

**KLASIFIKASI KELOMPOK BIDANG KEAHLIAN (KBK)  
BERDASARKAN JUDUL DAN ABSTRAK SKRIPSI MENGGUNAKAN  
ALGORITMA SUPPORT VECTOR MACHINE**

**SKRIPSI**

**OLEH  
MERCYANO DANDI HIDAYAT  
NIM 190535646051**



**UNIVERSITAS NEGERI MALANG  
FAKULTAS TEKNIK  
PROGRAM STUDI TEKNIK INFORMATIKA  
DESEMBER 2022**



**KLASIFIKASI KELOMPOK BIDANG KEAHLIAN (KBK)  
BERDASARKAN JUDUL DAN ABSTRAK SKRIPSI MENGGUNAKAN  
ALGORITMA SUPPORT VECTOR MACHINE**

**SKRIPSI**

diajukan kepada

Universitas Negeri Malang

untuk memenuhi salah satu persyaratan

dalam menyelesaikan program Sarjana

Teknik Informatika

**OLEH**

**MERCYANO DANDI HIDAYAT**

NIM 190535646051

**UNIVERSITAS NEGERI MALANG**

**FAKULTAS TEKNIK**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**DESEMBER 2022**

## LEMBAR PERSETUJUAN

Skripsi oleh Mercyano Dandi Hidayat ini telah diperiksa dan disetujui untuk diujikan.

Malang, 6 Desember 2022

Pembimbing I,

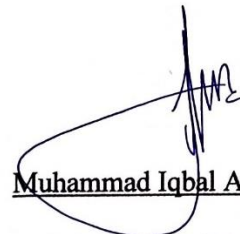


Harits Ar Rosyid, S.T., M.T., Ph.D.

NIP 198108112009121003

Malang, 6 Desember 2022

Pembimbing II,



Muhammad Iqbal Akbar, S.ST., M.MT

NIP 198810242015041002

## LEMBAR PENGESAHAN

Skripsi oleh Mercyano Dandi Hidayat ini telah di pertahankan di depan dewan penguji pada tanggal 28 Desember 2022

Dewan Penguji

Dr. Ir. Triyanna Widiyaningtyas, M.T.

Ketua

NIP 197412152008122002

Harits Ar Rosyid, S.T., M.T., Ph.D.

Anggota

NIP 198108112009121003

Muhammad Iqbal Akbar, S.ST., M.MT

Anggota

NIP 198810242015041002

Mengesahkan,

Dekan Fakultas Teknik

Prof. Dr. Andoko, S.T., M.T.

NIP 196508121991031005

Mengetahui,

Ketua Departemen Teknik Elektro

Dr. Ir. Triyanna Widiyaningtyas, M.T.

NIP 197412152008122002

## PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama : Mercyano Dandi Hidayat

NIM : 190535646051

Jurusan/Program Studi: Teknik Elektro/S1 Teknik Informatika

Fakultas/Program : Fakultas Teknik/Sarjana

Menyatakan dengan sesungguhnya bahwa **skripsi** yang saya tulis ini benar-benar tulisan saya, dan bukan merupakan plagiasi/falsifikasi/fabrikasi baik sebagian atau seluruhnya.

Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa **skripsi** ini hasil plagiasi/falsifikasi/fabrikasi, baik sebagian atau seluruhnya, maka saya bersedia menerima sanksi atas perbuatan tersebut sesuai dengan ketentuan yang berlaku.

Malang, ..1 Januari 2023.....

Yang membuat pernyataan



Mercyano Dandi Hidayat

NIM 190535646051

## RINGKASAN

Hidayat, Mercyano Dandi. 2022. *Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine*. Skripsi, Jurusan Teknik Elektro, Fakultas Teknik, Universitas Negeri Malang. Pembimbing: (I) Harits Ar Rosyid, S.T., M.T., Ph.D., (II) Muhammad Iqbal Akbar, S.ST., M.MT.

**Kata Kunci:** SISINTA TEUM, Skripsi, Kelompok Bidang Keahlian, Klasifikasi, Support Vector Machine.

Sistem Informasi Skripsi dan Tugas Akhir Jurusan Teknik Elektro Universitas Negeri Malang (SISINTA TEUM) merupakan sebuah sistem informasi berbasis situs web yang berfungsi dalam melakukan pengajuan terhadap kegiatan-kegiatan yang berkaitan dengan skripsi dan tugas akhir pada Jurusan Teknik Elektro Universitas Negeri Malang. Pada saat proses pengajuan skripsi dan tugas akhir pada web SISINTA TEUM, pengguna harus memilih sendiri Kelompok Bidang Keahlian (KBK) yang sesuai dengan tema judul penelitian skripsi yang diajukan. Dimana, pengguna sering merasa kesulitan dalam menentukan KBK berdasarkan judul dan abstrak yang hendak diteliti. Untuk itu, diperlukan suatu sistem kecerdasan buatan berupa klasifikasi KBK berdasarkan judul dan abstrak skripsi pengguna secara otomatis. Sehingga, sistem dapat memberikan rekomendasi KBK dari judul atau abstrak yang dimasukkan oleh pengguna.

Metode yang digunakan dalam membuat sistem klasifikasi KBK berdasarkan judul dan abstrak skripsi ini yaitu menggunakan algoritma Support Vector Machine. Dimana, tujuan dari penelitian ini yaitu untuk mengimplementasikan dan menguji performa dari algoritma tersebut pada kasus klasifikasi KBK. Tahapan penelitian yang dilakukan yaitu (1) *data collection*, (2) *text preprocessing*, (3) pembobotan istilah, (4) *resampling*, (5) *training model*, dan (6) evaluasi.

Penelitian ini melakukan uji coba terhadap tiga skenario *input data* yaitu *input data* judul, abstrak, serta gabungan antara judul dan abstrak. Parameter yang

diuji pada model yang dibuat yaitu kernel dan parameter regularisasi. Proses *tuning* dari parameter tersebut dilakukan menggunakan metode Grid Search. Dari percobaan yang dilakukan, diperoleh bahwa sistem klasifikasi KBK dengan input data judul memberikan hasil yang optimal dan efisien.

Hasil penelitian ini diperoleh bahwa skenario terbaik dari percobaan yang dilakukan yaitu pada skenario data input judul. Skenario tersebut memberikan hasil yang lebih optimal dan efisien, dengan nilai akurasi, presisi, *recall*, dan *f1-score* secara berturut-turut yaitu 63,16%, 61,25%, 63,16%, dan 60,34%. Kemudian, skenario terbaik kedua yaitu pada data input gabungan judul dan abstrak, dengan nilai akurasi, presisi, *recall*, dan *f1-score* secara berturut-turut yaitu 62,89%, 60%, 62,89%, dan 60,34%. Sedangkan untuk skenario dengan performa terendah yaitu pada data input abstrak, dengan nilai akurasi, presisi, *recall*, dan *f1-score* secara berturut-turut yaitu 61,32%, 58,75%, 61,32%, dan 58,78%.



## SUMMARY

Hidayat, Mercyano Dandi. 2022. *Classification of Expertise Groups Based on Thesis Titles and Abstracts Using Support Vector Machine Algorithm*. Thesis, Department of Electrical Engineering, Faculty of Engineering, State University of Malang. Advisors: (I) Harits Ar Rosyid, S.T., M.T., Ph.D., (II) Muhammad Iqbal Akbar, S.ST., M.MT.

**Keywords:** SISINTA TEUM, Thesis, Expertise Groups, Classification, Support Vector Machine.

Thesis and Final Project Information System for the Department of Electrical Engineering, State University of Malang (SISINTA TEUM) is a web-based information system that functions in submitting activities related to theses and final assignments at the Department of Electrical Engineering, State University of Malang. During the process of submitting a thesis and final project on the SISINTA TEUM web, users must choose their own Field of Expertise Group in accordance with the theme of the proposed thesis research title. Where, users often find it difficult to determine the Group of Expertise based on the title and abstract to be researched. For this reason, an artificial intelligence system is needed in the form of classification of the Group of Expertise based on the title and abstract of the user's thesis automatically. Thus, the system can provide recommendations for the Field of Expertise Group from the title or abstract entered by the user.

The method used in making the classification system of the Group of Expertise based on the title and abstract of this thesis is using the Support Vector Machine algorithm. Where, the purpose of this research is to implement and test the performance of the algorithm in the case of classification of Expertise Group. The research stages carried out are (1) data collection, (2) text preprocessing, (3) term weighting, (4) resampling, (5) training model, and (6) evaluation.

This research conducted trials on three data input scenarios, namely title, abstract, and combined title and abstract data input. The parameters tested in the model are kernel and regularization parameters. The tuning process of these

parameters is carried out using the Grid Search method. From the experiments conducted, it was found that the classification system of Expertise Group with title data input gave optimal and efficient results.

The results of this study obtained that the best scenario from the experiments conducted is in the title input data scenario. The scenario provides more optimal and efficient results, with accuracy, precision, recall, and f1-score values of 63,16%, 61,25%, 63,16%, and 60,34%, respectively. Then, the second best scenario is on combined title and abstract input data, with accuracy, precision, recall, and f1-score values of 62,89%, 60%, 62,89%, and 60,34% respectively. Meanwhile, the scenario with the lowest performance is the abstract input data, with accuracy, precision, recall, and f1-score values of 61,32%, 58,75%, 61,32%, and 58,78%, respectively.

## KATA PENGANTAR

Pertama-tama, peneliti ucapkan Puji syukur ke hadirat Tuhan Yang Maha Esa. Skripsi ini dapat terselesaikan dengan tepat waktu berkat rahmat dan karunia-Nya. Selain itu, tentu saja tidak lepas dari bantuan berbagai pihak yang ikut serta dalam membantu proses pengerjaan skripsi ini.

Tidak lupa, peneliti ucapkan terima kasih kepada Bapak Harits Ar Rosyid, S.T., M.T., Ph.D. selaku pembimbing I dalam penelitian ini. Peneliti juga mengucapkan terima kasih kepada Bapak Muhammad Iqbal Akbar, S.ST., M.MT selaku pembimbing II yang juga membina dan mendampingi peneliti dalam pengerjaan skripsi ini.

Peneliti menyadari bahwa skripsi ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi perbaikan skripsi ini di masa yang akan datang. Diharapkan dengan adanya penelitian ini, dapat menambah wawasan dan memberi manfaat bagi pembaca.

Malang, Desember 2022



Peneliti

## DAFTAR ISI

	Halaman
HALAMAN SAMPUL .....	i
LEMBAR LOGO .....	ii
HALAMAN JUDUL.....	iii
LEMBAR PERSETUJUAN.....	iv
LEMBAR PENGESAHAN .....	v
PERNYATAAN KEASLIAN TULISAN .....	vi
RINGKASAN .....	vii
SUMMARY .....	ix
KATA PENGANTAR .....	xi
DAFTAR ISI.....	xii
DAFTAR TABEL.....	xv
DAFTAR GAMBAR .....	xvi
DAFTAR LAMPIRAN.....	xviii
BAB I PENDAHULUAN.....	1
1.1    Latar Belakang Masalah .....	1
1.2    Rumusan Masalah .....	2
1.3    Tujuan Penelitian.....	3
1.4    Batasan Masalah.....	3
1.5    Manfaat Penelitian.....	3
BAB II KAJIAN PUSTAKA .....	4
2.1    Kelompok Bidang Keahlian .....	4
2.2    Klasifikasi Teks .....	4

2.3	<i>Preprocessing</i> .....	5
2.3.1	<i>Text Cleaning</i> .....	5
2.3.2	<i>Tokenization</i> .....	6
2.3.3	<i>Stop Words Removal</i> .....	6
2.3.4	<i>Stemming</i> .....	6
2.4	Metode Pembobotan Kata: TF-IDF .....	7
2.5	Metode <i>Resampling: Synthetic Minority Oversampling Technique</i> (SMOTE) .....	7
2.6	Support Vector Machine .....	8
2.7	Evaluasi Kinerja .....	10
2.7.1	<i>Confusion Matrix</i> .....	10
2.8	Penelitian Terdahulu .....	11
BAB III METODE PENELITIAN .....		13
3.1	Rancangan Penelitian .....	13
3.2	<i>Data Collection</i> .....	14
3.3	<i>Text Preprocessing</i> .....	14
3.4	Pembobotan Istilah .....	16
3.5	<i>Resampling</i> .....	16
3.6	<i>Training Model</i> .....	17
3.7	Evaluasi .....	18
BAB IV HASIL DAN PEMBAHASAN .....		19
4.1	Hasil Data Collection .....	19
4.2	Hasil Text Preprocessing .....	19
4.2.1	Hasil Text Cleaning .....	19
4.2.2	Hasil Tokenization .....	21

