *Accuracy* yang cukup bagus dalam *training* data, namun dalam proses *testing* data menghasilkan nilai *Accuracy* yang lebih rendah atau bahkan buruk (Bashir et al., 2020). Overfitting terjadi apabila model dilatih sangat kompleks sehingga menghasilkan varian tinggi tetapi nilai bias yang dihasilkan rendah (Ghojogh & Crowley, 2019).

b) *Underfitting*

Underfitting merupakan keadaan model yang menghasilkan nilai *Accuracy* yang rendah dalam *training* data dan *testing* data (Bashir et al., 2020). Dengan kata lain, apabila model dilatih sangat sederhana sehingga varian estimasinya rendah tetapi nilai bias yang didapatkan tinggi (Ghojogh & Crowley, 2019).

2.2.2 *Text Mining*

Text Mining merupakan sebuah proses penggalian informasi terstruktur berkualitas tinggi dari teks atau dokumen yang tidak terstruktur kemudian diubah menjadi informasi yang berguna sebagai landasan untuk pengambilan keputusan (Pejić Bach et al., 2019). *Text Mining* bertujuan untuk mengatur, memahami, serta mengungkapkan pola semantic yang tersembunyi dalam dokumen. *Text Mining* dikenal sebagai *Intelligent Text Analysis* (ITA), *Text* Data *Mining* (TDM), atau *Knowledge Discovery in Text* (KDT) yang umumnya mengacu pada proses penggalian informasi dan pengetahuan yang menarik dan *non trivial* dari teks yang tidak terstruktur (Ferreira-Mello et al., 2019).

Berikut merupakan langkah - langkah dalam pemrosesan *Text Mining*:

1. *Text Preprocessing*

Text preprocessing merupakan langkah awal pada penambangan teks untuk mempersiapkan teks menjadi data yang dapat digunakan pada proses berikutnya. Langkah ini terdiri dari:

a) Segmentasi merupakan tahap memecah seluruh dokumen menjadi kalimat dengan mengelompokkan berdasarkan tanda baca.
b) *Tokenizing* merupakan tahap penguraian kalimat menjadi kata dengan menghapus spasi, koma, dsb.
c) *Stopwords* merupakan tahap membuang kata yang cenderung tidak memiliki arti seperti yang, dan, di, ke, dsb.
d) *Stemming* merupakan tahap mencari akar atau kata dasar dari tiap kata dengan menghilangkan infiks, sufiks, dan prefiks.

2. *Text Transformation*

Text transformation merupakan tahap yang digunakan untuk mengontrol kapitalisasi teks dengan cara *bag of words* dan *vector space*.

3. *Feature Extraction/Selection*

Pada di tahap ini proses *text mining* berubah menjadi data *mining* yang bertujuan untuk mengurangi kelebihan beban komputasi dalam pengklasifikasian data. Terdapat beberapa area penerapan *text mining* yaitu, *information extraction, information retrieval, clustering, text summarization, dan natural language processing* (NLP). NLP merupakan serangkaian teknik komputasi untuk analisis

otomatis dan melakukan berbagai tugas terkait bahasa alami (Young et al., 2018). Dengan kata lain NLP merupakan ilmu yang mencakup komputer, linguistik, teoritis, dan pembelajaran mesin untuk mempelajari cara komputer dan manusia berkomunikasi dalam bahasa alami.

2.2.3 *Text Classification*

Menurut (Bhavani & Santhosh Kumar, 2021) *Text Classification* merupakan teknik mengkategorisasikan teks atau tag dalam kelas yang terorganisir atau sesuai dengan konteks ataupun isinya. *Text Classification* disebut sebagai *text tagging* atau *text categorization* dan termasuk dalam salah satu bidang penelitian NLP dan *text mining* yang dipelajari secara luas. Menurut (Kowsari et al., 2019) proses *text classification* terdiri dari 4 proses yaitu:

1. *Feature extraction*

Proses *feature extraction* digunakan untuk merubah format data *text* menjadi numerik agar dapat diproses pada tahap selanjutnya. Pada proses ini terdiri dari 4 sub proses yaitu:

a) *Text cleaning* dan *preprocessing*, sub proses ini dilakukan untuk menghilangkan karakter dan kata yang tidak penting. Terdapat beberapa teknik yang digunakan seperti data *cleaning, tokenization, stopwords, case folding, slang words, stemming,* dan POS-*Tagging*.

b) *Syntactic word representation*, sub proses ini dilakukan untuk mengatasi hilangnya hubungan sintaksis dan semantik antar kata. Pada sub proses ini dapat dilakukan dengan menggunakan teknik *N-Gram*.

c) *Weighted words*, sub proses ini digunakan untuk memberikan bobot pada kata. Terdapat beberapa teknik yang digunakan seperti *Bag of Words* (BoW) dan *Term Frequency – Inverse Document Frequency* (TF-IDF).

d) Word embedding, sub proses ini digunakan untuk memetakan kata ke vektor dimensi N sekaligus memberikan bobot pada kata. Terdapat beberapa teknik yang digunakan seperti *Word2Vec*, *Global Vectors* (Glove), dan *FastText*.

2. *Dimension reduction*

Proses ini dilakukan untuk menghindari penurunan kinerja serta mengurangi tingkat kompleksitas waktu dan memori yang digunakan. Dalam

proses ini terdapat beberapa teknik yang digunakan seperti PCA, ICA, LDA, LSI, NMF, *Random Projection*, dan *Autoencoder*.

3. *Classifier selection*

Proses ini digunakan untuk membuat model *text classification* dengan menggunakan *machine learning*, *ensemble learning*, dan *deep learning*. Terdapat beberapa algoritma yang digunakan seperti *Rocchio, Boosting, Bagging, Logistic Regression*, NB, KNN, SVM, DT, Random Forest, CNN, RNN, dsb.

4. *Evaluation*

Dalam proses pemodelan *machine learning* diperlukan untuk model berkerja dengan baik, mendapatkan pola yang sesuai dan benar serta tidak menimbulkan banyak *noise*. *K-fold Cross validation* merupakan teknik yang digunakan untuk evaluasi kinerja data yang melibatkan pembagian data ke dalam beberapa *k* partisi. Penggunaan *k-fold cross validation* dalam evaluasi kinerja model terbukti dapat meningkatkan hasil akurasi model (A. Ramezan et al., 2019).

Sebelum dilakukan pembagian data dalam setiap partisi, diperlukan penentuan data *training* dan *testing* terlebih dahulu dengan perbandingan 80:20. Sehingga jika diibaratkan memiliki data sebanyak 1.800, maka pembagian data *training* sebanyak 1.440 data sedangkan data *testing* sebanyak 360. Penentuan partisi (k) dalam implementasi *k-fold cross validation* diibaratkan menggunakan k=10, maka artinya 1.440 dibagi 9 yaitu 160 data untuk *training* sedangkan 360 data untuk *testing* yang terdapat pada masing - masing partisi.

Tabel 2. 3 Konsep *K-fold Cross Validation*

Iterasi	**Keseluruhan Data**									
	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	**Test**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**	**Train**
K2		**Test**								
K3			**Test**							
K4				**Test**						
K5					**Test**					
K6						**Test**				
K7							**Test**			
K8								**Test**		
K9									**Test**	
K10										**Test**

Pada baris pertama Tabel 2.3, posisi data training berada di K2 hingga K10 sedangkan posisi data testing berada pada K1. Posisi data training dan data

testing tersebut akan berganti sesuai dengan jumlah iterasi setiap partisi yang telah ditentukan. Setelah melakukan proses *k-fold cross validation*, dilakukan pengukuran performa model klasifikasi dengan menggunakan *Recall, Precision, Accuracy, F-Measure, ROC AUC,* dan *Logloss*. Penjelasan lebih rinci terkait dengan pengukuran performa atau evaluasi sebagai berikut,:

a) *Recall,* disebut dengan *sensitivity* dan *True Positive Rate,* digunakan untuk mengukur tingkat keberhasilan model dalam menemukan sebuah informasi kembali (Basavegowda & Dagnew, 2020).
b) *Precision* digunakan untuk mengukur keakuratan model atau sistem dengan menilai proporsi contoh positif yang diidentifikasi dengan benar dari total contoh yang diidentifikasi sebagai positif (Fergus et al., 2018).
c) *Accuracy* digunakan untuk mengevaluasi kemampuan prediksi keseluruhan model dan mengukur tingkat kesesuaian antara hasil prediksi dengan nilai sebenarnya (Latha & Jeeva, 2019).
d) *F-Measure* digunakan untuk mendapatkan nilai *harmonic mean* yang mencerminkan tingkat kesesuaian prediksi dengan nilai sebenarnya.
e) *Receiver Operating Characteristic* (*ROC*) *& Area Under the Curve* (*AUC*) digunakan untuk . Apabila kurva ROC dan AUC sebuah model yang bergerak semakin ke pojok kiri atas, maka semakin baik pula akurasi klasifikasi yang dihasilkan (Wang et al., 2021). Dengan kata lain semakin tinggi nilai AUC yang dihasilkan, maka akan semakin baik pula performa model dalam mengklasifikasikan setiap kelas.
f) *Logloss* digunakan untuk mengukur tingkat kesalahan (*eror*) model dan mengukur kualitas probabilitas model dalam memprediksi. Model dikatakan baik apabila menghasilkan nilai *logloss* minimum (nol) atau memiliki *eror* yang sedikit (Basavegowda & Dagnew, 2020).

Tabel 2. 4 *Confusion Matrix*

Confusion Matrix		**Prediktif**	
		Positive	***Negative***
Aktual	***Positive***	*True Positive (TP)*	*False Negative (FN)*
	Negative	*False Positive (FP)*	*True Negative (TN)*

Pada Table 2.4 *confusion matrix* terdiri dari:

a) *True Positive* (TP) merupakan kumpulan data bernilai positif dan diprediksi benar sebagai positif.
b) *False Positive* (FP) merupakan kumpulan data bernilai negatif namun diprediksi sebagai positif.
c) *True Negative* (TN) merupakan kumpulan data yang bernilai negatif dan diprediksi benar sebagai positif.
d) *False Negative* (FN) merupakan kumpulan data yang bernilai positif tetapi diprediksi sebagai negatif.

Berdasarkan penjabaran Tabel 2.4 *confusion matrix* diatas, Pengukuran *Recall* pada model dapat dilakukan dengan persamaaan (6) berikut:

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

Pengukuran *Precision* pada model dapat dilakukan dengan persamaaan (7) berikut:

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

pengukuran performa dengan *Accuracy* dapat dilakukan dengan persamaan (8) berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Pengukuran *F-Measure* pada model dapat dilakukan dengan persamaaan (9) berikut:

$$F - Measure = 2\,x\,\frac{Precision\,x\,Recall}{Precision+Recall} \quad (9)$$

Pengukuran *Logloss* pada model dapat dilakukan dengan persamaaan (10) berikut:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (10)$$

2.2.4 *Term Frequency – Inverse Document Frequency* (TD-IDF)

Data penelitian yang digunakan berbentuk teks yang tidak terstruktur dan tidak dapat diproses oleh bahasa pemrograman. Hal tersebut dikarenakan komputer hanya dapat memproses data dalam bentuk angka. Oleh karena itu, diperlukan proses untuk mengubah kumpulan dokumen teks menjadi vektor

numerik (Kim & Gil, 2019). TF-IDF merupakan salah satu teknik yang sering digunakan pada proses pengolahan data tekstual dan dapat digunakan untuk mengubah data teks menjadi data numerik. TF-IDF digunakan untuk memberikan bobot pada kata dengan menggunakan persamaan (8) (Zhao et al., 2018). TF digunakan untuk memunculkan kata tertentu dalam sebuah dokumen, sedangkan IDF menyiratkan seberapa banyak kata tertentu yang muncul dalam sebuah dokumen.

$$TFIDF\ (t,d) = TF(t,d)\ x\ IDF(t) \quad (8)$$

Frekuensi kata T dalam dokumen D adalah TF yang digunakan untuk menghitung kemampuan kata dalam mendeskripsikan dokumen. Sementara itu, IDF merepresentasikan frekuensi dokumen D yang mengandung kata T dalam korpus yang digunakan untuk menghitung kemampuan kata membedakan dokumen tersebut. Jika frekuensi suatu kata tinggi di dokumennya sendiri tetapi rendah di dokumen lain, kata ini memiliki kemampuan yang kuat untuk membedakan dari dokumen lain dan diberi bobot yang tinggi.

2.2.5 Banjir

Berdasarkan laporan Badan Nasional Penanggulangan Bencana (BNPB) tahun 2022 kejadian bencana alam di Indonesia terjadi sebanyak 3.544 peristiwa yang didominasi oleh banjir. Banjir termasuk dari salah satu bencana alam yang dapat menyebabkan risiko dan mengganggu kehidupan manusia. Banjir terjadi dikarenakan debit aliran sungai melebihi kapasitas aliran sungai atau sistem drainase akibat curah hujan yang cukup tinggi dan berlangsung lama sehingga terjadi peluapan air yang dapat menggenangi wilayah daratan sekitarnya (Yuliantika and Kartika, 2022). Selain itu, Banjir disebabkan oleh faktor alami, non alami maupun manusia tanpa upaya pencegahan yang menyebabkan perubahan struktur bumi. Menurut website databoks.co.id (Annur, 2023), Salah satu provinsi di Indonesia yang sering mengalami banjir yaitu Jawa Timur dengan total kejadian sebanyak 400 kali.

Banjir dapat dikategorikan berdasarkan asal sumber air dan jenis air. Berdasarkan sumber air, banjir dibedakan menjadi dua macam yaitu banjir lokal dan banjir kiriman. Banjir lokal merupakan banjir yang terjadi akibat dari

tingginya curah hujan sehingga, terjadi peluapan air dari penampungan air seperti danau, waduk, dan sungai sedangkan banjir yang diakibatkan oleh kiriman dari daerah yang lebih tinggi disebut dengan banjir kiriman. Kejadian banjir memberikan dampak secara signifikan pada seluruh sektor. Agar dampak banjir tidak semakin parah, pendeteksian dini kejadian banjir diperlukan untuk menekan dampak dari banjir. Telah banyak penelitian yang membahas upaya penanggulangan banjir seperti deteksi tinggi muka air dengan menggunakan *water gauge*, sistem sensor deteksi banjir, dan penggunaan alat seperti *water level detection* (Utomo, Irawan and Alinra, 2021).

Bencana alam banjir dapat berdampak negatif terhadap lingkungan sekitar apabila curah hujan yang turun sangat deras dengan durasi waktu yang cukup lama. Berdasarkan jenisnya, banjir dikategorikan menjadi banjir bandang dan banjir rob. Banjir bandang disebut sebagai air bah yang merupakan banjir datang secara tiba-tiba dengan volume ketinggian air lebih dari standar dan biasanya berasal dari sungai, selokan, dan berbagai jenis tempat air yang meluap. Ketinggian banjir bandang dapat mencapai 3 meter lebih. Banjir yang diakibatkan oleh air laut yang pasang disebut dengan banjir rob. Banjir rob disebut dengan banjir pasang surut air laut yang dipengaruhi oleh gaya tarik bulan dan matahri terhadap massa air laut di bumi. Ketinggian banjir rob dapat mencapai 1 meter hingga 2 meter.

Berdasarkan Peraturan Kepala Badan Nasional Penanggulangan Bencana No. 02 Tahun 2012 tentang pedoman umum pengkajian risiko bencana, kategori tingkat keparahan banjir terdiri dari tiga kategori, diantaranya:

1) Tingkat Rendah, dalam tingkat ini kedalaman banjir kurang dari 100 centimeter.
2) Tingkat Sedang, dalam tingkat ini kedalaman banjir berada di antara 100 centimeter hingga kurang dari 300 centimeter.
3) Tingkat Tinggi, dalam tingkat ini kedalaman banjir berada di atas 300 centimeter.

2.2.6 Data *Twitter*

Di era perkembangan dan kemajuan teknologi saat ini, banyak media sosial yang mengalami perkembangan, salah satunya *Twitter*. *Twitter* merupakan

salah satu platform yang memungkinkan pengguna berbagi pesan gratis bermodal paket data. Bagi sejumlah pengguna aktif media sosial, *Twitter* merupakan salah satu aplikasi media sosial yang memiliki pengaruh dalam menyebarkan informasi. Sehingga, banyak penelitian terdahulu yang memanfaatkan data *Twitter* sebagai sumber informasi dalam penelitiannya. Pada penelitian (Aziz et al., 2019) menjelaskan bahwa, data *Twitter* memiliki peran penting dalam memberikan informasi secara cepat seperti ketika terjadinya bencana alam di suatu daerah, banyak masyarakat yang memberikan informasi melalui *Twitter*. Dalam penelitian tersebut, hasil analisis bencana alam menggunakan *Twitter* memberikan hasil sepuluh kejadian bencana alam yang berbeda.

Terdapat beberapa fitur dalam *Twitter*:

1. *Tweet* atau kicauan merupakan fitur utama dalam *Twitter* berfungsi sebagai pemberi informasi berupa tulisan, gambar maupun video yang diunggah oleh pengguna ke publik.
2. Nama pengguna dalam *Twitter* digunakan untuk mengidentifikasi akun satu pengguna dengan pengguna lain yang diawali dengan karakter "@" kemudian diikuti nama pengguna.
3. *Hashtag* atau tagar dalam *Twitter* digunakan untuk identifikasi pembicaraan yang sedang hangat dibicarakan (*trending topic*) dengan menggunakan karakter "#" pada *tweet*.
4. *Retweet* dalam *Twitter* berfungsi sebagai pemberi informasi ulang dengan cara membagikan kembali postingan pribadi maupun pengguna lain.
5. Fitur balasan dalam *Twitter* digunakan untuk memberi komentar atau tanggapan dalam sebuah *tweet*.

