Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 6, December 2023, pp. 1555-1566
DOI: https://doi.org/10.52436/1.jutif.2023.4.6.1536
p-ISSN: 2723-3863
e-ISSN: 2723-3871

IMPLEMENTATION OF SUPPORT VECTOR MACHINE METHOD IN CLASSIFYING SCHOOL LIBRARY BOOKS WITH COMBINATION OF TF-IDF AND WORD2VEC

Salsabila Nida Cahyani[*1], Galuh Wilujeng Saraswati[2]

[1,2]Informatics Engineering, Computer Science Faculty, Universitas Dian Nuswantoro, Indonesia
Email: [1]111202012982@mhs.dinus.ac.id, [2]galuhwilujeng@dsn.dinus.ac.id

(Article received: November 20, 2023; Revision: December 24, 2023; published: December 30, 2023)

Abstract

The development of technology in education is integral to enhancing its quality, such as implementing information technology in school libraries. Searching for books in school libraries is time-consuming due to conventional book classification, lacking organization based on classifications. Therefore, implementing information technology in school libraries is crucial to improve library management effectiveness. An innovative solution optimizing library management involves leveraging artificial intelligence, particularly machine learning. In applying machine learning to library book classification, Support Vector Machine acts as an algorithm understanding patterns and characteristics of book titles, categorizing them into Dewey Decimal Classification (DDC). The dataset comprises 10 classes aligned with DDC. Random data collection follows an 80:20 scale for training and testing data. Data preprocessing is an initial research stage, addressing imbalanced data through oversampling. Testing the SVM algorithm with a linear kernel and C = 1 parameter is conducted three times using different feature extraction methods: TF-IDF alone, Word2Vec alone, and a combination of TF-IDF and Word2Vec. Model performance evaluation employs K-Fold Cross-Validation. After the three objective tests, the most accurate book classification results were obtained using a combination of TF-IDF and Word2Vec feature extraction. It's concluded that SVM's book classification method can be applied, yielding the highest accuracy of 73% with the TF-IDF and Word2Vec feature extraction combination. This outperforms other feature extraction methods, with precision at 83%, recall at 72%, and an F1-Score of 76%.

Keywords: *classification, dewey decimal classification, library, support vector machine, TF-IDF, Word2Vec.*

PENERAPAN METODE SUPPORT VECTOR MACHINE PADA KLASIFIKASI BUKU PERPUSTAKAAN SEKOLAH DENGAN KOMBINASI TF-IDF DAN WORD2VEC

Abstrak

Perkembangan teknologi di bidang pendidikan sangat penting untuk meningkatkan kualitasnya, seperti menerapkan teknologi informasi di perpustakaan sekolah. Pencarian buku di perpustakaan sekolah memakan waktu lama karena klasifikasi buku yang konvensional, kurang terorganisir berdasarkan klasifikasi. Oleh karena itu, penerapan teknologi informasi di perpustakaan sekolah sangat penting untuk meningkatkan efektivitas pengelolaan perpustakaan. Solusi inovatif untuk mengoptimalkan pengelolaan perpustakaan melibatkan pemanfaatan kecerdasan buatan, khususnya *machine learning*. Dalam menerapkan *machine learning* untuk klasifikasi buku perpustakaan, *Support Vector Machine* berfungsi sebagai algoritma yang memahami pola dan karakteristik judul buku, mengelompokkannya ke dalam kategori *Dewey Decimal Classification* (DDC). Dataset terdiri dari 10 kelas yang disesuaikan dengan DDC. Pengambilan data secara acak mengikuti skala 80:20 untuk data latih dan uji. Tahap awal penelitian melibatkan preprocessing data, termasuk penanganan data tidak seimbang melalui *oversampling*. Pengujian algoritma SVM dengan *kernel linear* dan parameter C = 1 dilakukan sebanyak tiga kali menggunakan metode ekstraksi fitur yang berbeda: hanya TF-IDF, hanya *Word2Vec*, dan kombinasi TF-IDF dan *Word2Vec*. Evaluasi performa model menggunakan metode *K-Fold Cross-Validation*. Setelah dilakukan tiga kali pengujian secara objektif, mendapatkan hasil klasifikasi buku terakurat dengan menggunakan ekstraksi fitur kombinasi TF-IDF dan *Word2Vec*. Sehingga disimpulkan bahwa metode klasifikasi buku perpustakaan dengan SVM dapat diterapkan dan menghasilkan akurasi tertinggi sebesar 73% pada kombinasi ekstraksi fitur TF-IDF dan *Word2Vec*. Hal ini lebih tinggi dibandingkan dengan metode ekstraksi fitur lain, dengan *precision* sebesar 83%, *recall* sebesar 72%, dan *F1-Score* sebesar 76%.

Kata kunci: *dewey decimal classification, klasifikasi, perpustakaan, support vector machine, TF-IDF, Word2Vec.*

1. PENDAHULUAN

Perkembangan teknologi saat ini semakin pesat dengan kehadiran berbagai sistem yang dapat memudahkan manusia dalam menjalankan kegiatan sehari-hari. Kemajuan teknologi dan informasi telah membawa kemudahan dalam mengakses informasi secara cepat [1]. Teknologi ini telah mengubah cara manusia berinteraksi dengan informasi dan pengetahuan secara signifikan. Teknologi di bidang pendidikan merupakan bagian dari pengembangan teknologi untuk peningkatan mutu pendidikan, salah satunya menerapkan teknologi informasi dalam perpustakaan sekolah [2]. Pendidikan memegang peranan sentral dalam pembangunan suatu bangsa. Peningkatan mutu pendidikan menjadi tujuan yang harus terus diperjuangkan demi mencetak generasi yang kompeten, kreatif, dan inovatif. Salah satu elemen penting dalam sistem pendidikan adalah perpustakaan, yang dianggap sebagai jantungnya ilmu pengetahuan. Perpustakaan sekolah memiliki peran strategis dalam memberikan akses ke berbagai sumber daya belajar, literatur, dan informasi kepada siswa dan guru. Perpustakaan yang baik dan terkelola dengan baik akan memberikan manfaat besar bagi siswa dan guru dalam hal peningkatan mutu pendidikan. Namun, pengelolaan perpustakaan yang optimal menjadi tantangan yang tak bisa diabaikan.

SD Negeri Pledokan merupakan salah satu sekolah dasar negeri yang memiliki perpustakaan sekolah dengan koleksi buku yang cukup lengkap bagi sekolah dasar yaitu sebanyak 1.435 koleksi buku dengan berbagai kategori. SD Negeri Pledokan terletak di Kecamatan Sumowono, Kabupaten Semarang, Jawa Tengah. Sebagai lembaga pendidikan yang berkomitmen tinggi terhadap mutu pendidikan di era digital. Berbagai kendala dalam pengelolaan data perpustakaan masih menjadi isu yang perlu diselesaikan. Berdasarkan hasil observasi di lapangan, SD Negeri Pledokan masih menggunakan sistem konvensional dimana ribuan buku di data pengklasifikasiannya pada suatu buku catatan sehingga tidak efektif dan tidak efisien. Selain itu, jika ingin mencari sebuah buku membutuhkan waktu yang lama, karena pencarian klasifikasi buku berdasarkan buku catatan konvensional sehingga buku juga tidak tersusun berdasarkan klasifikasinya. Tak hanya itu, kurangnya pengetahuan dan keterampilan guru dalam memanfaatkan kemajuan teknologi menjadi salah satu tantangan yang dapat mempengaruhi akreditasi perpustakaan sekolah dan juga dapat mengakibatkan pelayanan sekolah menjadi kurang optimal. Sehingga penerapan teknologi informasi pada perpustakaan sekolah sangat diperlukan untuk meningkatkan efektivitas pengelolaan perpustakaan di sekolah [3].