4.2.3	Hasil Stop Words Removal.....	21
4.2.4	Hasil Stemming.....	22
4.3	Hasil Pembobotan Istilah Menggunakan TF-IDF .....	22
4.4	Hasil Resampling .....	23
4.6	Hasil Training Model .....	24
4.6.1	Skenario Judul.....	24
4.6.2	Skenario Abstrak.....	24
4.6.3	Skenario Judul dan Abstrak .....	25
4.7	Hasil Evaluasi.....	26
4.7.1	Skenario Judul.....	26
4.7.1	Skenario Abstrak.....	26
4.7.1	Skenario Judul dan Abstrak .....	27
4.8	Analisis Perbandingan Hasil Performa.....	27
BAB V KESIMPULAN DAN SARAN.....		30
5.1	Kesimpulan.....	30
5.2	Saran .....	30
DAFTAR RUJUKAN .....		31
LAMPIRAN.....		37
RIWAYAT HIDUP.....		40

## DAFTAR TABEL

Tabel	Halaman
1. Confusion Matrix .....	10

## DAFTAR GAMBAR

Gambar	Halaman
1. SVM .....	8
2. Klasifikasi Linear dan Non-Linear.....	9
3. Tahapan Metode Penelitian.....	13
4. Jumlah Baris pada Masing-Masing KBK .....	14
5. Contoh Hasil Data Collection .....	19
6. Distribusi Kelas pada Data Mentah.....	19
7. Contoh Hasil Text Cleaning.....	20
8. Jumlah Missing Values pada Setiap Kolom Dataset.....	20
9. Duplikasi pada Kolom Judul.....	20
10. Duplikasi pada Kolom Abstrak.....	20
11. Distribusi Kelas Setelah Text Preprocessing .....	20
12. Hasil Tokenization pada Kolom Judul.....	21
13. Hasil Tokenization pada Kolom Abstrak.....	21
14. Hasil Stop Words Removal pada Kolom Judul.....	21
15. Hasil Stop Words Removal pada Kolom Abstrak.....	22
16. Hasil Stemming pada Kolom Judul .....	22
17. Hasil Stemming pada Kolom Abstrak.....	22
18. Hasil Pembobotan Menggunakan TF-IDF .....	23
19. Distribusi Kelas Tidak Seimbang pada Dataset .....	23
20. Hasil Resampling Menggunakan SMOTE.....	23
21. Hasil Grid Search pada Skenario Judul.....	24
22. Hasil Grid Search pada Skenario Abstrak.....	25
23. Hasil Grid Search pada Skenario Judul dan Abstrak .....	25
24. Confusion Matrix pada Skenario Judul.....	26
25. Evaluasi pada Skenario Judul .....	26
26. Confusion Matrix pada Skenario Abstrak.....	26
27. Evaluasi pada Skenario Abstrak.....	27
28. Confusion Matrix pada Skenario Judul dan Abstrak .....	27
29. Evaluasi pada Skenario Judul dan Abstrak .....	27



30. Perbandingan Performa Masing-Masing Skenario .....	28
31. Contoh Kesalahan Penulisan pada Dataset .....	29

## DAFTAR LAMPIRAN

Lampiran	Halaman
1. Perhitungan Skala Likert Survey .....	37
2. Perbandingan Performa pada Masing-Masing Skenario.....	38
3. Demo Sistem .....	39

## **BAB I**

### **PENDAHULUAN**

#### **1.1 Latar Belakang Masalah**

Sistem Informasi Skripsi dan Tugas Akhir Jurusan Teknik Elektro Universitas Negeri Malang (SISINTA TEUM) merupakan suatu sarana informasi daring yang digunakan oleh mahasiswa Jurusan Teknik Elektro Universitas Negeri Malang untuk kegiatan yang berkaitan dengan skripsi dan tugas akhir. Pada situs web ini, mahasiswa dapat melakukan pengajuan judul, pendaftaran seminar, pendaftaran sidang skripsi atau tugas akhir, dan lain-lainnya terkait hal tersebut.

Pada proses pengajuan judul skripsi pada web SISINTA TEUM, mahasiswa harus memilih KBK yang sesuai dengan judul dan abstrak skripsi yang akan diajukan. Untuk itu, mahasiswa perlu mengetahui hubungan antara judul atau abstrak dengan KBK yang tepat. Dimana, pemilihan KBK dengan tepat merupakan suatu syarat untuk dapat memilih dosen pembimbing. Oleh karena itu, pemilihan KBK merupakan hal yang sangat penting dalam pengajuan penelitian judul skripsi.

Pada pemilihan KBK, tidak menutup kemungkinan bahwa mahasiswa memilih KBK yang tidak sesuai dengan tema dari judul penelitian skripsi yang diambil. Dimana, terdapat 13 KBK pada Jurusan Teknik Elektro Universitas Negeri Malang. Selain karena banyaknya jumlah KBK, kesalahan pemilihan KBK juga dapat disebabkan karena ruang lingkup masing-masing KBK yang luas. Luasnya ruang lingkup KBK dapat menyebabkan tema judul yang diambil berada pada irisan antara suatu KBK dengan KBK lainnya, sehingga dapat membingungkan mahasiswa untuk memilih KBK yang sesuai.

Dilakukan survei dengan pengukuran skala likert terhadap 25 mahasiswa aktif ataupun alumni Jurusan Teknik Elektro Universitas Negeri Malang yang pernah mengakses SISINTA TEUM. Berdasarkan survey tersebut, diperoleh hasil bahwa mereka merasa kesulitan dalam menentukan KBK yang sesuai. Selain itu, mereka juga mengatakan sangat setuju untuk dilakukan pengembangan sistem klasifikasi KBK otomatis pada web SISINTA TEUM. Sehingga, diperlukan suatu sistem yang dapat memberikan rekomendasi KBK berdasarkan judul ataupun abstrak yang diambil untuk mempermudah dalam proses penentuan KBK.

## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

Penelitian terhadap klasifikasi KBK berdasarkan judul skripsi sudah pernah dilakukan. Pada penelitian sebelumnya yang dilakukan oleh Pujiyanto, Widiyaningtyas, Prasetya, & Romadhon, dilakukan pengujian performa terhadap algoritma klasifikasi yaitu Naïve Bayes dengan menggunakan metode K-Fold Cross Validation (Pujiyanto dkk., 2019). Penelitian tersebut diuji 10 kali pada 1103 judul skripsi dan tugas akhir dengan rata-rata akurasi, presisi, dan skor *recall* masing-masing adalah 94%, 80%, dan 69%.

Di bidang pembelajaran tugas tunggal (*single-task learning*), SVM mendapat banyak perhatian akademis, karena landasan teoretisnya yang kuat dan kinerja yang efektif (Krell dkk., 2017). Namun, SVM tidak hanya dapat digunakan untuk masalah klasifikasi biner, tetapi juga masalah klasifikasi multi kelas dengan menerapkan strategi agregasi yang tepat (Nishitsuji & Nasseri, 2022). Sehingga, algoritma SVM layak untuk diuji dan diimplementasikan untuk sistem klasifikasi KBK, dimana merupakan tugas klasifikasi multi kelas.

Sehingga, berdasarkan latar belakang tersebut, penelitian ini mengangkat judul “Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine”. Penelitian dilakukan dengan tujuan untuk mencoba mengimplementasikan dan menguji performa dari algoritma SVM untuk klasifikasi KBK. Selain itu, dengan adanya sistem klasifikasi KBK yang dibuat, diharapkan dapat membantu mahasiswa dalam memilih KBK berdasarkan judul atau abstraknya secara otomatis dan efektif.

### 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan diatas, maka dapat dirumuskan beberapa rumusan masalah sebagai berikut:

1. Bagaimana implementasi sistem klasifikasi KBK berdasarkan judul dan abstrak skripsi Jurusan Teknik Elektro Universitas Negeri Malang menggunakan algoritma SVM?
2. Bagaimana performa dari sistem klasifikasi KBK berdasarkan judul dan abstrak skripsi Jurusan Teknik Elektro Universitas Negeri Malang menggunakan algoritma SVM?

### 1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan diatas, berikut ini merupakan tujuan dari penelitian ini:

1. Untuk mengimplementasikan sistem klasifikasi KBK berdasarkan judul dan abstrak skripsi Jurusan Teknik Elektro Universitas Negeri Malang menggunakan algoritma SVM.
2. Untuk mengetahui performa dari klasifikasi KBK berdasarkan judul dan abstrak skripsi Jurusan Teknik Elektro Universitas Negeri Malang menggunakan algoritma SVM.

### 1.4 Batasan Masalah

Adapun batasan-batasan masalah pada penelitian ini, yaitu:

1. Sistem klasifikasi berdasarkan judul dan abstrak skripsi di Jurusan Teknik Elektro Universitas Negeri Malang.
2. Data yang digunakan berupa judul, abstrak, beserta KBK yang sesuai dengan tema masing-masing penelitian skripsi di Jurusan Teknik Elektro Universitas Negeri Malang dari tahun 2020 hingga tahun 2022.

### 1.5 Manfaat Penelitian

Diharapkan dengan adanya penelitian ini dapat memberikan berbagai manfaat, yaitu antara lain sebagai berikut:

1. Bagi mahasiswa, dapat mempermudah proses pemilihan KBK berdasarkan judul dan abstrak skripsi secara otomatis dengan menggunakan sistem klasifikasi KBK yang dikembangkan.
2. Bagi peneliti, dapat memberikan pengetahuan mengenai pengembangan sistem klasifikasi KBK menggunakan algoritma SVM.
3. Bagi instansi lain, dapat menerapkan sistem klasifikasi KBK menggunakan algoritma SVM.

## BAB II

### KAJIAN PUSTAKA

#### 2.1 Kelompok Bidang Keahlian

Berdasarkan Peraturan Rektor Universitas Negeri Malang Nomor 4 Tahun 2015 tentang Pembentukan dan Pengembangan Kelompok Bidang Keahlian Dosen Universitas Negeri Malang, Kelompok Bidang Keahlian (KBK) merupakan pengelompokan dosen ke dalam bidang keilmuan berdasarkan pendidikan terakhirnya, keterlibatan dalam penelitian dan suatu pengabdian masyarakat, publikasi karya ilmiah asli atau karya cipta yang dilindungi hak kekayaan intelektual, dan pengajaran suatu mata kuliah selama periode waktu tertentu (Universitas Negeri Malang, 2015). KBK dibagi berdasarkan jurusan masing-masing di Universitas Negeri Malang. Sehingga, masing-masing jurusan memiliki KBK yang berbeda.

Pada Jurusan Teknik Elektro Universitas Negeri Malang terdapat 13 jenis KBK. Dimana, KBK tersebut terdiri dari: (1) Evaluasi dan Pengelolaan Pendidikan Kejuruan, (2) Kurikulum Pendidikan Teknologi dan Kejuruan, (3) Strategi Pembelajaran Teknologi dan Kejuruan, (4) Pengembangan Aplikasi dan Media Pembelajaran Teknologi dan Kejuruan, (5) Ketenagakerjaan Teknologi dan Kejuruan, (6) *Intelligent Power Electronics and Smart Grid* (IPESG), (7) *Intelligent Power and Advanced Energy System* (IPAES), (8) *Telematics IoT System and Devices*, (9) Sistem Dinamis, Kendali, dan Robotika (*Dynamic Systems, Control and Robotics*), (10) *Biomedic and Intelligent Assistive Technology* (TAT), (11) Rekayasa Pengetahuan dan Ilmu Data (*Knowledge Engineering and Data Science*), (12) Teknologi Digital Cerdas (*Ubiquitous Computing Technique*), dan (13) *Game Technology and Machine Learning Applications* (Kelompok Bidang Keahlian (KBK) / Teknik Elektro – UM, 2020).

#### 2.2 Klasifikasi Teks

Salah satu tugas paling umum dalam kecerdasan buatan yaitu permasalahan klasifikasi (Żurek & Pietroń, 2020). Klasifikasi teks merupakan suatu metode dalam mengelompokkan data berupa teks bahasa alami kedalam beberapa kelas

berbeda yang telah didefinisikan sebelumnya. Tujuan dilakukannya klasifikasi teks yaitu untuk mengelompokkan teks secara otomatis berdasarkan kategori yang dipelajari sebelumnya (Riza Adrianti Supono & Muhammad Azis Suprayogi, 2021). Di era digitalisasi, klasifikasi teks menjadi sangat penting karena saat ini ada sejumlah besar data teks yang diproduksi, dan berkembang pesat (Mutawalli dkk., 2019).

Algoritma yang dapat digunakan untuk klasifikasi teks bermacam-macam. Beberapa algoritma yang sering digunakan pada saat ini yaitu Support Vector Machine, Decision Tree, Naïve Bayes, K Nearest Neighbors, dan Hidden Markov Model (Aliwy & Ameer, 2017). Selain itu, terdapat juga algoritma klasifikasi teks dengan pendekatan *deep learning*. Performa dari algoritma dengan pendekatan *deep learning* sangat baik untuk tugas klasifikasi dengan kumpulan data yang banyak. Namun, kelemahan utama algoritma *deep learning* dibandingkan dengan metode *machine learning* konvensional yaitu seringkali membutuhkan lebih banyak data dan masalah klasifikasi yang melibatkan kumpulan data kecil sulit untuk diselesaikan (Lampinen & McClelland, 2018).