2.2.7 *Python*

Bahasa pemrograman python mendapatkan popularitas yang cukup luar biasa di kalangan ilmuwan data dan pengembang perangkat lunak (Hao & Ho, 2019). Python pertama kali dirilis pada tahun 1991 oleh Guido van Rossum. *Python Language Programming* merupakan bahasa pemrograman bersifat *open source, high level*, dan kuat dengan fitur yang canggih, sementara sintaks dasar python yang mudah dipelajari (Hill, 2020). Python dapat digunakan untuk *web applications, software development, data science,* dan *machine learning*. Dalam

penelitian ini terdapat beberapa *library* yang digunakan untuk mensupport *python language programming* seperti pada Tabel 2.5.

Tabel 2. 5 *Library Python*

No.	***Library Package***	**Keterangan**
1.	*Snscrape*	Digunakan untuk mendapatkan *tweet* dengan teknik *scraping* tanpa API *Twitter*.
2.	*Pandas*	*Package* digunakan untuk struktur dan manipulasi data.
3.	*Numpy*	*Package* dasar yang digunakan untuk komputasi numerik.
4.	*Matplotlib*	*Package* ini menyediakan fungsionalitas dasar *plotting*.
5.	*RegEx*	Modul ini digunakan untuk pre-processing dan filter tingkat keparahan banjir.
6.	*Sastrawi*	*Library* ini digunakan untuk proses stemming pada *preprocessing* teks agar kata berubah menjadi bentuk dasarnya.
7.	*Natural Language Toolkit Python*	*Toolkit* yang digunakan untuk pemrosesan bahasa.
8.	*Crf.tagger*	*Library* ini digunakan untuk proses POS-*Tagging*.
9.	*Scikit learn*	*Tools Machine Learning* pada analisis data.
10.	*Sklearn.svm.SVC*	*Library* untuk implementasi SVC (S*upport Vector Classification*).
11.	*Sklearn,feature extraction.text.TfidfVectorizer*	Digunakan untuk implementasi TF-IDF

Selain dengan menggunakan *package* dalam implementasi *python*, penelitian ini menggunakan beberapa *corpus* maupun *dictionary*. *Corpus* maupun *dictionary* tersebut digunakan untuk membantu dalam proses *preprocessing* agar dapat menyesuaikan dengan kaidah bahasa indonesia. Dalam penelitian ini, terdapat beberapa *corpus* maupun *dictionary* yang digunakan seperti pada Tabel 2.6.

Tabel 2. 6 *Corpus* dan *Dictionary*

No.	***Corpus* / *Dictionary***	**Keterangan**
1.	*Corpus* Colloquial Indonesian Lexicon (Salsabila, 2018)	Digunakan untuk proses handel *slang words*.
2.	*Dictionary slang words* (Owen, 2020)	
3.	*Dictionary slang words* (kirralabs, 2020)	
4.	*Dictionary slang words* (Ibrohim, 2019)	
5.	*Dictionary stopwords* (Devid, 2016)	Digunakan untuk proses menghapus *stopwords*.
6.	*Dictionary Location* Jawa timur	Digunakan untuk *filtering* lokasi Jawa

	(edwardsamuel, 2018)	timur.

2.3 Integrasi Keilmuan

Penelitian ini dilakukan untuk mendeteksi peristiwa dan lokasi banjir sesuai batasan dan metode yang telah diuraikan di bab pendahuluan. Berdasarkan wawancara yang telah dilakukan dengan seorang pakar bernama Maulana Muhammad Fahmi Faiz, S.H yang merupakan Gus di PP. Sumberjati, Pamekasan. Beliau sedang menempuh Pendidikan S2 di Kairo, Mesir. Beliau mengatakan bahwasannya terdapat dua korelasi penelitian ini dengan ayat Al-Qur'an dan Hadits yaitu, peninjauan teknologi dari pandangan islam dan hubungan antara teknologi dengan pemerintahan.

Korelasi yang pertama yaitu peninjauan teknologi dari sudut pandang islam. Adanya teknologi yang canggih saat ini patut disyukuri. Tanpa teknologi, manusia akan mengalami kesulitan dalam menjalankan kehidupannya. Teknologi yang dibuat oleh manusia berdasarkan ilmu pengetahuan, seperti dalam Al-Qur'an Surat Ar-Rahman Ayat 33, Allah berfirman:

يَا مَعْشَرَ الْجِنِّ وَالإِنْسِ إِنِ اسْتَطَعْتُمْ أَنْ تَنْفُذُوا مِنْ أَقْطَارِ السَّمَاوَاتِ وَالأَرْضِ فَانْفُذُوا لا تَنْفُذُونَ إِلا بِسُلْطَان

Artinya: "wahai jin dan manusia, jika kamu mampu menembus (melintasi) penjuru langit dan bumi, maka lintasilah, kamu tidak dapat menembusnya kecuali dengan kekuatan."

Dari ayat tersebut, beberapa ahli menjelaskan bahwa kata "Sulthan" dapat diartikan sebagai kekuatan, kekuasaan, kemampuan serta ilmu pengetahuan. Arti tersebut diperkuat kembali dalam buku Al-Muslimun wa al-Ilm al-Hadits karangan Abdul al-Razzaq Naufal yang mengartikan kata "Sulthan" dengan ilmu pengetahuan dan kemampuan teknologi.

Korelasi yang kedua yaitu, hubungan antara teknologi dengan pemerintahan. Adanya teknologi yang serba canggih dapat membantu pemerintahan dalam melakukan tugasnya seperti, memberikan informasi secara cepat dan akurat sehingga dapat dijadikan landasan pengambilan keputusan, Seperti dalam Al-Qur'an Surat An-Nisa ayat 59:

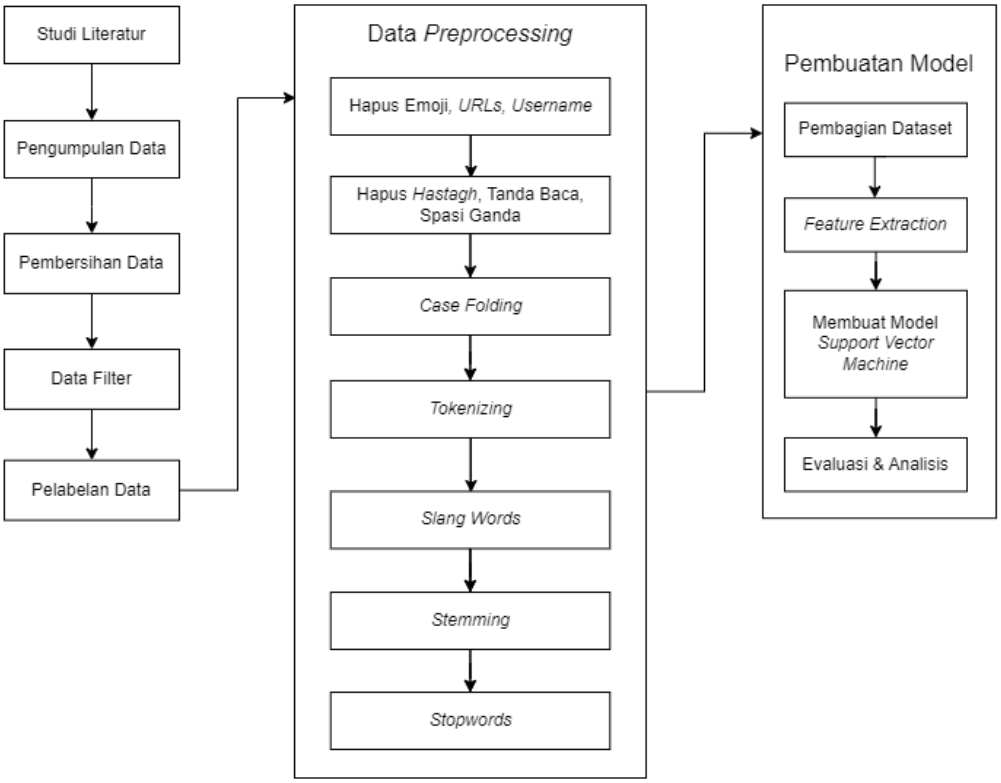
يَاأَيُّهَا الَّذِينَ آمَنُواْ أَطِيعُواْ اللّهَ وَأَطِيعُواْ الرَّسُولَ وَأُوْلِي الأَمْرِ مِنكُمْ فَإِن تَنَازَعْتُمْ فِي شَيْءٍ فَرُدُّوهُ إِلَى اللّهِ وَالرَّسُولِ إِن كُنتُمْ تُؤْمِنُونَ بِاللّهِ وَالْيَوْمِ الآخِرِ ذَلِكَ خَيْرٌ وَأَحْسَنُ تَأْوِيلاً

Artinya: “hai orang-orang yang beriman! Taatilah Allah dan Rasul (Muhammad), dan Ulil Amri (pemegang kekuasaan) di antara kamu. Maka jika kamu berbeda pendapat tentang sesuatu, maka kembalikanlah kepada Allah (Al-Qur’an) dan Rasul (sunnahnya), jika kamu beriman kepada Allah dan hari kemudian.”

Dari ayat tersebut diperintahkan untuk seluruh umat muslimin agar taat dan patuh kepada-Nya, kepada rasul-Nya, dan kepada orang yang memegang kekuasaan. Orang yang memegang kekuasaan dalam hal ini dapat diartikan sebagai pemerintah. Adanya teknologi ini, turut membantu tugas pemerintahan menjadi lebih mudah seperti masyarakat semakin mudah dalam mengakses informasi pemerintah terkait dengan kebijakan dan informasi penting. Dalam hal ini, informasi penting yang dimaksud mengenai informasi kejadian bencana alam khususnya banjir yang dibahas dalam skripsi ini, agar dapat mengantisipasi dan meminimalisir dampak yang ditimbulkan.

BAB III
METODOLOGI PENELITIAN

3.1 Desain Penelitian



Gambar 3. 1 Diagram Alir Metode Penelitian

Gambar 3.1 merupakan Diagram Alir desain penelitian yang digunakan pada penelitian ini. Pencarian penelitian terdahulu yang berkorelasi dengan penelitian ini merupakan tahap awal pada penelitian ini, kemudian mengumpulkan data *Twitter* dengan *library snscrape*. Setelah data didapatkan, proses selanjutnya yaitu melakukan pembersihan data, data filter, pelabelan data, dan data *preprocessing*. Proses data *preprocessing* diawali dengan menghapus emoji, *URLs, Username, Hashtag*, tanda baca dan spasi ganda, kemudian *case folding, tokenizing, slang words, stemming*, dan *stopwords.* Setelah data *preprocessing*, selanjutnya yaitu pembuatan model. Dalam proses pembuatan model dilakukan proses pembagian dataset terlebih dahulu, kemudian *feature extraction* dengan TFIDF, membuat model *text classification* dengan algoritma SVM, dan yang terakhir yaitu mengevaluasi model dan menganalisis hasil.

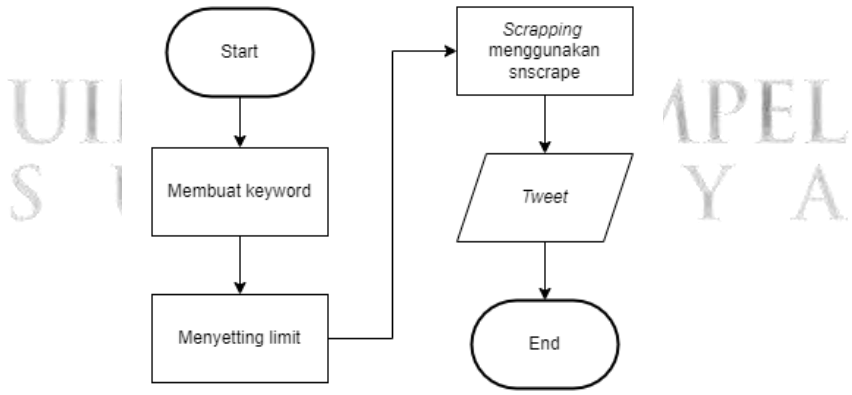
3.2 Uraian Desain Penelitian

3.2.1 Studi Literatur

Studi literatur merupakan tahapan utama dalam penelitian ini dengan cara mencari penelitian terdahulu yang memiliki korelasi dengan penelitian ini, khususnya pada deteksi kejadian bencana alam banjir dengan menggunakan data *Twitter*. Pencarian referensi penelitian terdahulu yang bersumber dari buku dan artikel yang berkaitan dengan NLP, *text classification*, *text mining*, serta SVM dengan menggunakan bantuan alat seperti *google scholar, science direct,* dan *researchgate*.

3.2.2 Pengumpulan Data

Data yang dipakai dalam penelitian ini berasal dari *Twitter* yang berkaitan dengan peristiwa banjir yang berlangsung di Indonesia. Pengumpulan data *Twitter* dilakukan dengan teknik *scraping* dan bantuan *library snscrape* pada *python language programming*. *Snscrape* tidak hanya dapat dilakukan pada *Twitter*, melainkan dapat digunakan untuk mengekstraksi media sosial seperti pada facebook, instagram, dan telegram. *Snscrape* dapat mengekstraksi *tweet* dengan kata kunci tertentu dan tanpa batasan jumlah.



Gambar 3. 2 Diagram Alir Pengumpulan Data

Berdasarkan laporan BMKG Jawa Timur, musim hujan di Jawa Timur terjadi pada bulan Oktober hingga Februari. Pengumpulan data *Twitter* pada penelitian ini dapat dilihat pada Gambar 3.2. Langkah pertama yaitu dengan membuat kata kunci (*keyword*) banjir yang terjadi di Indonesia, kemudian menentukan limit sebanyak 500.000 dengan rentang waktu yang dimulai pada

bulan Oktober 2022 hingga Februari 2023. Hasil yang didapatkan kemudian disimpan dalam format csv. Terdapat beberapa contoh *tweet* pada Tabel 3.1 yang didapatkan dari hasil *scraping* sesuai dengan diagram alir pada Gambar 3.2.

Tabel 3. 1 Contoh Dataset

User	***Data Create***	***Tweet***
korangala	2022-01-31 23:50:43+00:00	JAKARTA Terendam Banjir Besar, Puluhan Orang Meninggal Dunia, Ratusan Ribu Jiwa Mengungsi pada 1 Februari 2007 https://t.co/3ns4duXPQk
twelvebot	2022-01-31 19:18:37+00:00	Tinggal di Depok Terendam Banjir: Hujan deras membuat sejumlah rumah di Haurduni Maja, Tiga Pekerja Luka
arifunasiin	2022-01-31 16:27:39+00:00	Banjir di Jayapura, Relawan PKS Sigap Salurkan Makanan dan Kebutuhan Lainnya di Lokasi Banjir

3.2.3 Pembersihan Data

Pembersihan data dilakukan agar tidak terdapat data yang kosong (*missing value*) dan data duplikat. Proses pembersihan data ini dapat menghindarkan kesalahpahaman dalam menganalisis data. Berikut penjelasan detail dari proses pembersihan data:

a) Menghapus *missing value*

Missing value merupakan sebuah kondisi baris yang tidak memiliki nilai atau informasi yang terkandung pada suatu dataset. Apabila hal tersebut dibiarkan, maka akan memberikan *misinformation* sehingga diperlukan penangan secara khusus untuk *missing value*. *Missing value* yang terdapat pada dataset penelitian ini akan dihapus.

b) Menghapus data duplikat

Data duplikat merupakan kondisi pada suatu dataset yang memiliki dua atau lebih nilai yang sama. Kondisi tersebut dapat memberikan kesalahpahaman dalam suatu data sehingga perlunya penanganan data duplikat secara khusus. Data duplikat yang terdapat pada dataset penelitian ini akan dihapus.

3.2.4 Data Filter

Proses filter data dalam penelitian ini digunakan untuk menyaring dan menghapus data yang tidak termasuk daerah Jawa Timur, kejadian *non* banjir, tingkat keparahan dan angka selain angka tingkat keparahan. Proses filter lokasi

Jawa Timur pada penelitian ini menggunakan bantuan *dictionary* (edwardsamuel, 2018), sedangkan untuk menangani data *non* kejadian banjir, digunakan korpus yang dibuat secara mandiri pada penelitian ini. Proses filter pada penelitian ini sebagai berikut:

a) Filtering lokasi dilakukan dengan mengecek terlebih dahulu, apakah lokasi pada *tweet* termasuk kedalam *dictionary*. Jika lokasi termasuk dalam *dictionary*, maka *tweet* tersebut tidak akan dihapus, dan sebaliknya.
b) Filter *non* kejadian banjir dilakukan dengan mengecek *tweet* tersebut, apakah mengandung kejadian banjir yang termasuk dalam *dictionary*. Jika kejadian banjir tersebut terdeteksi, maka baris *tweet* akan dihapus, dan sebaliknya.
c) Filter tingkat keparahan dilakukan dengan mengecek *tweet* terlebih dahulu dengan cara mengambil angka yang diikuti dengan satuan "milimeter", "centimeter", dan "meter". Jika terdapat *tweet* yang mengandung tingkat keparahan, maka *tweet* tidak akan dihapus, dan sebaliknya.
d) Filter angka selain angka tingkat keparahan dilakukan dengan menghapus angka yang tidak diikuti oleh "milimeter", "centimeter", dan "meter". Jika terdapat angka yang tidak diikuti dengan satuan, maka angka tersebut akan dihapus.

Proses filter terakhir dalam penelitian ini yaitu filter angka selain angka tingkat keparahan. Setelah proses filter tersebut, kemudian dilakukan konversi satuan dari "milimeter" ke "centimeter" dan "meter" ke "centimeter". Proses konversi dilakukan untuk menyesuaikan dengan satuan kategori tingkat keparahan banjir yang telah dijelaskan pada Bab 2. Apabila nilai satuannya adalah milimeter, maka angka dalam nilai satuan milimeter tersebut akan dibagi dengan 10. Kemudian apabila nilai satuannya adalah meter, maka angka dalam nilai satuan meter tersebut akan dikali dengan 100. Hasil dari konversi tersebut akan dijadikan satu dengan hasil satuan yang lainnya, kemudian digunakan untuk pelabelan data.