Menanggapi masalah di atas, diperlukan solusi inovatif yang mampu mengoptimalkan pengelolaan perpustakaan yaitu dengan memanfaatkan teknologi *artificial intelligence*. *Artificial Intelligence* atau kecerdasan buatan merupakan salah satu cabang ilmu komputer yang dapat dimanfaatkan dalam berbagai bidang termasuk perpustakaan [4]. Salah satu teknologi *artificial intelligence* yang dimanfaatkan adalah teknologi *machine learning*. *Machine learning* merupakan sistem yang secara otomatis mempelajari model dari data untuk membuat keputusan yang lebih baik [5]. Machine learning dapat digunakan untuk mengklasifikasikan buku secara efisien yang mana hal tersebut sulit dilakukan [6]. Banyak penelitian *machine learning* yang menggunakan pembelajaran *supervised learning,* salah satunya adalah metode *support vector machine* [7]. *Support Vector Machines* (SVM) adalah suatu metode yang handal dalam menyelesaikan masalah klasifikasi data [8]. SVM memiliki prinsip dasar linear classifier yaitu kasus klasifikasi yang secara linier, namun SVM telah dikembangkan agar dapat bekerja pada problem non-linier dengan memasukkan konsep *kernel* pada ruang kerja berdimensi tinggi [9].

Pada klasifikasi buku perpustakaan, SVM berperan sebagai algoritma yang dapat memahami pola dan karakteristik judul buku sehingga dapat mengelompokkan judul tersebut ke dalam kategori *Dewey Decimal Classification* (DDC) yang sesuai dengan pola data pelatihan. Sistem ini memperhatikan prinsip pemanfaatan secara berulang semua jenis koleksi yang ada di perpustakaan, sehingga memerlukan suatu sistem yang sanggup menyimpan sebanyak mungkin data atau informasi, untuk kemudian bisa dipanggil kembali jika dibutuhkan [10]. Penggunaan model SVM akan mengolah data menjadi data latih dan data uji untuk pengklasifikasian judul buku berdasarkan *Dewey Decimal Classification* (DDC) dengan 10 kelas. Penggunaan SVM dengan tepat memungkinkan untuk dapat melakukan klasifikasi terhadap banyak kelas, karena itu dapat dimanfaatkan dalam pemberian multi label pada sebuah dokumen [11].

Sistem DDC akan digunakan sebagai acuan dalam pengklasifikasian buku. Sistem DDC atau pengelompokan persepuluh *dewey decimal classification* merupakan suatu aturan pengklasifikasian buku yang sering digunakan secara umum pada perpustakaan [12]. Klasifikasi DDC termasuk sistem klasifikasi fundamental yang menganut prinsip desimal untuk membagi semua bidang ilmu pengetahuan [13]. DDC menggunakan konsep angka yang mewakili bidang subjek sebagai angka tiga digit, seperti 500 untuk "Ilmu Pengetahuan" dan 510 untuk "Matematika". Biasanya, angka-angka dalam DDC memiliki berbagai kategori dengan tingkat yang berbeda dalam struktur hirarki [14].

Dalam implementasinya, pengklasifikasian dengan SVM telah dipraktikkan di beberapa bidang

dan pendekatan ini berhasil meningkatkan akurasi sistem klasifikasi. Salah satu contohnya yaitu pada penelitian sebelumnya yang telah dilakukan oleh Amalia dan Yustanti pada penelitian [15]. Penelitian tersebut berjudul "Klasifikasi Buku Menggunakan Metode Support Vector Machine pada Digital Library". Dalam penelitian ini menghasilkan bahwa metode SVM dalam klasifikasi teks pada digital library dapat digunakan dan menghasilkan akurasi tertinggi yang didapat dengan nilai 69,24% pada penggunaan *kernel linear* dibandingkan dengan kernel lainnya.

Penelitian selanjutnya yang dilakukan oleh Alwi, Putra dan Muriyatmoko pada penelitian [16] yang berjudul "Classification of Book Collections Based on DDC 23 Using Text Mining Algorithm at UNIDA Gontor Library". Penelitian ini menghasilkan bahwa klasifikasi koleksi buku dengan *Dewey Decimal Classification* edisi 23 (DDC 23) sebagai acuan sistem penomoran. Melalui penggunaan algoritma *support vector classifier* pada penelitian ini, memperoleh nilai akurasi tertinggi yaitu 72% dibandingkan dengan algoritma *multinomial nb* sebesar 59%, algoritma *random forest* sebesar 69% dan algoritma *logistic regression* sebesar 69%.

Penelitian selanjutnya yang dilakukan oleh Arifin, Enri dan Sulistiyowati pada penelitian [17] yang berjudul "Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification". Klasifikasi dengan menentukan jumlah minimal kata yang diproses dan fitur *N-Gram* berupa *Unigram* dan *Bigram* pada tahapan *transformation* dengan *Term Frequency Inverse Document Frequency* (TF-IDF). Penelitian ini menghasilkan algoritma *Support Vector Machine* (SVM) yang memiliki nilai performa terbaik pada kernel linear dengan skenario 4, pembagian 90:10 yang memiliki nilai akurasi sebesar 70%.

Tujuan dari penelitian ini yaitu mengubah sistem klasifikasi buku perpustakaan secara konvensional menjadi berbasis digital, memudahkan dalam pencarian buku dengan melakukan pengklasifikasian buku, serta menciptakan pelayanan sekolah yang lebih efektif dan efisien. Dalam melakukan klasifikasi buku akan menggunakan teknologi *machine learning* dengan algoritma *support vector machine*. Algoritma *support vector machine* yang diterapkan menggunakan tiga ekstraksi fitur yang akan di uji coba pada penelitian ini. Ketiga ekstraksi fitur tersebut yaitu, hanya TF-IDF, hanya *Word2Vec* serta kombinasi TF-IDF dan *Word2Vec*.

TF-IDF merupakan salah satu teknik ekstraksi fitur untuk melakukan pembobotan kata pada tiap dokumen dengan memperhatikan kelangkaan kata pada keseluruhan dokumen [18]. Cara kerja TF-IDF yaitu dengan menghitung frekuensi (TF) suatu kata dalam dokumen dan menggunakannya dengan *invers* dari frekuensi (IDF) dokumen. TF-IDF akan memberi bobot tinggi pada kata yang sering muncul dalam satu dokumen namun, jarang muncul di seluruh dokumen. Ekstraksi fitur TF-IDF memiliki kelebihan, dimana cocok untuk mengidentifikasi keunikan kata dalam satu dokumen dan memberikan bobot untuk kata-kata penting dalam konteks tertentu.

Word2Vec merupakan salah satu teknik ekstraksi fitur yang memungkinkan pengukuran dan pemahaman kemiripan semantik melalui representasi kata sebagai vektor berdimensi rendah. Dengan menggunakan data yang cukup, ekstraksi fitur *Word2Vec* dapat melakukan prediksi atau pengklasifikasian secara akurat berdasarkan riwayat kemunculannya [19]. Hal tersebut dapat digunakan untuk menentukan asosiasi dari sebuah kata dengan kata lainnya yang hampir mirip. Kelebihan dari ekstraksi fitur ini yaitu dapat memahami makna sebuah kata dan dapat menemukan hubungan semantik seperti sinonim, antonim dan hubungan lainnya.

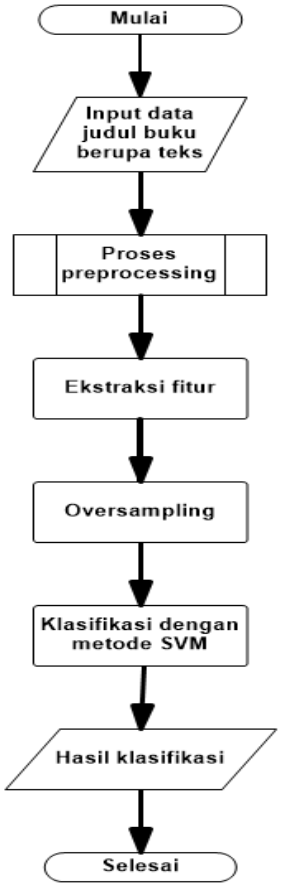
Penggunaan kombinasi ekstraksi fitur TF-IDF dan *Word2Vec* dilakukan untuk menggabungkan kelebihan dari kedua ekstraksi fitur tersebut, sehingga dapat menghasilkan klasifikasi yang akurat. Kombinasi ini dapat meningkatkan representasi semantik dokumen atau kata dengan menggabungkan kemampuan *Word2Vec* dalam menemukan makna kontekstual dengan keunggulan TF-IDF dalam memahami keunikan kata dan memberikan bobot kata. TF-IDF akan memberikan bobot kata, kemudian akan digunakan untuk menyesuaikan bobot vektor *Word2Vec*. Kombinasi ini bertujuan meningkatkan kemampuan model dan dapat meningkatkan keakuratan dalam klasifikasi buku di perpustakaan. Selain itu, penelitian ini juga bertujuan untuk memberikan kontribusi pada pemahaman teknologi dalam pengelolaan data perpustakaan di sekolah dasar. Penggunaan teknologi ini dalam jangka panjang dapat menjadi inovasi serta referensi dalam pengelolaan data perpustakaan di tingkat sekolah dasar lainnya terkhusus sekolah dasar di Kecamatan Sumowono.

