BAB II
LANDASAN TEORI

2.1 Tinjauan Pustaka

Tinjauan pustaka merupakan kumpulan dari penelitian-penelitian sebelumnya yang dapat digunakan untuk mendukung penelitian yang sedang dilakukan. Berikut adalah tinjauan literatur yang ditampilkan dalam Tabel 2.1.

Tabel 2.1 Tinjauan Literatur

No	Peneliti	Judul	Metode	Hasil
1	Isnin Apriyatin Ropikoh, Rijal Abdulhakim, Ultach Enri, Nina Sulistiyowati (2021)	Penerapan Algoritma Support Vector Machine (SVM) untuk Klasifikasi Berita Hoax Covid-19	Support Vector Machine (SVM) dan Knowledge Discovery in Database (KDD)	Hasil prediksi dengan kernel linear pada skenario 3 (80:20) sangat baik, dengan hasil 111 data hoax yang diprediksi hoax, ada 61 data hoax yang diprediksi bukan hoax. Sedangkan data bukan hoax yang diprediksi hoax ada 55 dan data bukan hoax diprediksi bukan hoax ada 1408. Akurasi tertinggi terdapat pada skenario 80:20 sebesar 92,90%. Akurasi terendah didapat oleh kernel RBF pada skenario (90:10) yaitu 90,46% dan model kurang baik menghasilkan data hoax yang diprediksi hoax ada 28 & data hoax yang diprediksi bukan hoax ada 75. Data bukan hoax yang diprediksi hoax ada 3 & data bukan hoax yang diprediksi bukan hoax ada 712.

Tabel 2.1 Tinjauan Literatur (Lanjutan)

No	Peneliti	Judul	Metode	Hasil
2	D. Maulina and R. Sagara (2018)	Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency-Inverse Document Frequency	Support Vector Machine dengan kernel Linear	Akurasi artikel hoax dan tidak hoax dengan data 108 artikel hoax dan 132 artikel tidak hoax adalah 95.8333%, dengan Cross Validation dengan Fold 10. Kernel linear sangat cocok digunakan untuk melakukan klasifikasi text dengan menggunakan pembobotan TF-IDF. SVM Linear memiliki kecepatan training 1.37 detik untuk 240 vector data.
3	M. R. A. Nasution and M. Hayaty (2019)	Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter	Support Vector Machine (SVM) dan K-Nearest Neighbor (K-NN)	SVM memiliki akurasi lebih tinggi dibandingkanK-NN, yaitu sebesar 89,70% tanpa validasi K-Fold Cross Validation dan sebesar 88,76% dengan validasi K-Fold Cross Validation. K-NN memiliki waktu proses yang lebih cepat daripada SVM, yaitu sebesar 0.0160s tanpa validasi K-Fold Cross Validation dan sebesar 0.1505s dengan validasi K-Fold Cross Validation.

Tabel 2.1 Tinjauan Literatur (Lanjutan)

No	Peneliti	Judul	Metode	Hasil
4	C. Juditha (2019)	Literasi Informasi Melawan Hoaks Bidang Kesehatan di Komunitas Online	Analisis Isi Kualitatif	Terbangunnya tujuh pilar literasi informasi dalam komunitas Indonesia Hoaxes. Literasi informasi setiap anggota terbentuk dengan sendirinya dengan bergabung di dalamnya, mulai dari proses identifikasi informasi, cakupan, perencanaan, pengumpulan informasi, evaluasi, pengelolaan, serta penyajian informasi. Admin memiliki tanggung jawab dalam penyebaran informasi yang benar dan meluruskan hoaks.
5	C. Juditha (2020)	Perilaku Masyarakat Terkait Penyebaran Hoaks Covid-19	Survei Dengan Pendekatan Kuantitatif.	Hasil penelitian menggambarkan bahwa pengetahuan responden tentang Covid-19, hoaks secara umum, dan hoaks tentang Covid-19 sangat memadai, dan dikategorikan sebagai pengetahuan level dua, yaitu memahami. Hal ini dibuktikan dengan kemampuan responden dalam mendefenisi kanapa itu virus Corona dan juga hoaks. Informasi terkait Covid-19 banyak diperoleh responden dari situs berita, media sosial, televisi, pesan instan, serta website resmi pemerintah.