3.2.5 Pelabelan Data

Proses pelabelan data digunakan untuk memberikan label pada data teks agar dapat dikategorikan sesuai dengan karakteristik yang sudah ditentukan. Pada penelitian ini, pelabelan data digunakan untuk memberikan label tingkat

keparahan yang terdiri dari rendah, sedang, dan tinggi. Setelah dilakukan proses pelabelan data, selanjutnya yaitu proses validasi label. Validasi label dilakukan agar label yang didapatkan lebih akurat dan terpercaya. Dalam hal ini, proses pelabelan data dan validasi label dilakukan oleh Badan Penanggulangan Bencana Daerah (BPBD) Jawa Timur.

3.2.6 Data *Preprocessing*

Preprocessing yang dilakukan dalam penelitian ini digunakan untuk mengubah data mentah yang diperoleh dari proses pengumpulan data menjadi struktur data yang lebih efisien dan siap untuk di *training* serta di *testing* pada proses berikutnya. Proses *preprocessing* ini dilakukan dengan bantuan *library* dan modul seperti NLTK, Regex, dsb. Tahap *preprocessing* penelitian ini dimulai dari hapus Emoji, *URLs*, *Username*, *Hashtag*, tanda baca, dan spasi ganda. Kemudian dilanjutkan proses *case folding*, *tokenizing*, *slang words*, *stemming*, dan *stopwords*.

1) Menghapus Emoji, *URLs*, dan *Username*

Pada proses ini dilakukan pada *tweets* yang didalamnya terdapat Emoji, *URLs* serta *tweets* yang terdapat *username* (mention/@). Proses tersebut terbukti dapat meningkatkan akurasi dari pembuata model klasifikasi (Keerthi Kumar & Harish, 2018).

2) Menghapus Hashtag, tanda baca, dan spasi ganda

Data tweet merupakan data yang terdiri dari berbagai karakter seperti *hashtag*, tanda baca, dan spasi ganda. Hal tersebut dapat berdampak pada proses pengklasifikasian dengan model sehingga perlunya penanganan untuk mengatasi hal tersebut (Hickman et al., 2022). Penghapusan tanda baca dan spasi ganda dapat mempengaruhi hasil akurasi model klasifikasi (Yang & Zhang, 2018). Proses penghapusan ini menggunakan bantuan modul regex.

3) *Case Folding*

Case folding merupakan proses yang digunakan untuk merubah dan menyamaratakan penggunaan huruf kapital dalam sebuah kata. Apabila terdapat kata yang teridentifikasi tidak berhuruf kapital, maka akan diganti dengan padanan huruf kapital dan begitupun sebaliknya. Penelitian ini melakukan proses *case folding* dengan mengatur semua kata yang diawali dengan huruf kapital

diubah menjadi huruf kecil agar dapat memudahkan proses kata. Proses *case folding* pada penelitian ini dimulai dengan mengidentifikasi apakah kata tersebut berhuruf kapital atau tidak, jika berhuruf kapital maka dilakukan penggantian dengan huruf kecil. Contoh penerapan *case folding* seperti pada Tabel 3.2 berikut.

Tabel 3. 2 Contoh *Case Folding*

Input	***Output***
JAKARTA Terendam Banjir Besar, Puluhan Orang Meninggal Dunia.	jakarta terendam banjir besar, puluhan orang meninggal dunia.
Tinggal di Depok Terendam Banjir: Hujan deras membuat sejumlah rumah.	tinggal di depok terendam banjir: hujan deras membuat sejumlah rumah.
Banjir di Jayapura, Relawan PKS Sigap Salurkan Makanan.	banjir di jayapura, relawan pks sigap salurkan makanan.

4) *Tokenizing*

Proses *tokenizing* merupakan pemecahan dokumen, paragraf, dan kalimat menjadi kata yang disebut dengan token. Dengan proses tersebut dapat memudahkan pembedaan antara pemisah kata atau bukan seperti tanda baca. Tahap *tokenizing* pada penelitian ini dilakukan dengan batasan karakter spasi. Contoh penerapan *tokenizing* seperti pada Tabel 3.3 berikut.

Tabel 3. 3 Contoh *Tokenizing*

Input	***Output***
jakarta terendam banjir besar.	jakarta terendam banjir besar
di depok terendam banjir.	di depok terendam banjir

5) *Slang Words*

Slang words merupakan kumpulan bahasa gaul atau alay pada Bahasa Indonesia. Pengendalian *slang words* dibutuhkan dalam proses ekstraksi informasi agar kata mudah terdeteksi pada proses selanjutnya. Proses menghandle *slang words* ini, membutuhkan bantuan sumber daya eksternal baik *dictionary* maupun *corpus*. Penelitian ini menggunakan kombinasi dari *dictionary* (Owen, 2020), *dictionary* (kirralabs, 2020), *dictionary* (Ibrohim, 2019), dan *corpus Colloquial Indonesian Lexicon* (Salsabila, 2018). Pada proses ini dilakukan dengan

mengecek terlebih dahulu, apakah kata yang diidentifikasi sesuai dengan *corpus slang words*. Jika kata sesuai maka akan diubah ke dalam bentuk formal dan begitupun seterusnya hingga tidak terdeteksi kembali *slang words*. Contoh proses handle slang words seperti pada Tabel 3.4 berikut.

Tabel 3. 4 Contoh *Slang Words*

Input	***Output***
Kebanjiran **sblm** pabrik gula arah **sby**	Kebanjiran **sebelum** pabrik gula arah **Surabaya**
Alow akuh anak **sby**, tolong **ambilin** air ya	**Halo aku** anak **Surabaya**, tolong **ambilkan** air ya

6) *Stemming*

Proses *handle slang words* menghasilkan kata formal yang merupakan input dari proses *stemming*. *Stemming* merupakan pencarian kata dasar atau akar kata dengan membuang imbuhan *prefiks*, *suffix,* dan *infix*. Pada penelitian ini, proses *stemming* membutuhkan bantuan dari *library stemming sastrawi python*. Proses *stemming* dalam penelitian ini dimulai dengan mendeteksi kata, apakah kata tersebut termasuk dalam kata dasar atau tidak, jika tidak maka akan diubah dan disesuaikan dengan bantuan *library* dan begitupun seterusnya hingga tidak terdeteksi kembali. Contoh proses *stemming* seperti pada Tabel 3.5 berikut.

Tabel 3. 5 Contoh Proses *Stemming*

Input	***Output***
Semalam	Malam
Anginnya	Angin
Menimbulkan	Timbul
Kerugian	Rugi

7) Stopwords

Proses *stopwords* merupakan menghilangkan kata yang tidak penting dan tidak memiliki informasi yang terlalu penting seperti “yang”, “dan”, “di”, “dari”, dan “ke” dalam sebuah kata. Penelitian ini menggunakan *stopwords* Bahasa Indonesia dengan total 758 yang dikumpulkan oleh masdevid dalam akun githubnya (Devid, 2023). Proses *stopwords* dalam penelitian ini dimulai dengan mencocokkan kata (token) yang diidentifikasi dengan *stopwords list*, jika terdapat token yang sesuai maka akan dihapus dan begitupun seterusnya hingga tidak

terdeteksi *stopwords*. Contoh dari proses *stopwords* seperti pada Tabel 3.6 berikut.

Tabel 3. 6 Contoh *Stopwords*

Input	Output
Manajemen pengetahuan **adalah** sebuah konsep baru **di** dunia bisnis	Manajemen pengetahuan konsep baru dunia bisnis
Maju tak gentar mencela **dan** membela **yang** bayar	Maju tak gentar mencela membela bayar

3.2.7 Pembuatan Model

1. Pembagian dataset

Sebelum masuk dalam proses pembuatan model, dilakukan pembagian dataset terlebih dahulu. Pada penelitian ini, pembagian proporsi *dataset* menjadi data *training* dan data *testing*. Pada umumnya jumlah data *training* lebih banyak agar dapat melatih dan menghasilkan model yang baik. Pada penelitian ini pembagian dataset menggunakan skenario 80:20 yaitu, data *training* sebanyak 80% sedangkan data *testing* 20%.

2. *Feature Extraction*

Setelah dilakukan proses pembagian dataset, proses selanjutnya yaitu mengubah format teks menjadi vektor numerik dengan menerapkan TF-IDF sesuai dengan persamaan (8). Proses *feature extraction* dalam penelitian ini menggunakan *library* python TF-IDF *Vectorizer* dan kemudian dinormalisasikan dengan menggunakan TF-IDF *Transformer*.

3. Pembuatan Model SVM

Pada umumnya SVM *multiclass* dilakukan dengan pendekatan *one versus one* (OVO) dan *one versus all* (OVA). Pendekatan OVO pada dasarnya dilakukan pada *dataset* yang memiliki dua kelas, sedangkan OVA lebih utama dilakukan pada *dataset* yang memiliki lebih dari dua kelas. Pendekatan OVA disebut dengan *one versus rest* (OVR). Pada penelitian ini label kelas yang dimiliki terdiri dari 3 kelas, sehingga implementasi SVM *multiclass* dengan menggunakan pendekatan OVA. Pendekatan OVA dilakukan dengan membandingkan sebuah kelas bernilai positif dengan semua kelas selain dirinya dalam satu kesatuan yang bernilai negatif (Delimayanti et al., 2021).

Terdapat tiga parameter yang digunakan dalam penelitian ini, yaitu:

a) *Kernel trick*

Kernel trick digunakan untuk memisahkan data secara linear dengan memindahkan data ke ruang dimensional yang lebih tinggi. Dalam penelitian ini menggunakan parameter kernel RBF pada *package kernlab*. Kernel RBF dengan menggunakan Gaussian memberikan hasil akurasi terbaik dari kernel yang lain (Nguyen et al., 2021).

b) Parameter regularisasi

Parameter regularisasi sering disebut dengan parameter C pada Sklearn. Parameter ini pada umumnya digunakan dalam semua kernel SVM ketika proses pemisahan kelas yang digunakan untuk membantu dalam menghindari kesalahan klasifikasi.

c) Parameter gamma

Parameter gamma digunakan untuk menentukan seberapa besar pengaruh yang dimiliki oleh satu sampel *dataset* pelatihan. Nilai parameter gamma yang rendah menunjukkan titik yang jauh juga dipertimbangkan untuk menentukan *hyperplane* berada, sedangkan nilai parameter gamma yang tinggi menunjukkan hanya titik yang dekat dipertimbangkan untuk menentukan *hyperplane* berada.

4. Evaluasi dan Analisis

Apabila model telah terbentuk maka dibutuhkan validasi model dengan menggunakan *cross validation*. Pembagian proporsi validasi model menggunakan teknik *cross validation* sebanyak 5 *fold* yang dapat dilihat pada Gambar 3.4.

Gambar 3. 3 Skenario *K-fold Cross Validation*

Detail pembagian proporsi *dataset* pada penelitian ini dilakukan sebagai berikut:

a. Jumlah partisi (k) yang digunakan pada penelitian ini yaitu 5 *fold*.

b. Pembagian *dataset* pada penelitian ini dengan proporsi 80% data *training* dan 20% data *testing*. Kemudian dilakukan iterasi dengan ketentuan setiap partisi berisi 80% data *training* (K2 - K5) dan 20% data *testing* (K1).
c. Setiap subsampel akan mengalami sebagai data *training* dan data *testing* sebanyak *k* kali.

Setelah proses validasi dengan *cross validation*, selanjutnya dilakukan evaluasi efisiensi kinerja model *text classification* dengan mengukur performa *Precision, Recall, Accuracy, F-Measure, ROC AUC,* dan *Logloss* yang dapat menunjukkan model berjalan baik atau sebaliknya. Perhitungan nilai *Precision* dapat disesuaikan dan dihitung dengan persamaan (6), *Recall* dengan persamaan (7), *Accuracy* dengan persamaan (8), *F-Measure* dengan persamaan (9), dan *Logloss* dengan persamaan (10). Setelah proses evaluasi model, dilanjutkan dengan analisis hasil *text classification* terkait dengan tingkat keparahan banjir.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

Hasil penelitian ini akan dijabarkan pada bab ini yang terdiri dari hasil pengumpulan data, pembersihan data, data filter, pelabelan data, data *preprocessing*, dan pembuatan model. Pembuatan model diawali dengan pembagian dataset, *feature extraction*, pembuatan model, dan evaluasi serta analisis hasil.

4.1.1 Pengumpulan Data

Pada penelitian ini, Pengumpulan data dilakukan dengan teknik *scraping Twitter* dan menggunakan library *snscrape* python. *Keyword* yang digunakan adalah "banjir" dalam bahasa Indonesia, dengan rentang waktu mulai dari bulan Oktober 2022 hingga Februari 2023. Hasil pengumpulan data yang diperoleh dalam penelitian sebanyak 537.340 dan dapat dilihat pada Tabel 4.1.

Tabel 4. 1 Hasil Pengumpulan Data

User	***Date Create***	***Tweet***
doelpaten	2023-02-28 23:59:51+00:00	@Yo_NTKTS @Candra4smara @pengarang_sajak @Miduk17 @kurawa @Dennysiregar7 Kalo Anies gak bikin itu, tentu bisa mengulangi sejarah kelam banjir Jakarta seperti era Jokowi Ahok...
safwan_yusoff	2022-10-01 00:00:41+00:00	Boleh la guna utk banjir nnti
lampungpost_	2023-02-25 03:38:24+00:00	Pembuatan sumur resapan dan embungÂ secara terstruktur atau terorganisir pada daerah-daerah rawan banjir dinilai dapat mencegah terjadinya banjir. https://t.co/J9e3yOVDuN
KompasTV	2023-02-25 03:36:46+00:00	Banjir hingga setinggi 1 meter, merendam permukiman warga, di Kebon Pala, Kecamatan Jatinegara, Jakarta Timur. Sejumlah warga memilih bertahan di lantai dua rumah mereka. Sementara sebagian lainnya, tetap beraktivitas seperti biasa.
Buletincoid	2023-02-13 14:53:37+00:00	Menjadi langganan banjir, warga kecamatan Ijen Bondowoso tagih janji pemerintah

Sesuai dengan Tabel 4.1 bahwa data telah berhasil dikumpulkan dan disimpan dalam format csv yang kemudian digunakan sebagai input pada proses

preprocessing. Data yang didapatkan terdiri dari tiga kolom dengan nama *user*, *date create*, dan *tweet*.

4.1.2 Pembersihan Data

Proses pembersihan data dilakukan agar tidak terjadi *misinformation* dalam sebuah data, sehingga perlu penanganan khusus untuk hal tersebut. Proses pembersihan data dalam penelitian ini digunakan untuk menangani *missing value* dan data duplikat dengan cara menghapus data tersebut. Jumlah *missing value* dalam penelitian ini sebanyak 13 data pada kolom date create, dan 15 data pada kolom tweet. Hasil *missing value* yang didapatkan akan dihapus, sehingga data yang didapatkan sebanyak 537.325. Setelah proses menghapus *missing value*, selanjutnya yaitu menangani data duplikat. Data duplikat yang terdapat pada penelitian ini sebanyak 15.120. Data duplikat tersebut akan dihapus dan menghasilkan data bersih sebanyak 522.205.

4.1.3 Data Filter

Proses data filter dalam penelitian ini digunakan untuk menyaring dan menghapus data yang tidak sesuai dengan kategori yang relevan, agar dapat menghasilkan model yang terbaik dan akurat. Terdapat empat tahap dalam proses *filtering* pada penelitian ini, yaitu:

1) Filter Lokasi Jawa Timur

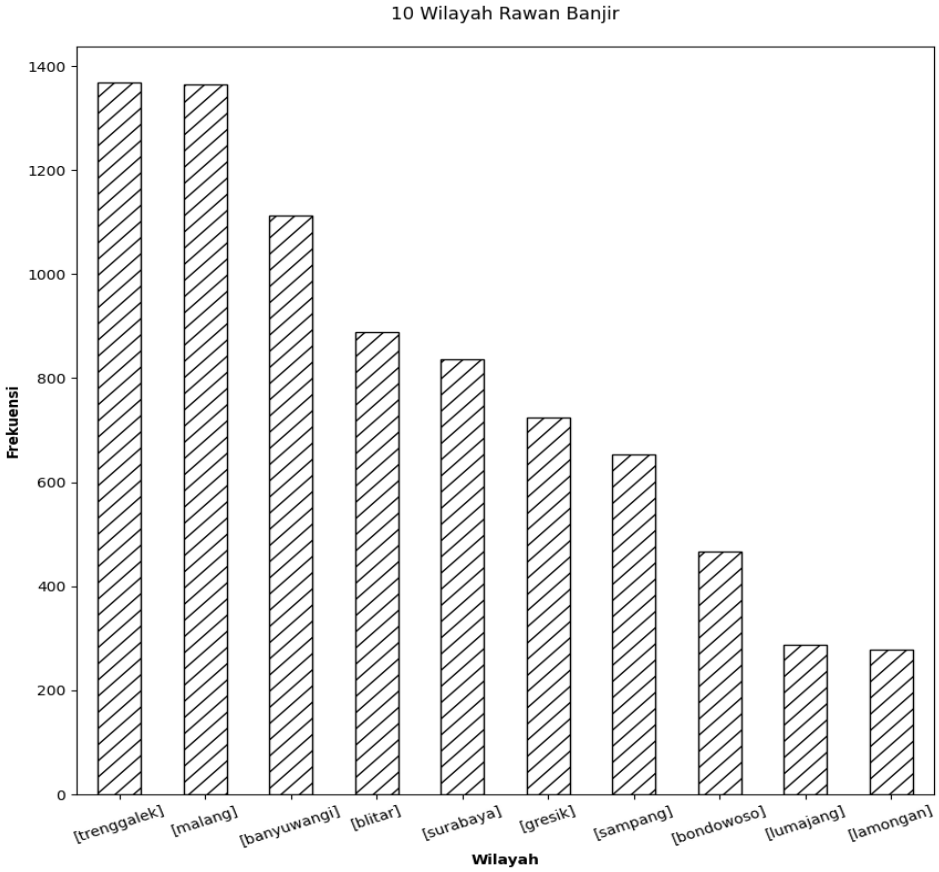
Proses filtrasi lokasi dalam penelitian ini melibatkan penggunaan corpus yang disesuaikan dengan Peraturan Kementerian Dalam Negeri (Permendagri) tahun 2020. Corpus ini mencakup 31 daerah di Jawa Timur, yang terdiri dari 29 kabupaten dan 2 kota. Deteksi lokasi dalam penelitian ini berdasarkan kata-kata yang terdapat dalam *tweet* yang menunjukkan daerah di Jawa Timur. Deteksi dilakukan dengan tanpa menggunakan kata kabupaten dan kota kecuali kota batu. Jika sebuah baris *tweet* mengandung kata yang terdapat dalam corpus, maka baris tersebut akan tetap dan sebaliknya. Hasil dari proses ini menghasilkan data *tweet* yang berasal dari lokasi di Jawa Timur dengan jumlah total 11.095 data.