### 2.3 *Preprocessing*

*Preprocessing* merupakan langkah pertama dalam tugas klasifikasi teks, dimana hal tersebut dilakukan agar data teks siap digunakan pada langkah selanjutnya (Khairunnisa dkk., 2021). Sebelum masuk ke tahap pelatihan model, data teks mentah harus diolah terlebih dahulu agar model dapat dilatih dengan benar. Pada *preprocessing* dalam klasifikasi teks, terdapat 5 hal dasar yang perlu dilakukan: (1) *Text Cleaning*, (2) *Tokenization*, (3) *Stemming*, (4) *Stop Words Removal*, dan (5) *Resampling*. Dengan melalui semua tahapan tersebut, data yang tidak penting akan berkurang, data menjadi lebih terstruktur, dan konsisten.

#### 2.3.1 *Text Cleaning*

*Text Cleaning* merupakan suatu proses yang dilakukan pada data teks mentah dengan tujuan untuk menghilangkan data yang tidak penting, menghilangkan *noise* di dalam teks, serta membuat teks menjadi konsisten. Tahapan utama yang menjadi dasar dalam *text cleaning* meliputi menghapus data



duplikat dan *missing values*, *case folding*, *trimming*, serta menghapus tanda baca, karakter spesial, dan spasi ganda. *Case folding* merupakan proses pengolahan data mentah yang masih mengandung huruf kapital menjadi huruf kecil semua (Normawati & Prayogi, 2021). Sedangkan *trimming* dilakukan dengan untuk menghapus spasi yang jika terdapat pada awal dan akhir teks.

### 2.3.2 Tokenization

*Tokenization* merupakan proses memenggal suatu teks menjadi *tokens* atau biasanya kata-kata, karena teks yang lengkap terlalu spesifik untuk dianalisis oleh komputer (Welbers dkk., 2017). *Tokens* yang dihasilkan merupakan bagian dari teks asli dan tidak mengalami perubahan sama sekali. Pertimbangan yang paling penting untuk proses *tokenization* yaitu memberikan efisiensi dan akurasi (A. Mullen dkk., 2018). *Tokenization* Tahapan ini wajib dilakukan sebelum melalui tahap *preprocessing* selanjutnya seperti *stop words removal* dan *stemming*.

### 2.3.3 Stop Words Removal

*Stop words removal* merupakan suatu proses menghapus kata-kata yang sering muncul dan hanya memberikan sedikit informasi pada teks (Ganeshkumar & Padmanabhan, 2022). Masing-masing bahasa memiliki *stop words* yang berbeda-beda. Dalam Bahasa Indonesia, contoh *stop words* yaitu “di”, “yang”, “dan”, “atau”, dan kata-kata umum lainnya yang tidak penting.

### 2.3.4 Stemming

*Stemming* adalah suatu proses menghilangkan semua imbuhan dari kata, termasuk akhiran, sisipan, awalan, serta kombinasi antara awalan dan akhiran sehingga mengubah varian morfologis kata menjadi bentuk dasar yang sama (Simarangkir, 2017). Dengan mengubah menjadi bentuk kata dasarnya, *stemming* bertujuan untuk meningkatkan kinerja *Information Retrieval* atau proses pengambilan informasi dari kumpulan teks (Novitasari, 2017). Contoh *stemming* dalam Bahasa Indonesia yaitu seperti kata “berbagi” dan “membagikan” yang memiliki makna yang sama akan diubah menjadi bentuk kata dasarnya yaitu “bagi”.



Sehingga, komputer tidak mengartikan kedua hal tersebut sebagai dua makna yang berbeda.

#### 2.4 Metode Pembobotan Kata: TF-IDF

Metode pembobotan kata dilakukan setelah semua tahap *preprocessing* telah dilakukan. *Term Frequency-Inverse Document Frequency* (TF-IDF) merupakan suatu metode pembobotan kata untuk merubah data berupa teks menjadi bentuk numerik atau bobot pada kata tersebut (Septian dkk., 2019). Perhitungan bobot pada kata dapat dilakukan dengan cara menghitung kemunculan setiap kata untuk memetakan suatu data teks menjadi angka. Bobot kata akan semakin meningkat sebanding dengan seberapa sering kata itu muncul pada sebuah dokumen dan menurun jika sering muncul pada banyak dokumen (Amrizal, 2018). Sehingga dengan menggunakan TF-IDF, memungkinkan komputer memahami konteks kata pada seluruh kumpulan dokumen dan bukan hanya dalam satu dokumen saja. Rumus dari *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) dijabarkan pada Persamaan 2.1 dan 2.2 (Fan & Qin, 2018).

$$TF = \frac{t}{s} \quad (2.1)$$

$$IDF = \log \left( \frac{M}{m} + 0.01 \right) \quad (2.2)$$

Dimana,

$t$  : jumlah kemunculan kata dalam *file*.

$s$  : jumlah kemunculan semua kata dalam *file*.

$M$  : jumlah total dokumen dalam korpus.

$m$  : jumlah dokumen yang berisi istilah fitur.

#### 2.5 Metode *Resampling: Synthetic Minority Oversampling Technique* (SMOTE)

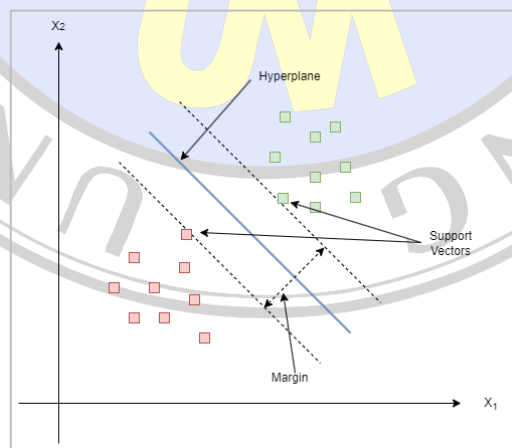
Synthetic Minority Oversampling Technique (SMOTE) merupakan metode *resampling* yang bertujuan untuk menyeimbangkan dataset dengan rasio yang sangat tidak seimbang dengan cara membuat sampel sintetis di kelas minoritas (W. Li dkk., 2022). Metode ini adalah teknik oversampling yang paling populer

(Czarnowski, 2022). Konsep dasar SMOTE yaitu menghasilkan sampel data baru di kelas minoritas dengan cara melakukan interpolasi antara sampel kelas yang berdekatan satu sama lain (D.-C. Li dkk., 2022). Sehingga, SMOTE menyeimbangkan kumpulan data dengan cara menambah jumlah sampel data pada kelas minoritas.

SMOTE menggunakan algoritma k-Nearest Neighbors (kNN) untuk menentukan tetangga dari sampel kelas minoritas dan kemudian secara acak memilih tetangga ke-k untuk membuat sampel baru (Mayabadi & Saadatfar, 2022). Pada metode SMOTE, terdapat tiga langkah utama yang dilakukan secara iteratif: (1) memilih sampel secara acak yang termasuk dalam kelas minoritas, (2) memilih K (secara *default* 5) tetangga terdekat dari sampel, dan (3) N atau jumlah *oversampling* yang diinginkan dari K tetangga tersebut dipilih secara acak untuk interpolasi dan menghasilkan sampel baru (Joloudari dkk., 2022).

## 2.6 Support Vector Machine

Support Vector Machine (SVM) merupakan suatu metode pembelajaran mesin terawasi (*supervised learning*) yang efektif digunakan untuk memecahkan masalah klasifikasi (Behmanesh dkk., 2021). Konsep dasar dari algoritma SVM yaitu mencari *hyperlane* atau *decision boundary* yang paling optimal. *Hyperplane* tersebut berfungsi untuk memisahkan antara kelas yang berbeda.



Gambar 1. SVM

Dapat dilihat pada Gambar 1, komponen utama dalam algoritma SVM yaitu *hyperplane*, *support vectors*, dan *margin*. Support vectors merupakan suatu data di masing-masing kelas yang memiliki jarak terdekat dengan *hyperlane*. Kemudian,

## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

jarak antara masing-masing *support vector* tersebut dinamakan margin. Sehingga, dengan menghitung margin dan menentukan titik maksimumnya, *hyperlane* yang optimal dapat ditemukan (Faraby, 2018). Rumus perhitungan hyperplane dapat dilihat pada Persamaan 2.3 (Listiana, 2017).

$$w_n \cdot X + b = 0 \quad (2.3)$$

Dimana,

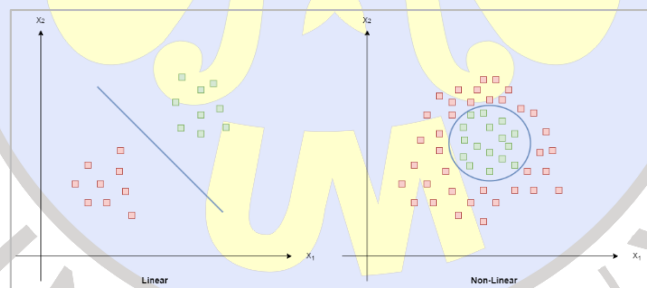
$w$  : bobot vektor

$n$  : jumlah atribut

$b$  : bias

$X$  : *training tuple*

Mulanya, SVM dikembangkan hanya untuk memecahkan masalah klasifikasi biner (dua) dengan pemisahan linier, tetapi sekarang juga dapat digunakan untuk masalah klasifikasi multi kelas dengan menerapkan fungsi agregasi yang tepat (Nishitsuji & Nasser, 2022). Pada algoritma SVM, terdapat fungsi kernel yang dapat digunakan untuk menangani klasifikasi multi kelas yaitu dengan metode pemisahan non-linier. Pada Gambar 2 menunjukkan perbedaan antara klasifikasi dengan pemisahan linier dan non-linier.



**Gambar 2.** Klasifikasi Linear dan Non-Linear

Pada klasifikasi non-linier, terdapat 3 kernel SVM yang sering digunakan yaitu kernel RBF, Polynomial, dan Sigmoid (Puspitasari dkk., 2018). Dimana, Persamaan fungsi ketiga kernel tersebut secara berturut-turut dapat dilihat pada Persamaan 2.4, 2.5, dan 2.6.

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (2.4)$$

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d \quad (2.5)$$

$$k(x, y) = \tanh(\alpha x^T y + c) \quad (2.6)$$

## 2.7 Evaluasi Kinerja

### 2.7.1 Confusion Matrix

Pada tugas klasifikasi, model hanya dapat mencapai dua hasil yaitu prediksi klasifikasi benar atau salah. Perhitungan jumlah dari masing-masing kedua hasil tersebut dapat digunakan untuk menguji performa suatu model, salah satunya yaitu *confusion matrix*. *Confusion matrix* merupakan suatu tabel khusus yang memungkinkan visualisasi kinerja algoritma dan terutama digunakan dalam *supervised learning* serta lebih khusus lagi untuk klasifikasi statistik (Haghighi dkk., 2018). Tabel pada *confusion matrix* menyajikan perbandingan jumlah antara data uji yang benar dan salah pada hasil klasifikasi (Normawati & Prayogi, 2021). Bentuk tabel dari *confusion matrix* dapat dilihat pada Tabel 1.

**Tabel 1.** Confusion Matrix

		<i>Actual Class</i>	
		<i>Positive</i>	<i>Negative</i>
<i>Predicted Class</i>	<i>Positive</i>	<i>True Positive</i> (TP)	<i>False Positive</i> (FP)
	<i>Negative</i>	<i>False Negative</i> (FN)	<i>True Negative</i> (TN)

Dimana,  
 TP = Jumlah prediksi benar pada kelas positif  
 FN = Jumlah prediksi salah pada kelas negatif  
 TN = Jumlah prediksi benar pada kelas negatif  
 FP = Jumlah prediksi salah pada kelas positif

Sebagai representasi dari hasil proses klasifikasi, terdapat empat istilah yang digunakan dalam pengukuran kinerja dengan menggunakan *confusion matrix*, yaitu True Negative (TN), False Positive (FP), True Positive (TP), dan False Negative (FN) (Hadianto dkk., 2019). Dengan menggunakan empat perhitungan tersebut, kemudian dapat diketahui metrik akurasi, *recall*, presisi, dan *f1-score* yang mana berfungsi untuk mengukur kinerja algoritma klasifikasi. Akurasi dihitung sebagai perbandingan antara prediksi benar dibagi dengan jumlah semua prediksi, dimana

rumus tersebut dijelaskan pada Persamaan 2.7. *Recall* mengukur kemampuan suatu model klasifikasi dalam menemukan semua kasus yang relevan dalam kumpulan data. Rumus *recall* dijabarkan pada Persamaan 2.8. Presisi digunakan untuk mengukur kemampuan model klasifikasi dalam mengidentifikasi titik data yang relevan. Rumus presisi dijelaskan pada Persamaan 2.9. Kemudian, *f1-score* merupakan gabungan antara *recall* dan presisi yang optimal, dengan rumus pada Persamaan 2.10 (Manning dkk., 2008).