Tabel 2.1 Tinjauan Literatur (Lanjutan)

No	Peneliti	Judul	Metode	Hasil
6	Wahyu Nugraha, Agung Sasongko (2022)	Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search	Supervised learning atau klasifikasi dan Tuning Hyperparameter dengan Grid Search	Grid Search menggunakan Cross Validation memberikan kemudahan dalam menguji coba setiap parameter model tanpa harus melakukan validasi manual satu persatu. Hasil eksperimen menunjukkan bahwa model XGBoost memperoleh nilai terbaik yaitu sebesar 0,772 sedangkan Decision tree memiliki nilai terendah yaitu 0,701.
7	Ranggi Praharaningtyas Aji , Sarmini (2019)	Pelatihan Identifikasi dan Pelaporan Berita Hoax melalui portal ‘ turnbackhoax.id ’ kepada Masyarakat Desa Kedungwringin	Metode Pendidikan Masyarakat	Kegiatan ini diikuti oleh 40 orang perwakilan warga desa Kedungwringin. Kegiatan pengabdian masyarakat ini dilakukan dengan penyuluhan mengenai hoax. Materi yang diberikan pada penyuluhan ini meliputi : 1. Prinsip bersosial media dan internet 2. Dampak bersosial media dan internet 3. Hal yang boleh dan tidak boleh pada sosial media dan internet 4. Contoh berita hoax pada media sosial dan internet.

Tabel 2.1 Tinjauan Literatur (Lanjutan)

No	Peneliti	Judul	Metode	Hasil
8	Agatha Deolika , Kusrini , Emha Taufiq Luthf (2019)	ANALISIS PEMBOBOTAN KATA PADA KLASIFIKASI TEXT MINING	Naïve bayes	Pembobotan TF.RF dengan klasifikasi Naïve bayes lebih baik dari pembobotan TF.IDF dan WIDF dengan nilai Accuracy 98,67%, Precision 93,81%, dan Recall 96,67%. Klasifikasi naïve bayes dapat digunakan untuk mengelompokan atau klasifikasi text mining dengan baik.
9	Aji Prasetya Wibawa, Muhammad Guntur Aji Purnama, Muhammad Fathony Akbar, Felix Andika Dwiyanto (2018)	Metode-metode Klasifikasi	Jaringan Saraf Tiruan, Naïve Bayes, Support Vector Machine, Decission Tree, dan Fuzzy	Setiap metode mempunyai karakteristik tertentu termasuk kelemahan dan kelebihan masing-masing pendekatan. Berdasarkan kelemahan dan kelebihan tersebut dapat dijadikan pertimbangan untuk memilih metode yang sesuai dengan macam data yang akan diolah.
10	Iwan Syarif , Adam Prugel-Bennett, Gary Wills (2016)	SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance	Support Vector Machine (SVM) dengan optimasi Grid Search dan Genetic Algorithm (GA)	Dalam 1 dataset (spambase), pencarian grid memiliki akurasi yang lebih baik dari GA. Namun, pada dataset madelon dan intrusi GA tidak dapat menjamin hasil yang baik untuk semua kernel karena kinerja klasifikasi yang tidak begitu baik (pada dataset madelon F-measure hanya 66,67% dan pada dataset intrusi F-measure hanya 61,31%).

2.2 Text Mining

Text mining adalah teknik yang digunakan untuk melakukan klasifikasi untuk menemukan pola-pola menarik dari kumpulan data tekstual yang banyak (Feldman and Sanger, 2007). Text mining merupakan salah satu dari bentuk variasi data mining. Mirip dengan data mining, kecuali teknik yang dirancang untuk bekerja pada data terstuktur dalam database di data mining. Text mining dapat bekerja pada data yang tidak dan semi-terstruktur seperti dokumen teks yang lengkap, kode halaman web, dan lainnya (Pratama and Atmi, 2020).

