BAB 2

TINJAUAN PUSTAKA

2.1 Penelitian Terkait

(Saputra et al., 2019) mengimplementasikan SVR pada AES dengan penelitian yang berjudul "*Penggunaan Support Vector Regression dalam Pemodelan Indeks Saham Syariah Indonesia dengan Algoritma Grid Search*" yang bertujuan untuk mengetahui faktor yang mempengaruhi nilai indeks saham dengan membandingkan beberapa *kernel*. Data yang digunakan oleh peneliti adalah data sekunder, yaitu data bulanan dan data mingguan Indeks Saham Syariah Indonesia (ISSI). Dengan algoritma *Grid Search* dan RMSE sebagai evaluasi performa algoritma, evaluasi model SVR terbaik mendapatkan nilai RMSE sebesar 2,289 dan nilai korelasi sebesar 0,873 untuk data uji dengan menggunakan *kernel liniear*.

(Bharata & Sulistyowati, 2020) melakukan penelitian AES yang berjudul "*Optimasi Sistem Penilaian Ujian Essay Online Menggunakan Support Vector Machine (SVM) dan Latent Semantic Analysis (LSA) dengan Bahasa R*" yang bertujuan untuk mengkaji permasalahan penilaian jawaban esai secara otomatis dengan menggabungkan SVM sebagai teknik klasifikasi teks dan LSA untuk menangani sinonim dan polisemi antar index *term* serta fitur generic yang berfungsi sebagai penguji model penilaian esai dengan pertanyaan yang berbeda. Penelitian ini menggunakan bahasa pemrograman R. hasil dari penelitian ini mendapatkan akurasi 96,23%, sehingga penggunaan metode SVM dan LSA mencapai tingkat akurasi yang cukup tinggi.

"*Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification*" (Arifin et al., 2021) Penelitian ini dilakukan untuk mengklasifikasikan artikel ilmiah ke dalam kategori sesuai dengan fokus dan ruang lingkup yang terdapat pada laman *Syntax* Jurnal Informatika secara otomatis dengan memanfaatkan proses text *mining* pada algoritma SVM dengan menggunakan empat *kernel*, yaitu, *kernel liniear*, *kernel* polynomial, *kernel*

sigmoid dan *kernel* RBF serta ekstraksi fitur TF-IDF dan *N-Gram*. Data dibagi menjadi empat skenario dan diuji terhadap model ukur dengan nilai *Accuracy*, *Precision*, *Recall* dan *F-measure*. Hasil terbaik yang didapat oleh peneliti adalah *Accuracy* sebesar 70%, *precision* sebesar 75%, *recall* sebesar 69% dan *F-measure* sebesar 71% pada skenario perbandingan 90:10 pada *kernel liniear*.

Pada tahun 2021, (Thamrin et al., 2021) melakukan penelitian yang berjudul "*Text Classification and Similarity Algorithms in Essay Grading*" dengan mengklasifikasikan data esai dari ujian Bahasa Indonesia dengan hasil 1648 jawaban. Algoritma yang digunakan adalah SVM dan KNN dengan ekstraksi fitur TF-IDF dan LSA serta RMSE yang berfungsi sebagai pengevaluasi performa algoritma yang digunakan. Hasil terbaik yang telah didapatkan oleh peneliti dengan menggunakan RMSE adalah 2,730.

(Verdikha et al., 2021) melakukan penelitian yang berjudul "*Regression and Oversampling Method for Indonesian Language Automated Essay Scoring*" dengan menggunakan data penelitian yang sama seperti Thamrin (2021) dan diperoleh secara *regresi* menggunakan tiga metode yaitu SVR, LR, dan MLP-R serta tambahan oversampling SMOTE agar data yang diperoleh dapat seimbang dan dapat dibandingkan performa yang dihasilkan menggunakan ekstraksi TF-IDF. Hasil dari penelitian ini dengan menggunakan metode SVR mendapatkan hasil RMSE dengan nilai 2,166.