Dengan mengintegrasikan teknologi machine learning pada sistem pengelolaan data perpustakaan diharapkan mampu menghasilkan sebuah sistem pengelolaan data perpustakaan yang dapat mengklasifikasikan judul buku ke dalam kategorinya yang tepat secara otomatis, sehingga dapat mempercepat akses terhadap dan informasi yang dibutuhkan. Namun, untuk menerapkan teknologi *machine learning* pada sistem pengelolaan data perpustakaan, diperlukan perancangan dan implementasi sistem yang matang. Sistem tersebut perlu dirancang agar mampu mengolah data dan informasi perpustakaan dengan lebih efektif dan efisien.

2. METODE PENELITIAN

Alur penelitian dalam melakukan klasifikasi buku perpustakaan secara otomatis akan berpanduan pada pengklasifikasian dengan sistem klasifikasi

Dewey Decimal Classification (DCC). Buku akan diklasifikasikan ke dalam 10 kelas dalam persepuluhan dewey dengan menggunakan algoritma *support vector machine*. Dalam proses pembuatannya, sistem dibuat dengan bahasa pemrograman python. Sedangkan untuk membantu dalam proses klasifikasi digunakan *library* seperti *Sklearn*, *NLTK*, *Sastrawi* dan *Gensim*. Alur penelitian ditunjukan pada gambar 1.



Gambar 1. Alur Penelitian

Alur penelitian ditunjukkan pada gambar 1, dimana penelitian dimulai dengan pengumpulan data sebagai data latih dan data uji, yang kemudian dapat dimulai dengan menginputkan judul buku baru untuk dilakukan klasifikasi. Setelah itu, data judul buku tersebut akan melalui proses *preprocessing* yang terdiri dari 4 tahap, yaitu *case folding*, *tokenizing*, *filtering* dan *stemming*. Setelah melalui proses *preprocessing*, data judul akan di ekstraksi fitur dengan tiga kali percobaan yaitu menggunakan TF-IDF, *Word2Vec* dan kombinasi TF-IDF dan *Word2Vec*. Setelah itu, untuk menangani *imbalanced data* (data tidak seimbang) maka digunakan teknik *oversampling* menggunakan metode SMOTE. Setelah proses *oversampling*, pemodelan klasifikasi dilakukan dengan menggunakan algoritma *Support Vector Machine* (SVM). Hasil klasifikasi dari pemodelan kemudian akan dievaluasi dengan menggunakan *cross validation*. Berikut ini penjelasan secara lebih rinci setiap tahapan dari alur penelitian yang akan dilakukan:

2.1. Pengumpulan Data

Data diperoleh berdasarkan observasi langsung ke perpustakaan SD Negeri Pledokan, Kabupaten Semarang sehingga data yang diolah dan diproses dalam penelitian ini merupakan data yang dikumpulkan satu persatu dan disusun berdasarkan persepuluhan dewey atau subjek umum pada sistem klasifikasi DDC. Dataset yang digunakan sejumlah 1.435 *record* data sebagai data latih dan data uji serta 552 *record* data sebagai *data rules* (aturan data). Aturan data atau *data rules* merupakan suatu aturan yang digunakan untuk mengelola atau memproses data sebelum digunakan dalam pelatihan atau pengujian model. Dimana data yang digunakan memiliki sepuluh kategori yang disesuaikan dengan *Dewey Decimal Classification* (DDC) sebagai berikut:

Tabel 1. Klasifikasi DDC

Kode DDC	Klasifikasi	Buku
000	Karya Umum	ilmu pengetahuan umum, komputer, bibliografi, ilmu perpustakaan, ensiklopedi umum, organisasi umum, museum, jurnalisme surat kabar, kumpulan karya umum, naskah-naskah dan buku-buku langka.
100	Filsafat dan Psikologi	filsafat, metafisika, epistemologi, fenomena paranormal, aliran filsafat khusus, psikologi, kepribadian, etika, logika, filsafat kuno, filsafat barat dan modern.
200	Agama	agama, agama alam, alkitab, teologi kristen, moral kristen, teologi kebaktian, gereja kristen setempat jemaat, teologi sosial kristen, sejarah gereja, denominasi, agama lain dan perbandingan agama.
300	Ilmu Sosial	ilmu sosial, statistik, ilmu politik, ilmu ekonomi, ilmu hukum, administrasi negara, perdagangan, komunikasi, pengangkutan, adat istiadat kebiasaan, dan etiket folkor
400	Bahasa	bahasa linguistik, bahasa indonesia, bahasa inggris, bahasa jerman, bahasa perancis, bahasa italia, bahasa spanyol, bahasa portugis, bahasa latin, bahasa yunani klasik, dan bahasa-bahasa lain.
500	Ilmu Murni dan Alam	ilmu murni, matematika, astronomi, fisika, kimia, ilmu pengetahuan tentang bumi, paleontologi, ilmu tentang kehidupan, ilmu tentang tumbuh-tumbuhan, dan ilmu tentang hewan.
600	Ilmu Terapan dan Teknologi	teknologi, ilmu kedokteran, ilmu teknik, pertanian, kesejahteraan rumah tangga, manajemen, teknologi kimia, pabrik-pabrik, pembuatan produk untuk penggunaan khusus, dan bangunan.
700	Kesenian, Hiburan dan Olahraga	kesenian, seni dekorasi, seni perkotaan dan pertamanan, arsitektur, seni plastik, seni pahat patung, menggambar dan seni dekorasi, seni lukis dan lukisan, seni grafika, cetakan, fotografi dan foto, musik, seni rekreasi dan pertunjukan.

800	Kesusasteraan	kesusasteraan, kesusasteraan indonesia, kesusasteraan inggris, kesusasteraan jerman, kesusasteraan perancis, kesusasteraan spanyol, kesusasteraan portugis, kesusasteraan latin, kesusasteraan yunani, dan kesusasteraan bahasa-bahasa lain.
900	Geografi dan Sejarah Umum	geografi umum, sejarah umum, geografi umum perjalanan, biografi umum, silsilah, sejarah dunia purba, sejarah umum asia, sejarah umum eropa, sejarah umum afrika, sejarah umum amerika utara, sejarah umum amerika selatan, dan sejarah umum bagian lain dari bumi.

Dataset yang digunakan merupakan sekumpulan data buku yang akan diklasifikasikan berdasarkan judul buku dalam bentuk teks. Data tersebut telah dilabeli sebelumnya dengan salah satu dari sepuluh kategori pada tabel 1. Sepuluh kategori ini nantinya akan menjadi kelas dalam proses klasifikasi buku secara otomatis. Setiap kelas tidak diisi dengan jumlah data yang sama di setiap kelasnya. Jumlah data pada setiap kelas dipresentasikan pada tabel 2.

Tabel 2. Jumlah Data Setiap Kelas

Kode DDC	Klasifikasi	Jumlah
000	Karya Umum	60 Data
100	Filsafat dan Psikologi	34 Data
200	Agama	74 Data
300	Ilmu Sosial	184 Data
400	Bahasa	18 Data
500	Ilmu Murni dan Alam	244 Data
600	Ilmu Terapan dan Teknologi	249 Data
700	Kesenian, Hiburan dan Olahraga	167 Data
800	Kesusasteraan	345 Data
900	Geografi dan Sejarah Umum	60 Data
TOTAL		**1.435 Data**

2.2. Pre-Processing

Proses *pre-processing* merupakan salah satu bagian awal dimana data yang awalnya tidak terstruktur menjadi terstruktur dengan melakukan 4 tahapan yaitu, *case folding*, *tokenizing*, *filtering* dan *stemming*. Melalui proses *pre-processing* ini, diharapkan dapat meningkatkan hasil pengukuran pada tahap evaluasi model. Tahap pertama yaitu *case folding. Case folding* bertujuan untuk mengubah seluruh kalimat yang semula menggunakan *uppercase* (huruf kapital) menjadi *lowercase* (huruf kecil). Pada tahap ini juga menghapus setiap kalimat yang mengandung *noise* yang berupa angka, tanda baca ataupun simbol lainnya. Tahap kedua yaitu *tokenizing*. Pada tahap ini, setiap kalimat akan dipecah menjadi satu kata yang disebut sebagai token. Kemudian tahap ketiga yaitu *filtering*. Tahap *filtering* disebut juga stopwords dimana pada tahap ini akan menghapus kata-kata umum yang sering muncul namun tidak memberikan makna dan tidak berpengaruh dalam sebuah kalimat. Setiap token akan dicek dengan kata-kata yang ada pada *stopwords*, jika token sesuai dengan salah satu kata, maka token tersebut akan dihapus. Tahap terakhir yaitu *stemming*. Pada tahap ini, setiap kata hasil *stopwords* akan dihilangkan semua imbuhan kata dan diubah ke dalam kata bentuk dasar menggunakan bantuan *library sastrawi*. *Library sastrawi* dapat memudahkan pengolahan kata dalam bahasa Indonesia dan membantu mempersiapkan data sebelum digunakan.