Text mining diartikan sebagai penemuan informasi baru yang sebelumnya tidak diketahui komputer yang secara otomatis mengekstraksi informasi dari berbagai sumber. Menggabungkan informasi yang berhasil diekstrak dari berbagai sumber merupakan kunci dari proses ini (Rahmawati *et al.*, 2020). Tujuan dari text mining adalah untuk menemukan katakata yang dapat merepresentasikan isi dokumen yang nantinya dapat dianalisis.

Terdapat beberapa proses yang dapat dilakukan dalam melakukan text mining untuk mendapatkan informasi yang terdapat dalam data teks (Purba and Situmorang, 2017), diantaranya :

1. Preprocessing

   Tahap ini adalah tahapan dimana teks yang disiapkan untuk data harus dibersihkan terlebih dahulu dan disisakan kalimat pentingnya saja.
2. Transformation

   Merupakan tahapan dimana kata diubah menjadi bentuk dasarnya dan dimensi kata yang ada dalam dokumen dikurangi.
3. Feature Selection

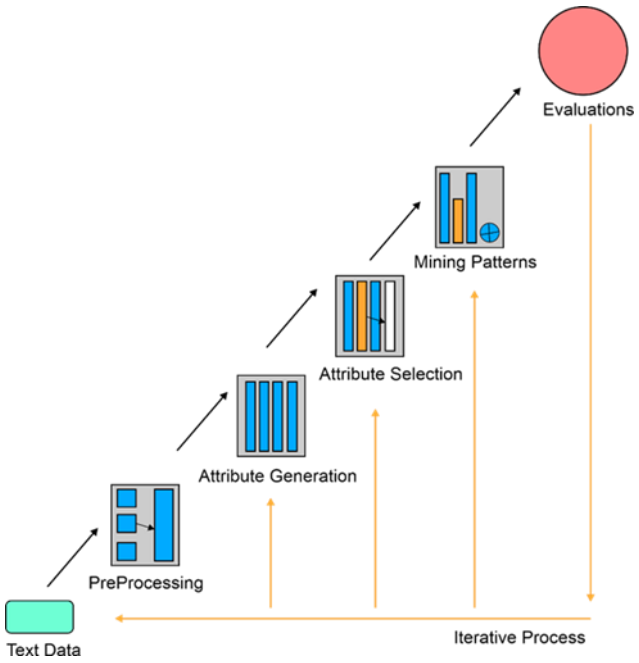
   Adalah proses penghapusan kata yang dianggap tidak deskriptif dan pemilahan kata-kata yang dianggap penting. Ide dasarnya adalah menghapus kata-kata yang terlalu sedikit atau terlalu banyak muncul.
4. Pattern Discovery

   Tahapan penting untuk menemukan pola dalam keseluruhan teks. Input awal berupa data teks akan menghasilkan output berupa pola yang akan

menjadi hasil dari interpretasi. Jika output tidak sesuai, maka evaluasi dilanjutkan dengan melakukan iterasi ke tahapan sebelumnya.

5. Evaluation

   Tahapan dimana hasil dievaluasi apakah dapat digunakan atau harus dibuang.



Gambar 2.1 Tahapan Text Mining

Pada tahap Preprocessing, text mining dapat menghasilkan banyak fitur untuk merepresentasikan dokumen. Terdapat empat fitur yang sering digunakan pada tahap ini, yaitu :

a) Karakter

   Merupakan fitur yang paling sederhana diantara fitur lainnya. Fitur ini jarang digunakan dalam text mining dikarenakan keterbatasan jumlah dari karakter itu sendiri. Terdiri dari huruf (a,b,c,….,z), angka (0,1,2,….,9), tanda baca (!,?, dan lain sebagainya), serta simbol khusus (@,#,$,%, dan lain sebagainya).

b) Words (Kata)

   Merupakan kumpulan dari beberapa karakter yang mengandung suatu makna. Fitur ini lebih banyak digunakan dibandingkan karakter karena

memiliki variasi yang lebih banyak sebagai pembeda. Contohnya adalah kata “aku” yang terdiri dari gabungan karakter a, k, dan u.

c) Terms

Dapat diartikan mirip seperti words, namun pembedanya lebih signifikan dibanding words. Terdiri dari satu atau lebih words phrase yang memiliki arti yang baru.