Tabel 4. 2 Hasil Filter Lokasi

Tweet	**Hasil Filter Lokasi**
Banjir 130 centimeter di Bondowoso Meluas Hingga 20 RT. https://t.co/IX2dx7jIDe	Bondowoso

rek kalau dari surabaya mau ke cerme gresik lewat mana ya yang banjir?	Surabaya Gresik
Luapan sungai berulang-ulang membanjiri Gresik dan Lamongan di Jawa Timur terutama di musim hujan. Banjir menjadi rutinitas di musim hujan selama prasarana pengairan tidak mampu menampung dan mengendalikan curahan hujan.	Gresik Lamongan
Warga Situbondo harap siap-siap kiriman air dari Bondowoso..	Situbondo Bondowoso
Banjir 30 centimeter Rendam Sekolah hingga Permukiman di Trowulan Mojokerto https://t.co/IBFtxY2E6r	Mojokerto

Dalam proses deteksi lokasi, seperti yang tercantum pada Tabel 4.2, terdapat beberapa baris yang menghasilkan lebih dari dua lokasi yang berbeda. Untuk menangani hal tersebut, jika baris *tweet* tersebut mengandung lebih dari dua lokasi yang berbeda, maka baris *tweet* tersebut akan dihapus untuk menghindari ambiguitas dalam data. Setelah proses tersebut, data akhir yang dihasilkan adalah sebanyak 10.673 data.



Gambar 4. 1 Hasil Filter Lokasi

Berdasarkan Gambar 4.1 terkait dengan hasil filter lokasi menunjukkan bahwa, Wilayah Trenggalek merupakan wilayah yang paling sering mengalami banjir. Namun, perlu diperhatikan bahwa dalam konteks ini, masih terdapat kejadian yang bukan banjir seperti "banjir hadiah", "banjir tangis", "banjir rejeki", dsb. Oleh karena itu, diperlukan filter lanjutan untuk mengatasi hal tersebut.

2) Filter *Non* Kejadian Banjir

Pada data *tweet*, seringkali terdapat penggunaan kata "banjir" yang tidak berkaitan dengan kejadian banjir sebenarnya. Oleh karena itu, perlu dilakukan penghapusan kata-kata yang tidak relevan dengan kejadian banjir sebenarnya, misalnya "banjir hadiah", "banjir tangis", "banjir *comment*", dan sebagainya.

Tabel 4. 3 *Dictionary* Kejadian *Non* Banjir

Corpus		
banjir air mata	banjir gol	banjir pesanan
banjir apresiasi	banjir hadiah	banjir pemesanan
banjir bacot	banjir haru	banjir peserta
banjir bahagia	banjir hujat	banjir prestasi
banjir buzzer	banjir hujatan	banjir promo
banjir buzzerp	banjir kabar	banjir puji
banjir cebi	banjir keringat	banjir pujian
banjir diskon	banjir kritik	banjir ritek
banjir dukung	banjir lahar dingin	banjir selamat
banjir dukungan	banjir maaf	banjir sholawat
banjir duren	banjir ngecor	banjir simpati
banjir durian	banjir omong kosong	banjir sindiran
banjir fatamorgana	banjir orderan	banjir tangis
banjir tertawa	tahu banjir	

Deteksi kejadian non banjir dilakukan dengan menggunakan bantuan *dictionary* yang telah dibuat sebelumnya. *Dictionary* tersebut terdiri dari 40 kejadian *non* banjir seperti pada Tabel 4.3. Jika suatu baris *tweet* mengandung kata-kata yang terdapat dalam *dictionary* tersebut, maka baris tersebut akan dihapus. Setelah proses ini, data yang dihasilkan sebanyak 10.502 data. Hasil dari proses filter *non* kejadian banjir dapat dilihat pada Tabel 4.4 berikut.

Tabel 4. 4 Hasil Filter *Non* Kejadian Banjir

Tweet	**Hasil Filter *Non* Kejadian Banjir**
Berkat Dipromosikan di Lapak Ganjar, UMKM Kota Malang Ini Banjir Pesanan! https://t.co/OHHt3wcfaO	Banjir Pesanan

Goyang Bar Bar, 4 Video Ukhti Pasuruan Goyang Dipundak Orang di TikTok Banjir Hujatan https://t.co/WiTFE7zbm5	Banjir Hujatan
Detik-detik Banjir Lahar Dingin di Lumajang Datang Tiba-tiba	Banjir Lahar Dingin
Pasrepan kabupaten Pasuruan Jawa timur Indonesia banjir durian https://t.co/xFrazA2UrY	Banjir Durian
MASHA ALLAH, MADIUN BANJIR AIR MATA LIAT ELU @.JONGHO https://t.co/JkXIjq4MK7	Banjir Air Mata

3) Filter Tingkat Keparahan

Proses filter terakhir melibatkan penyaringan data berdasarkan tingkat keparahan banjir sesuai dengan kriteria yang telah dijelaskan pada bab 2. Proses ini dilakukan dengan mengambil angka yang diikuti oleh kata-kata "milimeter", "centimeter", dan "meter". Jika angka tidak terdeteksi dan diikuti oleh ketiga kata tersebut, maka baris tersebut akan dianggap sebagai nilai kosong seperti pada Tabel 4.5. Kemudian, nilai kosong tersebut akan dihapus dari data. Setelah proses ini, data yang dihasilkan sebanyak 3.820.

Tabel 4. 5 Hasil Filter Tingkat Keparahan

Tweet	**Hasil Filter Tingkat Keparahan Banjir**
Dapet info dari temen2 kantor lama, macet banget jalanan gresik gegara banjir 40 centimeter. Keluar kantor jam 5an baru sampe rumah jam 9. Udah gila ðŸ¤¯	40 centimeter
Terdampak Banjir, Warga di Lamongan dapat Bantuan Pengobatan dan Bengkel Gratis. https://t.co/3biGiwI5eL	*Nan*
tercatat 10 desa di 4 kecamatan gresik dilanda banjir dengan ketinggian hingga 1 meter.	1 meter
Sehari sebelumnya masyarakat terdampak banjir di Lamongan juga mendapatkan bantuan dari Gubernur Jawa Timur, Khofifah Indar Parawansa. https://t.co/k37UBlXHl6	*Nan*
@syahidatunaimah LAmongan langganan banjir sampe 130 centimeter yoo wkwkwk	130 centimeter
Menurut Eri, menyelesaikan persoalan banjir di Surabaya bukan hanya tugas Pemkot, namun juga semua warga. https://t.co/zTrhcBbAaz	*Nan*

4) Filter angka selain angka tingkat keparahan

Pada Tabel 4.5, terdapat beberapa angka yang bukan termasuk angka tingkat keparahan. Oleh karena itu diperlukan proses filter angka selain angka tingkat keparahan. Proses filter ini dilakukan dengan cara menghapus angka yang

tidak diikuti oleh satuan “milimeter”, “centimeter”, dan “meter”. Hasil proses filter angka selain angka tingkat keparahan dapat dilihat pada Tabel 4.6 berikut.

Tabel 4. 6 Hasil Filter Angka Selain Angka Tingkat Keparahan

Tweet	**Hasil Filter**
Dapet info dari temen2 kantor lama, macet banget jalanan gresik gegara banjir 40 centimeter. Keluar kantor jam 5an baru sampe rumah jam 9. Udah gila ðŸ¤¯	Dapet info dari temen kantor lama, macet banget jalanan gresik gegara banjir 40 centimeter. Keluar kantor jam 5an baru sampe rumah jam. Udah gila ðŸ¤¯
tercatat 10 desa di 4 kecamatan gresik dilanda banjir dengan ketinggian hingga 1 meter.	tercatat desa di kecamatan gresik dilanda banjir dengan ketinggian hingga 1 meter.
Banjir di Lamongan Rendam hingga 135 centimeter pada 6 Ribu Rumah dan 22 Tempat Ibadah https://t.co/KTqcgmwp2v	Banjir di Lamongan Rendam hingga 135 centimeter pada Ribu Rumah dan Tempat Ibadah https://t.co/KTqcgmwp2v
BPBD Bojonegoro Catat 216 Rumah dan 3 Fasilitas Umum Terendam Banjir hingga 135 centimeter https://t.co/5vYE0gLmLy	BPBD Bojonegoro Catat Rumah dan Fasilitas Umum Terendam Banjir hingga 135 centimeter https://t.co/5vYE0gLmLy
Hujan lebat 4 jam, Jember dihantam banjir 40 centimeter dan tanah longsor #beritaindo #beritaterkini #indonews https://t.co/jMkCnwRZr0	Hujan lebat jam, Jember dihantam banjir 40 centimeter dan tanah longsor #beritaindo #beritaterkini #indonews https://t.co/jMkCnwRZr0

Setelah dilakukan proses filter angka selain angka tingkat keparahan banjir, proses selanjutnya adalah konversi satuan. Proses konversi dilakukan untuk menyesuaikan dengan satuan kategori tingkat keparahan banjir yang telah dijelaskan pada Bab 2. Apabila nilai satuannya adalah “milimeter”, maka angka dalam nilai satuan “milimeter” tersebut akan dibagi dengan 10. Kemudian apabila nilai satuannya adalah “meter”, maka angka dalam nilai satuan “meter” tersebut akan dikali dengan 100. Hasil dari konversi tersebut akan dijadikan satu dengan hasil satuan yang lainnya, kemudian digunakan untuk pelabelan data. Hasil konversi satuan pada penelitian dapat dilihat pada Tabel 4.7 berikut.

Tabel 4. 7 Hasil Konversi Satuan

Tweet	**Satuan**	**Hasil Konversi Satuan**
Pemprov Jatim membangun tiga jembatan di Trenggalek karena konstruksi jembatan sebelumnya hancur akibat tersapu banjir bandang dengan kedalaman 2500 milimeter https://t.co/MtQTDdn	2500 milimeter	250 centimeter

banjir kabupaten nganjuk 1090 milimeter	1090 milimeter	109 centimeter
Banjir Bandang 4 meter, Jalur Utama Jember Macet Jam #banjir https://t.co/etUrhCge	4 meter	400 centimeter
Warga Banyuwangi Dipaksa Mengungsi, Banjir dengan ketinggian mencapai 1 meter Rendam Puluhan Rumah https://t.co/XLfGkPss	1 meter	100 centimeter
trenggalek landa banjir 3 meter	3 meter	300 centimeter

4.1.4 Pelabelan Data

Proses pelabelan tingkat keparahan banjir dalam penelitian ini melibatkan bantuan dari ahli banjir, yaitu Badan Penanggulangan Bencana Daerah (BPBD) Jawa Timur. Ahli banjir tersebut akan memberikan label berdasarkan tingkat keparahan banjir dalam tiga kategori, yaitu rendah, sedang, dan tinggi. Dikatakan sebagai kategori rendah apabila ketinggian banjir kurang dari 100 centimeter, sedangkan ketinggian kategori sedang berada diantara 100 hingga kurang dari 300 centimeter, dan ketinggian kategori tinggi lebih dari 300 centimeter. Hasil dari proses pelabelan data ini dapat dilihat pada Tabel 4.8.

Tabel 4. 8 Hasil Pelabelan Data

Tweet	Label
hujan lebat moga banjir 120 centimeter surut cepat gresik	Sedang
banjir 50 centimeter rendam ratus rumah gedung sekolah dasar negeri bondowoso	Rendah
bmkg hadap masyarakat jember waspada cuaca ekstrem cuaca picu hujan lebat potensi banjir 340 centimeter wilayah jawa timur masuk jember	Tinggi
banjir 45 centimeter landa blitar mari bersamasama ikan dukung warga dampak	Rendah
kali motor overthinking gemeteran nerjang banjir 160 centimeter macet blitar doa motor mogok tau mogok	Sedang

4.1.5 Data *Preprocessing*

Proses *preprocessing* dilakukan untuk mengubah data menjadi normal agar dapat digunakan pada proses selanjutnya. Data p*reprocessing* pada penelitian ini terdiri dari beberapa proses yaitu, hapus Emoji, *URLs*, *Username*, *Hashtag*, tanda baca, dan spasi ganda. Kemudian dilanjutkan proses *case folding*, *tokenizing*, *slang word*, *stemming*, dan *stopwords*.

1) Menghapus Emoji, *URLs*, dan *Username*

Pada proses ini dilakukan identifikasi emoji pada data *tweet*. Kemudian emoji yang teridentifikasi akan dihapus dengan menggunakan karakter non-ASCII dari suatu string yang berupa emoji. Dalam hal ini, data tweet akan diubah menjadi bentuk byte dengan metode encoding ASCII. Karakter non-ASCII yang terdeteksi biasanya akan berbentuk tanda tanya "?" yang kemudian akan dihapus. hasil proses ini dapat dilihat pada Tabel 4.9 berikut.

Tabel 4. 9 Hasil Menghapus Emoji, *URLs*, dan *Username*

Tweet	**Hasil Menghapus Emoji, *URLs*, dan *Username***
banjir 130 centimeter di gresik kurleb sudah hari. saya membawa balita. belum ada tanda air surut. 😞😫😭😭😭😭😪😥 #menjelangpanenraya #kebanjiran #petani @PetaniRasional https://t.co/gayKHuHqM	banjir 130 centimeter di gresik kurleb sudah hari. saya membawa balita. belum ada tanda air surut.#menjelangpanenraya #kebanjiran #petani
Kontrakan tahun pertama gw di Malang, sering banjir sampe 120 centimeter karna posisinya di bawah jalanan turunan gtu pun sebelahnya selokan, untungnya kamar gw lante 😬😅 https://t.co/gFmAobYzi	Kontrakan tahun pertama gw di Malang, sering banjir sampe 120 centimeter karna posisinya di bawah jalanan turunan gtu pun sebelahnya selokan, untungnya kamar gw lante
@JackMusi @VIVAcoid Dalam keadaan sulit akibat banjir 60 centimeter, mari bersatu dan bekerja sama untuk membangun kembali rumah-rumah di Bondowoso. 😭😪😭😞	Dalam keadaan sulit akibat banjir 60 centimeter, mari bersatu dan bekerja sama untuk membangun kembali rumah-rumah di Bondowoso.
Innalillahi, banjir 90 centimeter di Banyuwangi, Jawa Timur, menelan Korban Jiwa. Seorang remaja berkebutuhan khusus tewas, Setelah terjatuh dari atas jembatan 😭 , saat tengah bermain sepeda. #NewsSCTV #LiputanSiang Berita lainnya: https://t.co/qblWVzxl https://t.co/CSOrtlgdO	Innalillahi, banjir 90 centimeter di Banyuwangi, Jawa Timur, menelan Korban Jiwa. Seorang remaja berkebutuhan khusus tewas, Setelah terjatuh dari atas jembatan, saat tengah bermain sepeda. #NewsSCTV #LiputanSiang Berita lainnya:
Bhabinkamtibmas Polsek Bojonegoro menutup akses jalan saat terjadi banjir yang mencapai 45 centimeter agar tidak di lalui oleh kendaraan Roda maupun Roda . @divhumaspolri @bidhumaspoldajatim #kapolresbojonegoro #polresbojonegoro #bolokuompolisi #polripedulibencanabanjir https://t.co/ACRxm	Bhabinkamtibmas Polsek Bojonegoro menutup akses jalan saat terjadi banjir yang mencapai 45 centimeter agar tidak di lalui oleh kendaraan Roda maupun Roda. #kapolresbojonegoro #polresbojonegoro #bolokuompolisi #polripedulibencanabanjir

2) Menghapus *Hashtag*, tanda baca, dan spasi ganda

Proses penghapusan *hashtag*, tanda baca, dan spasi ganda pada penelitian ini menggunakan bantuan modul regex. Hasil dari proses ini dapat dilihat pada Tabel 4.10 berikut.