2.3. Ekstraksi Fitur

Ekstraksi fitur pada klasifikasi buku menggunakan SVM merupakan suatu proses untuk menemukan fitur-fitur penting dari teks buku yang akan digunakan sebagai basis untuk klasifikasi. Fitur-fitur ini mewakili karakteristik dari setiap klasifikasi buku dan membedakannya dari klasifikasi buku yang lainnya. Proses ekstraksi fitur dapat dilakukan dengan beberapa metode, contohnya adalah dengan menggunakan metode TF-IDF dan *Word2Vec*.

TF-IDF merupakan metode untuk menentukan bobot kata pada teks berdasarkan frekuensi kemunculan kata dan bobot invers dari frekuensi kemunculan kata-kata tersebut di seluruh kalimat. Metode ini mempertimbangkan beberapa aspek dalam pemilihan fitur, seperti relevansi dan keterulangan kata-kata dalam kumpulan kalimat. Sementara itu, *Word2Vec* merupakan metode *deep learning* yang digunakan untuk mencari relasi antara kata-kata dalam teks berdasarkan distribusi kemunculan kata-kata tersebut dalam kalimat. Metode ini memodelkan setiap kata dalam bentuk vektor yang merepresentasikan makna dan keterkaitannya dengan kata-kata lainnya.

Persamaan (1) merupakan persamaan untuk mencari *Term Frequency Inverse Document Frequency* (TF-IDF).

Pada persamaan (1), $f_{(t,d)}$ merupakan frekuensi kemunculan pada kata t dalam kalimat d. Sedangkan N merupakan jumlah kalimat dan Nt merupakan jumlah kalimat yang mengandung kata t.

$$TF - IDF_{(t,d)} = f_{(t,d)} * log(\frac{N}{Nt}) \quad (1)$$

Persamaan (2) merupakan persamaan untuk mencari vektor kata menggunakan ekstraksi fitur Word2Vec.

Pada persamaan (2), n merupakan jumlah kata pada kalimat. Sedangkan v_t merupakan vektor dari kata t.

$$V_{(t)} = \frac{1}{n} \Sigma_{i=1}^{n} v \quad (2)$$

Sehingga dari kedua persamaan diatas, dapat disimpulkan menjadi sebuah persamaan kombinasi antara TF-IDF dan Word2Vec dalam Persamaan (3).

Pada persamaan (3), $TF - IDF_{(t,d)}$ merupakan hasil pembobotan kata dalam proses TF-IDF. Sedangkan $V_{(t)}$ merupakan hasil vektor dari kata.

$$Kombinasi_{(t,d)} = TF - IDF_{(t,d)} * V_{(t)} \quad (3)$$

Proses ekstraksi fitur dilakukan dengan mengkombinasikan kedua metode tersebut untuk menemukan fitur penting dari teks judul buku dan memodelkannya dalam bentuk vektor. Setelah fitur-fitur ditemukan, SVM digunakan untuk melakukan klasifikasi buku berdasarkan fitur-fitur yang sudah diekstrak tadi. Metode ekstraksi fitur yang tepat dapat membantu mengidentifikasi karakteristik khas dari setiap klasifikasi buku sehingga memudahkan dalam klasifikasi yang akurat. Dengan demikian, penggunaan kombinasi TF-IDF dan *Word2Vec* pada ekstraksi fitur dapat meningkatkan akurasi dalam klasifikasi buku menggunakan SVM.

2.4. Penanganan Data Tidak Seimbang

Data yang tidak seimbang atau *imbalanced data* mengacu pada situasi dimana jumlah sampel dalam salah satu kelas jauh lebih banyak dibandingkan dengan kelas yang lain. Kondisi imbalanced data dapat mempengaruhi kinerja model, sehingga membuat akurasi model menurun. Penanganan data yang tidak seimbang dapat dilakukan dengan menggunakan teknik *oversampling*. Teknik *oversampling* dapat digunakan untuk menangani *imbalanced data* dengan membuat duplikasi atau membuat sampel sintetis dari kelas lain sebanyak jumlah sampel dari kelas dengan sampel yang terbanyak, sehingga jumlah sampel dapat seimbang dengan kelas yang memiliki jumlah sampel terbanyak tersebut. Salah satu metode yang digunakan adalah *Synthetic Minority Over-Sampling Technique* (SMOTE). *Oversampling* dengan SMOTE efektif untuk meningkatkan akurasi model. Selain itu, teknik ini juga dapat mencegah *overfitting* terhadap data latih dan mengurangi tingkat kesalahan prediksi pada kelas dengan nilai paling sedikit yang sering terabaikan.

2.5. Pemodelan

Data yang telah dibersihkan melalui proses *pre-processing*, kemudian di ekstraksi fitur dan di *oversampling* akan dimasukkan ke dalam model. Pemodelan klasifikasi buku dengan menggunakan *Support Vector Machine* (SVM) berdasarkan kernel linear dan parameter C = 1 memiliki tujuan untuk mengklasifikasikan buku ke dalam salah satu dari 10 kelas yang ada. SVM adalah algoritma *machine learning* yang populer untuk klasifikasi dan regresi. Ketika digunakan untuk klasifikasi, SVM mencari hyperplane yang memiliki jarak maksimum dengan dua kelas yang berbeda. *Kernel linear* merupakan salah satu tipe kernel yang umum digunakan pada SVM. Kernel ini digunakan untuk memetakan data ke dalam ruang fitur baru secara *linear*, sehingga mempermudah SVM untuk menemukan *hyperplane* pemisah di antara kelas-kelas yang berbeda. Dalam klasifikasi buku, *kernel linear* dapat digunakan untuk memetakan data buku ke dalam fitur-fitur yang berkaitan dengan logika *linear*, seperti jumlah kata, frekuensi kata tertentu dan sebagainya. Kemudian, SVM dengan *kernel linear* dapat digunakan untuk mengklasifikasikan data buku ke dalam salah satu dari 10 kelas yang berbeda dengan margin maksimum. Dalam klasifikasi buku juga menggunakan parameter C. Parameter C adalah parameter penalti yang digunakan untuk mengontrol penyeimbangan antara mencapai margin maksimum dan mengurangi kesalahan klasifikasi.

Langkah pertama dalam pemodelan yaitu membagi data buku menjadi data pelatihan (data training) dan data pengujian (data testing). Pembagian yang digunakan adalah skala 80:20, dimana 80% data digunakan untuk pelatihan dan 20% data digunakan untuk pengujian. Pembagian ini bertujuan agar model dapat dilatih dengan baik dan diuji secara objektif. Setelah data dibagi, langkah selanjutnya yaitu melakukan *labeling* pada buku berdasarkan klasifikasinya. Jika ada 10 kelas yang akan diklasifikasikan, maka setiap judul buku memiliki label klasifikasi yang sesuai dengan salah satu dari 10 kelas tersebut.

Model SVM dengan kernel linear dilatih untuk mengklasifikasikan sampel data latih ke dalam salah satu dari 10 kelas yang berbeda. Model yang telah dilatih akan dievaluasi menggunakan data uji untuk mengevaluasi performa model dalam mengklasifikasikan sampel data larih ke dalam 10 kelas yang berbeda. Matriks evaluasi seperti akurasi, presisi, recall dan F1-score akan digunakan untuk mengukur keberhasilan model dalam mengklasifikasikan buku.