2.2 Teori Dasar Penelitian

Teori Dasar Penelitian adalah beberapa kumpulan metode, tahapan dan pemrosesan yang menyangkut semua pembahasan dalam sebuah penelitian.

2.2.1 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) adalah salah satu bagian dari bahasa alami yang merupakan cabang dari ilmu komputer yang membahas tentang interaksi antara manusia dengan komputer menggunakan bahasa manusia (Aditama, 2020). Dalam hal ini agar komputer dapat memahami Bahasa alami, komputer harus memiliki pengetahuan tentang Bahasa alami itu sendiri baik dari

segi kata yang digunakan, arti dari kata tersebut, fungsi kata dari sebuah kalimat dan bagaimana kata-kata tersebut dapat terbentuk sebuah kalimat.

2.2.2 *Automated Essay Scoring* (AES)

AES merupakan sistem penilaian soal ujian uraian secara otomatis dengan dengan menggunakan teknologi komputer. Penilaian AES dapat diperoleh dengan cara menghitung kemiripan teks antara kunci jawaban dan jawaban peserta ujian (Arfandy & Musdar, 2020). Dengan adanya penelitian AES, sangat membantu Pendidikan terutama pada saat pembelajaran secara online karena dapat memberikan penilaian secara otomatis.

2.2.3 *Term Frequency-Inverse Document Frequency* (TF-IDF)

TF-IDF merupakan pembobotan yang digunakan dalam pengambilan suatu informasi dalam *text mining*. Bobot TF-IDF digunakan untuk mengevaluasi seberapa pentingnya sebuah kata di dalam sekumpulan data. Adapun pembobotan TF-IDF sendiri terdiri dari dua faktor, yaitu:

1. ***Term Frequency*** **(TF)**

   Term Frequency (TF) adalah suatu kemunculan fl di dalam sebuah dokumen dj yang di bandingkan dengan fl yang sering muncul pada dokumen itu (Dalimunthe & Hayadi, 2022).

2. ***Inverse Document Frequency*** **(IDF)**

   Inverse Document Frequency (IDF) adalah *Frequency* kemunculan suatu istilah fi di dalam seluruh dokumen. Jika jumlah keseluruhan dokumen pada sebuah data dinyatakan dengan nilai N dan jumlah dokumen yang memiliki istilah fi dinyatakan dengan ni, maka nilai IDFi-nya dapat dinyatakan dengan:

$$idf(t) = log \frac{1+n}{1+df(t)} + 1 \quad (2.1)$$

Keterangan :

idf = *invers document frekuency*

t = *term* (kata)

n = total *document*

Untuk menentukan nilai antara TF dan IDF diperlukan penggabungan perhitungan antara TF dengan IDF, contohnya sebagai berikut:

$$tf - idf(t,d) = tf(t,d) \times idf(t) \quad (2.2)$$

Keterangan:

tf = *term Frequency*

idf = *inverse document frequency*

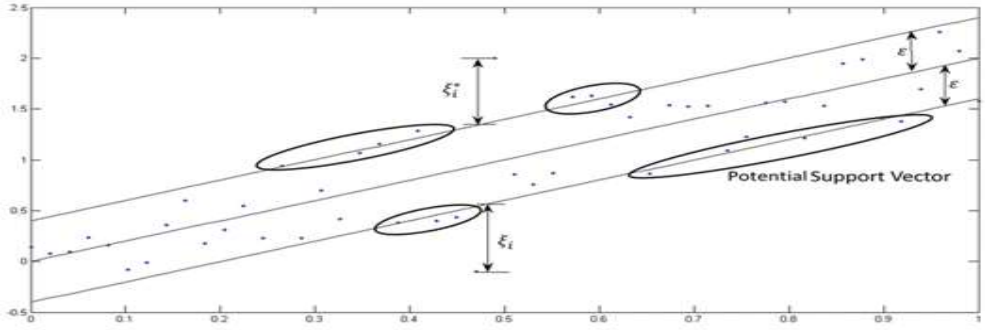
t = *term*

d = *document*

Hasil dari perhitungan TF IDF selanjutnya akan dilakukan tahapan normalisasi, Menurut (Maulana et al., 2019), Normalisasi memiliki tujuan untuk mendapatkan data dengan ukuran yang lebih kecil yang dapat mewakili data yang asli tanpa kehilangan karakteristiknya sendiri.