Tabel 4. 10 Hasil Menghapus *Hashtag*, tanda baca, dan spasi ganda

Hasil Menghapus Emoji, *URLs*, dan *Username*	**Hasil Menghapus *Hashtag*, tanda baca, dan spasi ganda**
banjir 130 centimeter di gresik kurleb sudah hari. saya membawa balita. belum ada tanda air surut.#menjelangpanenraya #kebanjiran #petani	banjir 130 centimeter di gresik kurleb sudah hari saya membawa balita belum ada tanda air surut menjelangpanenraya kebanjiran petani
Kontrakan tahun pertama gw di Malang, sering banjir sampe 120 centimeter karna posisinya di bawah jalanan turunan gtu pun sebelahnya selokan, untungnya kamar gw lante	Kontrakan tahun pertama gw di Malang sering banjir sampe 120 centimeter karna posisinya di bawah jalanan turunan gtu pun sebelahnya selokan untungnya kamar gw lante
Dalam keadaan sulit akibat banjir 60 centimeter, mari bersatu dan bekerja sama untuk membangun kembali rumah-rumah di Bondowoso.	Dalam keadaan sulit akibat banjir 60 centimeter mari bersatu dan bekerja sama untuk membangun kembali rumah rumah di Bondowoso
Innalillahi, banjir 90 centimeter di Banyuwangi, Jawa Timur, menelan Korban Jiwa. Seorang remaja berkebutuhan khusus tewas, Setelah terjatuh dari atas jembatan, saat tengah bermain sepeda. #NewsSCTV #LiputanSiang Berita lainnya:	Innalillahi banjir 90 centimeter di Banyuwangi Jawa Timur menelan Korban Jiwa Seorang remaja berkebutuhan khusus tewas Setelah terjatuh dari atas jembatan saat tengah bermain sepeda NewsSCTV LiputanSiang Berita lainnya
Bhabinkamtibmas Polsek Bojonegoro menutup akses jalan saat terjadi banjir yang mencapai 45 centimeter agar tidak di lalui oleh kendaraan Roda maupun Roda. #kapolresbojonegoro #polresbojonegoro #bolokuompolisi #polripedulibencanabanjir	Bhabinkamtibmas Polsek Bojonegoro menutup akses jalan saat terjadi banjir yang mencapai 45 centimeter agar tidak di lalui oleh kendaraan Roda maupun Roda kapolresbojonegoro polresbojonegoro bolokuompolisi polripedulibencanabanjir

3) *Case Folding*

Setelah proses data *cleaning*, proses selanjutnya yaitu membuat data menjadi satu bentuk berupa huruf kecil. Proses tersebut disebut dengan *case folding* yang dapat merubah huruf besar *uppercase* menjadi huruf kecil *lowercase*. Hasil pada proses ini dapat dilihat pada Tabel 4.11.

Tabel 4. 11 Hasil *Case Folding*

Hasil Menghapus *Hashtag*, tanda baca, dan spasi ganda	**Hasil *Case Folding***
banjir 130 centimeter di gresik kurleb sudah hari saya membawa balita belum ada tanda air surut menjelangpanenraya kebanjiran petani	banjir 130 centimeter di gresik kurleb sudah hari saya membawa balita belum ada tanda air surut menjelangpanenraya kebanjiran petani
Kontrakan tahun pertama gw di Malang sering banjir sampe 120 centimeter karna posisinya di bawah jalanan turunan gtu pun sebelahnya selokan untungnya kamar gw lante	kontrakan tahun pertama gw di malang sering banjir sampe 120 centimeter karna posisinya di bawah jalanan turunan gtu pun sebelahnya selokan untungnya kamar gw lante
Dalam keadaan sulit akibat banjir 60 centimeter mari bersatu dan bekerja sama untuk membangun kembali rumah rumah di Bondowoso	dalam keadaan sulit akibat banjir 60 centimeter mari bersatu dan bekerja sama untuk membangun kembali rumah rumah di bondowoso
Innalillahi banjir 90 centimeter di Banyuwangi Jawa Timur menelan Korban Jiwa Seorang remaja berkebutuhan khusus tewas Setelah terjatuh dari atas jembatan saat tengah bermain sepeda NewsSCTV LiputanSiang Berita lainnya	innalillahi banjir 90 centimeter di banyuwangi jawa timur menelan korban jiwa seorang remaja berkebutuhan khusus tewas setelah terjatuh dari atas jembatan saat tengah bermain sepeda newssctv liputansiang berita lainnya
Bhabinkamtibmas Polsek Bojonegoro menutup akses jalan saat terjadi banjir yang mencapai 45 centimeter agar tidak di lalui oleh kendaraan Roda maupun Roda kapolresbojonegoro polresbojonegoro bolokuompolisi polripedulibencanabanjir	bhabinkamtibmas polsek bojonegoro menutup akses jalan saat terjadi banjir yang mencapai 45 centimeter agar tidak di lalui oleh kendaraan roda maupun roda kapolresbojonegoro polresbojonegoro bolokuompolisi polripedulibencanabanjir

4) Hasil *Tokenizing*

Pada tahap ini dilakukan pemecahan teks *tweet* berdasarkan spasi menjadi kata (token) agar dapat diolah pada proses selanjutnya. Hasil pada proses *tokenizing* pada penelitian ini dapat dilihat pada Tabel 4.12.

Tabel 4. 12 Hasil *Tokenizing*

Hasil *Case Folding*	**Hasil *Tokenizing***
banjir 130 centimeter di gresik kurleb sudah hari saya membawa balita belum ada tanda air surut menjelangpanenraya kebanjiran petani	['banjir', '130', 'centimeter', 'di', 'gresik', 'kurleb', 'sudah', 'hari', 'saya', 'membawa', 'balita', 'belum', 'ada', 'tanda', 'air', 'surut', 'menjelangpanenraya', 'kebanjiran', 'petani']
kontrakan tahun pertama gw di malang sering banjir sampe 120 centimeter karna posisinya di bawah jalanan turunan gtu pun sebelahnya selokan untungnya kamar gw lante	['kontrakan', 'tahun', 'pertama', 'gw', 'di', 'malang', 'sering', 'banjir', 'sampe', '120', 'centimeter', 'karna', 'posisinya', 'di', 'bawah', 'jalanan', 'turunan', 'gtu', 'pun', 'sebelahnya', 'selokan', 'untungnya', 'kamar', 'gw', 'lante']
dalam keadaan sulit akibat banjir 60	['dalam', 'keadaan', 'sulit', 'akibat', 'banjir',

centimeter mari bersatu dan bekerja sama untuk membangun kembali rumah rumah di bondowoso	'60', 'centimeter', 'mari', 'bersatu', 'dan', 'bekerja', 'sama', 'untuk', 'membangun', 'kembali', 'rumahrumah', 'di', 'bondowoso']
innalillahi banjir 90 centimeter di banyuwangi jawa timur menelan korban jiwa seorang remaja berkebutuhan khusus tewas setelah terjatuh dari atas jembatan saat tengah bermain sepeda newssctv liputansiang berita lainnya	['innalillahi', 'banjir', '90', 'centimeter', 'di', 'banyuwangi', 'jawa', 'timur', 'menelan', 'korban', 'jiwa', 'seorang', 'remaja', 'berkebutuhan', 'khusus', 'tewas', 'setelah', 'terjatuh', 'dari', 'atas', 'jembatan', 'saat', 'tengah', 'bermain', 'sepeda', 'newssctv', 'liputansiang', 'berita', 'lainnya', ':']
bhabinkamtibmas polsek bojonegoro menutup akses jalan saat terjadi banjir yang mencapai 45 centimeter agar tidak di lalui oleh kendaraan roda maupun roda kapolresbojonegoro polresbojonegoro bolokuompolisi polripedulibencanabanjir	['bhabinkamtibmas', 'polsek', 'bojonegoro', 'menutup', 'akses', 'jalan', 'saat', 'terjadi', 'banjir', 'yang', 'mencapai', '45', 'centimeter', 'agar', 'tidak', 'di', 'lalui', 'oleh', 'kendaraan', 'roda', 'maupun', 'roda', 'kapolresbojonegoro', 'polresbojonegoro', 'bolokuompolisi', 'polripedulibencanabanjir']

5) Hasil *Slang Words*

Proses *slang word* dilakukan setelah proses *tokenizing*. *Slang word* digunakan untuk menghapus kata yang tidak sesuai dengan kaidah bahasa Indonesia seperti “kalo” menjadi “kalau”, “gue” menjadi “saya”, “jkt” menjadi “jakarta”. Dalam penelitian ini menggunakan kumpulan kombinasi dari tiga *dictionary* dan satu *corpus.* Hasil dari handel *slang word* pada penelitian ini sesuai dengan Tabel 4.13.

Tabel 4. 13 Hasil *Slang Words*

Hasil *Tokenizing*	**Hasil *Slang Words***
['banjir', '130', 'centimeter', 'di', 'gresik', 'kurleb', 'sudah', 'hari', 'saya', 'membawa', 'balita', 'belum', 'ada', 'tanda', 'air', 'surut', 'menjelangpanenraya', 'kebanjiran', 'petani']	['banjir', '130', 'centimeter', 'di', 'gresik', 'kurang lebih', 'sudah', 'hari', 'saya', 'membawa', 'balita', 'belum', 'ada', 'tanda', 'air', 'surut', 'menjelangpanenraya', 'kebanjiran', 'petani']
['kontrakan', 'tahun', 'pertama', 'gw', 'di', 'malang', 'sering', 'banjir', 'sampe', '120', 'centimeter', 'karna', 'posisinya', 'di', 'bawah', 'jalanan', 'turunan', 'gtu', 'pun', 'sebelahnya', 'selokan', 'untungnya', 'kamar', 'gw', 'lante']	['kontrakan', 'tahun', 'pertama', 'saya', 'di', 'malang', 'sering', 'banjir', 'sampai', '120', 'centimeter', 'karena', 'posisinya', 'di', 'bawah', 'jalanan', 'turunan', 'gitu', 'pun', 'sebelahnya', 'selokan', 'untungnya', 'kamar', 'saya', 'lantai']
['dalam', 'keadaan', 'sulit', 'akibat', 'banjir', '60', 'centimeter', 'mari', 'bersatu', 'dan', 'bekerja', 'sama', 'untuk', 'membangun', 'kembali', 'rumahrumah', 'di', 'bondowoso']	['dalam', 'keadaan', 'sulit', 'akibat', 'banjir', '60', 'centimeter', 'mari', 'bersatu', 'dan', 'bekerja', 'sama', 'untuk', 'membangun', 'kembali', 'rumahrumah', 'di', 'bondowoso']
['innalillahi', 'banjir', '90', 'centimeter', 'di', 'banyuwangi', 'jawa', 'timur', 'menelan', 'korban', 'jiwa', 'seorang', 'remaja', 'berkebutuhan', 'khusus', 'tewas', 'setelah',	['innalillahi', 'banjir', '90', 'centimeter', 'di', 'banyuwangi', 'jawa', 'timur', 'menelan', 'korban', 'jiwa', 'seorang', 'remaja', 'berkebutuhan', 'khusus', 'tewas', 'setelah',

'terjatuh', 'dari', 'atas', 'jembatan', 'saat', 'tengah', 'bermain', 'sepeda', 'newssctv', 'liputansiang', 'berita', 'lainnya', ':']	'terjatuh', 'dari', 'atas', 'jembatan', 'saat', 'tengah', 'bermain', 'sepeda', 'newssctv', 'liputansiang', 'berita', 'lainnya', ':']
['bhabinkamtibmas', 'polsek', 'bojonegoro', 'menutup', 'akses', 'jalan', 'saat', 'terjadi', 'banjir', 'yang', 'mencapai', '45', 'centimeter', 'agar', 'tidak', 'di', 'lalui', 'oleh', 'kendaraan', 'roda', 'maupun', 'roda', 'kapolresbojonegoro', 'polresbojonegoro', 'bolokuompolisi', 'polripedulibencanabanjir']	['bhayangkara pembina keamanan dan ketertiban masyarakat', 'kepolisian sektor', 'bojonegoro', 'menutup', 'akses', 'jalan', 'saat', 'terjadi', 'banjir', 'yang', 'mencapai', '45', 'centimeter', 'agar', 'tidak', 'di', 'lalui', 'oleh', 'kendaraan', 'roda', 'maupun', 'roda', 'kapolresbojonegoro', 'polresbojonegoro', 'bolokuompolisi', 'polripedulibencanabanjir']

6) Hasil *Stemming*

Tahap *stemming* ini digunakan untuk merubah kata menjadi bentuk dasarnya seperti “menurut” menjadi “turut”, “meminta” menjadi “minta”, “permasalahan” menjadi “masalah”, dsb. Dalam penelitian ini proses stemming menggunakan *library stemming* sastrawi python. Hasil dari proses *stemming* dapat dilihat pada Tabel 4.14.

Tabel 4. 14 Hasil *Stemming*

Hasil *Slang Words*	**Hasil *Stemming***
['banjir', '130', 'centimeter', 'di', 'gresik', 'kurang lebih', 'sudah', 'hari', 'saya', 'membawa', 'balita', 'belum', 'ada', 'tanda', 'air', 'surut', 'menjelangpanenraya', 'kebanjiran', 'petani']	['banjir', '130', 'centimeter', 'di', 'gresik', 'kurang', 'lebih', 'sudah', 'hari', 'saya', 'bawa', 'balita', 'belum', 'ada', 'tanda', 'air', 'surut', 'menjelangpanenraya', 'banjir', 'tani']
['kontrakan', 'tahun', 'pertama', 'saya', 'di', 'malang', 'sering', 'banjir', 'sampai', '120', 'centimeter', 'karena', 'posisinya', 'di', 'bawah', 'jalanan', 'turunan', 'gitu', 'pun', 'sebelahnya', 'selokan', 'untungnya', 'kamar', 'saya', 'lantai']	['kontra', 'tahun', 'pertama', 'saya', 'di', 'malang', 'sering', 'banjir', 'sampai', '120', 'centimeter', 'karena', 'posisi', 'di', 'bawah', 'jalan', 'turun', 'gitu', 'pun', 'belah', 'selokan', 'untung', 'kamar', 'saya', 'lantai']
['dalam', 'keadaan', 'sulit', 'akibat', 'banjir', '60', 'centimeter', 'mari', 'bersatu', 'dan', 'bekerja', 'sama', 'untuk', 'membangun', 'kembali', 'rumahrumah', 'di', 'bondowoso']	['dalam', 'ada', 'sulit', 'akibat', 'banjir', '60', 'centimeter', 'mari', 'satu', 'dan', 'kerja', 'sama', 'untuk', 'bangun', 'kembali', 'rumahrumah', 'di', 'bondowoso']
['innalillahi', 'banjir', '90', 'centimeter', 'di', 'banyuwangi', 'jawa', 'timur', 'menelan', 'korban', 'jiwa', 'seorang', 'remaja', 'berkebutuhan', 'khusus', 'tewas', 'setelah', 'terjatuh', 'dari', 'atas', 'jembatan', 'saat', 'tengah', 'bermain', 'sepeda', 'newssctv', 'liputansiang', 'berita', 'lainnya', ':']	['innalillahi', 'banjir', '90', 'centimeter', 'di', 'banyuwangi', 'jawa', 'timur', 'telan', 'korban', 'jiwa', 'orang', 'remaja', 'butuh', 'khusus', 'tewas', 'telah', 'jatuh', 'dari', 'atas', 'jembatan', 'saat', 'tengah', 'main', 'sepeda', 'newssctv', 'liputansiang', 'berita', 'lain']
['bhayangkara pembina keamanan dan ketertiban masyarakat', 'kepolisian sektor',	['bhayangkara', 'bina', 'aman', 'dan', 'tertib', 'masyarakat', 'polisi', 'sektor',

'bojonegoro', 'menutup', 'akses', 'jalan', 'saat', 'terjadi', 'banjir', 'yang', 'mencapai', '45', 'centimeter', 'agar', 'tidak', 'di', 'lalui', 'oleh', 'kendaraan', 'roda', 'maupun', 'roda', 'kapolresbojonegoro', 'polresbojonegoro', 'bolokuompolisi', 'polripedulibencanabanjir']	'bojonegoro', 'tutup', 'akses', 'jalan', 'saat', 'jadi', 'banjir', 'yang', 'capai', '45', 'centimeter', 'agar', 'tidak', 'di', 'lalu', 'oleh', 'kendara', 'roda', 'maupun', 'roda', 'kapolresbojonegoro', 'polresbojonegoro', 'bolokuompolisi', 'polripedulibencanabanjir']

7) Hasil *Stopwords*

Setelah kata diubah kedalam bentuk dasarnya, proses selanjutnya yaitu menghapus kata yang tidak memiliki makna penting (*stopword*) seperti “lah”, “untuk”, “yang”, dsb. Pada penelitian ini menggunakan bantuan *dictionary stopwords* untuk mendeteksi kata *stopwords* pada *tweet*, apabila kata *tweet* mengandung kata *stopwords*, maka kata tersebut akan dihapus dari *tweet*. Hasil dari proses penghapusan *stopwords* pada penelitian ini dapat dilihat pada Tabel 4.15.

Tabel 4. 15 Hasil *Stopwords*

Hasil *Stemming*	**Hasil *Stopwords***
['banjir', '130', 'centimeter', 'di', 'gresik', 'kurang', 'lebih', 'sudah', 'hari', 'saya', 'bawa', 'balita', 'belum', 'ada', 'tanda', 'air', 'surut', 'menjelangpanenraya', 'banjir', 'tani']	['banjir', '130', 'centimeter', 'gresik', 'bawa', 'balita', 'tanda', 'air', 'surut', 'menjelangpanenraya', 'banjir', 'tani']
['kontra', 'tahun', 'pertama', 'saya', 'di', 'malang', 'sering', 'banjir', 'sampai', '120', 'centimeter', 'karena', 'posisi', 'di', 'bawah', 'jalan', 'turun', 'gitu', 'pun', 'belah', 'selokan', 'untung', 'kamar', 'saya', 'lantai']	['kontra', 'malang', 'banjir', '120', 'centimeter', 'posisi', 'jalan', 'turun', 'gitu', 'belah', 'selokan', 'untung', 'kamar', 'lantai']
['dalam', 'ada', 'sulit', 'akibat', 'banjir', '60', 'centimeter', 'mari', 'satu', 'dan', 'kerja', 'sama', 'untuk', 'bangun', 'kembali', 'rumahrumah', 'di', 'bondowoso']	['sulit', 'akibat', 'banjir', '60', 'centimeter', 'mari', 'kerja', 'bangun', 'rumahrumah', 'bondowoso']
['innalillahi', 'banjir', '90', 'centimeter', 'di', 'banyuwangi', 'jawa', 'timur', 'telan', 'korban', 'jiwa', 'orang', 'remaja', 'butuh', 'khusus', 'tewas', 'telah', 'jatuh', 'dari', 'atas', 'jembatan', 'saat', 'tengah', 'main', 'sepeda', 'newssctv', 'liputansiang', 'berita', 'lain']	['innalillahi', 'banjir', '90', 'centimeter', 'banyuwangi', 'jawa', 'timur', 'telan', 'korban', 'jiwa', 'orang', 'remaja', 'butuh', 'khusus', 'tewas', 'jatuh', 'jembatan', 'main', 'sepeda', 'newssctv', 'liputansiang', 'berita']
['bhayangkara', 'bina', 'aman', 'dan', 'tertib', 'masyarakat', 'polisi', 'sektor', 'bojonegoro', 'tutup', 'akses', 'jalan', 'saat', 'jadi', 'banjir', 'yang', 'capai', '45', 'centimeter', 'agar', 'tidak', 'di', 'lalu', 'oleh', 'kendara', 'roda', 'maupun', 'roda', 'kapolresbojonegoro', 'polresbojonegoro', 'bolokuompolisi', 'polripedulibencanabanjir']	['bhayangkara', 'bina', 'aman', 'tertib', 'masyarakat', 'polisi', 'sektor', 'bojonegoro', 'tutup', 'akses', 'jalan', 'banjir', 'capai', '45', 'centimeter', 'kendara', 'roda', 'roda', 'kapolresbojonegoro', 'polresbojonegoro', 'bolokuompolisi', 'polripedulibencanabanjir']

4.1.6 Pembuatan Model

1. Pembagian Dataset

Pembagian dataset pada penelitian ini menggunakan skenario 80:20, 80 untuk data *training* dan 20 untuk data *testing*. Jumlah data *training* pada penelitian ini sebanyak 3.056, sedangkan data *testing* sebanyak 764.