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.7)$$

$$Recall = \frac{TP}{TP+FN} \quad (2.8)$$

$$Presisi = \frac{TP}{TP+FP} \quad (2.9)$$

$$F1\ Score = 2 \cdot \frac{Presisi \cdot Recall}{Presisi + Recall} \quad (2.10)$$

## 2.8 Penelitian Terdahulu

Penelitian ini dibuat dengan tujuan untuk menemukan solusi dari *research gap* pada penelitian-penelitian terdahulu. Untuk itu, *State of The Art* ini dibuat untuk menentukan posisi pada penelitian ini dengan penelitian-penelitian relevan terdahulu. Penelitian ini membandingkan beberapa penelitian terdahulu yang berhubungan dengan klasifikasi teks menggunakan algoritma SVM yaitu sebagai berikut:

- “Penerapan Algoritma Naïve Bayes Classifier untuk Klasifikasi Judul Skripsi dan Tugas Akhir Berdasarkan Kelompok Bidang Keahlian (Pujianto dkk., 2019)”. Penelitian tersebut melakukan percobaan terhadap algoritma naïve bayes untuk melakukan klasifikasi KBK berdasarkan judul skripsi dan tugas akhir. Pengujian dilakukan menggunakan metode K-Fold Cross Validation dengan 10 kali pengujian. Performa model terbaik yang didapatkan dari penelitian tersebut yaitu dengan nilai rata-rata akurasi, presisi, dan *recall* masing-masing sebesar 94%, 80%, dan 69%.
- “Klasifikasi Jenis Pantun dengan Metode Support Vector Machine (SVM) (Irmada & Ria Astriratma, 2020)”. Penelitian tersebut mengimplementasikan sistem klasifikasi terhadap tiga jenis pantun yaitu

## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

pantun tua, pantun muda, dan pantun anak dengan menggunakan algoritma SVM. Dari penelitian tersebut didapatkan kinerja terbaik yaitu dengan nilai akurasi, presisi, dan recall masing-masing sebesar 81,91%, 90,63%, dan 87,88%.

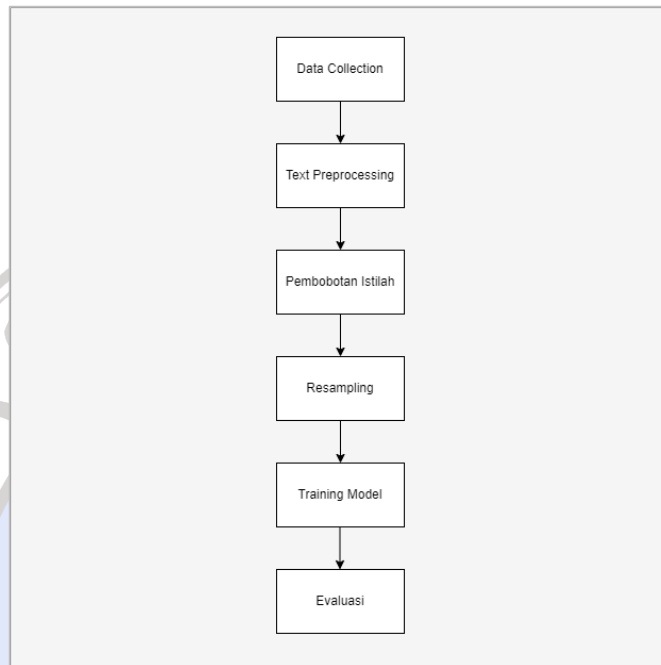
- “Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine (Oryza Habibie Rahman dkk., 2021)”. Penelitian ini melakukan percobaan klasifikasi terhadap 5 jenis ujaran kebencian pada Twitter yaitu ras, suku, agama, antar golongan, dan netral dengan menggunakan metode SVM. Didapatkan hasil terbaik yaitu pada penggunaan kernel RBF. Model tersebut memiliki nilai akurasi, presisi, *recall*, dan *f-measure* masing-masing sebesar 93%, 84%, 86%, dan 83%.



## BAB III

### METODE PENELITIAN

#### 3.1 Rancangan Penelitian

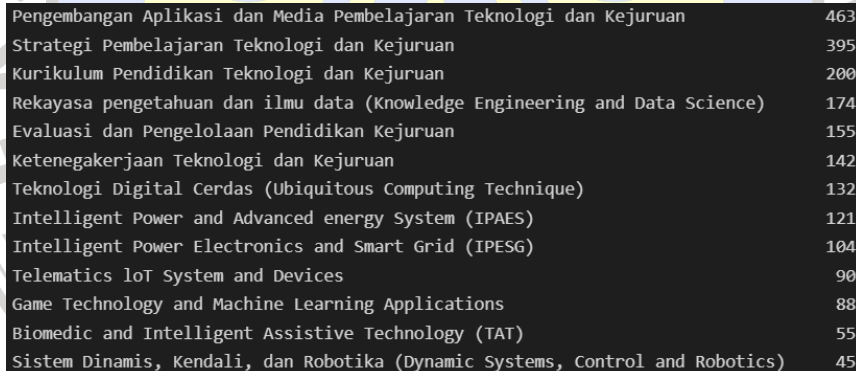


**Gambar 3.** Tahapan Metode Penelitian

Dapat dilihat pada Gambar 3 mengenai tahapan metode penelitian, tahapan pertama yang dilakukan yaitu *data collection* atau pengumpulan data. Pengumpulan data dilakukan dari database SISINTA TEUM. Data yang diperoleh masih merupakan data mentah. Sehingga, perlu dilakukan tahap *text preprocessing*. Tahapan tersebut bertujuan agar data teks menjadi lebih bersih, konsisten dan dapat dilakukan proses pengolahan selanjutnya. Setelah dilakukan *text preprocessing*, data masih dalam format teks dan harus diubah ke dalam format yang didukung oleh algoritma pembelajaran mesin. Untuk itu, tahapan selanjutnya yaitu dilakukan pembobotan terhadap istilah atau kata pada teks, sehingga menghasilkan format berupa vektor berisi bobot kata. Dalam menghindari *overfitting* karena *imbalanced class*, dilakukan proses *resampling*. Setelah itu, kumpulan data dapat dilakukan proses *training* atau pelatihan model. Tahap terakhir yaitu evaluasi model yang dilakukan untuk menghitung performa dari model yang telah dilatih.

### 3.2 Data Collection

Pada tahap ini, dilakukan pengumpulan data yang akan digunakan untuk proses pemodelan. Data yang digunakan sebagai proses pelatihan model dalam penelitian ini yaitu berupa teks judul, abstrak, dan KBK skripsi mahasiswa Teknik Elektro Universitas Negeri Malang. Sumber data didapat dari *database* web SISINTA TEUM dengan rentang waktu dari tanggal 13 April 2020 hingga 7 Oktober 2022. Semua data tersebut berjumlah 2164 baris dan 3 kolom. Kolom judul dan abstrak merupakan variabel independen atau fitur. Sedangkan kolom KBK sebagai variabel dependen atau label, dimana terdapat 13 kelas yang berbeda. Jumlah baris pada masing-masing kolom KBK ditunjukkan pada Gambar 4. Data didapatkan dalam format *.SQL (Structured Query Language)*, kemudian diubah menjadi format *.CSV (comma-separated values)* untuk mempermudah dalam pengolahan data. Kemudian, untuk memuat *dataset* yang telah diperoleh tersebut, digunakan *library* Pandas yang merupakan *library* untuk bahasa pemrograman Python. Selain itu, Pandas juga digunakan untuk melakukan analisis dan pengolahan data dalam penelitian ini.



Pengembangan Aplikasi dan Media Pembelajaran Teknologi dan Kejuruan	463
Strategi Pembelajaran Teknologi dan Kejuruan	395
Kurikulum Pendidikan Teknologi dan Kejuruan	200
Rekayasa pengetahuan dan ilmu data (Knowledge Engineering and Data Science)	174
Evaluasi dan Pengelolaan Pendidikan Kejuruan	155
Ketenagakerjaan Teknologi dan Kejuruan	142
Teknologi Digital Cerdas (Ubiquitous Computing Technique)	132
Intelligent Power and Advanced energy System (IPAES)	121
Intelligent Power Electronics and Smart Grid (IPESG)	104
Telematics IoT System and Devices	90
Game Technology and Machine Learning Applications	88
Biomedic and Intelligent Assistive Technology (TAT)	55
Sistem Dinamis, Kendali, dan Robotika (Dynamic Systems, Control and Robotics)	45

Gambar 4. Jumlah Baris pada Masing-Masing KBK

### 3.3 Text Preprocessing

Pada tahap *text preprocessing*, dilakukan pemrosesan awal terhadap data teks mentah sebelum data tersebut dapat digunakan untuk proses pemodelan teks klasifikasi. Terdapat 6 metode pemrosesan yang dilakukan dalam tahap ini: *text cleaning*, *remove missing values*, *remove duplicate data*, *tokenization*, *stop word removal*, dan *stemming*.



Tahapan pertama dalam *text preprocessing* dalam penelitian ini yaitu *text cleaning*. Proses ini dilakukan untuk menghilangkan karakter yang tidak penting, *noise* pada teks, dan membuat teks menjadi konsisten. Dalam penelitian ini, *text cleaning* dilakukan menggunakan *module regular expression* (RE) dari bahasa pemrograman Python. Terdapat beberapa teknik dalam *text cleaning* yang dilakukan: *tag removal*, *case folding*, *trim text*, penghapusan tanda baca, karakter spesial, spasi ganda, dan angka. Karena data diambil dari *database* web SISINTA TEUM, maka *tag removal* perlu dilakukan untuk menghilangkan *tag* HTML pada data teks. Kemudian, dilakukan proses *case folding* yaitu mengubah semua huruf kapital pada teks menjadi huruf kecil. Proses tersebut bertujuan untuk mencegah komputer dalam mengartikan kata yang sama dengan makna yang berbeda. Selanjutnya yaitu dilakukan proses *trim text* yang bertujuan untuk menghilangkan spasi jika terdapat pada awal dan akhir pada teks. Kemudian, dilakukan proses penghapusan terhadap tanda baca (*punctuation*), karakter spesial, spasi ganda, dan angka. Sehingga, setelah semua teknik *text cleaning* dilakukan, *noise* atau karakter yang tidak penting dalam teks akan menjadi berkurang.

Kemudian, tahap kedua dalam *text preprocessing* yaitu *remove missing values*. Pada tahap ini, dilakukan penghapusan terhadap baris yang kosong atau NaN (*Not a Number*). Penghapusan tersebut dilakukan untuk mengurangi bias pada data. Tahap selanjutnya dalam *text preprocessing* adalah *remove duplicate data*. Dalam proses tersebut, jika terdapat duplikasi data pada baris akan dihapus. Hal tersebut dilakukan untuk menghindari *overfitting*. Karena, kemungkinan data duplikat dapat terisi pada *training set* dan *test set*.

Tahapan selanjutnya yaitu dilakukan proses *tokenization*. Dalam tahap ini, teks akan dipisahkan menjadi *tokens* atau biasanya berupa kata-kata (Welbers dkk., 2017). *Tokenization* harus dilakukan untuk dapat melalui proses selanjutnya dalam *text preprocessing* yaitu *stop words removal* dan *stemming*. Tahap *tokenizing* dalam penelitian ini dilakukan menggunakan Natural Language Toolkit (NLTK) dari bahasa pemrograman Python, khususnya *package nltk.tokenize*. *Stop words removal* adalah tahapan yang dilakukan selanjutnya. Tahap ini dilakukan penghapusan kata-kata atau *tokens* yang sering muncul dan tidak memiliki makna penting dalam teks. Dalam proses ini juga dilakukan menggunakan NLTK dengan

*package* yaitu *nlk.corpus*. Dalam bahasa indonesia, stop words dapat berupa kata-kata umum yang tidak penting seperti “yang”, “dan”, “atau”, “di”, dan lain sebagainya. Setelah itu, data akan masuk pada tahap terakhir dalam *text preprocessing* yaitu *stemming*. Dalam penelitian ini, proses stemming dilakukan menggunakan *library stemmer* khusus Bahasa Indonesia yaitu Sastrawi. Semua imbuhan dalam kata seperti akhiran, sisipan, awalan, serta kombinasi antara awalan dan akhiran akan dihilangkan dalam proses *stemming*. Hal tersebut bertujuan untuk membuat suatu kata yang terbentuk dari kata dasar yang sama tidak akan diartikan oleh komputer sebagai makna yang berbeda.