2.2.4. *Support Vector Regression* (SVR)

Support Vector Regression (SVR) adalah penerapan dari SVM yang digunakan untuk kasus *regresi* yang mana outputnya berupa bilangan riil atau kontinu. Ide dasar dari SVR sendiri adalah untuk menentukan *dataset* yang mana datanya akan di bagi menjadi *data training* dan *data testing*. Kemudian dari dua data tersebut akan dilakukan suatu fungsi *regresi* dengan Batasan definisi tertentu sehingga dapat memiliki prediksi yang memiliki nilai aktual (Lestari et al., 2021). Pada metode SVM, penerapannya di kasus klasifikasi dengan menghasilkan nilai bulat, sedangkan pada metode SVR, penerapannya di kasus *regresi* yang menghasilkan bilangan riil.



Gambar 2.1 Ilustrasi SVR

Pada Gambar 2.1 di atas menjelaskan bahwa, garis tengah yang biasa disebut *hyperplane* diapit dengan dua garis pembatas yang memiliki nilai "− & +" seperti gambar diatas, adapun simbol ε di atas sebagai jarak antar *hyperplane* dengan pembatas, Titik-titik yang dilingkari adalah *potential support vectors* dan titik titik yang berada pada garis *hyperplane* dan garis pembatas maupun yang berada diluar garis pembatas adalah *data points* yang dapat menjadi calon pembatas, sehingga keseluruhan *data points* dapat masuk menjadi satu klaster, sehingga tetap dapat meminimalisir nilai ε nya. Jika divisualisasikan, garis *hyperplane* sebisa mungkin melewati semua titik-titik *data points* tersebut.

2.2.5. *Kernel*

Dalam penelitian (Asyiva, 2019) Fungsi dari *kernel* sangat penting dalam penggunaan metode SVR, *Kernel* sendiri sangat membantu mengatasi permasalahan *non-linear* pada dimensi tinggi yang harus dilakukan yaitu mengganti inner product ($xi\ dan\ xj$) dengan fungsi *kernel* yang telah dipilih, Karena kinerja dari metode SVR ditentukan oleh jenis fungsi *kernel* dan parameter yang akan digunakan, adapun *kernel* yang digunakan dalam penelitian ini adalah:

1) *Kernel liniear*

Kernel liniear merupakan fungsi *kernel* yang paling sederhana. digunakan ketika data yang dianalisis sudah terpisah secara *linear*. *kernel Linear* cocok ketika terdapat banyak fitur dikarenakan pemetaan ke ruang dimensi yang lebih tinggi tidak benar – benar meningkatkan kinerja seperti pada klasifikasi teks (Julianto et al., 2022).

$$K(x, y) = x.y \tag{2.3}$$

Keterangan:

$K(x, y)$ = nilai *kernel* dari data x dan data y

x = nilai data 1

y = nilai data 2

2) *Kernel RBF*

Kernel RBF merupakan fungsi *kernel* yang biasa digunakan dalam analisis ketika data tidak terpisah secara *linear*. RBF *kernel* memiliki dua parameter yaitu *Gamma* dan *Cost*. Parameter *Cost* atau biasa disebut sebagai C merupakan parameter yang bekerja sebagai pengoptimalan SVM untuk menghindari misklasifikasi di setiap sampel dalam training *dataset*. Parameter *Gamma* menentukan seberapa jauh pengaruh dari satu sampel training *dataset* dengan nilai rendah berarti "jauh", dan nilai tinggi berarti "dekat" (Julianto et al., 2022).

Fungsi *kernel* RBF menghitung antara dua vektor, rumus persamaan yang digunakan menggunakan rumus persamaan 2.4 sebagai berikut:

$$k(x, y) = exp(-\gamma \parallel x - y \parallel^2 \quad (2.4)$$

Dimana x dan y adalah input vektor. γ dikenal sebagai kemiringan. *Kernel* RBF dikenal juga *kernel* varian *Gaussian* (Barupal & Fiehn, 2019).

