

**KLASIFIKASI TEMUAN INSPEKSI MUTU
HASIL PERIKANAN MENGGUNAKAN
*SUPPORT VECTOR MACHINE***

PROPOSAL TESIS

Disusun oleh:
Weldy Sujarmanto
NIM: 216150101111008



**PROGRAM MAGISTER ILMU KOMPUTER
JURUSAN TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS BRAWIJAYA
MALANG
TAHUN 2022**

PENGESAHAN

KLASIFIKASI TEMUAN INSPEKSI MUTU HASIL PERIKANAN MENGGUNAKAN *SUPPORT VECTOR MACHINE*

PROPOSAL TESIS

Disusun oleh:
Weldy Sujarmanto
NIM: 216150101111008

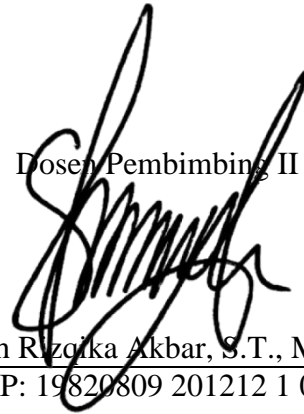
Proposal Tesis ini,
Telah diperiksa dan disetujui oleh:

Dosen Pembimbing I



Prof. Wayan Firdaus Mahmudy, S.Si., MT., Ph.D.
NIP: 19720919 199702 1 001

Dosen Pembimbing II



Sabriansyah Rizaika Akbar, S.T., M.Eng., Ph.D
NIP: 19820809 201212 1 004

BAB 1 PENDAHULUAN

1.1.Latar Belakang

Indonesia merupakan negara kepulauan terbesar di dunia dan mempunyai wilayah perairan yang luas, sehingga mempunyai potensi hasil perikanan sangat besar. Globalisasi perdagangan hasil pertanian yang termasuk di dalamnya hasil perikanan sudah mulai meluas ke berbagai negara, dan kehadirannya tidak dapat dihindarkan. Konsumen telah menyadari bahwa mutu khususnya keamanan pangan hasil perikanan tidak dapat dijamin hanya dengan hasil uji produk akhir dari laboratorium (Saifullah, 2018). Industri pengolahan hasil perikanan dituntut untuk menerapkan sistem jaminan mutu dan keamanan pangan agar produk makanan yang dihasilkan dapat diterima dan memenuhi regulasi yang telah ditetapkan oleh masing-masing negara (Vatria, 2022).

Sistem *Hazard Analysis Critical Control Point* (HACCP) adalah suatu metode manajemen keamanan hasil perikanan yang bersifat sistematis dan didasarkan pada prinsip-prinsip yang telah dikenal, yang ditujukan untuk mengidentifikasi bahaya (*hazard*) yang kemungkinan dapat terjadi pada setiap tahapan dari rantai persediaan makanan (Permen KP No 10/2021 Tentang Standar Kegiatan Usaha & Produk Pada Penyelenggaraan Perizinan Berusaha Berbasis Risiko Sektor KP, 2021). Sebagai bentuk pengawasan keamanan pangan hasil perikanan, Pemerintah melalui Kementerian Kelautan dan Perikanan (KKP) menerbitkan Sertifikat Penerapan HACCP.

Menteri Kelautan dan Perikanan (MKP) melimpahkan kewenangan kepada unit eselon I Badan Karantina Ikan, Pengendalian Mutu, dan Keamanan Hasil Perikanan (BKIPM) untuk menerbitkan Sertifikat Penerapan HACCP (Sertifikat HACCP) bagi Unit Pengolahan Ikan (UPI) yang memasarkan produk perikanan ke luar negeri, Kepala BKIPM memberikan tugas kepada Pegawai KKP yang telah ditetapkan sebagai Inspektur Mutu (IM), untuk melakukan inspeksi terhadap penerapan sistem jaminan mutu dan keamanan hasil perikanan pada UPI, IM menyampaikan hasil inspeksi berupa Laporan IM kepada Tim Sertifikasi (Tim) yang dibentuk oleh Kepala BKIPM untuk melakukan evaluasi, berdasarkan hasil evaluasi tersebut, Kepala BKIPM menerbitkan Sertifikat Penerapan HACCP (Permen KP No 51/2018 Tentang Persyaratan & Tata Cara Penerbitan Sertifikat Penerapan HACCP, 2018).

Laporan IM adalah dasar untuk menentukan suatu UPI memenuhi syarat atau tidak diberikan Sertifikat HACCP. Laporan IM berupa daftar temuan inspeksi mutu hasil perikanan (Temuan) yang diberikan kepada UPI jika pada proses pengolahan ikan melanggar sistem HACCP. Setiap Temuan tersebut diberikan empat tingkat kriteria dari yang paling bahaya sampai yang paling aman yaitu kritis, serius, mayor dan minor (Permen KP No 51/2018 Tentang Persyaratan & Tata Cara Penerbitan Sertifikat Penerapan HACCP, 2018).

IM berperan untuk menentukan suatu Temuan masuk ke dalam salah satu kriteria berdasarkan ketentuan yang berlaku. Namun, dalam menentukan kriteria tersebut, Tim Sertifikasi sebagai pihak yang berperan melakukan evaluasi Laporan IM, masih banyak menemukan IM salah dalam menentukan kriteria untuk suatu Temuan.

KKP sebagai salah satu institusi pemerintah yang mendukung Sistem Pemerintahan Berbasis Elektronik (SPBE), oleh karena itu KKP memanfaatkan teknologi informasi untuk memberikan pelayanan kepada masyarakat, termasuk pada pelayanan penerbitan Sertifikat HACCP yang sudah menggunakan suatu sistem informasi berbasis web bernama HACCP Online System (Honest). Honest dibangun dan dikembangkan oleh unit teknis KKP yang bertanggungjawab pada bidang Teknologi Informasi yaitu Pusat Data, Statistik dan Informasi (Pusdatin).

Berdasarkan permasalahan dengan masih banyaknya terjadi kesalahan IM dalam menentukan kriteria Temuan, maka penelitian ini mencoba membangun suatu model yang mampu melakukan klasifikasi Temuan tersebut ke dalam kriteria tertentu secara akurat dengan menggunakan data Temuan pada *database* Honest pada periode 2019 sampai dengan 2022 sebagai acuan *ground truth*.

Temuan berupa kalimat dalam bahasa Indonesia, agar bisa diproses oleh model ML maka perlu menggunakan *word embedding* (WE) untuk mengubah setiap kata dalam kalimat tersebut menjadi vektor, selain itu representasi teks ke dalam bentuk vektor bertujuan untuk mengoptimalkan algoritma klasifikasi (Nugroho, Bachtiar, et al., 2022).

Banyak jenis teknik WE yang telah diusulkan dalam berbagai penelitian dan yang populer yang digunakan adalah Word2Vec, FastText dan GloVe, ketiga teknik tersebut banyak digunakan untuk tugas khusus pada *natural language processing* (NLP), seperti analisis atau bahkan dalam tugas yang lebih kompleks, seperti deteksi spam dan tanya jawab (Kanakaris et al., 2022). Pada penelitian ini dilakukan perbandingan antara ketiganya untuk menemukan teknik dengan akurasi terbaik.

Vektor dari hasil representasi teks bisa mencapai 300 dimensi untuk Word2Vec (Gupta et al., 2022), GloVe (Muhammad et al., 2021), dan FastText (Khasanah, 2021), maka untuk menyederhanakan komputasi dan performa algoritma ML pada penelitian ini mengusulkan penerapan reduksi dimensi pada vektor hasil WE.

Ada begitu banyak teknik untuk reduksi dimensi, *Autoencoder* (AE) dan *Principal Component Analysis* (PCA) sangat dikenal di antara teknik tersebut, PCA umumnya bekerja dengan baik pada dataset kecil, karena AE dasarnya adalah *neural network*, maka AE membutuhkan banyak data dibandingkan dengan PCA (Pramoditha, 2022), sehingga pada penelitian ini AE lebih sesuai digunakan untuk melakukan reduksi dimensi.

Kemudian, untuk melakukan klasifikasi penelitian ini mengusulkan menggunakan algoritma ML yang umum digunakan. Ada enam algoritma ML yang banyak digunakan untuk melakukan klasifikasi yaitu *Logistic Regression* (LR), *Decision Tree* (DT), *Random Forest* (RF), *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN) dan *Naive Bayes* (NB) (Gong, 2022), pada penelitian ini enam algoritma tersebut digunakan untuk dibandingkan satu sama lain, sehingga pada hasil akhir penelitian diharapkan ditemukan kombinasi terbaik di antara teknik WE dan algoritma ML yang paling akurat untuk melakukan klasifikasi Temuan.

1.2. Rumusan Masalah

Rumusan Masalah dari tesis ini adalah sebagai berikut:

1. Bagaimana tingkat akurasi setiap kombinasi WE, AE dan ML dalam klasifikasi kriteria Temuan?
2. Manakah kombinasi WE, AE dan ML yang terbaik untuk klasifikasi kriteria Temuan?

1.3. Tujuan

Tujuan tesis ini adalah sebagai berikut:

1. Menguji tingkat akurasi setiap kombinasi WE, AE dan ML dalam klasifikasi kriteria Temuan;
2. Mengetahui kombinasi WE, AE dan ML yang terbaik untuk klasifikasi kriteria Temuan.

1.4.Manfaat

Manfaat tesis ini adalah:

1. Memberi bantuan pilihan kriteria pada Temuan secara otomatis dan akurat bagi IM saat melakukan kegiatan inspeksi mutu hasil perikanan;
2. Memberikan cetak biru model ML untuk klasifikasi berdasarkan teks pada lingkup KKP;
3. Menambah referensi perbandingan performa teknik WE untuk representasi teks ke dalam bentuk vektor dan algoritma ML untuk klasifikasi berdasarkan vektor hasil reduksi dimensi dari AE.

1.5.Batasan Masalah

Batasan masalah pada tesis ini mencakup:

1. Perbandingan hanya dilakukan pada tiga jenis teknik WE yaitu Word2Vec, GloVe dan FastText;
2. Teknik reduksi dimensi yang digunakan adalah AE dengan target keluaran berupa 16 dimensi;
3. Algoritma ML yang digunakan untuk klasifikasi hanya menggunakan enam algoritma yaitu LR, DT, RF, SVM, KNN dan NB;
4. Data Temuan yang digunakan merupakan periode hasil inspeksi tahun Januari 2019 s.d September 2022.

1.6.Sistematika Pembahasan

Tesis ini disusun dengan sistematika sebagai berikut: bagian pertama berisi pendahuluan, selanjutnya bagian kedua terkait landasan pustaka, kemudian bagian ketiga dijelaskan metode penelitian dan data yang akan digunakan, dilanjutkan bagian empat tentang hasil pelaksanaan metode, pada bagian lima tentang pembahasan mengenai pemahaman baru yang diperoleh dari penelitian, dan pada bagian akhir yaitu bagian enam adalah penutup berisi kesimpulan dan saran.