2. *Feature Extraction*

Proses *feature extraction* pada penelitian ini menggunakan TF-IDF *Vectorizer*. Proses TF-IDF dilakukan untuk mengubah data teks menjadi data numerik, agar dapat diproses untuk pembuatan model pada proses selanjutnya. Pada penelitian ini, proses TF-IDF menggunakan modul *TfidfVectorizer* pada yang disediakan oleh *Scikit-Learn*.

Gambar 4. 2 TF-IDF Data *Training* dan Data *Testing*

Hasil proses *extraction feature* dengan TF-IDF *Vectorizer* dapat dilihat pada Gambar 4.2. Berdasarkan hasil tersebut dengan menggunakan *wordcloud*, bahwa kata surabaya banjir dan 40 centimeter merupakan kata yang sering muncul dalam data training, sedangkan kata yang sering muncul dalam data testing adalah banjir capai dan 50 centimeter. Kedua kata tersebut merupakan bagian dari kategori rendah yang merupakan kategori tingkat keparahan yang paling banyak.

3. Pembuatan Model SVM

Pembuatan model dilakukan setelah proses pembagian data. Model SVM yang digunakan berdasarkan pendekatan OVA/OVR, sedangkan kernel SVM yang digunakan adalah RBF. Terdapat dua parameter yang digunakan untuk membuat model SVM yaitu, parameter C dan parameter gamma. Nilai parameter

C dan gamma yang digunakan merujuk pada artikel (Shuai et al., 2018) dengan nilai parameter C sebesar 100 dan parameter gamma sebesar 0.01. Nilai parameter dalam penelitian tersebut sudah dioptimalkan dengan menggunakan *GridSearchCV*. Setelah nilai parameter dan kernel yang digunakan sudah ditentukan, proses pembuatan model (*fit* model) akan dilakukan dengan menggunakan data *train*. Model yang telah dilatih dengan data train akan dilakukan prediksi dengan menggunakan data *testing*.

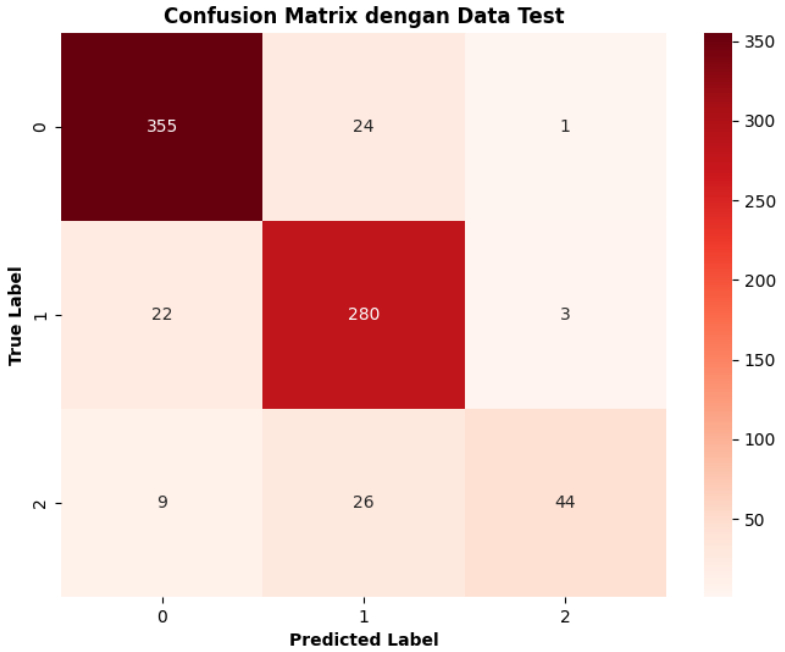
4. Evaluasi dan Analisis

Model yang dihasilkan akan divalidasi menggunakan teknik *cross validation* dengan jumlah *fold* yang digunakan adalah 5 *fold*. Hasil *Accuracy* dengan *cross validation* terbaik, terdapat pada *fold* 1 dengan nilai *Accuracy* sebesar 90.19%. Hasil validasi dengan 5 *fold cross validation* pada penelitian ini dapat dilihat pada Tabel 4.16 berikut.

Tabel 4. 16 Hasil *Accuracy* Skenario *K-fold Cross Validation*

Fold ke-	**Hasil *Accuracy* Prediksi Algoritma *Support Vector Machine* dengan *Cross Validation***
1	90.19%
2	89.54%
3	86.92%
4	89.54%
5	88.15%

Berdasarkan Tabel 4.6 tersebut, didapatkan nilai rata-rata dari *Accuracy* SVM dengan 5 *fold cross validation* menghasilkan nilai sebesar 88.87% yang selanjutnya akan dibulatkan menjadi 89%. Hasil tersebut menunjukkan, bahwasannya model SVM yang dibuat dapat mengklasifikasikan dengan baik. Setelah dilakukan proses validasi model dengan 5 *fold cross validation*, proses selanjutnya adalah evaluasi model dengan menggunakan bantuan dari tabel *confusion matrix*. Tabel *confusion matrix* digunakan untuk mengukur kinerja model dengan memberikan informasi rinci terkait dengan prediksi. Dengan kata lain, *confusion matrix* memberikan representasi tabular dari prediksi yang dibuat oleh model klasifikasi dan label sebenarnya dari data.



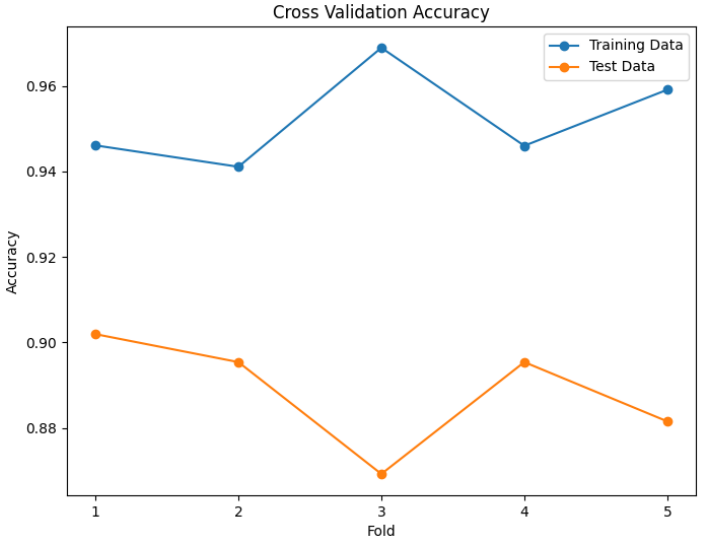
Gambar 4. 3 *Confusion Matrix* Hasil Evaluasi Model

Hasil evaluasi model dengan bantuan tabel *confusion matrix* pada penelitian ini dapat dilihat pada Gambar 4.3. Pada Gambar 4.3 tersebut, didapatkan jumlah data yang secara benar diprediksi (TP) pada label 0 (rendah) sebanyak 355, label 1 (sedang) sebanyak 280, dan label 2 (tinggi) sebanyak 44. Selain TP, terdapat nilai TN, FN, dan FP. Keseluruhan dari nilai tersebut digunakan untuk mengukur performa model dengan *Accuracy*, *Precision*, *Recall*, maupun *F1-Score*.

Tabel 4. 17 Hasil Evaluasi

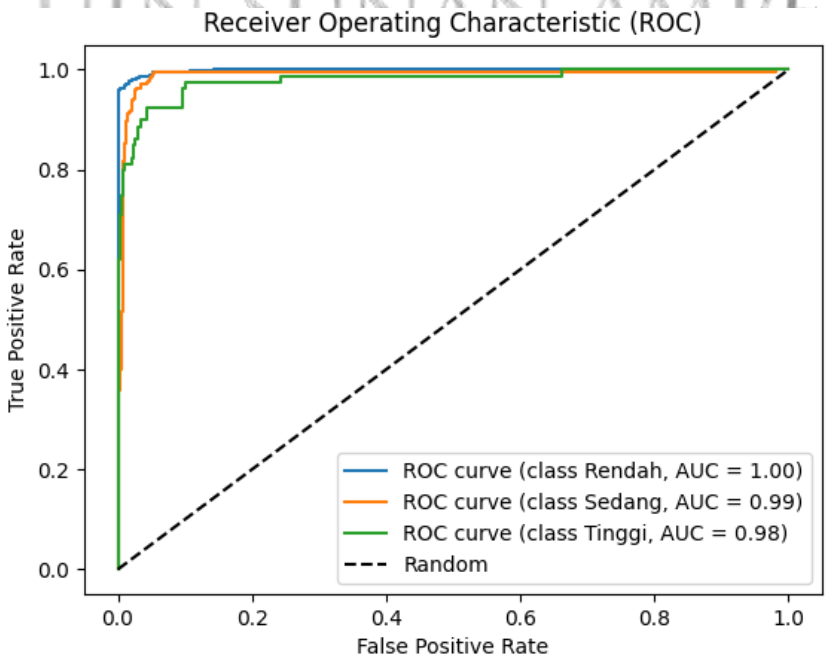
Jenis Pengukuran	Hasil
Accuracy	89%
Precision	89%
Recall	80%
F1 Score	83%

Hasil dari pengukuran performa dengan *Accuracy*, *Precision*, *Recall*, dan *F1-Score* dapat dilihat pada Tabel 4.16. berdasarkan tabel tersebut, diperoleh nilai *Accuracy* 89%, *Precision* 89%, *Recall* 80%, dan *F1-Score* sebesar 83%. Berdasarkan nilai *Accuracy*, *Precision*, *Recall*, dan *F1-Score* tersebut, dapat disimpulkan bahwa model yang telah dibuat memiliki kemampuan yang baik dalam mengklasifikasikan tingkat keparahan.



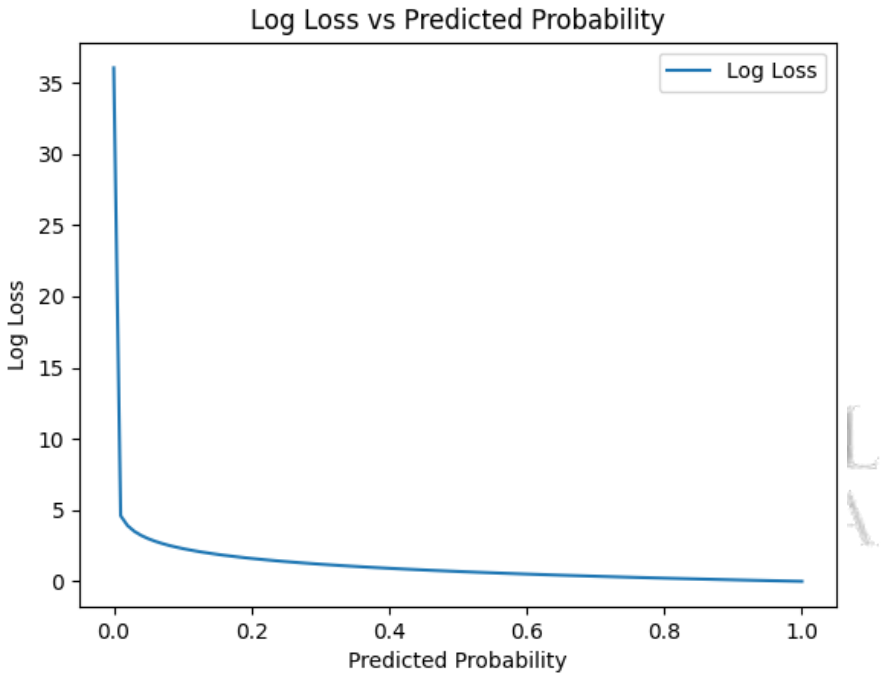
Gambar 4. 4 Visualisasi Hasil *Accuracy train* dan *test* dengan *Cross Validation*

Bentuk visualisasi nilai *Accuracy* yang dihasilkan oleh data *train* dan data *test* dengan 5 *fold cross validation* dapat dilihat pada Gambar 4.4. Pada Gambar 4.4 tersebut menunjukkan, bahwa *Accuracy* data *train* menghasilkan *Accuracy* yang lebih baik daripada data *test*. Berdasarkan hal tersebut, model SVM yang dibuat mengalami permasalahan *overfitting*, sehingga dibutuhkan pengukuran performa lanjutan untuk mengevaluasi model. Pengukuran performa dalam penelitian tidak hanya dengan menggunakan *Precision, Recall, F1-Score*, dan *Accuracy* saja, namun juga menggunakan ROC AUC dan *Logloss*.



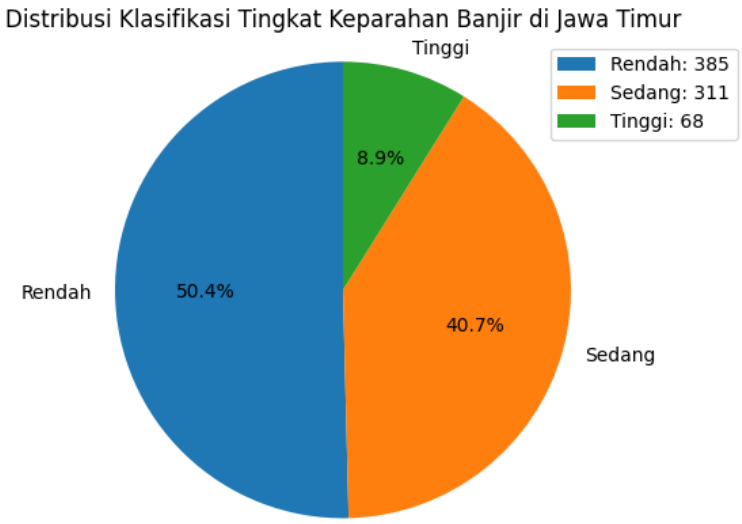
Gambar 4. 5 Visualisasi ROC dan AUC

Dalam penelitian ini, evaluasi yang dilakukan juga menerapkan ROC dan AUC yang digunakan untuk mengevaluasi kinerja model dan menilai seberapa kuat dalam memprediksi suatu data. Sebuah model klasifikasi yang baik akan menghasilkan ROC lebih dekat ke bagian pojok kiri atas, hal tersebut sesuai dengan hasil visualisasi ROC dan AUC pada penelitian ini yang terdapat pada Gambar 4.5. Pada gambar tersebut menunjukkan nilai TPR yang lebih tinggi dan FPR yang lebih rendah dalam berbagai pengaturan *threshold*. Namun, dalam *class* tinggi hasil AUC yang dihasilkan cukup rendah dibandingkan dengan *class* rendah maupun sedang. Hal tersebut dikarenakan dalam *class* tinggi, jumlah data yang digunakan cukup sedikit daripada *class* rendah dan juga *class* sedang. Ketidakseimbangan data tersebut yang menyebabkan nilai AUC dalam *class* tinggi paling rendah diantara *class* yang lainnya.