### 3.4 Pembobotan Istilah

Data yang telah diproses pada *text preprocessing* masih berbentuk teks. Algoritma pembelajaran mesin tidak dapat bekerja dengan format tersebut. Sehingga, data harus diubah terlebih dahulu menjadi bentuk numerik. Pembobotan istilah dilakukan untuk mengatasi hal tersebut, dimana menghasilkan keluaran berupa vektor berisi bobot kata. Selain itu, pembobotan istilah juga bertujuan untuk memberikan bobot pada setiap kata yang menunjukkan penting atau tidaknya kata tersebut dalam dokumen. Pada penelitian ini, metode pembobotan istilah yang digunakan yaitu Term Frequency-Inverse Document Frequency (TF-IDF). Metode TF-IDF dilakukan dengan memberikan bobot frekuensi munculnya suatu kata atau istilah dalam dokumen. Ketika sebuah kata semakin sering muncul dalam sebuah dokumen, maka bobotnya meningkat secara proporsional, sedangkan jika kata itu muncul lebih sering di banyak dokumen, bobotnya menurun (Amrizal, 2018). Metode TF-IDF dalam penelitian ini dilakukan menggunakan *library* scikit-learn, khususnya *package* *feature\_extraction*.

### 3.5 Resampling

Setelah dilakukan tahap-tahap *text preprocessing*, *dataset* yang digunakan masih memiliki distribusi kelas yang tidak seimbang (*imbalanced class*). Hal ini dapat menyebabkan bias pada data. Dimana, data yang bias menaikkan jumlah prediksi salah dan menyebabkan *overfitting* (Umer dkk., 2021). Tahapan resampling dilakukan untuk mengatasi hal tersebut.

Metode resampling yang digunakan dalam penelitian ini yaitu Synthetic Minority Oversampling Technique (SMOTE). Alasan digunakannya metode oversampling ketimbang undersampling yaitu karena metode undersampling dapat menyebabkan berkurangnya informasi penting yang didapat dari kumpulan data (Maldonado dkk., 2019). Konsep dasar dari metode SMOTE yaitu menambahkan lebih banyak sampel data dari kelas minoritas untuk menyeimbangkan suatu kumpulan data. Sehingga, jumlah masing-masing kelas pada kumpulan data akan seimbang. Metode SMOTE dalam penelitian ini menggunakan *library* scikit-learn, khususnya *package* imbalanced-learn.

### 3.6 *Training Model*

Algoritma yang digunakan untuk proses *training model* dalam penelitian ini yaitu Support Vector Machine (SVM). SVM merupakan algoritma yang efektif digunakan untuk menyelesaikan permasalahan klasifikasi (Behmanesh dkk., 2021). Tahap training model pada penelitian ini menggunakan *library* scikit-learn. Dimana, proses pelatihan model dilakukan dengan persentase *training set* sebesar 70% dan *test set* sebesar 30% dari keseluruhan *dataset* yang digunakan. Penentuan perbandingan tersebut tergantung pada jumlah data yang digunakan, tetapi penggunaan *training set* 70% dan *test set* 30% merupakan standar yang baik untuk digunakan (Chollet, 2021).

Penelitian ini bertujuan untuk memecahkan masalah klasifikasi multi kelas, karena di dalam kumpulan data yang digunakan terdapat lebih dari dua kelas berbeda. Saat ini, SVM telah dikembangkan untuk dapat mengatasi klasifikasi multi kelas yaitu dengan cara menggunakan fungsi yang sesuai (Nishitsuji & Nasser, 2022). Dalam metode SVM, fungsi tersebut dikatakan sebagai fungsi kernel.

Masalah klasifikasi multi kelas dapat diselesaikan menggunakan fungsi kernel dengan metode pemisahan non-linier. Dalam penelitian ini, dilakukan pengujian terhadap 3 skenario fungsi kernel yang umum digunakan untuk klasifikasi non-linier, yaitu RBF, polynomial, dan sigmoid (Puspitasari dkk., 2018). Selain itu, dilakukan juga skenario terhadap pemilihan atribut yang digunakan untuk proses pelatihan model. Selain itu, terdapat 4 skenario nilai parameter regularisasi (C) yaitu dengan nilai 0,1, 1, 10, dan 100. Serta terdapat 3 skenario

pengujian atribut yang dilakukan: judul, abstrak, serta gabungan antara judul dan abstrak. Sehingga, tujuan dari penelitian ini yaitu untuk menemukan hasil terbaik dari kombinasi gabungan skenario yang dilakukan.

Dalam mencari kombinasi parameter terbaik digunakan metode tuning parameter yaitu Grid Search. Metode tuning ini bertujuan untuk menemukan kombinasi parameter dari model yang menghasilkan prediksi yang paling optimal. Pada penelitian ini, metode Grid Search dilakukan menggunakan *library* scikit-learn, khususnya *package* model\_selection. Tahapan Grid Search juga sudah termasuk proses *k-fold cross validation* di dalamnya. Dimana, digunakan 10 lipatan atau nilai  $k=10$  pada penelitian ini.

### 3.7 Evaluasi

Tahap evaluasi dilakukan untuk mengukur performa dari model klasifikasi yang telah dibuat. Dalam penelitian ini, digunakan metode evaluasi yaitu *confusion matrix*. *Confusion Matrix* digunakan mencari nilai akurasi, presisi, *recall* dan *f1-score*. Sehingga, dengan perhitungan nilai-nilai tersebut dapat diketahui kinerja dari model yang dibuat.

## BAB IV

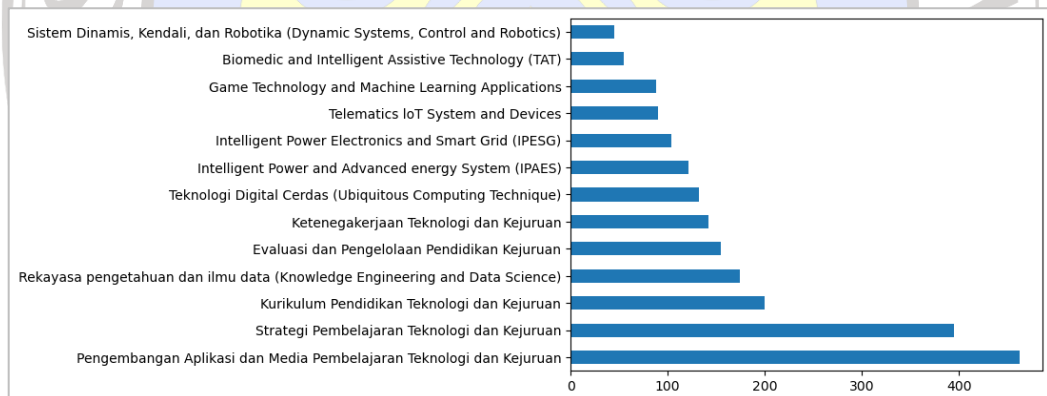
### HASIL DAN PEMBAHASAN

#### 4.1 Hasil Data Collection

Dalam penelitian ini, data yang dikumpulkan berupa tabel yang berjumlah 2164 baris. Terdapat 3 kolom pada data tabel tersebut yaitu kolom judul, abstrak, dan KBK. Data tersebut bertipe data *string*. Pada kolom judul dan abstrak, data masih terdapat *tag* HTML yang perlu dibersihkan. Sebagaimana contoh hasil data collection dapat dilihat pada Gambar 5. Selain itu, distribusi jumlah kelas pada data mentah ini ditunjukkan pada Gambar 6.

judul	abstrak	kbk
<p>Pengembangan Sistem Pendukung Keputusan unt...	<p>Sistem Pendukung Keputusan (SPK) merupakan ...	Pengembangan Aplikasi dan Media Pembelajaran T...
<p>HUBUNGAN EFIKASI DIRI DENGAN KESIAPAN KERJA...	<p>Pandemi covid-19 yang melanda dunia, teruta...	Ketenegakerjaan Teknologi dan Kejuruan
<p>Alat Bantu Penyandang Tunanetra Berbasis De...	<p>Tujuan dilakukannya penelitian ini untuk me...	Biomedic and Intelligent Assistive Technology ...
<p class="MsoNormal" style="margin-left:35.45p...	<p><span style="font-size:12.0pt;line-height:1...	Intelligent Power Electronics and Smart Grid (...)
<p class="MsoNormal" align="center" style="tex...	<p class="MsoNormal" style="text-align:justify...	Pengembangan Aplikasi dan Media Pembelajaran T...

**Gambar 5.** Contoh Hasil Data Collection



**Gambar 6.** Distribusi Kelas pada Data Mentah

#### 4.2 Hasil Text Preprocessing

##### 4.2.1 Hasil Text Cleaning

Tahapan text cleaning dilakukan proses *tag removal*, *case folding*, *trim text*, penghapusan tanda baca, karakter spesial, spasi ganda, dan angka. Contoh hasil pemrosesan tahap ini dapat dilihat pada Gambar 7.



## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

judul	abstrak	kbk
pengembangan sistem pendukung keputusan untuk ...	sistem pendukung keputusan spk merupakan suatu...	Pengembangan Aplikasi dan Media Pembelajaran T...
hubungan efikasi diri dengan kesiapan kerja lu...	pandemi covid yang melanda dunia terutama indo...	Ketenegakerjaan Teknologi dan Kejuruan
alat bantu penyanggah tuetra berbasis deteksi ...	tujuan dilakukannya penelitian ini untuk memba...	Biomedic and Intelligent Assistive Technology ...
analisis thermovisi penghantar akibat transmis...	gardu induk waru merupakan sub transmisi listr...	Intelligent Power Electronics and Smart Grid (...)
pengembangan modul berbasis production based ed...	mata pelajaran dasar desain grafis merupakan m...	Pengembangan Aplikasi dan Media Pembelajaran T...

**Gambar 7.** Contoh Hasil Text Cleaning

Tahap selanjutnya yaitu dilakukan penghapusan terhadap *missing values*. Dalam *dataset*, terdapat 4 baris *missing values* pada kolom judul dan 896 baris *missing values* pada kolom abstrak. Dimana, jumlah *missing values* dalam *dataset* dapat dilihat pada Gambar 8.

judul	4
abstrak	896
kbk	0

**Gambar 8.** Jumlah Missing Values pada Setiap Kolom Dataset

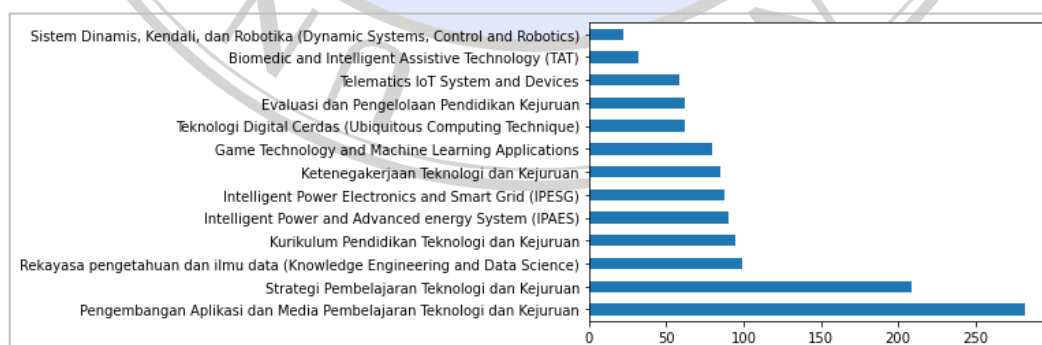
Selain itu, dilakukan juga penghapusan terhadap duplikasi baris pada data. Pada kolom judul dan abstrak, masing-masing terdapat 1 duplikasi data yang mana dapat dilihat pada Gambar 9 dan Gambar 10. Kemudian, jumlah distribusi kelas setelah dilakukan semua tahapan *text preprocessing* dapat dilihat pada Gambar 11.

judul	abstrak	kbk
pengembangan sistem pendukung keputusan untuk ...	proses penentuan dosen pembimbing skripsi atau...	Game Technology and Machine Learning Applications

**Gambar 9.** Duplikasi pada Kolom Judul

judul	abstrak	kbk
pengembangan alat peraga pemindah barang berba...	industri dan perkembangannya tentu tidak lepas...	Pengembangan Aplikasi dan Media Pembelajaran T...

**Gambar 10.** Duplikasi pada Kolom Abstrak



**Gambar 11.** Distribusi Kelas Setelah Text Preprocessing

#### 4.2.2 Hasil Tokenization

Tahap *tokenization* dilakukan untuk memisahkan teks menjadi tokens atau kata-kata. Kemudian, untuk contoh hasil perubahan dari proses *tokenization* pada kolom judul dan abstrak dapat dilihat pada Gambar 12 dan Gambar 13. Sedangkan untuk kolom KBK tidak perlu dilakukan *tokenization* karena kolom tersebut akan digunakan sebagai label pada *dataset*.

judul	judul_tokens
pengembangan sistem pendukung keputusan untuk ...	[pengembangan, sistem, pendukung, keputusan, u...
hubungan efikasi diri dengan kesiapan kerja lu...	[hubungan, efikasi, diri, dengan, kesiapan, ke...
alat bantu penyanggah tuetra berbasis deteksi ...	[alat, bantu, penyanggah, tuetra, berbasis, de...
analisis thermovisi penghantar akibat transmis...	[analisis, thermovisi, penghantar, akibat, tra...
pengembangan modulberbasis production based ed...	[pengembangan, modulberbasis, production, base...