2.6. Validasi Silang

Validasi silang atau *cross validation* adalah metode evaluasi performa suatu model dengan membagi dataset menjadi k data yang berbeda, dimana setiap data digunakan secara bergantian sebagai data uji dan data latih dalam k iterasi berbeda. *Cross validation* digunakan dalam *machine learning* untuk mencegah kebocoran data yang dapat terjadi ketika dataset dibagi menjadi data latih dan data uji secara acak. Dalam divisi acak, kemungkinan terjadinya tumpang tindih pada dara yang sama diantara dataset latih dan uji sangatlah besar. Kondisi tersebut dapat mempengaruhi performa model secara tidak adil dan mengakibatkan *overfitting* atau *underfitting* pada model. Dengan menggunakan *cross validation*, dapat meningkatkan keandalan dan akurasi performa model dengan mengurangi dampak dari kebocoran data dan dapat memberikan estimasi terbaik untuk performa rata-rata model pada semua subset data.

Salah satu metode *cross validation* yang paling umum digunakan pada *machine learning* yaitu metode *K-Fold Cross Validation*. Metode ini memungkinkan untuk mendapatkan hasil evaluasi performa model yang obyektif dan akurat. Dengan membagi dataset menjadi k data, dan melakukan evaluasi pada setiap data secara bergantian sebagai

data uji dan data latih, metode *K-Fold Cross Validation* akan memastikan bahwa model telah diuji pada semua data, tanpa kecenderungan perubahan secara acak yang mungkin terjadi saat melakukan divisi acak pada dataset. Dalam metode ini, dataset dibagi menjadi k data yang sama besar, umumnya k = 5 atau k = 10. Pada setiap iterasi, satu data digunakan sebagai dataset uji dan data yang lainnya menjadi dataset latih. Model dilatih pada data latih dan dinilai pada data uji. Performa model dihitung dengan mengambil nilai rata-rata hasil semua iterasi.

3. HASIL DAN PEMBAHASAN

3.1. Data

Data yang telah dikumpulkan dari perpustakaan SD Negeri Pledokan merupakan data awal, dimana terdiri dari beberapa banyak kolom. Kemudian data yang dibutuhkan dalam penelitian adalah kolom judul dan kategori. Sebelum melalui tahapan *pre-processing*, data terlebih dahulu dilakukan pengecekan data kosong atau *NaN*. Selanjutnya, data pada kolom judul akan melalui tahapan *pre-processing*. Setelah dilakukan tahap *preprocessing*, data akan melalui tahap ekstraksi fitur dengan menggunakan kombinasi TF-IDF dan *Word2Vec*. Data pada kolom kategori akan melalui tahap labeling data sebanyak 10 kelas sebelum masuk ke tahap pemodelan. Kolom kategori ini merupakan kolom yang akan menentukan hasil klasifikasi buku, dimana telah disesuaikan dengan kaidah persepuluhan *Dewey Decimal Classification* (DDC).

Pada penelitian ini menggunakan 1.435 *record* data ditambah dengan 552 *record data rules*, sehingga keseluruhan data yang diproses berjumlah 1.987 *record* data. Dimana data tersebut akan dipisahkan sebagai data latih dan data uji. Pembagian data dilakukan secara acak dengan kombinasi skala 80:20, sehingga data latih berjumlah 1.590 *record* data dan data uji berjumlah 397 *record* data.

3.2. Hasil Pengujian Dengan Ekstraksi Fitur TF-IDF

Pada penelitian ini, pengujian pertama menggunakan *TF-IDF* sebagai metode ekstraksi fitur. Metode ini memperhitungkan frekuensi kata tertentu dalam sebuah dokumen dan invers dari frekuensi kata tersebut di seluruh kumpulan dokumen untuk menyatakan seberapa besar kontribusi kata tertentu dalam sebuah dokumen terhadap makna keseluruhan. Penggunaan *TF-IDF* sebagai metode ekstraksi fitur bertujuan untuk menyorot kata kunci yang memiliki kepentingan tinggi, menangkap keunikan setiap dokumen, dan memberikan representasi fitur yang menjadi dasar analisis lebih lanjut. Hasil pengujian dapat dilihat pada tabel 3.

Hasil pengujian dengan menggunakan ekstraksi fitur TF-IDF yang dipresentasikan pada tabel 3. Hasil pengujian mendapatkan nilai akurasi sebesar 64%, nilai *precision* sebesar 72%, nilai *recall* sebesar 62%, dan nilai *F1-Score* sebesar 65%. Adapun pada Tabel 3 merupakan hasil dari perhitungan dengan ekstraksi fitur TF-IDF dan dimodelkan dengan algoritma *Support Vector Machine*.

Tabel 3. Hasil Uji dengan TF-IDF

Kelas	Precision	Recall	F1-Score
000 - Karya Umum	86%	89%	87%
100 - Filsafat dan Psikologi	65%	50%	57%
200 - Agama	79%	37%	50%
300 - Ilmu Sosial	75%	63%	69%
400 - Bahasa	100%	75%	86%
500 - Ilmu Murni dan Alam	70%	60%	65%
600 - Ilmu Terapan dan Teknologi	67%	65%	66%
700 - Kesenian, Hiburan dan Olahraga	71%	69%	70%
800 - Kesusasteraan	42%	74%	54%
900 - Geografi dan Sejarah Umum	62%	36%	45%
AVG	**72%**	**62%**	**65%**
Accuracy		**64%**	

3.3. Hasil Pengujian Dengan Ekstraksi Fitur Word2Vec

Pada tahap pengujian kedua ini, pendekatan berbeda terhadap metode ekstraksi fitur menggunakan *Word2Vec*. *Word2Vec* adalah teknologi pemrosesan bahasa alami yang mengubah kata menjadi vektor dalam ruang berdimensi rendah, memungkinkan ekspresi semantik yang lebih kaya. Dalam penelitian ini, *Word2Vec* digunakan untuk menghasilkan representasi vektor kata yang mencerminkan hubungan semantik antar kata dalam dokumen. Metode ini membantu mengatasi beberapa keterbatasan metode lain, terutama dalam menangkap makna kontekstual dan hubungan semantik yang lebih dalam antar kata. Pengujian kedua ini bertujuan untuk membandingkan dan melengkapi hasil pengujian pertama dengan TF-IDF. Hasil pengujian dapat dilihat pada tabel 4.

Tabel 4. Hasil Uji dengan Word2Vec

Kelas	Precision	Recall	F1-Score
000 - Karya Umum	53%	33%	41%
100 - Filsafat dan Psikologi	22%	19%	20%
200 - Agama	12%	40%	19%
300 - Ilmu Sosial	30%	18%	22%
400 - Bahasa	80%	50%	62%
500 - Ilmu Murni dan Alam	60%	5%	9%
600 - Ilmu Terapan dan Teknologi	36%	39%	38%
700 - Kesenian, Hiburan dan Olahraga	71%	33%	45%
800 - Kesusasteraan	28%	50%	36%
900 - Geografi dan Sejarah Umum	50%	21%	30%
AVG	**44%**	**31%**	**32%**
Accuracy		**30%**	

Hasil pengujian dengan menggunakan ekstraksi fitur *Word2Vec* yang dipresentasikan pada tabel 4. Hasil pengujian mendapatkan nilai akurasi sebesar

30%, nilai *precision* sebesar 44%, nilai *recall* sebesar 31%, dan nilai *F1-Score* sebesar 32%. Adapun pada Tabel 4 merupakan hasil dari perhitungan dengan ekstraksi fitur Word2Vec dan dimodelkan dengan algoritma Support Vector Machine, dimana hasil akurasi yang didapatkan lebih rendah 34% dari tingkat akurasi pada pengujian dengan ekstraksi fitur TF-IDF. Perbedaan tingkat akurasi disebabkan oleh karakteristik khusus dari kumpulan data yang digunakan atau kompleksitas hubungan semantik yang sulit ditangkap dengan model Word2Vec. Faktor-faktor tersebut dapat mempengaruhi kinerja algoritma SVM dalam mengklasifikasikan data meskipun Word2Vec memberikan representasi vektor kata yang komprehensif.