2.3 Data Mining

Data mining atau penambangan data adalah proses untuk mengumpulkan data yang nantinya akan diolah dengan beberapa metode. Data mining juga sering disebut penemuan pengetahuan dalam database. Tujuannya adalah untuk memanfaatkan data dan mengolahnya menjadi sesuatu atau informasi yang baru dan berguna (Fitrani and Novarika, 2019). Data mining dapat diartikan sebagai proses menemukan pola dan pengetahuan menarik dari suatu data dalam jumlah besar.

Proses data mining terbagi menjadi beberapa tahap (Rahmawati *et al.*, 2020), yaitu :

1. Data Cleaning

   Proses pembersihan noice dan data yang tidak penting.
2. Data Integration

   Penggabungan data dari berbagai sumber yang masih terkait.
3. Data Selection

   Pengambilan data yang relevan untuk digunakan yang ada di database.
4. Data Transformation

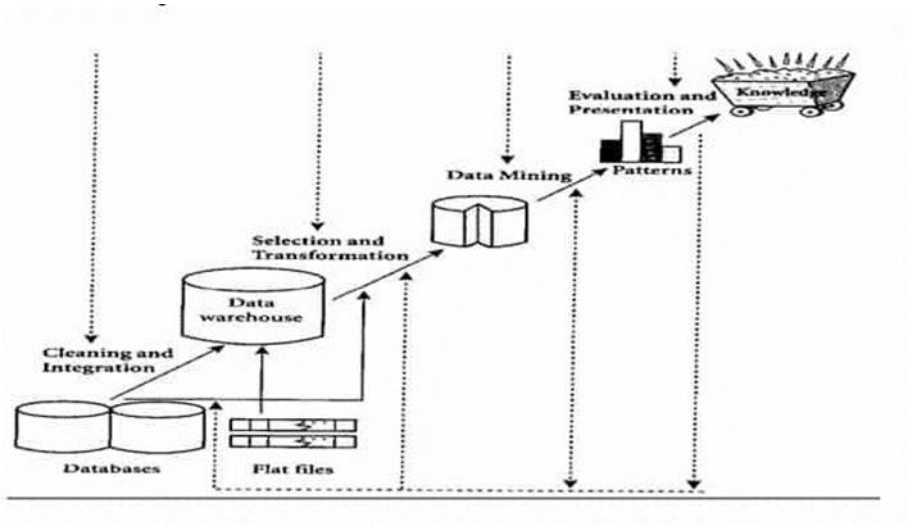
   Data diubah dan dikonsolidasikan menjadi bentuk yang sesuai untuk melakukan operasi ringkasan dan agregasi.
5. Data Mining

   Proses dimana metode cerdas dipakai untuk mengekstraksi pola data.
6. Pattern Evaluation

Mengidentifikasi pola yang merepresentasikan pengetahuan berdasarkan ukuran ketertarikan.

7. Knowledge Presentation

   Proses visualisasi dan representasi pengetahuan untuk menyajikan pengetahuan yang telah diproses kepada pengguna.



Gambar 2.2 Alur Data Mining

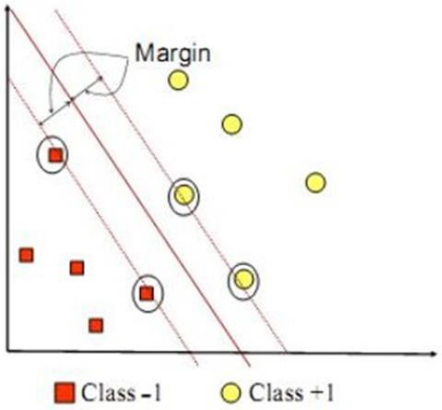
2.4 Support Vector Machine

SVM adalah metode learning machine yang bekerja atas prinsip Structural Risk Minimization (SRM) dengan tujuan menemukan hyperplane terbaik yang memisahkan dua buah class pada input space. Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangkaian harmonis konsep-konsep unggulan dalam bidang pattern recognition. Sebagai salah satu metode pattern recognition, usia SVM terbilang masih relatif muda. Walaupun demikian, evaluasi kemampuannya dalam berbagai aplikasinya menempatkannya sebagai state of the art dalam pattern recognition (Cortes and Vapnik, 1995).