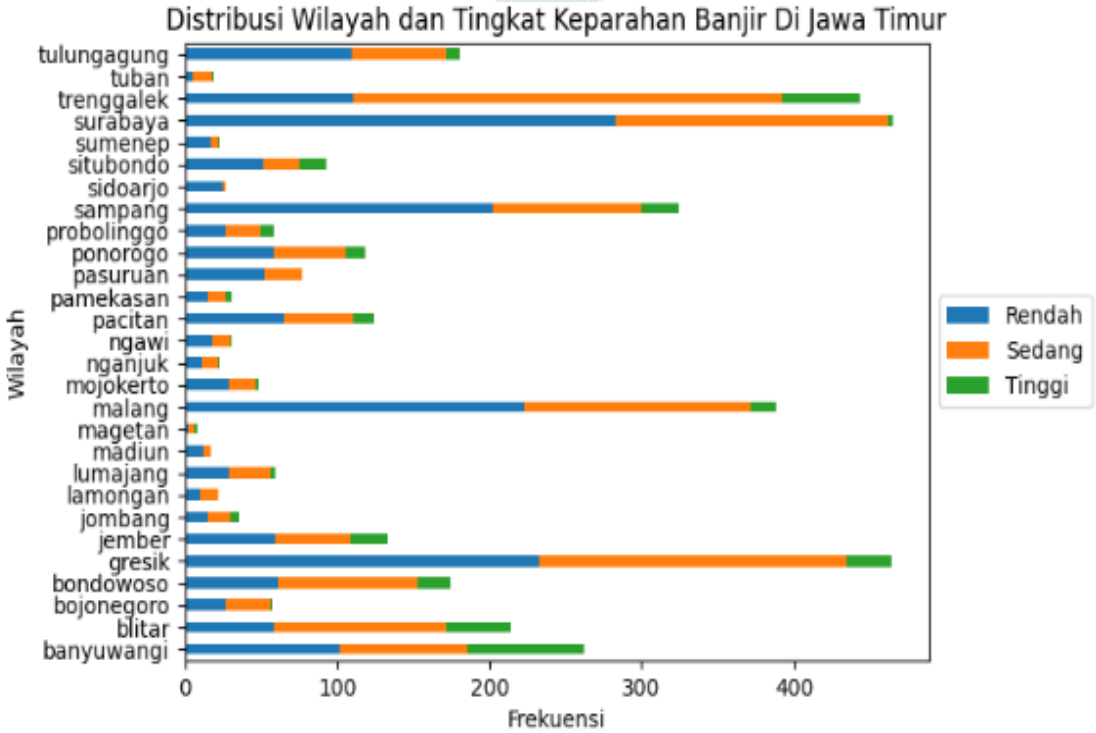
Gambar 4. 6 Visualisasi *Losslog*

Metrik evaluasi lainnya yang digunakan dalam penelitian ini adalah *Logloss*. *Logloss* digunakan untuk mengukur kualitas prediksi probabilitas dari suatu model klasifikasi dengan membandingkannya dengan label yang sebenarnya. Gambar 4.6 merupakan visualisasi *Losslog* yang dihasilkan dalam penelitian ini dengan nilai *Logloss* sebesar 0.1160. Semakin kecil nilai *Logloss*, maka akan semakin baik pula kualitas prediksi yang dihasilkan dari model klasifikasi seperti pada penelitian ini yang menghasilkan nilai *Losslog* yang kecil.



Gambar 4. 7 Distribusi Tingkat Keparahan Banjir Di Jawa timur

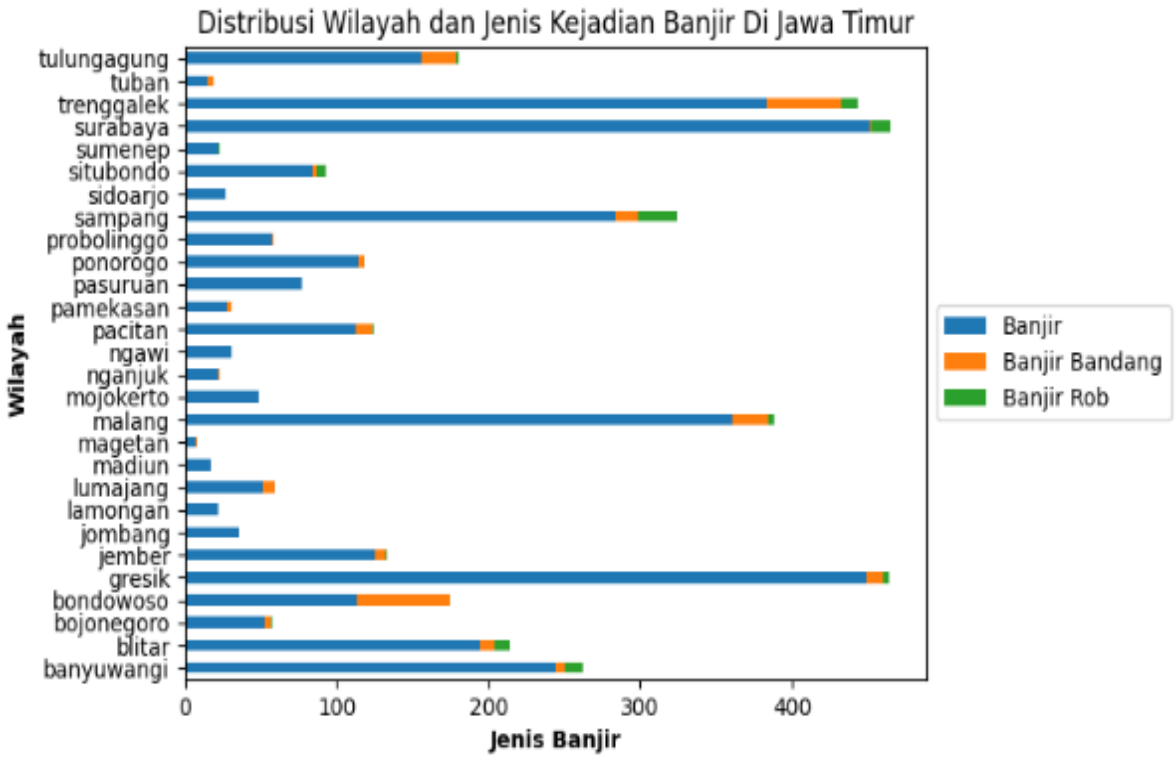
Berdasarkan hasil pembuatan model dan evaluasi model klasifikasi tingkat keparahan banjir pada data *tweet*, diperoleh distribusi tingkat keparahan banjir yang terjadi di Jawa Timur. Kategori tingkat keparahan banjir di Jawa Timur terbanyak adalah kategori rendah. Dari hasil tersebut, kategori tingkat keparahan rendah sebanyak 50,4%, sedang 40,7%, dan tinggi 8,9%. Rincian lengkap dari hasil distribusi tersebut dapat dilihat pada Gambar 4.7.



Gambar 4. 8 Distribusi Wilayah dan Tingkat Keparahan Banjir di Jawa Timur

Gambar 4.8 merupakan distribusi wilayah dan tingkat keparahan banjir terjadi di Jawa Timur. Berdasarkan Gambar 4.8 tersebut, terdapat beberapa informasi yang didapatkan diantaranya:

a. Wilayah yang sering terjadi banjir adalah Surabaya dengan tingkat keparahan banjir yang sering terjadi yaitu rendah, kemudian sedang dan tinggi.
b. Wilayah Magetan merupakan daerah yang jarang terjadi banjir dengan tingkat keparahan yang sering terjadi adalah sedang, kemudian tinggi, dan rendah.
c. Wilayah dengan kategori tingkat keparahan banjir "Rendah" terbanyak yaitu Surabaya, Gresik, dan Malang.
d. Wilayah dengan kategori tingkat keparahan banjir "Sedang" terbanyak yaitu Trenggalek, kemudian Gresik dan yang terakhir yaitu Surabaya.
e. Wilayah dengan kategori tingkat keparahan banjir "Tinggi" yang paling sering yaitu Banyuwangi kemudian Trenggalek dan yang terakhir yaitu Blitar.

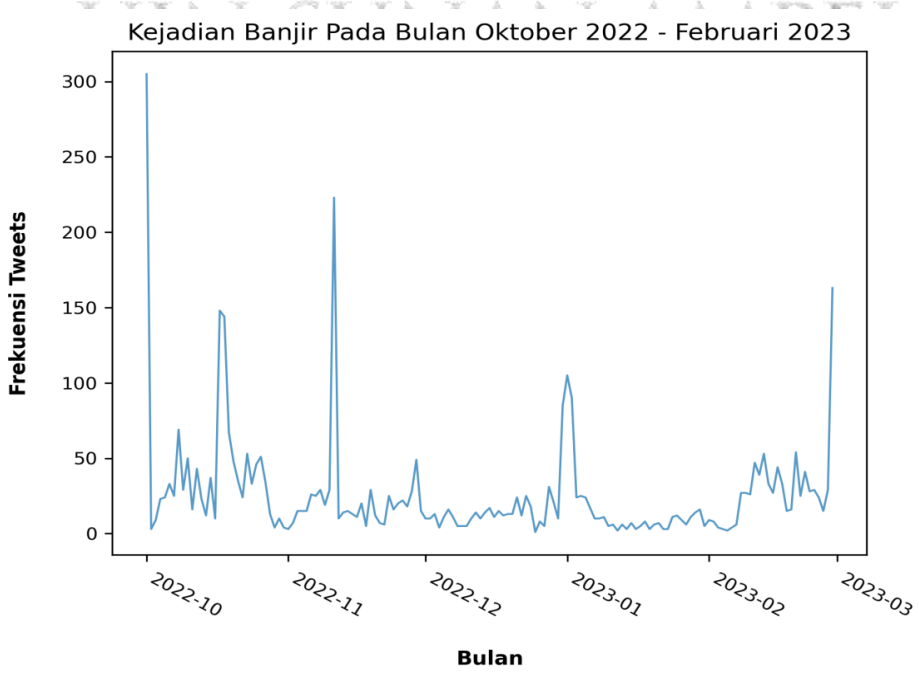


Gambar 4. 9 Distribusi Wilayah dan Jenis Kejadian Banjir Di Jawa Timur

Kejadian banjir yang terjadi di Jawa Timur didominasi oleh kejadian banjir biasa, kemudian banjir bandang dan banjir rob. Berdasarkan Gambar 4.9, Wilayah yang rawan akan terjadi banjir rob adalah wilayah Sampang dikarenakan wilayah Sampang merupakan wilayah pesisir yang dekat dengan air laut khususnya wilayah Sampang bagian utara dan selatan. Selain rawan banjir rob, terdapat wilayah yang rawan banjir bandang. Dalam hal ini wilayah yang rawan terjadi banjir bandang adalah wilayah Bondowoso. Hal tersebut dikarenakan kondisi topografi wilayah Bondowoso yang bervariasi, mulai dari dataran sampai berbukti dan bergunung, sehingga berbentuk cekungan besar (basin) (Arifianto, 2019).

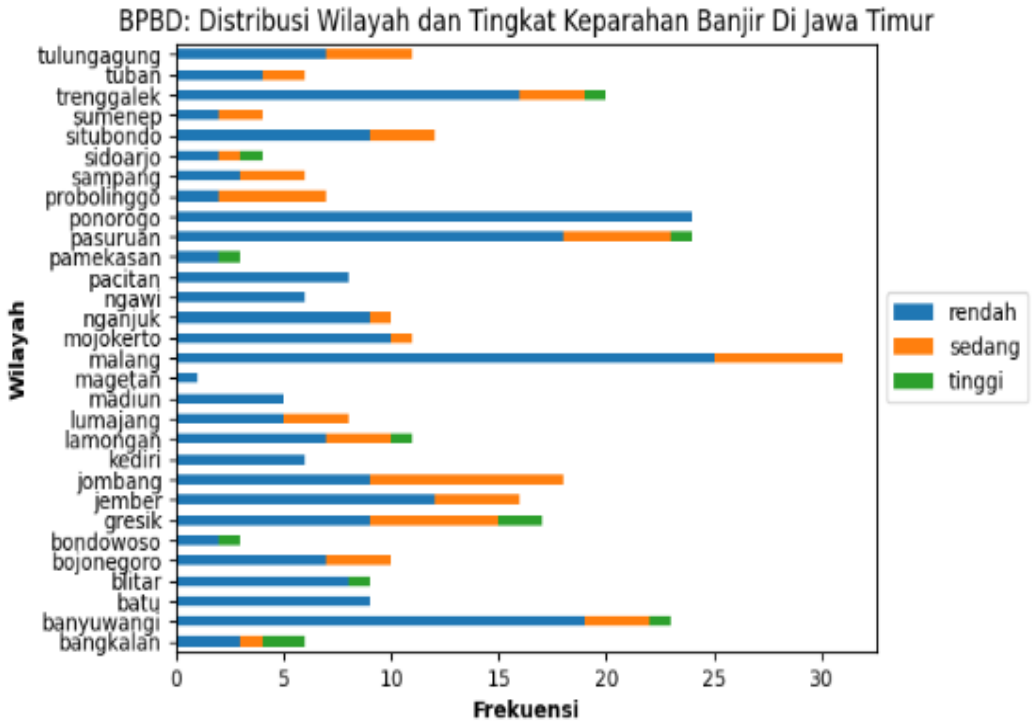
4.2 Pembahasan

Kejadian banjir yang terjadi pada penelitian ini dimulai dari bulan Oktober 2022 hingga bulan Februari 2023. Bulan tersebut menurut Badan Penanggulangan Bencana Daerah (BPBD) Jawa Timur merupakan musim hujan. Sesuai dengan Gambar 4.10 terkait dengan kejadian banjir yang terjadi di Jawa Timur, pada awal bulan Oktober terjadi pelonjakan yang cukup tinggi. Hal tersebut dikarenakan pada bulan Oktober merupakan awal dari musim hujan yang terjadi di Jawa Timur. Namun perlonjakan tersebut tidak hanya terjadi pada awal bulan Oktober saja, Akan tetapi juga terjadi pada bulan November.



Gambar 4. 10 Kejadian Banjir Di Jawa Timur

Terjadinya pelonjakan tersebut merupakan hal yang wajar terjadi pada saat musim hujan, mengingat Indonesia merupakan negara yang dilalui oleh garis khatulistiwa dan memiliki intensitas curah hujan yang tak menentu. Kejadian bencana banjir yang terjadi di Indonesia khususnya Jawa Timur ditangani oleh BPBD Jawa Timur. Tugas BPBD sendiri yaitu melakukan penanggulangan bencana yang terjadi di Provinsi maupun Kabupaten/Kota. Berdasarkan data kejadian banjir yang diperoleh dari BPBD Jawa Timur pada bulan Oktober 2022 hingga Februari 2023, kejadian banjir yang terjadi di provinsi Jawa timur sebanyak 329.



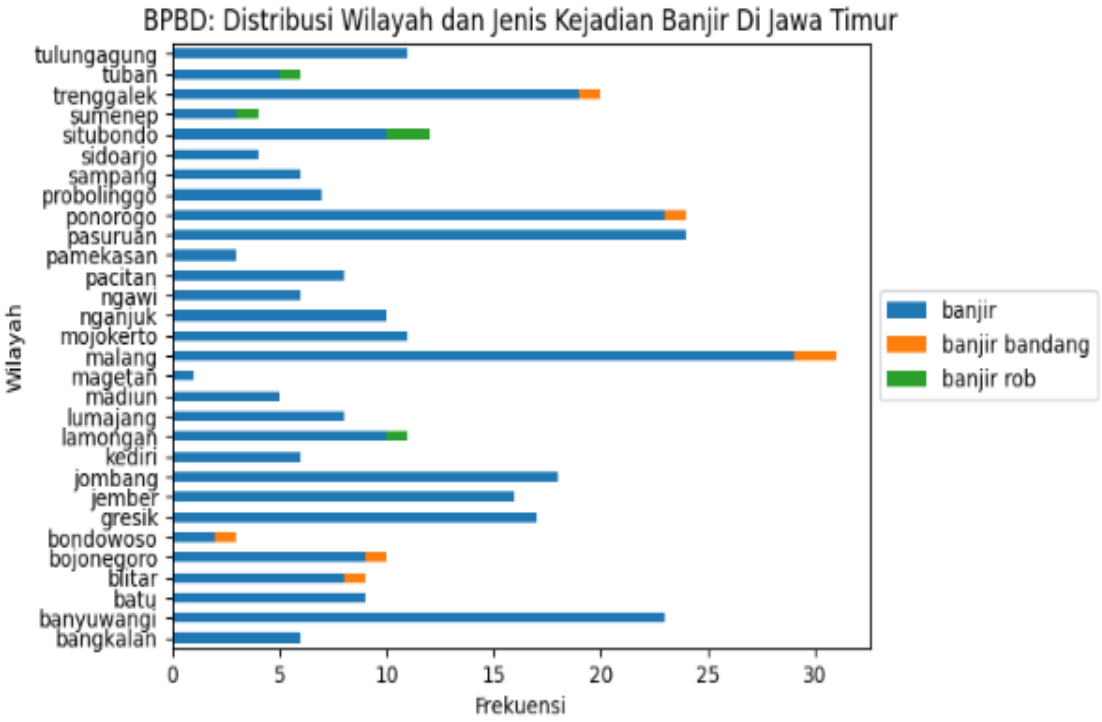
Gambar 4. 11 BPBD: Distribusi Wilayah dan Tingkat Keparahan Banjir Di Jawa Timur
(sumber data: https://smartpb.bpbd.jatimprov.go.id/public)

Berdasarkan data yang diperoleh dari laman *website* BPBD Jawa Timur, distribusi wilayah dan tingkat keparahan banjir di Jawa Timur dapat dilihat pada Gambar 4.11. Pada gambar tersebut menunjukkan bahwa wilayah malang merupakan wilayah yang sering terjadi banjir. Hal ini sudah sesuai dengan proses filter lokasi yang dilakukan pada penelitian ini sesuai dengan Gambar 4.1. Namun setelah dilakukan proses filter tersebut, wilayah yang sering terjadi banjir bukan wilayah Malang, melainkan wilayah Surabaya. Hal tersebut terjadi dikarenakan

dalam *tweet* yang siap untuk proses pembuatan model sudah bersih dari kejadian non banjir seperti “banjir tangis”, “banjir air mata”, “banjir kabar”, dsb. Sehingga hasil akhir yang didapatkan, bahwa wilayah yang sering terjadi banjir adalah wilayah Surabaya sesuai dengan Gambar 4.8. Selain itu, faktor lain yang menyebabkan wilayah surabaya sebagai wilayah yang rawan terjadi banjir dikarenakan penduduk wilayah tersebut sering menggunakan media *Twitter* sebagai tempat keluh kesah dan berbagi informasi meskipun berita banjir yang diinformasikan berkategori rendah.

Gambar 4. 12 BPBD: Tingkat Keparahan Banjir
(sumber data: https://smartpb.bpbd.jatimprov.go.id/public)

Tingkat keparahan banjir yang dihasilkan dari data banjir pada laman *website* BPBD Jawa Timur dapat dilihat pada Gambar 4.12. Pada gambar tersebut, kategori rendah merupakan kategori yang paling dominan kemudian disusul kategori sedang dan yang terakhir kategori tinggi. Hal tersebut selaras dengan hasil proses klasifikasi tingkat keparahan banjir dengan algoritma SVM yang dapat dilihat pada Gambar 4.7.