3.4. Hasil Pengujian Dengan Kombinasi TF-IDF dan Word2Vec

Pada tahap pengujian ketiga ini, pendekatan inovatif diterapkan menggunakan TF-IDF yang dikombinasikan dengan ekstraksi fitur *Word2Vec*. Kombinasi ekstraksi fitur ini, bertujuan memanfaatkan kelebihan kedua metode tersebut dengan harapan dapat meningkatkan kemampuan model dalam mengekstraksi dan memahami informasi dari kata-kata. TF-IDF memberikan bobot kata yang mencerminkan makna sebuah kata dalam dokumen, sedangkan *Word2Vec* menyediakan representasi vektor kata yang kaya akan konteks semantik. Kombinasi ini diharapkan dapat mengatasi keterbatasan dari masing-masing metode. Hasil pengujian dapat dilihat pada tabel 5.

Tabel 5. Hasil Uji dengan Kombinasi TF-IDF dan Word2Vec

Kelas	Precision	Recall	F1-Score
000 - Karya Umum	100%	85%	92%
100 - Filsafat dan Psikologi	80%	46%	59%
200 - Agama	95%	64%	77%
300 - Ilmu Sosial	79%	67%	72%
400 - Bahasa	90%	90%	79%
500 - Ilmu Murni dan Alam	83%	72%	77%
600 - Ilmu Terapan dan Teknologi	71%	78%	75%
700 - Kesenian, Hiburan dan Olahraga	90%	70%	79%
800 - Kesusasteraan	51%	84%	63%
900 - Geografi dan Sejarah Umum	93%	64%	76%
AVG	**83%**	**72%**	**76%**
Accuracy		**73%**	

Hasil pengujian dengan menggunakan kombinasi dari ekstraksi fitur TF-IDF dan *Word2Vec* yang dipresentasikan pada tabel 5, dimana hasil pengujian mendapatkan nilai akurasi sebesar 73%, nilai *precision* sebesar 83%, nilai *recall* sebesar 72%, dan nilai *F1-Score* sebesar 76%. Adapun pada Tabel 5 merupakan hasil dari perhitungan dengan kombinasi ekstraksi fitur TF-IDF dan *Word2Vec* serta dimodelkan dengan algoritma *Support Vector Machine*. Saat ekstraksi fitur diuji secara kombinasi, nilai akurasi yang diperoleh menunjukkan performa model yang lebih unggul dibandingkan saat menggunakan ekstraksi fitur secara tunggal yaitu hanya TF-IDF ataupun hanya *Word2Vec*. Hal ini disebabkan oleh kemampuan model dalam memanfaatkan kelebihan dari kedua metode ekstraksi fitur tersebut. Pemahaman arti kata *Word2Vec* dan penyorotan kata kunci *TF-IDF* bekerja sama untuk meningkatkan kinerja model dalam mengklasifikasikan judul buku. Namun, pada hasil pengklasifikasian terdapat sedikit kesalahan klasifikasi pada kelas filsafat dan psikologi yang diklasifikasikan menjadi kelas kesusasteraan.

3.5. Hasil Evaluasi Cross Validation

Pada penelitian ini, metode *cross validation* atau validasi silang digunakan untuk mengevaluasi performa model. Teknik ini memberikan pandangan yang lebih objektif tentang seberapa baik model menggeneralisasi data yang tidak terlihat selama pelatihan. Dengan menerapkan teknik validasi silang, dapat mengurangi kemungkinan terjadinya *overfitting* atau *underfitting* yang mungkin terjadi selama konstruksi model. Oleh karena itu, hasil evaluasi menggunakan teknik validasi silang memberikan pemahaman yang lebih lengkap tentang seberapa andal model tersebut dan sejauh mana model tersebut dapat digeneralisasi untuk menangani data baru yang belum pernah ada sebelumnya. Hasil evaluasi *cross validation* sebanyak lima kali iterasi dapat dilihat pada tabel 6.

Tabel 6. Hasil Evaluasi

Iterasi	Accuracy	Precision	Recall	F1 - Score
Iterasi ke-1	85%	87%	85%	86%
Iterasi ke-2	85%	86%	85%	85%
Iterasi ke-3	84%	86%	84%	85%
Iterasi ke-4	85%	86%	85%	85%
Iterasi ke-5	84%	85%	84%	84%
Rata-Rata		**85%**		

Evaluasi model dengan kombinasi ekstraksi fitur TF-IDF dan *Word2Vec* dilakukan menggunakan metode *K-Fold Cross Validation* dengan nilai k = 5. Pada iterasi ke-1 menghasilkan nilai akurasi 85%, *precision* 87%, *recall* 85% dan *F1-Score* 86%. Kemudian, pada iterasi ke-2 menghasilkan nilai akurasi 85%, *precision* 86%, *recall* 85% dan *F1-Score* 85%. Kemudian, pada iterasi ke-3 menghasilkan nilai akurasi 84%, *precision* 86%, *recall* 84% dan *F1-Score* 85%. Kemudian, pada iterasi ke-4 menghasilkan nilai akurasi 85%, *precision* 86%, *recall* 85% dan *F1-Score* 85% dan terakhir pada iterasi ke-5 menghasilkan nilai akurasi 84%, *precision* 85%, *recall* 84% dan *F1-Score* 84%.

Akurasi rata-rata model sebesar 85% selama lima iterasi. Hal ini menunjukkan bahwa model memiliki konsistensi yang baik dalam mengklasifikasikan judul buku. Selain itu, dengan melihat hasil dari setiap iterasi, menunjukkan bahwa model mempunyai kemampuan klasifikasi yang stabil dan reliabel. Secara keseluruhan, evaluasi ini

memberikan bukti bahwa kombinasi ekstraksi fitur TF-IDF dan *Word2Vec* secara efektif dapat meningkatkan kemampuan model untuk mengklasifikasikan data teks. Tingkat akurasi yang tinggi dan nilai konsisten membuktikan bahwa model ini dapat digunakan dalam sistem klasifikasi buku otomatis, memberikan keyakinan akan keandalan dan kinerja model ketika berhadapan dengan keragaman dan kompleksitas data teks.

4. DISKUSI

Pada penelitian ini, dihasilkan sebuah sistem yang dapat mengklasifikasikan buku berdasarkan judul buku tersebut dengan menerapkan algoritma *machine learning Support Vector Machin*e (SVM). Dimana sistem dibangun dengan membagi buku ke dalam 10 kelas yang berbeda dengan berpedoman pada kaidah persepuluhan *Dewey Decimal Classification* (DDC). Sepuluh kelas tersebut antara lain, 000 - karya umum, 100 - filsafat dan psikologi, 200 - agama, 300 - ilmu sosial, 400 - bahasa, 500 - ilmu murni dan alam, 600 - ilmu terapan dan teknologi, 700 - kesenian, hiburan dan olahraga, 800 - kesusasteraan, serta 900 - geografi dan sejarah umum.

Proses pelatihan dan pengujian sistem dilakukan sebanyak 3 kali untuk mencari metode ekstraksi fitur terbaik yang dapat menangani klasifikasi teks 10 kelas dengan algoritma SVM. Metode ekstraksi fitur tersebut yaitu yang pertama hanya TF-IDF, yang kedua hanya *Word2Vec* dan yang ketiga kombinasi TF-IDF dan *Word2Vec*. Metode ekstraksi fitur dengan nilai akurasi terbaik dari ketiga metode tersebut, akan dilakukan validasi silang dengan *K-Fold Cross Validation* untuk mengevaluasi performa model. Setelah dilakukan penelitian, diperoleh tahapan ekstraksi fitur terbaik dengan kombinasi TF-IDF dan *Word2Vec* yang mencapai akurasi terbaik sebesar 73% dengan nilai *precision* 83%, *recall* 72%, dan *F1-Scor*e 76%.

Penelitian sebelumnya terkait dengan penggunaan algoritma SVM untuk klasifikasi, salah satunya berjudul "Klasifikasi Buku Menggunakan Metode Support Vector Machine pada Digital Library" [15]. Menggunakan data latih dan data uji sejumlah 1.000 *record* data yang dibagi menjadi 10 kelas dan metode ekstraksi fitur yang digunakan yaitu TF-IDF. Menghasilkan bahwa metode SVM dalam klasifikasi teks pada digital library mencapai akurasi tertinggi sebesar 69,24% dengan nilai *precision* 71%, *recall* 61 % dan *F1-Score* 64% pada penggunaan kernel linear. Hasil penelitian tersebut masih dibawah pencapaian akurasi dari penelitian yang telah dilakukan dan jumlah data lebih sedikit dari yang digunakan.