BAB 2 LANDASAN KEPUSTAKAAN

2.1. Inspektur Mutu Hasil Perikanan

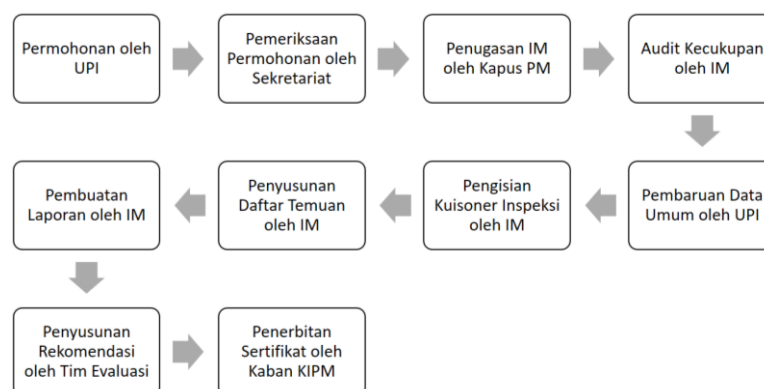
Inspektur Mutu Hasil Perikanan adalah Pegawai Negeri Sipil yang diberi tugas, tanggung jawab, wewenang, dan hak secara penuh oleh Pejabat yang Berwenang untuk melakukan kegiatan pengendalian mutu dan keamanan hasil perikanan pada KKP, sesuai dengan ketentuan peraturan perundang-undangan. Pengendalian Mutu dan Keamanan Hasil Perikanan adalah semua kegiatan yang meliputi inspeksi, verifikasi, surveilan, audit, dan pengambilan contoh dalam rangka memberikan jaminan mutu dan keamanan hasil perikanan (Permen PAN RB No 9/2018 Tentang Jafung Inspektur Mutu Hasil Perikanan, 2018). IM berkedudukan di kantor pusat dan di daerah pada Unit Pelaksana Teknis (UPT), sehingga pada setiap provinsi terdapat IM dan bertanggung jawab kepada Kepala BKIPM (SK Kaban KIPM No 13, 2020).

2.2. Temuan Inspeksi Mutu Hasil Perikanan

Temuan inspeksi mutu hasil perikanan merupakan salah satu hasil keluaran dari proses penerbitan Sertifikat HACCP yang meliputi beberapa tahap (Gambar 2.1). Temuan disusun ke dalam Laporan IM (Gambar 2.2a). Pada laporan tersebut, IM harus menentukan satu kriteria dari empat kriteria yaitu kritis, serius, mayor dan minor untuk setiap Temuan.

Selanjutnya, laporan tersebut dievaluasi oleh Tim yang dibentuk Kepala BKIPM. Tim melakukan pemeriksaan terhadap kesesuaian kriteria dan Temuan pada Laporan IM, jika Tim banyak menemukan Temuan dengan kriteria yang sesuai dengan ketentuan, maka kemampuan IM untuk menentukan kriteria secara tepat sudah baik dan sebaliknya.

Hasil pemeriksaan dari Tim berupa penetapan kesesuaian Temuan dengan kriteria sesuai ketentuan yang berlaku dan disusun ke dalam Form Rekomendasi Hasil Inspeksi (Gambar 2.2b). Kemudian, berdasarkan form tersebut Kepala BKIPM menerbitkan Sertifikat HACCP (Gambar 2.3) (Permen KP No. 51, 2018). Semua tahapan tersebut dalam implementasinya, mulai dari IM Pusat, IM UPT sampai dengan pihak UPI telah menggunakan aplikasi Honest yang berbasis web dengan alamat situs di haccp.bkipm.kkp.go.id (Gambar 2.4), baik untuk proses data ataupun menghasilkan keluaran berupa dokumen.



Gambar 2.1 Proses Penerbitan Sertifikat HACCP (Permen KP No 51/2018 Tentang Persyaratan & Tata Cara Penerbitan Sertifikat Penerapan HACCP, 2018)

KEMENTERIAN KELAUTAN DAN PERIKANAN
BADAN KARANTINA IKAN PENGENDALIAN MUTU DAN KEAMANAN HASIL PERIKANAN
OTORITAS KOMPETEN

Jl. Merdeka Timur No. 16, Jakarta 10115, Telp. (021) 3618076(Luas), Faks. (021) 3600148, Kotak Pos 4135, JRP 10041

DAFTAR TEMUAN KETIDAKSESUAIAN (NON-CONFORMITIES)
LAPORAN INSPEKTUR MUTU

Nama UPI				Status UPI	LAMA
Alamat					
Telepon		Faks			
Jenis Produk	1. Shrimp Powder			Naik Grade	
	Receiving, Processing, Packing/Labeling, Storing, Stuffing				
	2. Shrimp Extract			Naik Grade	
	Receiving, Processing, Packing/Labeling, Storing, Stuffing				
Pimpinan UPI					
Tim Inspeksi					
Ketua	Agung Santoso S.Si., M.P			345/Insp/06	
Anggota	Nugroho Ari Cahyono S.St.Pi., M.Eng			769/Insp/15	
Tanggal Inspeksi	01 April 2022				
Ketidaksesuaian	Kritis = 0	Serius = 0	Mayor = 5	Minor = 2	

No	Temuan Ketidaksesuaian (Problem, Location, Objective, Reference)	Acuan	Kriteria
A. Pelaksanaan GMP-SSOP			
1.	Pelaksanaan pemberian keterangan persetujuan pemasok bahan baku (Approved supplier) berdasarkan kategori Low Risk melalui Penilaian Mandiri Supplier dan High Risk melalui proses re-audit, namun belum ada faktor apa saja yang dapat menentukan kategori Low Risk atau High Risk. Surat keterangan approved supplier tidak ada tanggal dan jangka masa berlakunya sehingga tidak terupdate	Permen KP No. 10 Tahun 2021	Minor
2.	Pelaksanaan kebersihan dan Kesehatan karyawan : a. Pada ruang masuk karyawan ke ruang proses ditetapkan prosedur penggunaan Roll untuk membersihkan pakaian karyawan, namun dalam pelaksanaannya belum ada petugas khusus yang menangani pelaksanaan Roll tersebut sehingga pelaksanaannya masih dilakukan secara mandiri oleh karyawan yang masuk keruang proses b. Form pengaturan personel mencantumkan kondisi kebersihan dan pemakaian APD dan perhiasan, namun pada Form tersebut belum dijelaskan maksud dari perhiasan tersebut sehingga dapat menyebabkan perbedaan dalam penafsiran maksud tujuan form tersebut c. UPI sudah memiliki SOP pengendalian Covid-19 dan menetapkan tim gugus tugas. Namun tim tersebut belum diperbaharui karena terdapat salah satu penanggung jawab yang sudah tidak bekerja di	Permen KP No. 10 Tahun 2021	Mayor

(a)

KEMENTERIAN KELAUTAN DAN PERIKANAN
BADAN KARANTINA IKAN PENGENDALIAN MUTU DAN KEAMANAN HASIL PERIKANAN
OTORITAS KOMPETEN

Jl. Merdeka Timur No. 16, Jakarta 10115, Telp. (021) 3618076(Luas), Faks. (021) 3600148, Kotak Pos 4135, JRP 10041

FORM REKOMENDASI HASIL INSPEKSI

Nama UPI				Status UPI	LAMA
Alamat					
No. Register	-				
Ketua	Agung Santoso S.Si., M.P			345/Insp/06	
Anggota	Nugroho Ari Cahyono S.St.Pi., M.Eng			769/Insp/15	
UPT Pembina	Balai KIPM Jakarta II				
PJ Pusat	Christien Natalia Therik S.St.Pi., M.Si				
Tgl. Inspeksi	01 April 2022				
Tgl. Rencana Perbaikan	01 Mei 2022				

No	Ruang Lingkup	Jenis Inspeksi	Nilai Sebelumnya	Kriteria Ketidaksesuaian				Rekom. Koor- dinator
				Mn	My	Sr	Kr	
1	Shrimp Powder	Naik Grade	B	2	5	0	0	A
2	Shrimp Extract	Naik Grade	B	2	5	0	0	A

No.	Nama	Catatan Tim Teknis	Tgl	Tanda Tangan
1	Anggun Ratnawulan	Dapat diproses lebih lanjut	04-04-22	
2	Anis Sasono	Dapat diproses lebih lanjut	04-04-22	
3	Dede Ratnasari	NC B2 dapat dipertimbangkan kriteria minor Dapat diproses lebih lanjut	04-04-22	

Tgl: 05-04-22

Tanda Tangan:

Hendami Mulyani

Ketua Tim Teknis,

Widodo Sumiyanto

Pengarah,

Hari Maryadi

(b)

Gambar 2.2 (a) Laporan IM (b) Form Rekomendasi Hasil Inspeksi

KEMENTERIAN KELAUTAN DAN PERIKANAN
MINISTRY OF MARINE AFFAIRS AND FISHERIES
REPUBLIK INDONESIA
REPUBLIC OF INDONESIA

BADAN KARANTINA IKAN PENGENDALIAN MUTU DAN KEAMANAN HASIL PERIKANAN
FISH QUARANTINE AND INSPECTION AGENCY (QIA)

SERTIFIKAT
CERTIFICATE

PENERAPAN PROGRAM MANAJEMEN MUTU TERPADU BERDASARKAN KONSEP HACCP
IMPLEMENTATION OF INTEGRATED QUALITY MANAGEMENT PROGRAMME BASED ON HACCP CONCEPT

No. _____

Berdasarkan Peraturan Pemerintah Nomor 57 Tahun 2015 tentang Sistem Jaminan Mutu dan Keamanan Hasil Perikanan serta Peningkatan Nilai Tambah Produk Hasil Perikanan
Having regard to the Government Regulation No. 57 of 2015 laying down Quality and Safety Assurance System and Value Added Development of Fishery Products

Menyatakan bahwa:

To Certify that:

Unit Pemrosesan Ikan
Fish Processing Plant

Alamat
Address

Jenis Produk
Type of Product

Tahapan Pemrosesan
Processing Steps

Estimasi
Rate

Tanggal Inspeksi
Date of Inspection

Unit Pengolahan Ikan ini telah menerapkan dan memenuhi persyaratan Sistem Jaminan Mutu dan Keamanan Hasil Perikanan sesuai dengan ketentuan peraturan perundang-undangan
The Establishment has effectively implemented and fulfilled The Requirements of Quality and Safety Assurance System in accordance with prevailing laws and regulations

Dikeluarkan di
Issued in

Tanggal
Date

Berlaku sampai dengan
Valid until

Diketahui dan
Known and

Ditandatangani
Signed

Ir. Hari Nugroho M.Si
Plt. Kepala Badan Karantina Ikan, Pengendalian Mutu dan Keamanan Hasil Perikanan
Acting Director General of Fish Quarantine and Inspection Agency

Gambar 2.3 Sertifikat HACCP



Gambar 2.4 Aplikasi Honest

2.3.Text-Preprocessing

Text-Preprocessing adalah proses mengubah data tidak terstruktur menjadi data terstruktur sesuai kebutuhan untuk proses *text mining*, seperti *sentiment analysis*, *summarizes*, *document groupings*, dan sebagainya (Jatnika et al., 2019). Dataset yang tidak terstruktur dengan baik perlu *Text-Preprocessing* (Nugroho, Bachtiar, et al., 2022).