**Gambar 12.** Hasil Tokenization pada Kolom Judul

abstrak	abstrak_tokens
sistem pendukung keputusan spk merupakan suatu...	[sistem, pendukung, keputusan, spk, merupakan,...
pandemi covid yang melanda dunia terutama indo...	[pandemi, covid, yang, melanda, dunia, terutama...
tujuan dilakukannya penelitian ini untuk memba...	[tujuan, dilakukannya, penelitian, ini, untuk,...
gardu induk waru merupakan sub transmisi listr...	[gardu, induk, waru, merupakan, sub, transmisi...
mata pelajaran dasar desain grafis merupakan m...	[mata, pelajaran, dasar, desain, grafis, merup...

**Gambar 13.** Hasil Tokenization pada Kolom Abstrak

#### 4.2.3 Hasil Stop Words Removal

Tahap stop word removal dilakukan untuk menghapus kata atau tokens yang sering muncul dan tidak memiliki makna penting dalam teks. Untuk contoh hasil perubahan dari proses stop word removal pada kolom judul dan abstrak dapat dilihat pada Gambar 14 dan Gambar 15.

judul	judul_tokens
pengembangan sistem pendukung keputusan untuk ...	[pengembangan, sistem, pendukung, keputusan, m...
hubungan efikasi diri dengan kesiapan kerja lu...	[hubungan, efikasi, kesiapan, kerja, lulusan, ...
alat bantu penyanggah tuetra berbasis deteksi ...	[alat, bantu, penyanggah, tuetra, berbasis, de...
analisis thermovisi penghantar akibat transmis...	[analisis, thermovisi, penghantar, akibat, tra...
pengembangan modulberbasis production based ed...	[pengembangan, modulberbasis, production, base...

**Gambar 14.** Hasil Stop Words Removal pada Kolom Judul

## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

abstrak	abstrak_tokens
sistem pendukung keputusan spk merupakan suatu...	[sistem, pendukung, keputusan, spk, sistem, me...
pandemi covid yang melanda dunia terutama indo...	[pandemi, covid, melanda, dunia, indonesia, da...
tujuan dilakukannya penelitian ini untuk memba...	[tujuan, dilakukannya, penelitian, membantu, p...
gardu induk waru merupakan sub transmisi listr...	[gardu, induk, waru, sub, transmisi, listrik, ...
mata pelajaran dasar desain grafis merupakan m...	[mata, pelajaran, dasar, desain, grafis, mata, ...

**Gambar 15.** Hasil Stop Words Removal pada Kolom Abstrak

### 4.2.4 Hasil Stemming

Tahap stemming dilakukan untuk menghapus semua imbuhan dalam kata seperti akhiran, sisipan, awalan, serta kombinasi antara awalan dan akhiran. Untuk contoh hasil perubahan dari proses stemming pada kolom judul dan abstrak dapat dilihat pada Gambar 16 dan Gambar 17.

judul	judul_tokens
pengembangan sistem pendukung keputusan untuk ...	[kembang, sistem, dukung, putus, tentu, dosen, ...
hubungan efikasi diri dengan kesiapan kerja lu...	[hubung, efikasi, kesiap, kerja, lulus, smk, n...
alat bantu penyandang tuetra berbasis deteksi ...	[alat, bantu, sandang, tuetra, bas, deteksi, o...
analisis thermovisi penghantar akibat transmis...	[analisis, thermovisi, hantar, akibat, transmi...
pengembangan modulberbasis production based ed...	[kembang, modulberbasis, production, based, ed...

**Gambar 16.** Hasil Stemming pada Kolom Judul

abstrak	abstrak_tokens
sistem pendukung keputusan spk merupakan suatu...	[sistem, dukung, putus, spk, sistem, milik, ke...
pandemi covid yang melanda dunia terutama indo...	[pandemi, covid, landa, dunia, indonesia, damp...
tujuan dilakukannya penelitian ini untuk memba...	[tuju, laku, teliti, bantu, sandang, tuetra, g...
gardu induk waru merupakan sub transmisi listr...	[gardu, induk, waru, sub, transmisi, listrik, ...
mata pelajaran dasar desain grafis merupakan m...	[mata, ajar, dasar, desain, grafis, mata, ajar...

**Gambar 17.** Hasil Stemming pada Kolom Abstrak

### 4.3 Hasil Pembobotan Istilah Menggunakan TF-IDF

Dalam tahapan ini, dilakukan pembobotan terhadap data teks menjadi vektor berisi bobot dari masing-masing kata. Dimana, pada data *train set* menghasilkan vektor 884 x 2300. Sedangkan pada *test set* menghasilkan vektor 380 x 2300. Contoh hasil pembobotan istilah menggunakan TF-IDF dapat dilihat pada Gambar 18.



## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

term	rank
ajar	74.310959
siswa	34.673976
smk	33.582109
kembang	32.745617
kelas	31.869886
...	...
plant	0.220052
doubly	0.220052
sarima	0.200231
komoditas	0.200231
arima	0.200231

**Gambar 18.** Hasil Pembobotan Menggunakan TF-IDF

### 4.4 Hasil Resampling

Dapat dilihat pada Gambar 19, dataset yang digunakan memiliki distribusi kelas yang tidak seimbang (imbalanced class). Sehingga, hasil distribusi kelas setelah dilakukan oversampling menggunakan metode Synthetic Minority Oversampling Technique (SMOTE) dijelaskan pada Gambar 20.

Pengembangan Aplikasi dan Media Pembelajaran Teknologi dan Kejuruan	194
Strategi Pembelajaran Teknologi dan Kejuruan	147
Kurikulum Pendidikan Teknologi dan Kejuruan	72
Intelligent Power and Advanced energy System (IPAES)	68
Rekayasa pengetahuan dan ilmu data (Knowledge Engineering and Data Science)	66
Intelligent Power Electronics and Smart Grid (IPESG)	64
Ketenagakerjaan Teknologi dan Kejuruan	54
Game Technology and Machine Learning Applications	53
Evaluasi dan Pengelolaan Pendidikan Kejuruan	47
Telematics IoT System and Devices	44
Teknologi Digital Cerdas (Ubiquitous Computing Technique)	41
Biomedic and Intelligent Assistive Technology (TAT)	19
Sistem Dinamis, Kendali, dan Robotika (Dynamic Systems, Control and Robotics)	15
Name: kbk, dtype: int64	

**Gambar 19.** Distribusi Kelas Tidak Seimbang pada Dataset

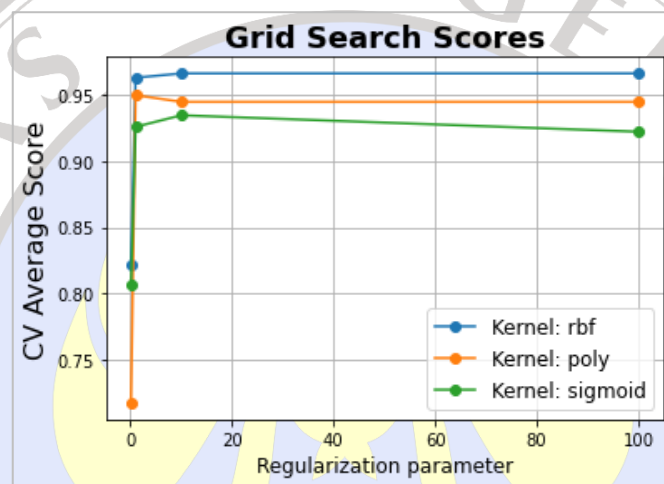
Pengembangan Aplikasi dan Media Pembelajaran Teknologi dan Kejuruan	194
Telematics IoT System and Devices	194
Sistem Dinamis, Kendali, dan Robotika (Dynamic Systems, Control and Robotics)	194
Evaluasi dan Pengelolaan Pendidikan Kejuruan	194
Ketenagakerjaan Teknologi dan Kejuruan	194
Rekayasa pengetahuan dan ilmu data (Knowledge Engineering and Data Science)	194
Biomedic and Intelligent Assistive Technology (TAT)	194
Kurikulum Pendidikan Teknologi dan Kejuruan	194
Intelligent Power and Advanced energy System (IPAES)	194
Teknologi Digital Cerdas (Ubiquitous Computing Technique)	194
Intelligent Power Electronics and Smart Grid (IPESG)	194
Strategi Pembelajaran Teknologi dan Kejuruan	194
Game Technology and Machine Learning Applications	194
Name: kbk, dtype: int64	

**Gambar 20.** Hasil Resampling Menggunakan SMOTE

## 4.6 Hasil Training Model

### 4.6.1 Skenario Judul

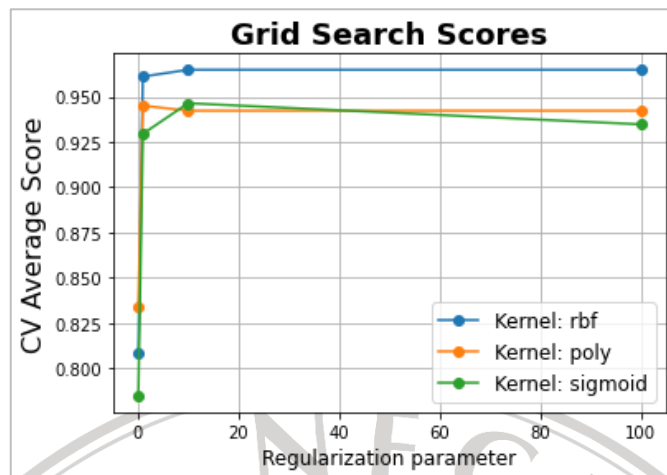
Dengan menggunakan metode Grid Search diperoleh kombinasi parameter terbaik pada dataset train judul. Hasil metode Grid pada skenario ini Search dapat dilihat pada Gambar 21. Pada gambar tersebut menunjukkan bahwa kombinasi parameter dengan nilai regularisasi atau  $C=10$  dan menggunakan kernel RBF merupakan parameter paling optimal yaitu dengan nilai rata-rata 10-fold *cross validation* sebesar 96,66%. Sedangkan kombinasi parameter yang paling tidak optimal yaitu pada nilai  $C=0,1$  dan menggunakan kernel *polynomial*.



Gambar 21. Hasil Grid Search pada Skenario Judul

### 4.6.2 Skenario Abstrak

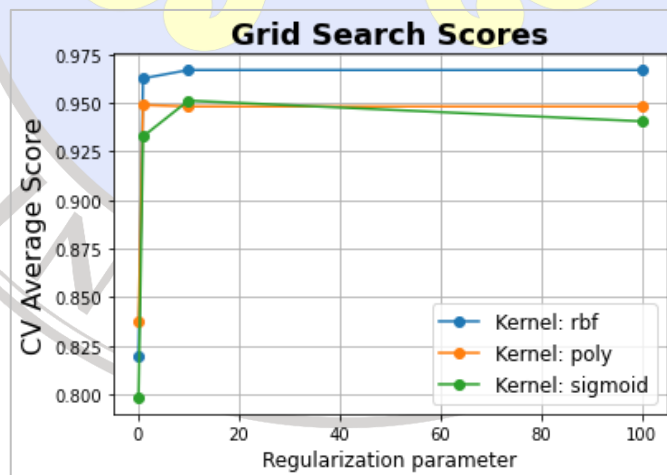
Hasil metode Grid Search pada skenario ini dapat dilihat pada Gambar 22. Pada gambar tersebut menunjukkan bahwa kombinasi parameter dengan nilai regularisasi atau  $C=10$  dan menggunakan kernel RBF merupakan parameter paling optimal yaitu dengan nilai rata-rata 10-fold *cross validation* sebesar 96,49%. Sedangkan kombinasi parameter yang paling tidak optimal yaitu pada nilai  $C=0,1$  dan menggunakan kernel *sigmoid*.



Gambar 22. Hasil Grid Search pada Skenario Abstrak

#### 4.6.3 Skenario Judul dan Abstrak

Hasil metode Grid Search pada skenario ini dapat dilihat pada Gambar 23. Pada gambar tersebut menunjukkan bahwa kombinasi parameter dengan nilai regularisasi atau  $C=10$  dan menggunakan kernel RBF merupakan parameter paling optimal yaitu dengan nilai rata-rata *10-fold cross validation* sebesar 96,69%. Sedangkan kombinasi parameter yang paling tidak optimal yaitu pada nilai  $C=0,1$  dan menggunakan kernel *sigmoid*.