2.2.6. *Root Mean Square Error* (RMSE)

RMSE digunakan untuk menghitung keakuratan nilai yang direkomendasikan oleh sistem. Semakin banyak nilai rekomendasi sistem yang sama dengan nilai sebenarnya (penilaian oleh pengajar) menunjukkan semakin akuratnya hasil dari sistem. Semakin kecil nilai yang dihasilkan maka semakin bagus pula hasil peramalan yang dilakukan (Cholis et al., 2019) Adapun rumusan dari RMSE sebagai berikut:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (2.5)$$

Keterangan :

Ŷi = Nilai hasil peramalan

yi = Nilai aktual / Nilai sebenarnya

n = Jumlah data

I = Urutan data pada dokumen

Terdapat beberapa penelitian yang telah diteliti sebelumnya dan menjadi referensi dari penelitian ini yang terangkum pada Tabel 2.1.

Tabel 2.1 Referensi Penelitian

Penulis Tahun	**Judul**	**Metode**	**Hasil**
(Bharata & Sulistyowati, 2020)	Optimasi Sistem Penilaian Ujian *Essay* Online Menggunakan *Support Vector Machine* (SVM) dan *Latent Semantic Analysis* (LSA) dengan Bahasa R	SVM & LSA	Hasil akurasi 96,23
(Saputra et al., 2019)	Penggunaan *Support Vector Regression* dalam Pemodelan Indeks Saham Syariah Indonesia dengan Algoritma *Grid Search*	SVR	Nilai RMSE 2,289 dan nilai korelasi sebesar 0,873
(Arifin et al., 2021)	Penerapan Algoritma *Support Vector Machine* (SVM) dengan TF-IDF *N-Gram* untuk *Text Classification*	SVM & TF-IDF	Nilai *Accuracy* 70%, *precision* 75%, *recall* 69% dan *F-measure* 71% pada skenario perbandingan 90:10 pada *kernel liniear*.
(Julianto et al., 2022)	Analisis Sentimen Ulasan Restoran Menggunakan Metode *Support Vector Machine*	SVM & TF-IDF	Hasil akurasi 93% dan f1-score 93%. Dari penelitian ini juga menunjukkan bahwa untuk meningkatkan nilai akurasi dan f1- score, model klasifikasi SVM membutuhkan parameter TF-IDF min_df sebesar 0.05, max_df sebesar 0.75, norm l2,

			n-gram (1, 2), *kernel* SVM *linear* dengan C sebesar 1.
(Pangarkar et al., 2020)	*Assessment of the Different Machine Learning Models for Prediction of Cluster Bean* (Cyamopsis tetragonoloba L. Taub.) Yield	KNN & SVR dengan evaluasi RMSE, MAE, dan RRSE	hasil nilai variabel (98,31%) dan validitas global (71%).
(Ritonga & Purwaningsih, 2018)	Penerapan *Support Vector Machine* (SVM) dalam Klasifikasi kualitas Pengelasan SMAW (*SHIELD METAL ARC WELDING*)	SVM	Hasil pengujian l sebesar 96,2%, dan pengujian menggunakan data uji menunjukkan hasil sebesar 98%
(Dinyanti, 2021)	Peramalan Curah Hujan Menggunakan Metode *Support Vector Regression* (SVR) di Kabupaten Manggarai Barat	SVR	Hasil proses training yang diperoleh menunjukan bahwa hasil terbaik berada pada *kernel liniear* dengan domain grid 10 x 10
(Thamrin et al., 2021)	*Text Classification and Similarity Algorithms in Essay Grading*	SVM & kNN	Hasil terbaik yang menggunakan RMSE adalah 2,73.
(Verdikha et al., 2021)	*Regression and Oversampling Method for Indonesian Language Automated Essay Scoring*	SVR, LR, dan MLP-R	Hasil dari penelitian menggunakan SVR dengan *Kernel* RBF mendapatkan hasil RMSE dengan nilai 2,16.