Tahap *Text-Preprocessing* terdiri dari *cleaning*, *removing stopwords*, dan *case folding* (Awalina et al., 2022). Dalam proses *cleaning* menghapus kata-kata yang bukan termasuk dalam kata alfabet maupun angka, sedangkan *remove stopwords* adalah menghapus beberapa kata yang dianggap kurang penting dalam deteksi, dan *case folding* berupa pengubahan seluruh karakter huruf menjadi huruf kecil (Awalina et al., 2022).

2.4.Word Embedding

Secara umum, *word embedding* (WE) adalah istilah yang digunakan untuk representasi kata untuk analisis teks, biasanya dalam bentuk vektor bernilai nyata yang mengkodekan makna kata sedemikian rupa sehingga setiap kata yang lebih dekat dalam ruang vektor diharapkan memiliki makna yang serupa (Kanakaris et al., 2022). Menerapkan WE dapat meningkatkan efisiensi dalam pengambilan informasi yang bermanfaat (Khatua et al., 2019). WE merupakan topik yang masih menjadi perhatian dalam penelitian dan industri NLP karena kemampuannya untuk menghasilkan ruang vektor dengan mempertahankan pengetahuan bahasa dan dapat digunakan kembali pada bagian hilir suatu sistem (Mikolov et al., 2017).

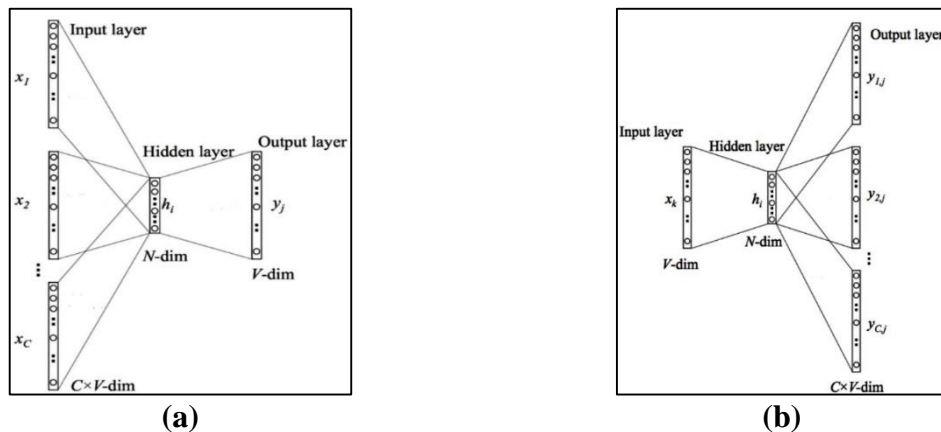
WE salah satu aplikasi *unsupervised learning* menggunakan *Deep Neural Network* serta dapat dibuat langsung dari dataset yang dimiliki atau menggunakan *pre-trained* WE yang telah tersedia (Nurdin et al., 2020). Sudah banyak teknik WE yang dibahas pada berbagai literatur,

dan yang paling banyak digunakan yaitu Word2Vec, GloVe, serta FastText, semua teknik tersebut telah dimanfaatkan secara luas dalam sejumlah besar tugas NLP yang khas, seperti *text classification* dan *sentiment analysis*, atau bahkan dalam tugas yang lebih kompleks, seperti *spam detection* dan *question-answering* (Kanakaris et al., 2022).

2.5. Word2Vec

Word2vec adalah seperangkat algoritma untuk menghasilkan WE berupa vektor numerik, ide utamanya adalah menggunakan konteks kata yang berdekatan dan mengidentifikasi kata serupa berdasarkan representasi mereka dalam ruang vektor, konsep ini sangat berguna untuk menjelajahi dokumen dan mengidentifikasi konten (Sabri et al., 2021). Word2Vec akan menghasilkan vektor yang mirip pada kata yang memiliki makna yang mirip (Juwiantho et al., 2020).

Terdapat dua algoritma Word2vec yaitu *Continuous Bag-of-Word* (CBOW) dan *Skip-Gram*. CBOW menggunakan konteks untuk memprediksi target kata. CBOW memiliki waktu training lebih cepat dan memiliki akurasi yang sedikit lebih baik untuk frequent words, sedangkan Skip-Gram menggunakan sebuah kata untuk memprediksi target konteks dan dapat bekerja dengan baik pada data pelatihan yang jumlahnya sedikit dan dapat merepresentasikan kata-kata yang dianggap langka (Nurdin et al., 2020). Arsitektur CBOW dan Skip-Gram bisa dilihat pada Gambar 2.1. Pada penelitian ini menggunakan algoritma Skip-Gram.



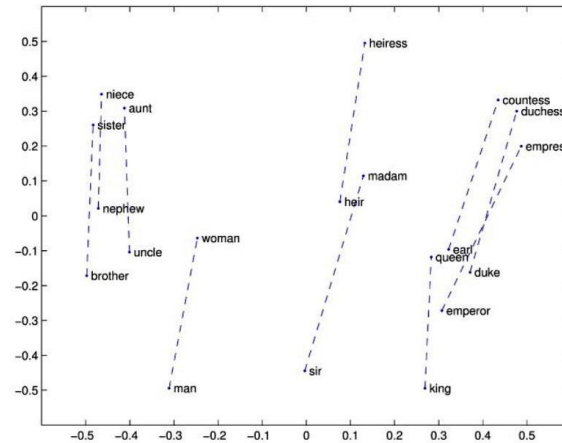
Gambar 2.5 (a) CBOW (b) Skip-Gram (Nawangarsi et al., 2019)

2.6. GloVe

GloVe mempelajari hubungan kata-kata dengan menghitung seberapa sering kata-kata muncul bersama satu sama lain dalam sebuah korpus yang diberikan, berbeda dengan word2vec yang hanya mengandalkan informasi lokal dari kata dengan *local context window* (CBOW dan Skip-gram), algoritma GloVe juga menggabungkan informasi *co-occurrence* kata atau statistik global untuk memperoleh hubungan semantik antara kata dalam korpus (Nurdin et al., 2020). GloVe merupakan pendekatan *unsupervised learning* untuk representasi kata, dan fokus utama penerapan GloVe adalah untuk memaksimalkan probabilitas konteks dari sebuah kata yang diperoleh dari korpus (Singh et al., 2021).

Word2vec adalah model *feedforward neural network* sehingga sering disebut sebagai *neural word embeddings*, sedangkan GloVe adalah model *log-bilinear* atau secara sederhana dapat disebut sebagai model berbasis hitungan, Glove mempelajari hubungan kata-kata dengan

menghitung seberapa sering kata-kata muncul bersama satu sama lain dalam sebuah korpus yang diberikan, rasio probabilitas kemunculan kata-kata memiliki potensi untuk mengkodekan beberapa bentuk makna serta membantu meningkatkan kinerja pada permasalahan (Nurdin et al., 2020). Hasil embedding GloVe mampu menunjukkan substruktur linier yang menarik antara kata dalam ruang vektor, sebagai contoh seperti pada Gambar 2.6 (Chawla, 2018).



Gambar 2.6 Substruktur Linier Hasil Embedding GloVe (Chawla, 2018)

2.7. FastText

FastText adalah metode WE yang merupakan pengembangan dari Word2Vec, mempelajari representasi kata dengan mempertimbangkan informasi subword, sehingga setiap kata direpresentasikan sebagai sekumpulan karakter n-gram, dengan demikian dapat membantu menangkap arti kata-kata yang lebih pendek dan memungkinkan *embedding* untuk memahami sufiks dan prefiks dari kata (Nurdin et al., 2020).

FastText ditujukan untuk mempelajari representasi kata-kata berkualitas tinggi sambil mempertimbangkan morfologinya (Ait Hammou et al., 2020), sehingga kata-kata yang paling mirip maka terkait juga secara morfologis (Kirschenbaum, 2022). Selain itu FastText merupakan *one-hidden layer neural network* dengan dua fitur utama yaitu untuk representasi teks dan klasifikasi teks (Sreelakshmi et al., 2020), serta mampu menangani *out of vocabulary*, karena fastText dapat memprediksi kata-kata yang tidak ada dalam kamus kosakata (Atikah et al., 2022).

FastText memiliki kinerja yang baik, dapat melatih model pada dataset yang besar dengan cepat dan dapat memberikan representasi kata yang tidak muncul dalam data latih. Jika kata tidak muncul selama pelatihan model, kata tersebut dapat dipecah menjadi n-gram untuk mendapatkan embedding vektornya (Nurdin et al., 2020). FastText mewakili setiap kata sebagai karakter n-gram, misalnya jika $n = 3$ dalam kata “artificial”, maka FastText akan merepresentasikan kata ini seperti dalam bentuk tanda kurung siku berikut <ar, art, rti, tif, ifi, fic, ici, ial, al> (Rahmadzani, 2021), contoh lain bisa dilihat pada Tabel 2.1.

Tabel 2.1 Penerapan N-Grams pada FastText (Chaudhary, 2020)

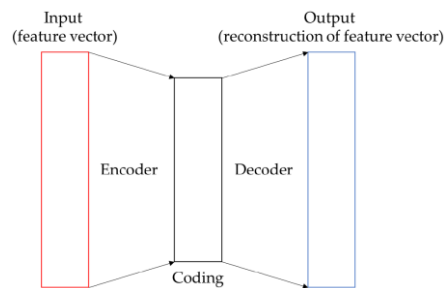
Word	Length(n)	Character n-grams
eating	3	<ea, eat, ati, tin, ing, ng>
eating	4	<eat, eati, atin, ting, ing>
eating	5	<eati, eatin, ating, ting>
eating	6	<eatin, eating, ating>

2.8.Autoencoder

Autoencoder (AE) yaitu neural network yang dapat merepresentasikan data kemudian merekonstruksinya kembali, ide utama AE yaitu mengaproksimasi/mengompresi data asli menjadi bentuk lebih kecil (*coding*), kemudian operasi pada bentuk *coding* merepresentasikan operasi pada data sebenarnya, AE terdiri dari *encoder* (sebuah *neural network*) dan *decoder* (sebuah *neural network*), *encoder* merubah input ke dalam bentuk dimensi lebih kecil (dapat dianggap sebagai kompresi), sedangkan *decoder* berusaha merekonstruksi *coding* menjadi bentuk aslinya, secara matematis, kita dapat menulis autoencoder sebagai persamaan 2.1, dimana \mathbf{d} melambangkan *decoder*, \mathbf{enc} melambangkan *encoder* dan \mathbf{x} adalah input, *encoder* diberikan pada persamaan 2.2 yang berarti melewati input pada suatu layer di *neural network* untuk menghasilkan representasi \mathbf{x} berdimensi rendah, disebut *coding* \mathbf{c} , \mathbf{U} dan α melambangkan weight matrix dan bias (Putra, 2019). Ilustrasi AE sederhana dapat dilihat pada Gambar 2.7.

$$f(\mathbf{d}, \theta) = \text{dec}(\text{enc}(\mathbf{x})) \quad (2.1)$$

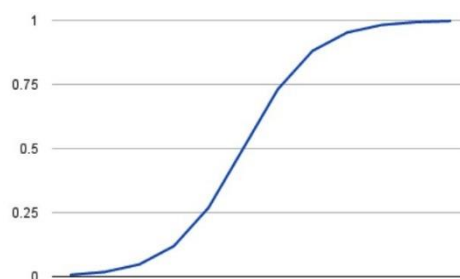
$$\mathbf{c} = \text{enc}(\mathbf{x}) = g(\mathbf{x}, \mathbf{U}, \alpha) \quad (2.2)$$



Gambar 2.7 Autoencoder Sederhana (Putra, 2019)

2.9.Logistic Regression

Logistic Regression (LR) adalah salah satu teknik bidang statistik yang digunakan oleh ML. LR digunakan pada permasalahan klasifikasi biner (masalah dengan dua nilai kelas). LR dinamakan berdasarkan fungsi utama yang menjadi inti dari metode LR yaitu *logistic function* atau juga disebut sebagai *sigmoid function*, dikembangkan oleh ahli statistik untuk mendeskripsikan properti pertumbuhan penduduk dalam ekologi yang meningkat dengan cepat dan menghabiskan daya dukung lingkungan, deskripsi tersebut dalam bentuk kurva berbentuk S (Gambar 2.8) yang dapat menggunakan semua bilangan real dan memetakannya menjadi nilai antara 0 dan 1, tetapi tidak pernah tepat pada batas tersebut (Brownlee, 2020b).