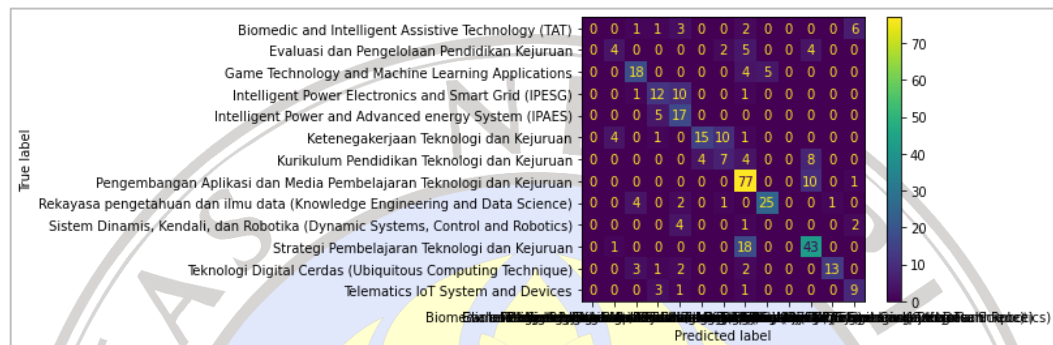
Gambar 23. Hasil Grid Search pada Skenario Judul dan Abstrak

## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

### 4.7 Hasil Evaluasi

#### 4.7.1 Skenario Judul

Hasil confusion matrix pada skenario ini ditunjukkan pada Gambar 24. Dimana, dari tabel confusion matrix tersebut menghasilkan nilai akurasi, presisi, recall, dan f1-score. Hasil perhitungan keempat metrik tersebut dapat dilihat pada Gambar 25.



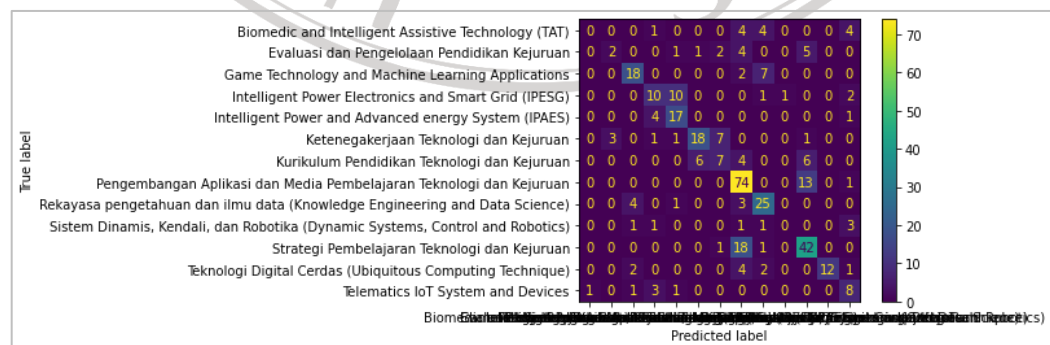
**Gambar 24.** Confusion Matrix pada Skenario Judul

Accuracy score : 0.631578947368421  
Precision score : 0.6124500297794848  
Recall score : 0.631578947368421  
F1 score : 0.6097028606103326

**Gambar 25.** Evaluasi pada Skenario Judul

#### 4.7.1 Skenario Abstrak

Hasil confusion matrix pada skenario ini ditunjukkan pada Gambar 26. Dimana, dari tabel confusion matrix tersebut menghasilkan nilai akurasi, presisi, recall, dan f1-score. Hasil perhitungan keempat metrik tersebut dapat dilihat pada Gambar 27.



**Gambar 26.** Confusion Matrix pada Skenario Abstrak

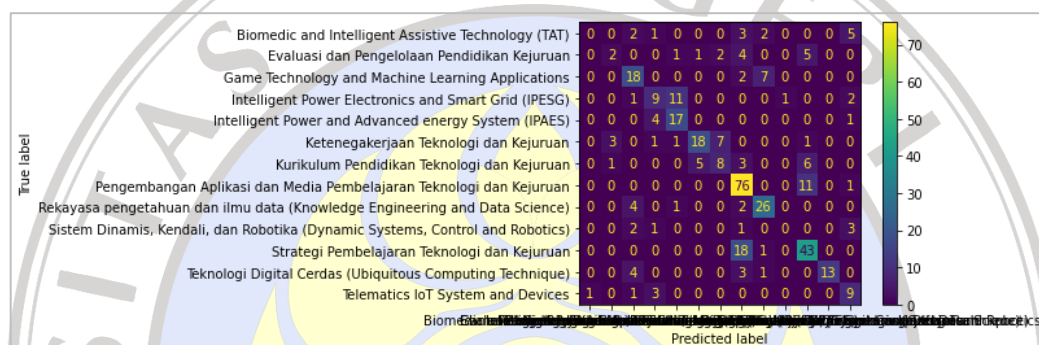
## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

Accuracy score : 0.6131578947368421  
 Precision score : 0.5875206366723188  
 Recall score : 0.6131578947368421  
 F1 score : 0.5877528761539266

Gambar 27. Evaluasi pada Skenario Abstrak

### 4.7.1 Skenario Judul dan Abstrak

Hasil confusion matrix pada skenario ini ditunjukkan pada Gambar 28. Dimana, dari tabel confusion matrix tersebut menghasilkan nilai akurasi, presisi, recall, dan f1-score. Hasil perhitungan keempat metrik tersebut dapat dilihat pada Gambar 29.



Gambar 28. Confusion Matrix pada Skenario Judul dan Abstrak

Accuracy score : 0.6289473684210526  
 Precision score : 0.599977318133797  
 Recall score : 0.6289473684210526  
 F1 score : 0.6033992279947701

Gambar 29. Evaluasi pada Skenario Judul dan Abstrak

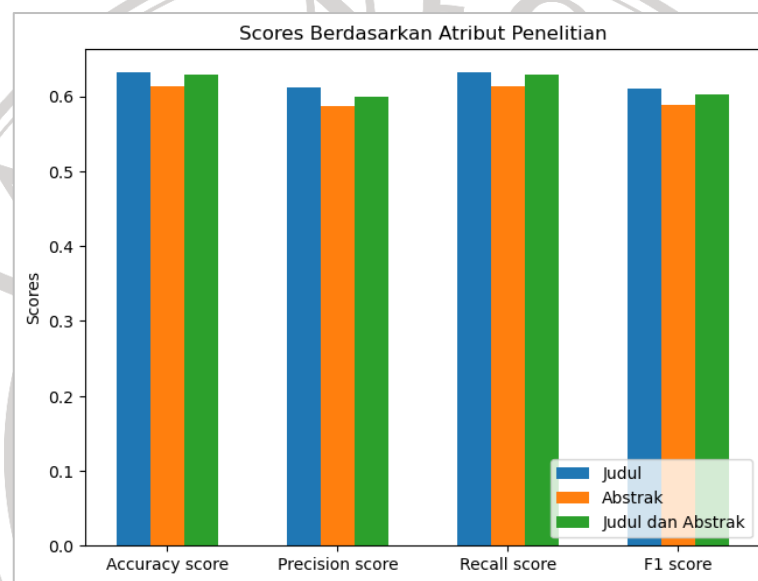
### 4.8 Analisis Perbandingan Hasil Performa

Dari hasil *tuning parameter* yang dilakukan menggunakan metode Grid Search, parameter paling optimal dalam semua skenario yaitu pada penggunaan kernel RBF dan nilai regularisasi = 10. Kemudian, didapatkan hasil evaluasi dari masing-masing skenario menggunakan parameter paling optimal, dimana perbandingan hasil metrik evaluasi divisualisasikan pada Gambar 30. Proses evaluasi tersebut diuji pada *test set* yaitu sebesar 30% dari seluruh data. Dari hasil perbandingan evaluasi tersebut, diperoleh bahwa skenario dengan menggunakan *input data* judul merupakan skenario terbaik dan optimal. Dimana, hasil *cross*

## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

*validation* menggunakan *train set* pada skenario tersebut menghasilkan nilai rata-rata sebesar 96,66%.

Hasil evaluasi pada skenario *input data* judul memiliki perbedaan yang tidak terlalu signifikan dibanding dengan skenario *input data* gabungan judul dan abstrak. Namun, skenario dengan *input data* judul jauh lebih efisien dibandingkan dengan skenario tersebut. Karena, dengan menggunakan data judul sebagai variabel input, membutuhkan lebih sedikit fitur untuk pemrosesan. Sehingga, hal tersebut dapat mengurangi waktu komputasi.



**Gambar 30.** Perbandingan Performa Masing-Masing Skenario

Hasil nilai metrik evaluasi dari masing-masing skenario dapat dikatakan relatif kecil. Hal tersebut kemungkinan dipengaruhi oleh beberapa faktor, yaitu terdapat kesalahan penulisan seperti galat tipografi (*typo*), kata bergandengan, *human error*, dan kurangnya proses validasi untuk memeriksa kesalahan tersebut. Dimana, contoh kesalahan penulisan yang terdapat pada *dataset* ditunjukkan pada Gambar 31.



## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

judul
pengembangan sistem pendukung keputusan untuk ...
hubungan efikasi diri dengan kesiapan kerja lu...
alat bantu penyanggah <u>tuetra</u> berbasis deteksi ...
analisis thermovisi penghantar akibat transmis...
pengembangan <u>modul</u> berbasis production based ed...

**Gambar 31.** Contoh Kesalahan Penulisan pada Dataset

Selain itu, terdapat irisan (*overlapping classes*) topik riset yang besar antarkelas KBK. Dimana, kombinasi antara masalah *class overlapping* dan *class imbalance* membuat masalah yang lebih rumit (Xiong dkk., 2010). Contohnya yaitu pada KBK “*Knowledge Engineering and Data Science*” dengan “*Game Technology and Machine Learning Applications*”. Penelitian dengan topik riset *data science* kemungkinan besar dapat masuk juga kedalam topik riset *machine learning*. Karena, kebanyakan dari topik riset *data science* menggunakan algoritma *machine learning* dalam menyelesaikan tugasnya.

Adapun terdapat kemungkinan solusi yang dapat dilakukan untuk meningkatkan performa sistem klasifikasi KBK yang dibuat yaitu dengan memperbanyak data yang digunakan. Pada penelitian ini menggunakan data KBK dari tahun 2020 hingga 2022. Sehingga, dengan memperbanyak dan memperluas rentang data tersebut, kemungkinan dapat memberikan hasil yang lebih baik.

Selain itu, diperlukan penambahan metode *n-gram* pada tahap *text preprocessing*. Dengan menggunakan *n-gram*, akan diperoleh lebih banyak informasi dari ekstraksi fitur yang dilakukan. Namun, penggunaan *n-gram* dapat menyebabkan bengkaknya fitur pada *input data*. Sehingga, untuk mengatasi hal tersebut, juga diperlukan algoritma dengan pendekatan *Deep Learning* yang kompleks.

Dari penelitian yang dilakukan, kelebihan dari sistem klasifikasi yang dibuat yaitu sistem dapat memberikan rekomendasi KBK hanya dengan menggunakan *input data* judul skripsi. Namun, kekurangannya yaitu sistem masih belum dapat mengatasi masalah kesalahan penulisan seperti galat tipografi dan huruf bergandengan. Solusi yang mungkin dapat mengatasi hal tersebut yaitu dengan menambahkan fitur pemeriksaan ejaan (*spell check*).

## BAB V

### KESIMPULAN DAN SARAN

#### 5.1 Kesimpulan

Dari penelitian yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut:

1. Algoritma Support Vector Machine dapat digunakan untuk mengimplementasikan sistem klasifikasi KBK berdasarkan judul dan abstrak skripsi Jurusan Teknik Elektro Universitas Negeri Malang.
2. Skenario terbaik dari penelitian ini yaitu pelatihan model dengan variabel input berupa data judul. Dengan menggunakan data judul sebagai input variabel dinilai optimal dan efisien, karena hanya menggunakan sedikit fitur. Skenario tersebut menghasilkan performa akurasi, presisi, *recall*, dan *f1-score* berturut-turut yaitu 63,16%, 61,25%, 63,16%, dan 60,34%. Selain itu, rata-rata hasil *10-fold cross validation* pada skenario tersebut yaitu sebesar 96,66%.

#### 5.2 Saran

Berdasarkan keterbatasan dari penelitian yang dilakukan, terdapat beberapa saran yang dapat ditindaklanjuti sebagai pengembangan untuk penelitian selanjutnya:

1. Memperbanyak kumpulan data yang digunakan. Dalam penelitian ini, dataset yang dikumpulkan berasal dari tahun 2020 hingga 2022. Dengan memperluas data yang digunakan, kemungkinan dapat memberikan hasil yang berbeda.
2. Menambahkan metode *n-gram* pada *text preprocessing*. *N-gram* dapat meningkatkan model dalam melakukan representasi teks dengan lebih baik. Sehingga, kemungkinan hal ini dapat meningkatkan performa dari model.
3. Menggunakan algoritma dengan pendekatan *Deep Learning*. Karena, penggunaan *n-gram* menyebabkan bengkaknya fitur data. Sehingga, metode *Deep Learning* dinilai dapat mengatasi hal tersebut.