Penelitian selanjutnya yang berjudul "Classification of Book Collections Based on DDC 23 Using Text Mining Algorithm at UNIDA Gontor Library" [16]. Menggunakan data latih dan data uji sejumlah 1.015 *record* data yang dibagi menjadi 8 kelas dan metode ekstraksi fitur yang digunakan yaitu TF-IDF. Pengujian dilakukan sebanyak 4 kali dengan algoritma yang berbeda yaitu *Multinomial Nb*, *Random Forest*, *Logistic Regression*, dan *Support Vector Classifier*. Menghasilkan bahwa klasifikasi koleksi buku dengan *Dewey Decimal Classification* edisi 23 (DDC 23) melalui penggunaan algoritma *support vector classifier* memperoleh nilai akurasi tertinggi yaitu 72% dibandingkan dengan algoritma lainnya dan mencapai nilai *precision* 71%, *recal*l 71%, dan *F1-Score* 70%. Hasil penelitian tersebut masih dibawah pencapaian akurasi dari penelitian yang telah dilakukan dan menggunakan jumlah kelas yang lebih sedikit.

Penelitian selanjutnya yang berjudul "Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification" [17]. Klasifikasi dengan menentukan jumlah minimal kata yang diproses dan fitur *N-Gram* berupa *Unigram* dan *Bigram* pada tahapan *transformation* dengan *Term Frequency Inverse Document Frequency* (TF-IDF). Menggunakan 2 kelas sebagai pembagi data menghasilkan algoritma *Support Vector Machine* (SVM) yang memiliki nilai performa terbaik pada *kernel linear* dengan skala pembagian 90:10 yang memiliki nilai akurasi sebesar 70%. Hasil penelitian tersebut masih dibawah pencapaian akurasi dari penelitian yang telah dilakukan dan menggunakan jumlah kelas yang lebih sedikit.

Penelitian selanjutnya yang berjudul "Classification of Book Types Using the Support Vector Machine (SVM) Method" [20]. Menggunakan data latih dan data uji sejumlah 600 *record* data yang dibagi menjadi 3 kelas dan metode ekstraksi fitur yang digunakan yaitu TF-IDF. Penelitian ini menggunakan tiga kali kombinasi data latih dan data uji yang berbeda. Proses pemodelan pada penelitian ini menggunakan empat kernel yang berbeda, yaitu *kernel linear, kernel RBF, kernel polynomial*, dan *kernel sigmoid*. Menghasilkan bahwa metode SVM dalam klasifikasi tipe buku mencapai akurasi pada *kernel linear* sebesar 67,24%, *kernel RBF* sebesar 71,57%, *kernel polynomial* sebesar 58,56% dan *kernel sigmoid* sebesar 68,02%. Hasil penelitian tersebut masih dibawah pencapaian akurasi dari penelitian yang telah dilakukan dengan jumlah data lebih sedikit meskipun penelitian tersebut menggunakan algoritma SVM dengan ekstraksi fitur hanya TF-IDF.

Penelitian selanjutnya yang berjudul "Analisis Dan Implementasi Support Vector Machine Dengan String Kernel Dalam Melakukan Klasifikasi Berita Berbahasa Indonesia" [21]. Penelitian mengklasifikasikan berita berdasarkan topiknya kedalam 3 kelas. Menghasilkan bahwa metode SVM dalam klasifikasi mendapat nilai akurasi tertinggi dengan kombinasi *stopword, tokenizing, term frequency & chi-square* sebesar 47,43%. Hasil penelitian tersebut masih dibawah pencapaian akurasi

dari penelitian yang telah dilakukan dengan jumlah kelas yang lebih sedikit.

Penelitian selanjutnya yang berjudul "Klasifikasi Topik Ayat Al-Qur'an Terjemahan Berbahasa Inggris Menggunakan Metode Support Vector Machine Berbasis Vector Space Model dan Word2Vec" [22]. Dataset yang digunakan yaitu terjemahan Bahasa Inggris Al-Qur'an Al-Jalain, berisi 780 *record* ayat yang terbagi menjadi 3 kelas kategori. Penelitian menggunakan 8 komposisi data latih dan data uji yang berbeda. Pada proses pemodelan menggunakan TF-IDF sebagai pembobotan kata, kemudian vektor yang dihasilkan melalui TF-IDF akan dilakukan pemetaan kata secara semantic berdasarkan keterkaitan kata pada dokumen menggunakan *Word2Vec*. Menghasilkan bahwa metode SVM dalam klasifikasi topik terjemahan ayat Al-Qur'an mencapai nilai akurasi paling optimal pada komposisi 70:30 pada *kernel linear* sebesar 64%. Meskipun penelitian tersebut menggunakan kombinasi TF-IDF dan *Word2Vec*, hasil akurasi masih dibawah penelitian yang telah dilakukan dengan jumlah kelas yang lebih sedikit.

Kelima penelitian sebelumnya menunjukkan bahwa penelitian yang telah dilakukan ini terbukti efektif dalam mengklasifikasikan data dengan kelas yang kompleks sebanyak 10 kelas. Hasil akurasi pada penelitian ini mencapai 73%, dimana angka tersebut lebih tinggi dibandingkan dengan penelitian-penelitian sebelumnya yang juga menggunakan algoritma yang sama. Model klasifikasi buku yang dihasilkan mempunyai kemampuan klasifikasi yang stabil dan reliabel terbukti dengan nilai rata-rata evaluasi model menggunakan *cross validation* dengan 5 kali iterasi sebesar 85%.

Pada penelitian sebelumnya yang sama-sama menggunakan kombinasi TF-IDF dan *Word2Vec* dalam klasifikasi yaitu penelitian [22]. Penelitian tersebut menggunakan jumlah dataset dan jumlah kelas yang lebih sedikit daripada penelitian yang telah dilakukan. Penelitian tersebut menggunakan 780 *record* dengan 3 kelas sedangkan penelitian ini menggunakan 1.435 *record* dengan 10 kelas. Jika sama-sama menggunakan *kernel linear* dengan komposisi data latih dan data uji 80:20, maka penelitian tersebut menghasilkan nilai akurasi sebesar 57% sedangkan penelitian yang telah dilakukan ini menghasilkan nilai akurasi sebesar 73%. Penelitian yang telah dilakukan ini terbukti lebih unggul 16% dibanding penelitian sebelumnya yang serupa dengan jumlah dataset dan kelas yang dua kali lebih kompleks.

Dari hasil perbandingan tersebut, dapat disimpulkan bahwa metode yang telah diterapkan untuk mengklasifikasikan judul buku dengan 10 kelas klasifikasi yang berpedoman pada DDC dapat berjalan dengan baik. Selain itu, metode klasifikasi judul buku menggunakan algoritma SVM dengan mengkombinasikan ekstraksi fitur TF-IDF dan *Word2Vec* juga menghasilkan akurasi yang lebih tinggi. Hal tersebut menjadi keunggulan penelitian dibandingkan dengan penelitian yang telah dilakukan sebelumnya. Meskipun masih terdapat beberapa kesalahan dalam hasil klasifikasi yang tidak sesuai karena ketidakseimbangan jumlah dataset di setiap kelas klasifikasi. Pada penelitian mendatang, dapat dilakukan pengujian terhadap dataset yang lebih besar dengan kategori yang lebih banyak dan seimbang. Hal ini untuk mengetahui dan membuktikan efektivitas algoritma S*upport Vector Machine* dengan kombinasi metode ekstraksi fitur TF-IDF dan *Word2Vec* dalam mengklasifikasikan teks secara kompleks.

5. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dimana penelitian menggunakan 1.435 *record* data buku ditambah dengan 552 *record* d*ata rules* sebagai data latih dan data uji dengan komposisi skala 80:20, terdiri dari 10 kelas yang telah disesuaikan dengan kaidah persepuluhan *Dewey Decimal Classification* (DDC). Tujuan dari penelitian ini yaitu mengubah sistem klasifikasi buku perpustakaan secara konvensional menjadi berbasis digital, memudahkan dalam pencarian buku dengan melakukan pengklasifikasian buku otomatis, serta menciptakan pelayanan sekolah yang lebih efektif dan efisien. Setelah dilakukan pengujian algoritma *Support Vector Machine* (SVM) sebanyak 3 kali dengan menggunakan TF-IDF saja, *Word2Vec* saja dan kombinasi TF-IDF dan *Word2Vec*, dapat disimpulkan bahwa algoritma SVM cocok untuk digunakan pada sistem klasifikasi judul buku otomatis. Nilai akurasi tertinggi yaitu 73% terdapat pada metode klasifikasi SVM dengan menggunakan kombinasi ekstraksi fitur TF-IDF dan *Word2Vec* dibandingkan dengan metode ekstraksi fitur lainnya, dimana memiliki nilai *precision* 83%, *recall* 72%, dan *F1-Score* sebesar 76%.

6. UCAPAN TERIMA KASIH

Terima kasih kepada SD Negeri Pledokan, Kecamatan Sumowono, Kabupaten Semarang yang telah memberikan izin dan kerjasamanya dalam mengumpulkan data dan sampel.

Terima kasih kepada Universitas Dian Nuswantoro yang telah memberikan kesempatan, dukungan dan fasilitas yang luar biasa dalam penulisan jurnal ini. Dukungan akademik dan teknis yang diberikan oleh Universitas Dian Nuswantoro telah memberikan kontribusi yang signifikan dalam keberhasilan penelitian ini.

DAFTAR PUSTAKA

[1] S. Nanda, "Perkembangan Trend Terbaru Dalam Temu Kembali Informasi Bagi Mahasiswa Pascasarjana Uin Sunan Kalijagayogyakarta," *shaut*, vol. 11, no. 2, pp. 198–209, Jan. 2020, doi:

10.37108/shaut.v11i2.251.

[2] Martinus Maslim and Stephanie Pamela Adithama, "Pembangunan Sistem Informasi Perpustakaan Sekolah Dasar Berbasis Web," *dinamisia*, vol. 3, no. 2, pp. 350–360, Jan. 2020, doi: 10.31849/dinamisia.v3i2.3073.

[3] D. Anggoro and A. Hidayat, "Rancang Bangun Sistem Informasi Perpustakaan Sekolah Berbasis Web Guna Meningkatkan Efektivitas Layanan Pustakawan," *EDUMATIC*, vol. 4, no. 1, pp. 151–160, Jun. 2020, doi: 10.29408/edumatic.v4i1.2130.

[4] A. O. P. Dewi, "Kecerdasan Buatan sebagai Konsep Baru pada Perpustakaan," *Anuva*, vol. 4, no. 4, pp. 453–460, Nov. 2020, doi: 10.14710/anuva.4.4.453-460.

[5] A. Le Glaz *et al.*, "Machine Learning and Natural Language Processing in Mental Health: Systematic Review," *J Med Internet Res*, vol. 23, no. 5, p. e15708, May 2021, doi: 10.2196/15708.

[6] P. Shiroya, "Book Genre Categorization Using Machine Learning Algorithms (K-Nearest Neighbor, Support Vector Machine and Logistic Regression) using Customized Dataset," *IJCSMC*, vol. 10, no. 3, pp. 14–25, Mar. 2021, doi: 10.47760/ijcsmc.2021.v10i03.002.

[7] A. A. Hermawan, G. W. Saraswati, and E. Kartikadarma, "Metode MICE Support Vector Machine (MICE-SVM) untuk Klasifikasi Performance Mahasiswa Merdeka Belajar Kampus Merdeka," *MIB*, vol. 7, no. 4, pp. 1686–1697, Oct. 2023, doi: 10.30865/mib.v7i4.6821.

[8] A. A. Kasim and M. Sudarsono, "Algoritma Support Vector Machine (SVM) untuk Klasifikasi Ekonomi Penduduk Penerima Bantuan Pemerintah di Kecamatan Simpang Raya Sulawesi Tengah," in *Seminar Nasional APTIKOM (SEMNASTIK) 2019*, Semarang, Indonesia: APTIKOM Universitas Dian Nuswantoro, Nov. 2019, pp. 568–573.

[9] F. Pratama, M. Nasir, and S. Sauda, "Implementasi Metode Klasifikasi Dengan Algoritma Support Vector Machine Untuk Menentukan Stok Persediaan Barang Pada Koperasi Karyawan Pangan Utama," *Journal-SEA*, vol. 1, no. 2, pp. 71–81, May 2020, doi: 10.51519/journalsea.v1i2.46.

[10] Musrifah, "Strategi Pengembangan Sistem Temu Kembali Informasi Berbasis Gambar (Content Based Image Retrieval System) Di Perpustakaan Perguruan Tinggi Kedokteran," *JIPI*, vol. 3, no. 1, pp. 1–20, 2018, doi: 10.30829/jipi.v3i1.1486.

[11] K. I. Gunawan and J. Santoso, "Multilabel Text Classification Menggunakan SVM dan Doc2Vec Classification Pada Dokumen Berita Bahasa Indonesia," *INSIGHT*, vol. 3, no. 01, pp. 29–38, Apr. 2021, doi: 10.37823/insight.v3i01.126.

[12] Mardianto, Maryaningsih, and R. Supardi, "Book Classification Application Using Dewey Decimal Classification Method (DDC) Case Study Library SMAN 4 Kaur," *JKOMITEK*, vol. 1, no. 2, pp. 290–298, 2021.

[13] D. B. Anggraeni, Widyastuti, F. P. Rahmawati, and M. G. Aditama, "Pengembangan Sistem Klasifikasi Kepustakaan dengan Dewey Decimal Classification (DDC)," *Buletin KKN Pendidikan*, vol. 3, no. 2, pp. 152–160, Dec. 2021, doi: 10.23917/bkkndik.v3i2.15734.

[14] K. Puritat and K. Intawong, "Development of an Open Source Automated Library System with Book Recommedation System for Small Libraries," in *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, Pattaya, Thailand: IEEE, Mar. 2020, pp. 128–132. doi: 10.1109/ECTIDAMTNCON48261.2020.9090753.

[15] D. H. Amalia and W. Yustanti, "Klasifikasi Buku Menggunakan Metode Support Vector Machine pada Digital Library," *JINACS*, vol. 3, no. 01, pp. 55–61, Aug. 2021, doi: 10.26740/jinacs.v3n01.p55-61.

[16] Muhammad Alwi, Oddy Virgantara Putra, and Dihin Muriyatmoko1, "Classification of Book Collections Based on DDC 23 Using Text Mining Algorithm at UNIDA Gontor Library," *PELS*, vol. 2, Nov. 2021, doi: 10.21070/pels.v2i0.1164.

[17] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification," *STRING*, vol. 6, no. 2, p. 129, Dec. 2021, doi: 10.30998/string.v6i2.10133.

[18] A. T. Ni'mah and A. Z. Arifin, "Perbandingan Metode Term Weighting terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis," *Rekayasa*, vol. 13, no. 2, pp. 172–180, Aug. 2020, doi: 10.21107/rekayasa.v13i2.6412.

[19] E. Suryati, Styawati, and A. A. Aldino, "Analisis Sentimen Transportasi Online Menggunakan Ekstraksi Fitur Model Word2vec Text Embedding Dan Algoritma Support Vector Machine (SVM)," *JUTISI*, vol. 4, no. 1, pp. 96–106, Mar. 2023, doi:

10.33365/jtsi.v4i1.2445.

[20] F. Riandari, H. T. Sihotang, T. Tarigan, and M. Rafli, "Classification of Book Types Using the Support Vector Machine (SVM) Method," *Mantik*, vol. 6, no. 1, pp. 42–49, Mar. 2022.

[21] Honakan, Adiwijaya, and S. Al Faraby, "Analisis Dan Implementasi Support Vector Machine Dengan String Kernel Dalam Melakukan Klasifikasi Berita Berbahasa Indonesia," in *eProceedings of Engineering*, in 1, vol. 5. Bandung, Indonesia: Telkom University Open Library, Mar. 2018, pp. 1701–1710.

[22] A. Salama, Adiwijaya, and S. Al Faraby, "Klasifikasi Topik Ayat Al-Qur'an Terjemahan Berbahasa Inggris Menggunakan Metode Support Vector Machine Berbasis Vector Space Model dan Word2Vec," in *e-Proceeding of Engineering*, in 2, vol. 6. Bandung, Indonesia: Telkom University Open Library, Aug. 2019, pp. 9133–9140..