Gambar 2.8 Logistic Function (Brownlee, 2020b)

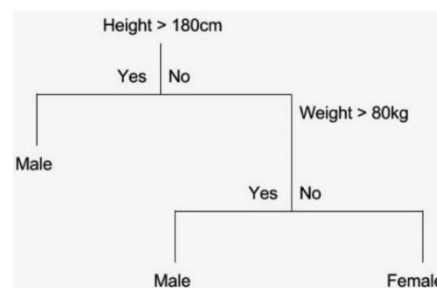
Pada awalnya LR tidak dapat digunakan untuk tugas klasifikasi yang memiliki lebih dari dua label kelas atau lebih dikenal sebagai *multi-class classification*, salah satu pendekatan populer untuk mengadaptasi LR ke masalah tersebut adalah dengan membagi *multi-class classification problems* menjadi *multiple binary classification problems* dan menyesuaikan model standar LR pada setiap *subproblem*, sehingga LR mampu melakukan prediksi untuk *multi-class classification* secara langsung (Brownlee, 2020b).

LR dibagi menjadi tiga jenis, pertama binary yaitu LR yang memberikan keluaran hanya berupa dua kategori, kedua multinomial yaitu LR yang menghasilkan keluaran tiga atau lebih kategori tanpa berurutan, dan ordinal yaitu LR yang sama dengan multinomial hanya kategori berurutan (Swaminathan, 2018). Pada penelitian ini, penulis menggunakan jenis multinomial untuk melakukan klasifikasi.

2.10. Decision Tree

Decision Tree (DT) adalah alat pendukung keputusan yang menggunakan grafik atau model seperti pohon untuk menunjukkan keputusannya dan kemungkinan hasil yang digunakan untuk menentukan jalur yang harus diikuti dalam analisis keputusan (Koklu & Ozkan, 2020). DT membuat cabang pohon melalui pendekatan hierarki dan setiap cabang dapat dianggap sebagai pernyataan if-else, cabang tersebut dikembangkan dengan mempartisi dataset menjadi banyak subset berdasarkan fitur yang paling penting, kemudian klasifikasi akhir akan ditemukan pada daun dari DT tersebut (Gong, 2022).

Salah satu algoritma DT yang sederhana tapi mutakhir adalah *Classification And Regression Trees* (CART), yang direpresentasikan ke dalam bentuk *binary tree* seperti pada Gambar 2.9, setiap simpul akar mewakili satu variabel input (x) dan sebuah titik pemisah pada variabel tersebut (dengan asumsi variabel adalah numerik), pada setiap daun dari pohon merupakan variabel keluaran (y) yang digunakan untuk membuat prediksi (Brownlee, 2020a).

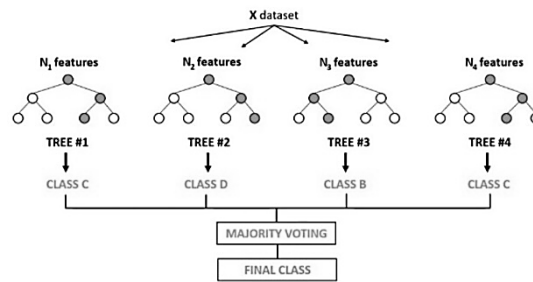


Gambar 2.9 Decision Tree (Brownlee, 2020a)

2.11. Random Forest

Random Forest (RF) pada dasarnya hanyalah sekumpulan DT (Neogi et al., 2021). RF merupakan salah satu pengklasifikasi terbaik yang banyak digunakan untuk tugas regresi dan klasifikasi, kesederhanaan algoritma menjadikannya pilihan yang menarik untuk klasifikasi teks, RF menggunakan sejumlah besar DT untuk pengambilan keputusan, untuk membuat keputusan akhir, menggunakan rata-rata atau rata-rata dari keluaran DT, sehingga memberikan hasil yang lebih akurat dibandingkan dengan DT (Jalal et al., 2022).

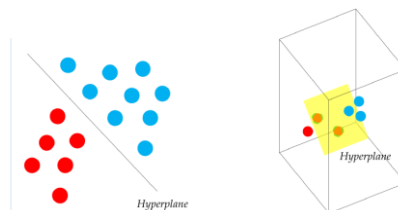
Ide RF adalah sebagai berikut, ambil banyak DT (menggunakan beberapa keacakan, jadi *tree*-nya berbeda) dan biarkan DT tersebut memilih, sehingga keputusan klasifikasi yang diambil oleh RF adalah keputusan yang diambil oleh kelompok DT yang paling banyak dalam kumpulan *tree* yang acak (Słowiński, 2021). Algoritma RF membangun banyak DT dari dataset yang diberikan (Nazeer et al., 2020) seperti yang ditunjukkan pada Gambar 2.10.



Gambar 2.10 Random Forest (Sumiran, 2022)

2.12. Support Vector Machine

Support Vector Machine (SVM) salah satu metode dalam *supervised learning* yang biasa digunakan untuk klasifikasi (*Support Vector Classification*) dan regresi (*Support Vector Regression*), SVM digunakan untuk mencari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas, *hyperplane* adalah fungsi yang dapat digunakan untuk memisahkan kelas, sebagai perbandingan, jika fungsi 2-D yang digunakan untuk kelas, maka klasifikasi disebut sebagai garis, sedangkan jika fungsi yang digunakan untuk klasifikasi kelas dalam 3-D, maka disebut bidang yang sama, sehingga fungsi yang digunakan untuk klasifikasi di ruang kelas dengan dimensi yang lebih tinggi, maka disebut *hyperplanes* (Gambar 2.11), kemudian objek data terluar yang paling dekat dengan *hyperplane* disebut *support vector*, yang merupakan objek paling sulit untuk diklasifikasikan karena posisinya yang hampir *overlap* dengan kelas lain (Suyoto et al., 2021).



Gambar 2.11 SVM Hyperplane (Putra, 2020)

SVM mendeskripsikan *hyperplane* dengan mengubah data didukung fungsi matematika yang disebut "Kernel", yang terdiri dari beberapa jenis yaitu *linier*, *sigmoid*, *RBF*, *non-linear*, *polinomial*, dan sebagainya, Kernel — "RBF" untuk masalah non-linier dan juga merupakan kernel umum yang digunakan ketika tidak ada pengetahuan sebelumnya tentang data, sedangkan Kernel — "linear" adalah untuk masalah yang dapat diselesaikan secara linier (Reddy, 2018).

SVM berada pada level untuk mewakili masalah yang kompleks dan tahan terhadap *overfitting*, walaupun awalnya dirancang untuk klasifikasi data linier kelas biner, metode ini kemudian dikembangkan untuk klasifikasi data kelas ganda dan non-linier, pengklasifikasi biner digunakan untuk menggeneralisasikan masalah kelas ganda dengan dua skema dasar: satu lawan satu dan satu lawan semua (Koklu & Ozkan, 2020).

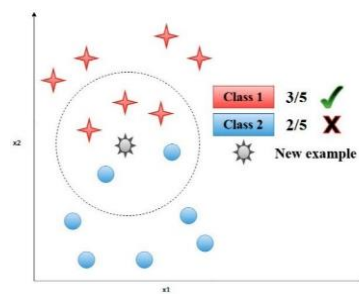
2.13. K-Nearest Neighbor

K-Nearest Neighbor (KNN) adalah pengklasifikasi dengan pendekatan non-parametrik, yang mengklasifikasikan titik data yang diberikan sesuai dengan mayoritas tetangganya, algoritma KNN menyelesaikan eksekusinya dalam dua langkah, pertama menemukan jumlah tetangga terdekat, dan kedua mengklasifikasikan titik data ke dalam kelas tertentu

menggunakan langkah pertama, untuk menemukan tetangga menggunakan metrik jarak seperti *euclidean distance* seperti yang diberikan dalam persamaan 2.3 (Bablani et al., 2018).

$$Distance(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (2.3)$$

KNN memilih **k** sampel terdekat dari set pelatihan, kemudian mengambil suara mayoritas kelas mereka di mana **k** harus menjadi bilangan ganjil untuk menghindari ambiguitas, pada Gambar 2.12 mengilustrasikan arsitektur klasifikasi KNN, terdapat 2 *class* yaitu *class* 1 dan *class* 2, tanda bintang merah menunjukkan *class* 1 dan lingkaran biru menunjukkan *class* 2, **k** yang dipilih adalah 5, dan diantara 5 tetangga terdekat, 3 sampel termasuk *class* 1 dan 2 sampel termasuk *class* 2, pengklasifikasi KNN bekerja berdasarkan prinsip memberikan sampel baru ke *class* dengan suara terbanyak di **k** yang ditentukan, jadi sampel baru ditugaskan ke *class* 1 (Bablani et al., 2018).



Gambar 2.12 Klasifikasi KNN (Bablani et al., 2018)

2.14. Naive Bayes

Naive Bayes (NB) adalah sebuah pengklasifikasi yang menggunakan teorema Bayes, NB memprediksi probabilitas keanggotaan untuk setiap kelas seperti probabilitas bahwa suatu data yang diberikan masuk ke dalam kelas tertentu, kelas dengan probabilitas tertinggi dianggap sebagai kelas yang paling mungkin, NB mengasumsikan bahwa semua fitur tidak berhubungan satu sama lain untuk mempelajari efek individualnya pada umpan balik, ada atau tidak adanya suatu fitur tertentu tidak mempengaruhi ada atau tidak adanya fitur lain, meskipun ada kemungkinan bergantung pada keberadaan fitur lain, semua fitur ini dianggap berkontribusi secara independen terhadap kemungkinan berupa umpan balik yang valid (Maitra et al., 2018). Teorema Bayes dirumuskan seperti pada persamaan 2.4.

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)} \quad (2.4)$$

Dengan menggunakan teorema Bayes, dapat mencari peluang terjadinya A, dengan syarat B telah terjadi, di sini B adalah bukti dan A adalah hipotesis, asumsi yang dibuat di sini adalah bahwa prediktor/fiturnya independen, artinya kehadiran satu fitur tertentu tidak mempengaruhi yang lain, sehingga disebut *naïve* (Gandhi, 2018).