### DAFTAR RUJUKAN

- A. Mullen, L., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, Consistent Tokenization of Natural Language Text. *Journal of Open Source Software*, 3(23), 655. <https://doi.org/10.21105/joss.00655>
- Aliwy, A. H., & Ameer, E. H. A. (2017). *Comparative Study of Five Text Classification Algorithms with their Improvements*. 12(14), 11.
- Amrizal, V. (2018). Penerapan Metode Term Frequency Inverse Document Frequency (TF-IDF) dan Cosine Similarity pada Sistem Temu Kembali Informasi untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim). *JURNAL TEKNIK INFORMATIKA*, 11(2), 149–164. <https://doi.org/10.15408/jti.v11i2.8623>
- Behmanesh, M., Adibi, P., & Karshenas, H. (2021). *Weighted Least Squares Twin Support Vector Machine with Fuzzy Rough Set Theory for Imbalanced Data Classification*. 13.
- Chollet, F. (2021). *Deep Learning with Python, Second Edition*. Simon and Schuster.
- Czarnowski, I. (2022). Weighted Ensemble with one-class Classification and Over-sampling and Instance selection (WECOI): An approach for learning from imbalanced data streams. *Journal of Computational Science*, 61, 101614. <https://doi.org/10.1016/j.jocs.2022.101614>
- Fan, H., & Qin, Y. (2018). Research on Text Classification Based on Improved TF-IDF Algorithm. *Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)*. 2018 International Conference on Network, Communication, Computer

Engineering (NCCE 2018), Chongqing, China.

<https://doi.org/10.2991/ncce-18.2018.79>

Faraby, S. A. (2018). *Analisis dan Implementasi Support Vector Machine dengan String Kernel dalam Melakukan Klasifikasi Berita Berbahasa Indonesia*. 10.

Ganeshkumar, P., & Padmanabhan, S. (2022). *Social Media Personal Event Notifier Using NLP and Machine Learning*. 4.

Hadianto, N., Novitasari, H. B., & Rahmawati, A. (2019). Klasifikasi Peminjaman Nasabah Bank Menggunakan Metode Neural Network. *Jurnal Pilar Nusa Mandiri*, 15(2), 163–170. <https://doi.org/10.33480/pilar.v15i2.658>

Haghighi, S., Jasemi, M., Hessabi, S., & Zolanvari, A. (2018). PyCM: Multiclass confusion matrix library in Python. *Journal of Open Source Software*, 3(25), 729. <https://doi.org/10.21105/joss.00729>

Irmada, H. N. & Ria Astriratma. (2020). Klasifikasi Jenis Pantun Dengan Metode Support Vector Machines (SVM). *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 4(5), 915–922. <https://doi.org/10.29207/resti.v4i5.2313>

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Sunday, S., & Hussain, S. (2022). *Effective Class-Imbalance learning based on SMOTE and Convolutional Neural Networks*. 43.

*Kelompok Bidang Keahlian (KBK) / Teknik Elektro – UM*. (2020). <http://elektro.um.ac.id/kelompok-bidang-keahlian-kbk/>

Khairunnisa, S., Adiwijaya, A., & Faraby, S. A. (2021). Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada

Media Sosial Twitter (Studi Kasus Pandemi COVID-19). *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 5(2), 406.  
<https://doi.org/10.30865/mib.v5i2.2835>

Krell, M. M., Wilshusen, N., Seeland, A., & Kim, S. K. (2017). Classifier Transfer with Data Selection Strategies for Online Support Vector Machine Classification with Class Imbalance. *Journal of Neural Engineering*, 14(2), 025003. <https://doi.org/10.1088/1741-2552/aa5166>

Lampinen, A. K., & McClelland, J. L. (2018). *One-Shot and Few-Shot Learning of Word Embeddings* (arXiv:1710.10280). arXiv.  
<http://arxiv.org/abs/1710.10280>

Li, D.-C., Wang, S.-Y., Huang, K.-C., & Tsai, T.-I. (2022). Learning Class-Imbalanced Data with Region-Impurity Synthetic Minority Oversampling Technique. *Information Sciences*, 607, 1391–1407.  
<https://doi.org/10.1016/j.ins.2022.06.067>

Li, W., Chen, J., Cao, J., Ma, C., Wang, J., Cui, X., & Chen, P. (2022). EID-GAN: Generative Adversarial Nets for Extremely Imbalanced Data Augmentation. *IEEE Transactions on Industrial Informatics*, 1–10.  
<https://doi.org/10.1109/TII.2022.3182781>

Listiana, E. (2017). *Penerapan Adaboost untuk Klasifikasi Support Vector Machine Guna Meningkatkan Akurasi pada Diagnosa Chronic Kidney Disease*. 7.

Maldonado, S., López, J., & Vairetti, C. (2019). An Alternative Smote Oversampling Strategy for High-Dimensional Datasets. *Applied Soft Computing*, 76, 380–389. <https://doi.org/10.1016/j.asoc.2018.12.024>

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mayabadi, S., & Saadatfar, H. (2022). Two Density-Based Sampling Approaches for Imbalanced and Overlapping Data. *Knowledge-Based Systems*, 241, 108217. <https://doi.org/10.1016/j.knosys.2022.108217>

Mutawalli, L., Zaen, M. T. A., & Bagye, W. (2019). Klasifikasi Teks Sosial Media Twitter Menggunakan Support Vector Machine (Studi Kasus Penusukan Wiranto). *Jurnal Informatika dan Rekayasa Elektronik*, 2(2), 43. <https://doi.org/10.36595/jire.v2i2.117>

Nishitsuji, Y., & Nasser, J. (2022). *Support-Vector-Machine with Bayesian Optimization for Lithofacies Classification Using Elastic Properties* (arXiv:2204.00081). arXiv. <http://arxiv.org/abs/2204.00081>

Normawati, D., & Prayogi, S. A. (2021). *Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter*. 5, 15.

Novitasari, D. (2017). Perbandingan Algoritma Stemming Porter dengan Arifin Setiono untuk Menentukan Tingkat Ketepatan Kata Dasar. *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, 1(2), 120. <https://doi.org/10.30998/string.v1i2.1031>

Oryza Habibie Rahman, Gunawan Abdillah, & Agus Komarudin. (2021). Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 5(1), 17–23. <https://doi.org/10.29207/resti.v5i1.2700>

Pujianto, U., Widiyaningtyas, T., Prasetya, D. D., & Romadhon, B. (2019).

Penerapan Algoritma Naïve Bayes Classifier untuk Klasifikasi Judul Skripsi dan Tugas Akhir Berdasarkan Kelompok Bidang Keahlian. *TEKNO*, 27(1), 79. <https://doi.org/10.17977/um034v27i1p79-92>

Puspitasari, A. M., Ratnawati, D. E., & Widodo, A. W. (2018). *Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine*. 9.

Riza Adrianti Supono & Muhammad Azis Suprayogi. (2021). Perbandingan Metode TF-ABS dan TF-IDF Pada Klasifikasi Teks Helpdesk Menggunakan K-Nearest Neighbor. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 5(5), 911–918. <https://doi.org/10.29207/resti.v5i5.3403>

Septian, J. A., Fachrudin, T. M., & Nugroho, A. (2019). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor. *Journal of Intelligent System and Computation*, 1(1), 43–49. <https://doi.org/10.52985/insyst.v1i1.36>

Simarangkir, M. S. H. (2017). Studi Perbandingan Algoritma-Algoritma Stemming untuk Dokumen Teks Bahasa Indonesia. *Jurnal Inkofar*, 1(1). <https://doi.org/10.46846/jurnalinkofar.v1i1.2>

Umer, M., Sadiq, S., Missen, M. M. S., Hameed, Z., Aslam, Z., Siddique, M. A., & Nappi, M. (2021). Scientific Papers Citation Analysis using Textual Features and Smote Resampling Techniques. *Pattern Recognition Letters*, 150, 250–257. <https://doi.org/10.1016/j.patrec.2021.07.009>

Universitas Negeri Malang. (2015). *Peraturan Rektor Universitas Negeri Malang Nomor 4 Tahun 2015 Tentang Pembentukan dan Pengembangan Kelompok Bidang Keahlian Dosen Universitas Negeri Malang.*

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures*, 11(4), 245–265.  
<https://doi.org/10.1080/19312458.2017.1387238>

Xiong, H., Wu, J., & Liu, L. (2010). Classification with ClassOverlapping: A Systematic Study. *Proceedings of the 2010 International Conference on E-Business Intelligence*. 2010 International Conference on E-Business Intelligence (ICEBI-2010), China. <https://doi.org/10.2991/icebi.2010.43>

Żurek, D., & Pietroń, M. (2020). *Training with Reduced Precision of A Support Vector Machine Model for Text Classification* (arXiv:2007.08657). arXiv. <http://arxiv.org/abs/2007.08657>



## LAMPIRAN

### Lampiran 1. Perhitungan Skala Likert Survey

**Pertanyaan 1:** Apakah Anda merasa kebingungan dan/atau kesulitan saat memilih KBK berdasarkan judul saat mengajukan judul skripsi di SISINTA?

Sentiment Level	Value Weight	Responses	Total
Sangat mudah	1	2	2
Mudah	2	11	22
Sulit	3	4	12
Sangat sulit	4	8	32

$$Rumus = \frac{Total\ Skor}{Jumlah\ Responden \times Jumlah\ Sentiment\ Level} = \frac{68}{25 \times 4} = 0,68$$

Hasil perhitungan skala likert yaitu 0,68, dimana berada pada rentang antara 0,5 dan 0,75. Sehingga, dapat diperoleh kesimpulan bahwa pengguna merasa sulit saat memilih KBK berdasarkan judul di SISINTA.

**Pertanyaan 2:** Apakah Anda setuju jika sistem informasi SISINTA memiliki rekomendasi KBK secara otomatis?

Sentiment Level	Value Weight	Responses	Total
Sangat tidak setuju	1	4	4
Tidak setuju	2	2	4
Setuju	3	5	15
Sangat setuju	4	14	56

$$Rumus = \frac{Total\ Skor}{Jumlah\ Responden \times Jumlah\ Sentiment\ Level} = \frac{79}{25 \times 4} = 0,79$$

Hasil perhitungan skala likert yaitu 0,79, dimana berada pada rentang antara 0,75 dan 1. Sehingga, dapat diperoleh kesimpulan bahwa pengguna sangat setuju jika web SISINTA memiliki sistem rekomendasi KBK secara otomatis.

**Lampiran 2. Perbandingan Performa pada Masing-Masing Skenario**

<b>Skenario</b>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
Judul	63,16%	61,25%	63,16%	60,34%
Abstrak	61,32%	58,75%	61,32%	58,78%
Judul dan Abstrak	62,89%	60%	62,89%	60,34%





## Klasifikasi Kelompok Bidang Keahlian (KBK) Berdasarkan Judul dan Abstrak Skripsi Menggunakan Algoritma Support Vector Machine

### Lampiran 3. Demo Sistem

Judul Skripsi: Klasifikasi KBK Menggunakan Algoritma Support Vector Machine		
	Kelompok Bidang Keahlian	Probability
8	Rekayasa pengetahuan dan ilmu data (Knowledge ...	0.538954
2	Game Technology and Machine Learning Applications	0.363559
3	Intelligent Power Electronics and Smart Grid (...)	0.019181
4	Intelligent Power and Advanced energy System (...)	0.017674
7	Pengembangan Aplikasi dan Media Pembelajaran T...	0.014713
11	Teknologi Digital Cerdas (Ubiquitous Computing...	0.010221
10	Strategi Pembelajaran Teknologi dan Kejuruan	0.008828
6	Kurikulum Pendidikan Teknologi dan Kejuruan	0.006526
12	Telematics IoT System and Devices	0.006311
5	Ketenagakerjaan Teknologi dan Kejuruan	0.004993
1	Evaluasi dan Pengelolaan Pendidikan Kejuruan	0.004919
0	Biomedic and Intelligent Assistive Technology ...	0.002432
9	Sistem Dinamis, Kendali, dan Robotika (Dynamic...	0.001691

### RIWAYAT HIDUP

Mercyano Dandi Hidayat menempuh pendidikan di Sekolah Menengah Atas Negeri 1 Cikarang Pusat dan selesai pada tahun 2019. Kemudian, melanjutkan pendidikannya menjadi Sarjana Teknik Informatika di Universitas Negeri Malang. Semasa menjadi mahasiswa, peneliti pernah mengikuti pelatihan dalam bidang *Artificial Intelligence* pada program “AI Mastery” di Orbit Future Academy.