Gambar 4. 13 BPBD: Distribusi Wilayah dan Jenis Kejadian Banjir Di Jawa Timur
(sumber data: https://smartpb.bpbd.jatimprov.go.id/public)

Kejadian banjir yang terjadi di Jawa Timur terdiri dari banjir biasa, banjir bandang, dan banjir rob. Gambar 4.13 merupakan distribusi wilayah dan jenis kejadin banjir yang terjadi di Jawa Timur berdasarkan data yang didapatkan pada laman BPBD Jawa Timur. Berdasarkan gambar tersebut, banjir bandang rawan terjadi di wilayah Malang, sedangkan banjir rob rawan terjadi pada wilayah Situbondo. Hasil kejadian ini tidak sesuai dengan hasil klasifikasi dengan SVM yang terdapat pada Gambar 4.9. Adanya ketidaksesuaian ini memungkinkan penduduk di wilayah Malang dan Situbondo jarang menggunakan aplikasi *Twitter* dalam membagikan informasi kejadian banjir yang terjadi. Namun, implementasi dengan menggunakan model memudahkan dalam mengklasifikasikan kejadian banjir dengan cepat dan akurat pada data yang banyak.

Implementasi Support Vector Machine pada penelitian ini digunakan untuk klasifikasi tingkat keparahan banjir dan menghasilkan nilai *Accuracy* 88%. Model klasifikasi *multiclass* yang dibuat dengan memanfaatkan pendekatan OVR untuk seperti pada penelitian (Delimayanti et al., 2021) yang menerapkan OVR dalam pembuatan model klasifikasi *multiclass* untuk pesan bencana banjir di

Twitter. Namun dalam penelitian tersebut, hasil *Accuracy* yang didapatkan 87,03%. Hal ini dapat disimpulkan bahwa pendekatan OVR pada algoritma *Support Vector Machine* dapat diterapkan dalam mengklasifikasikan tingkat keparahan banjir di Jawa Timur dan menghasilkan hasil *Accuracy* yang lebih baik.

Berdasarkan riset yang telah dilakukan oleh (Khalaf et al., 2018) yang menggunakan *Support Vector Machine* untuk memprediksi tingkat keparahan banjir menghasilkan *Accuracy* sebesar 78%. Data yang digunakan pada penelitian tersebut berasal dari kumpulan situs web lembaga lingkungan hidup. Pada penelitian ini algoritma *Support Vector Machine* digunakan untuk klasifikasi tingkat keparahan banjir dengan menggunakan data *Twitter* menghasilkan nilai *Accuracy* yang lebih baik. Hal tersebut membuktikkan bahwasannya algoritma *Support Vector Machine* dapat diterapkan untuk klasifikasi tingkat keparahan banjir.

Pada penelitian terdahulu, SVM telah digunakan dalam bidang kebencanaan, seperti penelitian (Sreenivasulu & Sridevi, 2020) yang menggunakan SVM untuk mengklasifikasikan kejadian gempa bumi. Dalam penelitian tersebut *extraction feature* yang digunakan adalah BOW, kemudian kernel SVM yang digunakan yaitu linear dan RBF. Hasil *Accuracy* yang didapatkan cukup bagus pada kernel RBF sebesar 77%. Pada domain kebencaan seperti banjir pada penelitian ini, menghasilkan nilai *Accuracy* yang jauh lebih baik. Model yang dibuat menggunakan kernel RBF dan *extraction feature* dengan implementasi *tfidfvectorizer*. Hal tersebut membuktikkan bahwasannya model SVM yang digunakan pada penelitian ini jauh lebih baik dari penelitian sebelumnya.

Penggunaan evaluasi AUC dalam model, dapat memberikan analisis lebih tajam terkait dengan performa model. Pada penelitian (Chowdhury, 2020) yang menggunakan AUC pada model SVM menghasilkan, bahwa model yang dibuat memiliki nilai *Accuracy* 96% dan AUC sebesar 90% meskipun pada penelitian tersebut tidak menggunakan teknik *cross validation*. Pada penelitian ini, teknik *cross validation* dengan evaluasi AUC menghasilkan Accuracy 88%, AUC *class* rendah 0.998, AUC *class* sedang 0.993, dan AUC *class* tinggi 0.983. Hal tersebut membuktikkan bahwa, meskipun hasil *Accuracy* yang didapatkan lebih rendah

namun AUC yang dihasilkan lebih tinggi dari penelitian sebelumnya. Semakin tinggi nilai AUC yang dihasilkan, maka akan semakin baik pula performa model dalam mengklasifikasikan setiap kelas. Sehingga model pada penelitian ini lebih baik dari pada penelitian sebelumnya.

Evaluasi metriks lainnya *seperti* Logloss yang digunakan untuk mengukur perbedaan antara probabilitas yang diprediksi dan nilai aktual. Semakin kecil nilai suatu *Logloss* yang dihasilkan, maka model memiliki performa yang sangat bagus. Pada penelitian (Lin et al., 2022) yang menggunakan *ensemble learning* menghasilkan nilai *Logloss* terbaik sebesar 0.3282, sedangkan pada penelitian ini *Logloss* yang dihasilkan sebesar 0.1160. Hasil tersebut menunjukkan bahwa model SVM dalam penelitian lebih baik daripada penelitian sebelumnya dalam proses klasifikasi.

BAB V
KESIMPULAN DAN SARAN

1.1 Kesimpulan

Berdasarkan hasil implementasi dari berbagai tahapan dalam penelitian ini, berikut merupakan kesimpulan yang didapatkan yaitu:

1. Penerapan algoritma *Support Vector Machine* untuk klasifikasi tingkat keparahan banjir di Jawa Timur dimulai dengan tahapan pengumpulan data *Twitter* dari bulan Oktober 2022 – Februari 2023 yang menghasilkan data sebanyak 537.340 baris. Data terkumpul akan dibersihkan dari *missing value* dan data duplikat yang kemudian digunakan untuk proses data filter. Proses selanjutnya yaitu pelabelan data oleh BPBD Jawa Timur dan dilanjutkan *preprocessing* dengan tahapan *cleaning, case folding, tokenizing, slangwords, stemming, stopwords* dan kemudian dilakukan pembuatan model dengan membagi dataset menjadi 80% *train* dan 20% *test*. Model SVM yang digunakan dengan pendekatan *multiclass* OVR, nilai paramater C 100, dan paramater gamma 0.01. Model yang dibuat kemudian divalidasi dengan teknik *cross validation* sebanyak 5 *fold* dan dievaluasi dengan *Precision, Recall, Accuracy, F1-Score,* ROC AUC, dan *Logloss*.
2. Performa algoritma *Support Vector Machine* dalam mengklasifikasi tingkat keparahan banjir di Jawa Timur menghasilkan nilai *Precision* 89%, *Recall* 80%, *Accuracy* 89%, *F1-Score* 83%, AUC *class* rendah 0.998, *class* sedang 0.993, dan *class* tinggi 0.983 serta rata-rata *Logloss* sebesar 0.114. Berdasarkan hasil tersebut dapat disimpulkan bahwa model yang dibuat memiliki kemampuan yang sangat baik dengan nilai *Accuracy* yang akurat dalam mengklasifikasikan tingkat keparahan banjir, meskipun model mengalami permasalahan *overfitting*.

1.2 Saran

Sesuai dengan hasil penelitian ini yang tidak luput dari kekurangan, oleh karena itu diperlukan perbaikan untuk penelitian selanjutnya. Terdapat beberapa saran yang dapat dilakukan pada penelitian selanjutnya, yaitu:

1. Deteksi tingkat keparahan pada penelitian ini dilakukan dengan menggunakan angka yang diikuti dengan satuannya. Sehingga pada penelitian selanjutnya, deteksi tingkat keparahan dapat dilakukan dengan menggunakan basis aspek seperti banjir setinggi lutut, banjir sepaha, banjir sepinggang, dsb.
2. Proses filter lokasi yang digunakan dalam penelitian ini dengan mengidentifikasi kata yang sesuai dengan corpus yang disesuaikan dengan Permendagri tahun 2020. Dalam hal ini proses filter lokasi dengan tanpa kata kabupaten maupun kota, sehingga masih terjadi pencampuran wilayah yang memiliki kabupaten dan kota dalam satu wilayah. Dalam penelitian lanjutan diharapkan dapat memperbaiki hal tersebut dan dapat mendetailkan kembali lokasi sampai kecamatan maupun desa. Selain itu pentingnya perluasan wilayah menjadi se-Jawa agar dapat membantu pemerintahan dalam analisis bencana banjir yang terjadi.
3. Proses klasifikasi yang dilakukan pada penelitian ini menggunakan *machine learning* dengan pendekatan *Support Vector Machine* sudah cukup menghasilkan nilai evaluasi yang baik dengan menggunakan evaluasi *Precision, Recall, Accuracy, F-Measure, ROC AUC,* dan *Logloss*. Namun, model yang dibuat memiliki permasalahn *overfitting* sehingga dibutuhkan penelitian lanjutan dengan menggunakan algoritma *hybrid machine learning, boosting* maupun *deep learning*.

DAFTAR PUSTAKA

A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. *Remote Sensing*, *11*(2), Article 2. https://doi.org/10.3390/rs11020185

Abu Taher, S., Afsana Akhter, K., & Azharul Hasan, K. M. (2018). N-Gram Based Sentiment Mining for Bangla Text Using Support Vector Machine. 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 1–5. https://doi.org/10.1109/ICBSLP.2018.8554716

Annur, C. M. (2023, January 4). Jawa Barat, Provinsi dengan Bencana Alam Terbanyak pada 2022 | Databoks. https://databoks.katadata.co.id/datapublish/2023/01/04/jawa-barat-provinsi-dengan-bencana-alam-terbanyak-pada-2022

Aziz, K., Zaidouni, D., & Bellafkih, M. (2019). Social Network Analytics: Natural Disaster Analysis Through *Twitter*. *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 1–7. https://doi.org/10.1109/ICDS47004.2019.8942337

Bhavani, A., & Santhosh Kumar, B. (2021). A Review of State Art of Text Classification Algorithms. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 1484–1490. https://doi.org/10.1109/ICCMC51019.2021.9418262

Chauhan, V. K., Dahiya, K., & Sharma, A. (2019). Problem formulations and solvers in linear SVM: A review. *Artificial Intelligence Review*, *52*(2), 803–855. https://doi.org/10.1007/s10462-018-9614-6

Delimayanti, M. K., Sari, R., Laya, M., & Faisal, M. R. (2021). *Pemanfaatan Metode Multiclass-SVM pada Model Klasifikasi Pesan Bencana Banjir di Twitter*.

Devid. (2023, April 16). Masdevid/ID-Stopwords. https://github.com/masdevid/ID-Stopwords (Original work published 2016)

Edwardsamuel. (2018, January 11). Wilayah Administratif Indonesia edwardsamuel. https://github.com/edwardsamuel/Wilayah-Administratif-Indonesia/tree/master/csv

Fariz, T. R., Suhardono, S., & Verdiana, S. (2021). Pemanfaatan Data *Twitter* Dalam Penanggulangan Bencana Banjir dan Longsor. *CogITo Smart Journal*, *7*(1), 135–147. https://doi.org/10.31154/cogito.v7i1.305.135-147

Fesseha, A., Xiong, S., Emiru, E. D., & Dahou, A. (2020). Text Classification of News Articles Using Machine Learning on Low-resourced Language: Tigrigna. *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 34–38. https://doi.org/10.1109/ICAIBD49809.2020.9137443

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348–361. https://doi.org/10.3102/1076998619832248

Hill, C. (2020). *Learning scientific programming with Python*. Cambridge University Press.

Ibrohim, M. O. (2019). *Id-multi-label-hate-speech-and-abusive-language-detection* [TeX]. https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection (Original work published 2019)

Kalcheva, N., Karova, M., & Penev, I. (2020). Comparison of the accuracy of SVM kemel functions in text classification. *2020 International Conference on Biomedical Innovations and Applications (BIA)*, 141–145. https://doi.org/10.1109/BIA50171.2020.9244278

Khalaf, M., Hussain, A. J., Al-Jumeily, D., Baker, T., Keight, R., Lisboa, P., Fergus, P., & Al Kafri, A. S. (2018). A data science methodology based on machine learning algorithms for flood severity prediction. *2018 IEEE Congress on Evolutionary Computation (CEC)*, 1–8.

Kim, S.-W., & Gil, J.-M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-Centric Computing and Information Sciences*, *9*, 1–21.

kirralabs. (2020). *Indonesian-NLP-resources*. kirralabs. https://github.com/kirralabs/indonesian-NLP-resources (Original work published 2018)

Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, & Brown. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, *106*, 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

Mosavi, A., Ozturk, P., & Chau, K. (2018). Flood Prediction Using Machine Learning Models: Literature Review. *Water*, *10*(11), Article 11. https://doi.org/10.3390/w10111536

Nguyen, H., Bui, X.-N., Choi, Y., Lee, C. W., & Armaghani, D. J. (2021). A Novel Combination of Whale Optimization Algorithm and Support Vector Machine with Different Kernel Functions for Prediction of Blasting-Induced Fly-Rock in Quarry Mines. *Natural Resources Research*, *30*(1), 191–207. https://doi.org/10.1007/s11053-020-09710-7

Owen, L. (2020, April 3). *Combined_slang_words*. https://github.com/louisowen6/NLP_bahasa_resources/blob/master/combined_slang_words.txt

Parapat, I. M., Furqon, M. T., & Sutrisno. (2018). Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *2*(10), 3163–3169.

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine Learning* (pp. 101–121). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Salsabila, N. A. (2018). *Kamus Alay—Colloquial Indonesian Lexicon*. https://github.com/nasalsabila/kamus-alay/blob/c710fa9f3f8d8727c24f3209f4f6a4b6ecaf6d8b/colloquial-indonesian-lexicon.csv

Schroeder, A. J., Gourley, J. J., Hardy, J., Henderson, J. J., Parhi, P., Rahmani, V., Reed, K. A., Schumacher, R. S., Smith, B. K., & Taraldsen, M. J. (2016). The development of a flash flood severity index. *Journal of Hydrology*, *541*, 523–532.

Sreenivasulu, M., & Sridevi, M. (2020). Comparative study of statistical features to detect the target event during disaster. *Big Data Mining and Analytics*, *3*(2), 121–130. https://doi.org/10.26599/BDMA.2019.9020021

Utami, R. C., & Tyas, W. P. (2021). *BENTUK KESIAPSIAGAAN MENGHADAPI BENCANA ALAM BANJIR BANDANG SUKU WANA, KABUPATEN MOROWALI UTARA, SULAWESI TENGAH*.

Yovellia Londo, G. L., Kartawijaya, D. H., Ivariyani, H. T., W.P., Y. S. P., Muhammad Rafi, A. P., & Ariyandi, D. (2019). A Study of Text Classification for Indonesian News Article. *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 205–208. https://doi.org/10.1109/ICAIIT.2019.8834611

Zhang, X.-D. (2020). Support Vector Machines. In X.-D. Zhang (Ed.), *A Matrix Algebra Approach to Artificial Intelligence* (pp. 617–679). Springer. https://doi.org/10.1007/978-981-15-2770-8_8

Arifianto, Y. D. (2019). Mitigasi Bencana Banjir Daerah Aliran Sungai (DAS) Sampean Berbasis Pemberdayaan Masyarakat, Di Wilayah Bondowoso Dan Situbondo, Provinsi Jawa Timur. *INFRASTRUKTUR*, 27.

Basavegowda, H. S., & Dagnew, G. (2020). Deep learning approach for microarray cancer data classification. *CAAI Transactions on Intelligence Technology*, *5*(1), 22–33.

Bashir, D., Montañez, G. D., Sehra, S., Segura, P. S., & Lauw, J. (2020). An information-theoretic perspective on overfitting and underfitting. *AI 2020: Advances in Artificial Intelligence: 33rd Australasian Joint Conference, AI 2020, Canberra, ACT, Australia, November 29–30, 2020, Proceedings 33*, 347–358.

Chowdhury, K. (2020). Spam identification on Facebook, Twitter and Email using machine learning. *CERES*, *19*.

Fergus, P., Montanez, C. C., Abdulaimma, B., Lisboa, P., Chalmers, C., & Pineles, B. (2018). Utilizing deep learning and genome wide association studies for epistatic-driven preterm birth classification in African-American women. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(2), 668–678.

Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial. *ArXiv Preprint ArXiv:1905.12787*.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, *25*(1), 114–146.

Keerthi Kumar, H. M., & Harish, B. S. (2018). Classification of Short Text Using Various Preprocessing Techniques: An Empirical Evaluation. In P. K. Sa, S. Bakshi, I. K. Hatzilygeroudis, & M. N. Sahoo (Eds.), *Recent Findings in Intelligent Computing Techniques* (pp. 19–30). Springer. https://doi.org/10.1007/978-981-10-8633-5_3

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.

Lin, S., Zheng, H., Han, B., Li, Y., Han, C., & Li, W. (2022). Comparative performance of eight ensemble learning approaches for the development of

models of slope stability prediction. *Acta Geotechnica*, *17*(4), 1477–1502. https://doi.org/10.1007/s11440-021-01440-1

Shuai, Y., Zheng, Y., & Huang, H. (2018). Hybrid Software Obsolescence Evaluation Model Based on PCA-SVM-GridSearchCV. *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, 449–453. https://doi.org/10.1109/ICSESS.2018.8663753

Wang, Y., Jin, J., Zhang, W., Yu, Y., Zhang, Z., & Wipf, D. (2021). Bag of tricks for node classification with graph neural networks. *ArXiv Preprint ArXiv:2103.13355*.

Yang, S., & Zhang, H. (2018). Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. *International Journal of Computer and Information Engineering*, *12*(7), 525–529.

Zhao, G., Liu, Y., Zhang, W., & Wang, Y. (2018). TFIDF based feature words extraction and topic modeling for short text. *Proceedings of the 2018 2nd International Conference on Management Engineering, Software Engineering and Service Sciences*, 188–191.