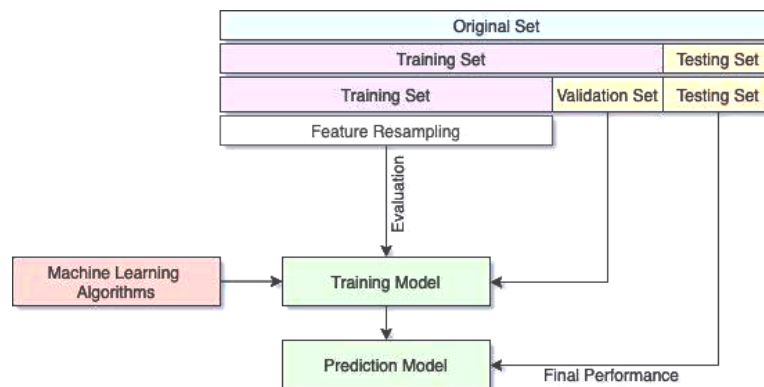
Jenis NB yaitu yang pertama adalah Multinomial yang banyak digunakan untuk masalah klasifikasi dokumen dengan fitur/prediktor adalah frekuensi kata yang ada dalam dokumen, kemudian jenis kedua adalah Bernoulli mirip dengan Multinomial tapi prediktornya adalah variabel boolean dengan hanya mengambil nilai ya atau tidak, dan jenis terakhir adalah Gaussian yang prediktor mengambil nilai kontinu dan tidak diskrit, dengan asumsi bahwa nilai-nilai tersebut diambil sampelnya dari distribusi gaussian (Gandhi, 2018).

2.15. Stratified K-Fold

Saat melakukan prosedur pemisahan data, masalah utama adalah harus dilakukan pembagian yang cukup pada dataset untuk *training set* dan *test set* sebagai representasi domain permasalahan (Nugroho, Sukmadewa, et al., 2022), kondisi ini bisa terjadi jika dataset berjumlah sedikit atau *unbalanced class distribution*.

Stratified K-Fold, seperti yang ditunjukkan pada, merupakan prosedur evaluasi yang sesuai jika dataset mempunyai permasalahan tersebut, sehingga mampu melakukan evaluasi akhir dari kinerja model yang diimplementasikan, setelah memisahkan dataset untuk *training set* dan *test set*, selanjutnya adalah membagi *training set* menjadi *validation set* untuk memvalidasi kinerja algoritma ML selama proses iterasi *K-Fold* (Nugroho, Sukmadewa, et al., 2022).

Pada penelitian ini digunakan *Stratified K-Fold* dengan $K = 10$ untuk mencegah *overfitting* (Nugroho, Bachtiar, et al., 2022), karena kondisi *unbalanced class distribution* pada dataset dengan salah satu *class* berjumlah diatas 5,000 sedangkan kelas lain berada diantara 1000 sampai dengan 1500.



Gambar 2.13 Stratified K-Fold (Nugroho, Sukmadewa, et al., 2022)

2.16. Evaluasi

Dalam klasifikasi, *confusion matrix* (CM) menggambarkan kinerja model dengan menghitung kelas mana yang diprediksi dengan benar dan salah dan jenis kesalahan apa yang dibuat, *true positive* (TP) didefinisikan sebagai data positif yang diprediksi benar, *true negative* (TN) didefinisikan sebagai data negatif yang diprediksi benar, *false positive* (FP) adalah data negatif yang diprediksi sebagai data positif, dan *false negative* (FN) adalah data positif yang diprediksi sebagai data negatif, selanjutnya untuk matrik performa yang paling sering digunakan berdasarkan CM untuk klasifikasi adalah *Accuracy* (Nugroho, Sukmadewa, et al., 2022).

Accuracy adalah rasio prediksi *true* (TP dan TN) dengan keseluruhan data yang menggambarkan tingkat kedekatan nilai prediksi dengan nilai sebenarnya, seperti yang ditunjukkan pada (2.5), pada proses pelatihan, *accuracy* diperoleh dari rata-rata setiap *accuracy* dari *fold* pada *cross-validation*, selain itu *standard deviation* juga dihitung untuk mengetahui variannya (Nugroho, Sukmadewa, et al., 2022).

$$Accuracy = \frac{TP+TN}{TP + FP+FN+TN} \quad (2.5)$$

Masalah yang bisa muncul dengan kondisi *unbalanced data* adalah data negatif menjadi kelas mayoritas dan data positif menjadi kelas yang lebih sedikit, sehingga untuk menginterpretasikan performansi model dengan data yang tidak seimbang, digunakan kurva *receiver operating characteristic* (ROC). Kurva ROC diperoleh dari *true positive rate* (TPR) seperti pada (2.6) dan *false positive rate* (FPR) seperti pada (2.7) (Nugroho, Sukmadewa, et al., 2022).

$$TPR = \frac{TP}{TP + FN} \quad (2.6)$$

$$FPR = \frac{TN}{TN + FP} \quad (2.7)$$

2.17. Penelitian Sebelumnya

Hingga saat ini penelitian tentang perbandingan akurasi antara WE telah banyak dilakukan. Pada penelitian Analisis Perbandingan Representasi WE dalam Kesastraan Bengali (Ahmed Chowdhury et al., 2018) dengan mengambil kasus *Authorship Attribution* yang memanfaatkan dataset dari 2,400 artikel blog berbahasa Bengali dari enam penulis, peneliti melakukan perbandingan teknik WE antara Word2Vec, FastText dan GloVe yang dikombinasikan dengan tiga jenis *neural network* yaitu *Convolutional Neural Network* (CNN), *Classical Multi-layered Perceptron* (NN), dan *Recurrent Neural Network with Long Short Term Memory* (RNN with LSTM) diperoleh hasil akurasi terbaik secara berurutan adalah FastText, Word2Vec dan GloVe, dengan kombinasi FastText(skip-gram) dan CNN mencapai akurasi 92.9%.

Penelitian tentang Evaluasi Perbandingan Teknik WE untuk Analisis Sentimen Twitter (Kaibi et al., 2019) yang menggunakan dataset Twitter berbahasa Arab untuk melakukan perbandingan akurasi teknik WE antara Word2Vec, FastText dan GloVe, dikombinasikan dengan enam algoritma ML yaitu GaussianNB, LinearSVC, NuSVC, LR, SGD dan RF untuk klasifikasi biner sentimen (positif atau negatif) memberikan hasil bahwa FastText yang dikombinasikan dengan NuSVC adalah yang terbaik mencapai F1-score sebesar 81.97% dibanding Word2Vec dan GloVe yang dikombinasikan dengan berbagai algoritma ML yang diuji.

Kemudian pada penelitian Pendekatan Berbasis WE untuk Identifikasi Keakuratan pada Aktivitas yang Saling Terkait (Shahzad et al., 2019) dalam rangka mencari teknik yang efisien untuk *Process Model Matching* (PMM) dengan menggunakan empat dataset berbahasa inggris berisi model proses dengan masing-masing domain yaitu *University Admissions*, *Birth Registration*, *Asset Management* dan *Domain Independent*, dengan hasil akhir menunjukkan bahwa akurasi FastText lebih baik mencapai F1-score sebesar 84% dibanding GloVe dan Word2Vec.

Pada penelitian lainnya yaitu PMCVec: Representasi Frasa Terdistribusi Untuk Pemrosesan Teks Biomedis (Gero & Ho, 2019), peneliti melakukan perbandingan antara Word2Vec, GloVe dan FastText untuk menemukan teknik WE terbaik yang bisa digunakan pada metode PMCVec dalam rangka memperoleh makna frasa pada teks biomedis dengan menggunakan lima dataset yang mengandung istilah medis dalam bahasa inggris serta mempunyai kesamaan dan keterkaitan antara kata, hasil penelitian tersebut menyimpulkan bahwa Word2Vec (CBOW) dengan *average similiarity score* mencapai 65% berkinerja lebih baik dibanding GloVe dan FastText.

Selain itu pada penelitian Perbandingan Kinerja WE Word2Vec, GloVe, dan FastText Pada Klasifikasi Teks (Nurdin et al., 2020) dengan menggunakan dataset dari 20 *newsgroup* dan *Reuters Newswire Topic Classification* dalam bahasa inggris yang diklasifikasikan dengan

algoritma *Convolutional Neural Network* (CNN), memberikan hasil bahwa FastText dengan F1-score sebesar 97.9% lebih unggul dibanding Word2Vec dan GloVe.

Selanjutnya pada penelitian Kategorisasi Dokumen Teks Bengali Berdasarkan *Very Deep Convolution Neural Network* (Hossain et al., 2021) yang menggunakan dataset yang berasal dari berbagai sumber seperti online newspapers, blogs dan e-book berbahasa Bengali, membuktikan bahwa GloVe mempunyai akurasi yang lebih baik dibanding dengan Word2Vec dan FastText dengan nilai 96.96%.

Kemudian pada penelitian Perbandingan *Pretrained Model Transformer* pada Deteksi Ulasan Palsu (Awalina et al., 2022) dengan dataset berasal dari 1,600 ulasan 20 hotel di Chicago, pada perbandingan Pretrained WE dengan dukungan CNN menunjukkan bahwa GloVe mempunyai performa lebih baik dibanding dengan Word2Vec dan FastText.

Dari beberapa penelitian yang telah melakukan perbandingan antara Word2Vec, GloVe dan FastText tersebut, menunjukkan bahwa satu sama lain bisa lebih unggul pada penelitian tertentu walaupun sering ditemukan FastText lebih unggul dibanding teknik lainnya, perbedaan hasil perbandingan pada setiap penelitian sangat bergantung pada dataset yang digunakan dan permasalahan yang ingin diselesaikan (Nurdin et al., 2020), kemudian dataset yang digunakan berupa bahasa Inggris, Arab dan Bengali, namun belum ada penelitian yang melakukan perbandingan WE secara khusus untuk dataset yang menggunakan bahasa Indonesia, selanjutnya dari penelitian tersebut juga diketahui bahwa belum ada yang menerapkan reduksi dimensi untuk meningkatkan kinerja pada algoritma ML untuk klasifikasi, sehingga penelitian ini mencoba mengisi berbagai ruang kosong tersebut.

BAB 3 METODOLOGI

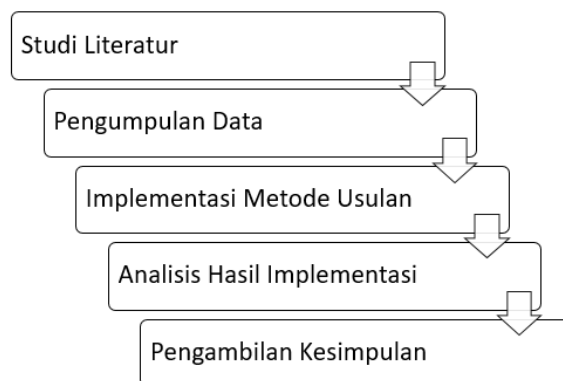
3.1. Tipe Penelitian

Penelitian ini berupa implementatif karena berupa pembangunan suatu model untuk melakukan klasifikasi Temuan yang dibangun berdasarkan hasil perbandingan performa dari kombinasi WE, AE dan ML sehingga diperoleh kombinasi terbaik dalam akurasi untuk melakukan klasifikasi Temuan tersebut.