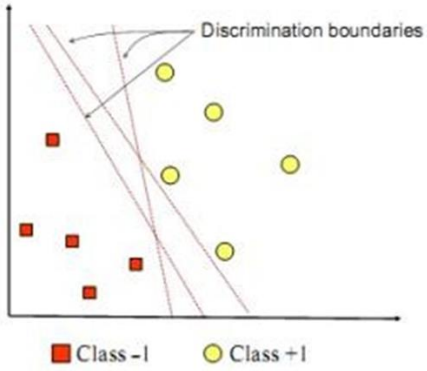
Gambar 2.3 memperlihatkan beberapa pattern yang merupakan anggota dari dua buah class: +1 dan –1. Pattern yang tergabung pada class –1 disimbolkan dengan warna merah (kotak), sedangkan pattern pada class +1, disimbolkan dengan warna kuning (lingkaran). Problem klasifikasi dapat

diterjemahkan dengan usaha menemukan garis (hyperplane) yang memisahkan antara kedua kelompok tersebut (Susilowati, Sabariah and Gozali, 2015).

Hyperplane pemisah terbaik antara kedua class dapat ditemukan dengan mengukur margin hyperplane tsb. dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane tersebut dengan pattern terdekat dari masing-masing class. Pattern yang paling dekat ini disebut sebagai support vector. Garis solid pada gambar menunjukkan hyperplane yang terbaik, yaitu yang terletak tepat pada tengah -tengah kedua class, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah support vector. Usaha untuk mencari lokasi hyperplane ini merupakan inti dari proses pembelajaran pada SVM (Feriyawan and Danoedoro, 2012).



Gambar 2.3 SVM berusaha untuk menemukan hyperplane terbaik yang memisahkan kedua kelas



Gambar 2. 4 Hyperplane terbentuk diantara kedua kelas

2.5 Grid Search

Algoritma Grid Search membagi jangkauan parameter yang akan dioptimalkan kedalam grid dan melintasi semua titik untuk mendapatkan parameter yang optimal. Dalam aplikasinya, algoritma grid search harus dipandu oleh beberapa metrik kinerja, biasanya diukur dengan cross-validation pada data training. Data training adalah data latih yang digunakan untuk melatih beberapa pasangan model, sedangkan data testing adalah data uji yang digunakan untuk menguji model terbaik yang diperoleh dari data training. Oleh karena itu disarankan untuk mencoba beberapa variasi pasangan parameter pada hyperplane SVM. Pasangan parameter yang menghasilkan akurasi terbaik yang didapatkan dari uji cross-validation merupakan parameter yang optimal. Parameter optimal tersebut yang selanjutnya digunakan untuk model SVM terbaik. Setelah itu, model SVM tersebut digunakan untuk memprediksi data testing untuk mendapatkan generalisasi tingkat akurasi model (Naufal, 2017).

2.6 K-fold Cross Validation

Salah satu pendekatan alternatif untuk “training dan testing” yang sering di adopsi dalam beberapa kasus (dan beberapa lainnya terlepas dari ukurannya) yang di sebut dengan k-fold cross validation, dengan cara menguji besarnya error pada data testing (Santosa, 2007). Pada penelitian ini digunakan k-1 sampel untuk training dan 1 sampel sisanya untuk testing. Misalnya ada 10 subset data, kita menggunakan 9 subset untuk training dan 1 subset sisanya untuk testing. Ada 10 kali training dimana pada masingmasing training ada 9 subset data untuk training dan 1 subset digunakan untuk testing. Kemudian di hitung rata-rata error dan standar deviasi error (Santosa, 2007). Setiap bagian k pada gilirannya digunakan sebagai ujian menetapkan dan k lainnya dan 1 bagian digunakan sebagai training set (Enri, 2018).