3.2. Strategi dan Rancangan Penelitian

3.2.1. Strategi secara Umum

Strategi yang digunakan pada penelitian ini meliputi beberapa tahap yaitu studi literatur, pengumpulan data, text-preprocessing, implementasi WE, implementasi AE, implementasi ML, evaluasi model ML, Analisis Hasil Evaluasi, dan Kesimpulan. Secara singkat tahapan penelitian bisa dilihat pada Gambar 3.1.



Gambar 3.1 Tahapan Penelitian

Studi literatur dilakukan dengan mencari dan mempelajari teori Temuan, WE, AE, ML, dan penelitian sebelumnya yang terkait perbandingan WE. Pengumpulan data dilakukan dengan wawancara narasumber yang kompeten dan observasi data dari aplikasi Honest. Implementasi Metode Usulan berupa pelaksanaan berbagai teknik yang telah direncanakan terhadap data yang telah dikumpulkan, dengan target keluaran berupa hasil pengujian dari berbagai model yang berhasil dibuat. Analisis Hasil Implementasi dengan membahas hasil pengujian model. Pengambilan kesimpulan untuk menentukan simpulan akhir yang diperoleh dari penelitian.

3.2.2. Subjek Penelitian

Subjek Penelitian ini mencakup beberapa pejabat struktural/fungsional yang terlibat dalam pembuatan Temuan, baik yang berasal dari unit eselon II BKIPM terkait mutu hasil perikanan yaitu Pusat Pengendalian Mutu (Pusat PM) maupun Unit Pelaksana Teknis (UPT) BKIPM di daerah. Selain itu, pegawai yang tergabung pada Tim Evaluasi yang dibentuk oleh Kepala BKIPM juga menjadi subjek dalam penelitian ini, sebagai pihak yang berwenang untuk memberi rekomendasi kriteria suatu Temuan.

3.2.3. Lokasi Penelitian

Lokasi pada penelitian ini meliputi:

1. Pusat PM, sebagai pihak yang melaksanakan penerbitan Sertifikat Penerapan HACCP, berlokasi di Kantor Pusat Kementerian Kelautan dan Perikanan, Gedung Mina Bahari II, Lantai 10, Jakarta Pusat, DKI Jakarta;
2. Pusdatin, sebagai pihak pengembang aplikasi Honest, berlokasi di Kantor Pusat Kementerian Kelautan dan Perikanan, Gedung Mina Bahari II, Lantai 16, Jakarta Pusat, DKI Jakarta;
3. Laboratorium Sistem Cerdas Fakultas Ilmu Komputer Universitas Brawijaya, sebagai fasilitas untuk melakukan pendalaman WE, AE dan ML, berlokasi di Fakultas Ilmu Komputer (FILKOM), Universitas Brawijaya, Jl. Veteran No. 8 Malang, Jawa Timur.

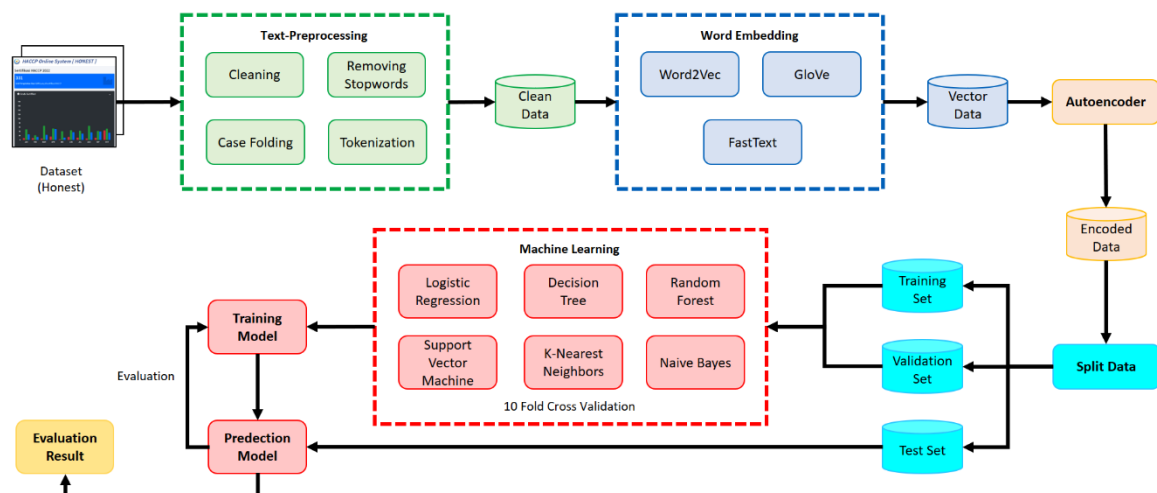
3.2.4. Metode Pengumpulan Data

Metode pengumpulan data pada penelitian ini, yaitu:

1. Wawancara
Wawancara dilakukan pada beberapa IM di Pusat PM dan UPT, sebagai pengguna yang menggunakan sistem untuk menentukan kriteria temuan. Target data yang ingin diperoleh adalah cara IM dalam menentukan kriteria suatu Temuan;
2. Survei
Survei dilakukan terhadap beberapa anggota Tim Evaluasi yang beranggotakan pegawai yang sudah menjadi pakar dalam bidang pengendalian mutu hasil perikanan. Data yang dikumpulkan adalah data terkait penilaian para pakar terhadap IM yang mempunyai kemampuan baik dalam menentukan kriteria Temuan;
3. Observasi
Observasi yaitu mengamati data Laporan IM dengan melakukan akses ke aplikasi Honest. Semua kegiatan inspeksi dilakukan melalui aplikasi tersebut, termasuk membuat Laporan IM, sehingga cara yang efektif dan efisien untuk memperoleh data tersebut adalah observasi berupa eksplorasi data terkait Laporan IM yang berisi Temuan beserta kriterianya yang tersimpan pada aplikasi Honest.

3.2.5. Metode Usulan

Untuk membandingkan performa teknik WE pada klasifikasi Temuan, maka dibuat sebuah metode usulan dalam penelitian ini seperti pada Gambar 3.2, metode tersebut dibagi menjadi beberapa bagian yaitu pembentukan Dataset yang merupakan hasil dari pengumpulan data, kemudian Text-Preprocessing untuk mendapatkan data dalam bentuk token, melakukan proses WE, selanjutnya hasil WE berupa data vektor diproses untuk reduksi dimensi dengan AE, hasilnya berupa vektor yang telah dienkoder, kemudian vektor tersebut dibagi menjadi training set dan validation set untuk training algoritma ML serta test set untuk prediksi, kemudian terakhir dilakukan evaluasi model yang telah dihasilkan. Subbagian berikutnya dijelaskan masing-masing bagian tersebut secara lebih rinci.



Gambar 3.2 Metode Usulan

1. Dataset

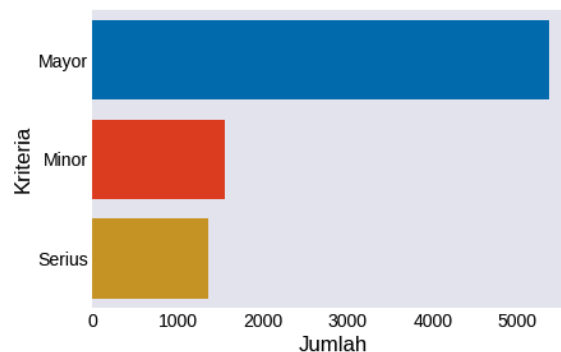
Dataset yang digunakan adalah data Temuan beserta kriterianya yang tersimpan pada database Honest dari periode Januari 2019 s.d September 2022, periode tersebut dipilih karena sesuai dengan mulai berlakunya Permen KP No. 51/2018 yang mulai berlaku 28 Desember 2018. Data Temuan seperti pada Tabel 3.1, merupakan hasil temuan yang dibuat oleh semua IM yang melaksanakan kegiatan inspeksi mutu ke UPI, sehingga perlu dilakukan pemeriksaan untuk memilih data temuan dari IM yang dinilai oleh Tim Evaluasi mampu untuk memberikan kriteria secara benar.

Tabel 3.1 Temuan dan Kriteria pada Database Honest

[illegible]

Berdasarkan wawancara dan survei diketahui ada 32 IM yang dinilai mampu, selanjutnya dilakukan observasi data yang berasal dari IM tersebut dan diperoleh sebanyak 8,290 data, dengan rincian kriteria sebagai berikut: *Serius* = 1,360; *Mayor* = 5,376; *Minor* = 1,554. Kriteria Kritis tidak ditemukan pada data yang terpilih, hal ini disebabkan kriteria tersebut jarang ditemukan pada UPI dan IM lebih mudah untuk menentukan suatu temuan masuk atau tidak ke dalam kriteria Kritis dibanding kriteria lainnya.

Berdasarkan distribusi kriteria, Dataset memiliki proporsi kriteria yang tidak seimbang, seperti yang ditunjukkan pada Gambar 3.3, sebagian besar Temuan masuk dalam kriteria *Mayor*, sedangkan kriteria *Minor* dan *Serius* dengan jumlah yang lebih sedikit. Pada penelitian ini tidak menangani *resampling* dalam menangani Dataset yang tidak seimbang.



Gambar 3.3 Distribusi Kriteria *Unbalanced*

2. *Text-Preprocessing*

Tahapan text-preprocessing pada penelitian ini meliputi:

A. *Cleaning*

Teks temuan yang tersimpan pada database *Honest* mengandung berbagai karakter, diantaranya adalah beberapa sintak HTML, hal ini disebabkan untuk kebutuhan untuk menampilkan data tersebut pada aplikasi *Honest*, sehingga proses *cleaning* harus dilakukan untuk menghapus berbagai jenis karakter yang tidak dibutuhkan sehingga vektor yang dihasilkan hanya berasal dari kata atau kalimat yang merupakan isi dari Temuan saja. Beberapa karakter yang dibersihkan dalam proses ini berupa : `<p>`, `</p>`, ` `, `!`, `#`, `%`, `$` dan sebagainya, contoh proses *cleaning* bisa dilihat pada Tabel 3.2.

Tabel 3.2 Proses *Cleaning*

Teks:
<code><p></code> Pelatihan Karyawan belum optimal, Terdapat Karyawan melakukan swabbing terlebih dahulu baru melakukan penimbangan II untuk produk Tuna Loin, Hal ini tidak sesuai dengan alur yang sudah ditetapkan oleh UPI <code></p></code>
Teks setelah <i>Cleaning</i> :
Pelatihan Karyawan belum optimal Terdapat Karyawan melakukan swabbing terlebih dahulu baru melakukan penimbangan II untuk produk Tuna Loin Hal ini tidak sesuai dengan alur yang sudah ditetapkan oleh UPI

B. *Case Folding*

Tujuan proses ini adalah mengubah semua karakter huruf, dari berbagai jenis huruf menjadi huruf kecil (*lower case*), kondisi tersebut dibutuhkan untuk memudahkan proses selanjutnya karena adanya keseragaman jenis huruf, sehingga untuk setiap kata yang sama susunan hurufnya akan mempunyai komposisi karakter yang sama juga, contoh proses *Case Folding* bisa dilihat pada Tabel 3.3.

Tabel 3.3 Proses *Case Folding*

Teks:
Pelatihan Karyawan belum optimal Terdapat Karyawan melakukan swabbing terlebih dahulu baru melakukan penimbangan II untuk produk Tuna Loin Hal ini tidak sesuai dengan alur yang sudah ditetapkan oleh UPI
Teks setelah <i>Case-Folding</i> :
pelatihan karyawan belum optimal terdapat karyawan melakukan swabbing terlebih dahulu baru melakukan penimbangan ii untuk produk tuna loin hal ini tidak sesuai dengan alur yang sudah ditetapkan oleh upi

C. *Tokenization*

Proses ini sangat penting karena membagi setiap kalimat Temuan menjadi kumpulan kata atau *Corpus*, setelah kalimat sudah diubah dalam bentuk *Corpus* maka proses pemeriksaan atau perubahan per kata pada suatu kalimat menjadi semakin mudah, pada merupakan contoh hasil dari proses tokenization yang merubah kalimat menjadi sekumpulan kata dalam bentuk *Corpus*.

Tabel 3.4 Proses *Tokenization*

Teks:
pelatihan karyawan belum optimal terdapat karyawan melakukan swabbing terlebih dahulu baru melakukan penimbangan ii untuk produk tuna loin hal ini tidak sesuai dengan alur yang sudah ditetapkan oleh upi
Teks setelah <i>Tokenization</i> :
['pelatihan', 'karyawan', 'belum', 'optimal', 'terdapat', 'karyawan', 'melakukan', 'swabbing', 'terlebih', 'dahulu', 'baru', 'melakukan', 'penimbangan', 'ii', 'untuk', 'produk', 'tuna', 'loin', 'hal', 'ini', 'tidak', 'sesuai', 'dengan', 'alur', 'yang', 'sudah', 'ditetapkan', 'oleh', 'upi']

D. *Removing Stopwords*

Proses ini berfungsi untuk menghilangkan kata yang mempunyai fungsi pada suatu kalimat tapi tidak mempunyai makna secara langsung pada kalimat tersebut, seperti: pada, dengan, yang, tersebut dan sebagainya. Pada penelitian ini *stopwords* yang digunakan berasal dari *Tala Stopwords Library* berisi 758 kata. Hasil dari proses ini menjadi hasil akhir dari Text-Preprocessing pada penelitian ini berupa *Clean Data*. Contoh data bisa dilihat pada Tabel 3.5.

Tabel 3.5 Proses *Removing Stopwords*

Teks:
['pelatihan', 'karyawan', 'belum', 'optimal', 'terdapat', 'karyawan', 'melakukan', 'swabbing', 'terlebih', 'dahulu', 'baru', 'melakukan', 'penimbangan', 'ii', 'untuk', 'produk', 'tuna', 'loin', 'hal', 'ini', 'tidak', 'sesuai', 'dengan', 'alur', 'yang', 'sudah', 'ditetapkan', 'oleh', 'upi']
Teks setelah <i>Tokenization</i> :
['pelatihan', 'karyawan', 'optimal', 'karyawan', 'swabbing', 'penimbangan', 'ii', 'produk', 'tuna', 'loin', 'sesuai', 'alur', 'ditetapkan', 'upi']

3. *Word Embedding*

Pada tahap ini mempunyai tujuan utama adalah mengubah kalimat Temuan menjadi vektor dengan memperhatikan kaitan antara kata dalam setiap kalimat. Setiap teknik WE pada penelitian ini digunakan untuk membuat masing-masing model WE berdasarkan hasil pelatihan terhadap semua kata dalam *Clean Data* yang dihasilkan dari *Text-Preprocessing*.

Setiap Model WE yang dihasilkan dari pelatihan tersebut mempunyai kemampuan untuk mengubah kata menjadi vektor, namun untuk mengubah kalimat yang terdiri dari beberapa kata maka digunakan fungsi *mean* setiap vektor kata dari kalimat tersebut, sehingga diperoleh suatu vektor dari kalimat Temuan berupa *Vector Data*. Konfigurasi parameter yang digunakan pada setiap teknik WE untuk membuat model bisa dilihat pada Tabel 3.6.

Tabel 3.6 Konfigurasi Paramater Word Embedding

Word Embedding	Paramater
Word2Vec	min_count = 5, size = 300, workers = 4, window = 5, sg = 1, iter = 100
GloVe	no_components=300, learning_rate=0.05, window=10, epochs=30 , no_threads=4, verbose= <i>True</i>
FastText	size=300, window=5, min_count=5, sample=1e-2, workers = 4, sg=1, iter=100

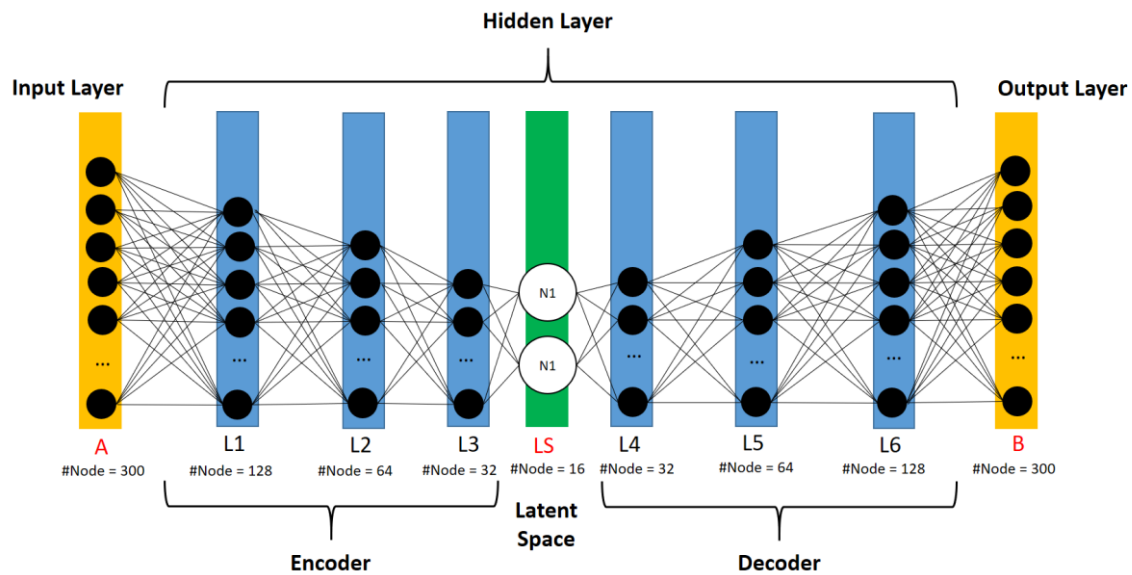
4. *Autoencoder*

Tujuan utama AE pada penelitian ini adalah melakukan reduksi dimensi dari data vektor yang dihasilkan dari WE. *Vector Data* yang dihasilkan dari setiap model WE terdiri dari 300 dimensi dengan tujuan untuk mendapatkan informasi yang lengkap dari setiap kata, namun semakin besar dimensi yang dihasilkan vektor bisa membebani model klasifikasi saat melakukan pelatihan, sehingga pada penelitian ini mengusulkan menggunakan AE untuk mereduksi dimensi vektor tersebut, dengan harapan proses pelatihan model menjadi lebih cepat dan tingkat akurasi tetap terjaga.

Arsitektur AE yang digunakan adalah berupa data input dan output dengan ukuran 300 dimensi sesuai dengan ukuran dimensi dari vektor kalimat Temuan, kemudian direduksi dengan target berupa *Encoded Data* berupa vektor 16 dimensi sesuai dengan ukuran pada *latent space layer* seperti pada Gambar 3.4. Konfigurasi hyperparameter untuk AE bisa dilihat pada Tabel 3.7.

Tabel 3.7 Konfigurasi Hyperparameter Autoencoder

Hyperparameter	Nilai
Fungsi Loss	<i>Binary Crossentropy</i>
Optimasi	Adam
Batch Size	100



Gambar 3.4 Arsitektur Autoencoder

5. Split Data

Encoded Data hasil dari reduksi dimensi oleh AE dibagi menjadi dua bagian, yaitu Train Set dan Test Set. Pada penelitian ini, data yang digunakan untuk *Train Set* sebesar 75% dan *Test Set* sebesar 25%.

Pemisahan data dilakukan dengan tujuan untuk tersedianya *Test Set* untuk kebutuhan prediksi data menggunakan model yang dihasilkan dari pelatihan menggunakan *Train Set* sehingga tingkat akurasi model bisa dievaluasi. Pada saat proses pelatihan, *Train Set* yang berupa 75% data akan dibagi lagi menjadi data yang peruntukan untuk data latih dan data validasi atau *Validation Set* yang pembagian berdasarkan *Stratified K Fold Cross Validation* dengan **K=10**.

6. Machine Learning

Algoritma ML digunakan untuk membangun model yang mampu melakukan klasifikasi Temuan, agar masuk ke dalam Kriteria yang sesuai berdasarkan proses pelatihan menggunakan *Train Set* dan dievaluasi dengan melakukan prediksi menggunakan *Test Set*.

Pada penelitiannya ini digunakan enam algoritma yang sudah diketahui mempunyai kinerja baik dalam klasifikasi yaitu LR, DT, RF, SVM, KNN dan NB, namun tidak dilakukan *hyperparameter tuning* untuk meningkatkan kinerja klasifikasi pada masing-masing model, karena fokus dari penelitian ini adalah untuk membandingkan performa dari teknik WE sebagai penghasil *Vector Data* yang digunakan oleh algoritma ML untuk membangun model.

7. Hasil Evaluasi

Hasil evaluasi dinilai dari model evaluasi yang digunakan yaitu *accuracy*, CM, dan kurva ROC. Semua nilai tersebut diperoleh dengan membandingkan hasil prediksi dengan Test Set yang telah ditentukan diawal percobaan. Setiap nilai pada model evaluasi dari setiap kombinasi teknik WE dan algoritma ML dibandingkan, dan berdasarkan perbandingan tersebut menjadi hasil evaluasi untuk mengetahui performa teknik WE terbaik.

3.2.6. Peralatan Pendukung

Peralatan yang digunakan selama penelitian adalah sebagai berikut:

1. Perangkat Keras

- Laptop dengan spesifikasi untuk programming aplikasi;
- Smartphone dengan spesifikasi untuk multimedia dengan tujuan dokumentasi, seperti untuk alat rekam dan foto kegiatan;
- Printer untuk kebutuhan pelaporan.

2. Perangkat Lunak

- Xampp, aplikasi untuk menjalankan web server dan database mySQL secara localhost untuk observasi data dari aplikasi Honest;
- SQLyog, aplikasi Client MySQL untuk pengolahan database;
- Google Colab untuk implementasi metode usulan menggunakan Python;
- Ms Word, Excel, Powepoint untuk kebutuhan pelaporan.

Daftar Pustaka

- Ahmed Chowdhury, H., Haque Imon, M. A., & Islam, M. S. (2018). A Comparative Analysis of Word Embedding Representations in Authorship Attribution of Bengali Literature. *2018 21st International Conference of Computer and Information Technology (ICCIT)*, 1–6. <https://doi.org/10.1109/ICCITECHN.2018.8631977>
- Ait Hammou, B., Ait Lahcen, A., & Mouline, S. (2020). Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics. *Information Processing and Management*, 57(1), 102122. <https://doi.org/10.1016/j.ipm.2019.102122>
- Atikah, L., Purwitasari, D., & Suciati, N. (2022). Deteksi Kejadian Lalu Lintas Pada Teks Twitter Dengan Pendekatan Klasifikasi Multi-Label Berbasis Deep Learning Multi-Label Classification Using Deep Learning Approach on Twitter. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 9(1), 87–96. <https://doi.org/10.25126/jtiik.202295206>
- Awalina, A., Bachtiar, F. A., Utaminingrum, F., & Korespondensi, P. (2022). *Perbandingan Pretrained Model Transformer Pada Deteksi Ulasan Palsu Comparison Of Pretrained Transformer Models On Spam Review Detection*. 9(3), 597–604. <https://doi.org/10.25126/jtiik.202295696>
- Bablani, A., Edla, D. R., & Dodia, S. (2018). Classification of EEG data using k-nearest neighbor approach for concealed information test. *Procedia Computer Science*, 143, 242–249. <https://doi.org/10.1016/j.procs.2018.10.392>
- SK Kaban KIPM No 13, (2020).
- Brownlee, J. (2020a, August 15). *Classification And Regression Trees for Machine Learning*. Machine Learning Mastery. <https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/>
- Brownlee, J. (2020b, August 15). *Logistic Regression for Machine Learning*. Machine Learning Mastery. <https://machinelearningmastery.com/logistic-regression-for-machine-learning/>
- Chaudhary, A. (2020, June 21). *A Visual Guide to FastText Word Embeddings*. Amitnss.Com. <https://amitnss.com/2020/06/fasttext-embeddings/>
- Chawla, J. S. (2018, April 24). *What is GloVe?* Analytics Vidhya. <https://medium.com/analytics-vidhya/word-vectorization-using-glove-76919685ee0b>
- Gandhi, R. (2018, May 5). *Naive Bayes Classifier*. Towards Data Science. <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>
- Gero, Z., & Ho, J. (2019). PMCVec: Distributed phrase representation for biomedical text processing. *Journal of Biomedical Informatics: X*, 3. <https://doi.org/10.1016/j.yjbinox.2019.100047>
- Gong, D. (2022). Top 6 machine learning algorithms for classification. In *Towards Data Science*. Towards Data Science. <https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501>
- Gupta, A. K., Sai Pardheev, C. G. V., Choudhuri, S., Das, S., Garg, A., & Maiti, J. (2022). A novel classification approach based on context connotative network (CCNet): A case of construction site accidents. *Expert Systems with Applications*, 202(October 2021), 117281. <https://doi.org/10.1016/j.eswa.2022.117281>
- Hossain, M. R., Hoque, M. M., Siddique, N., & Sarker, I. H. (2021). Bengali text document categorization based on very deep convolution neural network. *Expert Systems with Applications*, 184. <https://doi.org/10.1016/j.eswa.2021.115394>

- Jalal, N., Mehmood, A., Choi, G. S., & Ashraf, I. (2022). A novel improved random forest for text classification using feature ranking and optimal number of trees. *Journal of King Saud University - Computer and Information Sciences*, 34(6), 2733–2742. <https://doi.org/10.1016/j.jksuci.2022.03.012>
- Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2vec model analysis for semantic similarities in English words. *Procedia Computer Science*, 157, 160–167. <https://doi.org/10.1016/j.procs.2019.08.153>
- Juwiantho, H., Setiawan, E. I., Santoso, J., Purnomo, M. H., Informasi, D. T., Tinggi, S., & Surabaya, T. (2020). Sentiment Analysis Twitter Bahasa Indonesia Berbasis WORD2VEC Menggunakan Deep Convolutional Neural Network. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 7(1), 181–188. <https://doi.org/10.25126/jtiik.202071758>
- Kaibi, I., Nfaoui, E. H., & Satori, H. (2019). A Comparative Evaluation of Word Embeddings Techniques for Twitter Sentiment Analysis. *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 1–4. <https://doi.org/10.1109/WITS.2019.8723864>
- Kanakaris, N., Giarelis, N., Siachos, I., & Karacapilidis, N. (2022). Making personnel selection smarter through word embeddings: A graph-based approach. *Machine Learning with Applications*, 7(August 2021), 100214. <https://doi.org/10.1016/j.mlwa.2021.100214>
- Permen KP No 51/2018 tentang Persyaratan & Tata Cara Penerbitan Sertifikat Penerapan HACCP, (2018).
- Permen KP No 10/2021 tentang Standar Kegiatan Usaha & Produk pada Penyelenggaraan Perizinan Berusaha Berbasis Risiko Sektor KP, (2021).
- Permen PAN RB No 9/2018 tentang Jafung Inspektur Mutu Hasil Perikanan, 4 (2018).
- Khasanah, I. N. (2021). Sentiment Classification Using fastText Embedding and Deep Learning Model. *Procedia CIRP*, 189, 343–350. <https://doi.org/10.1016/j.procs.2021.05.103>
- Khatua, A., Khatua, A., & Cambria, E. (2019). A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. *Information Processing and Management*, 56(1), 247–257. <https://doi.org/10.1016/j.ipm.2018.10.010>
- Kirschenbaum, A. (2022). Unsupervised induction of inflectional families. *Computer Speech and Language*, 73(January 2019), 101324. <https://doi.org/10.1016/j.csl.2021.101324>
- Koklu, M., & Ozkan, I. A. (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, 174(May), 105507. <https://doi.org/10.1016/j.compag.2020.105507>
- Maitra, S., Madan, S., Kandwal, R., & Mahajan, P. (2018). Mining authentic student feedback for faculty using Naïve Bayes classifier. *Procedia Computer Science*, 132, 1171–1183. <https://doi.org/10.1016/j.procs.2018.05.032>
- Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *ArXiv Preprint ArXiv:1712.09405*.
- Muhammad, P. F., Kusumaningrum, R., & Wibowo, A. (2021). Sentiment Analysis Using Word2vec and Long Short-Term Memory (LSTM) for Indonesian Hotel Reviews. *Procedia Computer Science*, 179(2020), 728–735. <https://doi.org/10.1016/j.procs.2021.01.061>
- Nawangsari, R. P., Kusumaningrum, R., & Wibowo, A. (2019). Word2vec for Indonesian sentiment analysis towards hotel reviews: An evaluation study. *Procedia Computer Science*, 157, 360–366. <https://doi.org/10.1016/j.procs.2019.08.178>
- Nazeer, I., Rashid, M., Gupta, S. K., & Kumar, A. (2020). *Use of Novel Ensemble Machine Learning*

- Approach for Social Media Sentiment Analysis*. October, 16–28. <https://doi.org/10.4018/978-1-7998-4718-2.ch002>
- Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of Information Management Data Insights*, 1(2), 100019. <https://doi.org/10.1016/j.jjime.2021.100019>
- Nugroho, K. S., Bachtiar, F. A., & Mahmudy, W. F. (2022). Detecting Emotion in Indonesian Tweets: A Term-Weighting Scheme Study. *Journal of Information Systems Engineering and Business Intelligence*, 8(1), 61–70. <https://doi.org/10.20473/jisebi.8.1.61-70>
- Nugroho, K. S., Sukmadewa, A. Y., Vidiyanto, A., & Mahmudy, W. F. (2022). Effective predictive modelling for coronary artery diseases using support vector machine. *IAES International Journal of Artificial Intelligence*, 11(1), 345–355. <https://doi.org/10.11591/ijai.v11.i1.pp345-355>
- Nurdin, A., Anggo Seno Aji, B., Bustamin, A., & Abidin, Z. (2020). Perbandingan Kinerja Word Embedding Word2Vec, Glove, Dan Fasttext Pada Klasifikasi Teks. *Jurnal Tekno Kompak*, 14(2), 74. <https://doi.org/10.33365/jtk.v14i2.732>
- Pramoditha, R. (2022). *How Autoencoders Outperform PCA in Dimensionality Reduction*. Towards Data Science. <https://towardsdatascience.com/how-autoencoders-outperform-pca-in-dimensionality-reduction-1ae44c68b42f>
- Putra, J. W. G. (2019). Pengenalan konsep pembelajaran mesin dan deep learning. *Computational Linguistics and Natural Language Processing Laboratory*, 4, 1–235. <https://www.researchgate.net/publication/323700644>
- Putra, J. W. G. (2020). *Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.4* (1.4).
- Rahmadzani, R. F. (2021, January 26). *Cara Kerja Word Embedding FastText*. Rifqifai.Com. <https://rifqifai.com/cara-kerja-word-embedding-fasttext-catatan-penelitian-9/>
- Reddy, V. (2018, November 12). *Sentiment Analysis using SVM*. Medium.Com.
- Sabri, T., Beggar, O. El, & Kissi, M. (2021). Comparative study of Arabic text classification using feature vectorization methods. *Procedia Computer Science*, 198(2021), 269–275. <https://doi.org/10.1016/j.procs.2021.12.239>
- Saifullah. (2018). *Strategi Optimalisasi Kinerja Pelayanan Sertifikat HACCP (Hazard Analysis and Critical Control Point) di Kementerian Kelautan dan Perikanan*. Universitas Terbuka.
- Shahzad, K., Kanwal, S., Malik, K., Aslam, F., & Ali, M. (2019). A word-embedding-based approach for accurate identification of corresponding activities. *Computers and Electrical Engineering*, 78, 218–229. <https://doi.org/10.1016/j.compeleceng.2019.07.011>
- Singh, B., Desai, R., Ashar, H., Tank, P., & Katre, N. (2021). A Trade-off between ML and DL Techniques in Natural Language Processing. *Journal of Physics: Conference Series*, 1831(1). <https://doi.org/10.1088/1742-6596/1831/1/012025>
- Słowiński, G. (2021). Dry beans classification using machine learning. *CEUR Workshop Proceedings*, 2951, 166–173.
- Sreelakshmi, K., Premjith, B., & Soman, K. P. (2020). Detection of Hate Speech Text in Hindi-English Code-mixed Data. *Procedia Computer Science*, 171(2019), 737–744. <https://doi.org/10.1016/j.procs.2020.04.080>
- Sumiran. (2022, June 4). *Random Forest Ensemble Learning Technique*. Medium.Com. <https://medium.com/@sumiran182730/random-forest-6ba6b26494af>
- Suyoto, R. Z. H., Komarudin, M., Nama, G. F., & Yulianti, T. (2021). Classification of Civet and Canephora coffee using Support-Vector Machines (SVM) algorithm based on order-1 feature

extraction. *IOP Conference Series: Materials Science and Engineering*, 1173(1), 012006.
<https://doi.org/10.1088/1757-899x/1173/1/012006>

Swaminathan, S. (2018, March 15). *Logistic Regression — Detailed Overview*. Towards Data Science. <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>

Vatria, B. (2022). *Review : Penerapan Sistim Analysis And Critical Control Point (HACCP) Sebagai Jaminan Mutu dan Keamanan Pangan*. 104–113.