Jurnal Pendidikan Teknologi dan Kejuruan
Vol. 22, No. 1, Januari 2025
P-ISSN: 0216-3241 E-ISSN: 2541-0652

KLASIFIKASI JUDUL BERITA BAHASA INDONESIA MENGGUNAKAN SUPPORT VECTOR MACHINE DAN SELEKSI FITUR MUTUAL INFORMATION

I Putu Gede Hendra Suputra[1)], Linawati[2)], I Gede Sukadarmika[3)], Nyoman Putra Sastra[4)], I Gusti Bgs Darmika Putra [5)], I Wayan Trisna Wahyudi [6)]

[1,2,3,4] Fakultas Teknik, Universitas Udayana
[5,6] Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Email: hendra.suputra@unud.ac.id

ABSTRAK

Teknologi informasi dan komunikasi saat ini telah merubah tata cara berbagi informasi, mempengaruhi cara masyarakat mendapatkan dan menyampaikan berita. Jumlah berita digital yang terus meningkat setiap harinya oleh beberapa portal berita menimbulkan tantangan, di mana berita seringkali memiliki keterkaitan dengan lebih dari satu kategori. Dari permasalahan yang ada maka dilakukan penelitian klasifikasi judul berita online. Penelitian ini memanfaatkan metode SVM dengan seleksi fitur mutual information untuk melakukan klasifikasi judul berita online. Dataset yang digunakan yaitu judul berita dari detik.com menggunakan 6 kategori yaitu finance, travel, health, oto, food, dan sport dengan jumlah data perkategorinya yaitu 2000 data. Proses klasifikasi dimulai dari preprocessing text, term weighting menggunakan TF-IDF, selanjutnya seleksi fitur dengan multual information, dan terakhir klasifikasi dengan SVM. Hasil penelitian menunjukkan bahwa pengujian berbagai kernel SVM dan ambang batas mutual information (MI) dengan threshold 85% memberikan tingkat F1-score tertinggi pada mesin SVM dengan kernel RBF dan nilai C=10 yaitu sebesar 86,15%.

Kata kunci: Klasifikasi, Judul Berita, Support Vector Machine, Mutual Information

ABSTRACT

Current information and communication technology has changed the way information is shared, affecting the way people get and deliver news. The number of digital news that continues to increase every day by several news portals poses a challenge, where news is often related to more than one category. From the existing problems, a study was conducted on the classification of online news titles. This study uses the SVM method with mutual information feature selection to classify online news titles. The dataset used is the news title from detik.com using 6 categories, namely finance, travel, health, auto, food, and sport with the number of data per category being 2000 data. The classification process starts from text preprocessing, term weighting using TF-IDF, then feature selection with mutual information, and finally classification with SVM. The results of the study showed that testing various SVM kernels and mutual information (MI) thresholds with a threshold of 85% provided the highest level of F1-score on the SVM machine with the RBF kernel and a C value = 10, which was 86,15%.

Keywords: *Classification, News Title, Support Vector Machine, Mutual Information*

1. PENDAHULUAN

Teknologi informasi dan komunikasi saat ini telah merubah tata cara berbagi informasi, mempengaruhi cara masyarakat mendapatkan dan menyampaikan berita. Pada masa lalu, sumber utama informasi terletak pada koran atau berita televisi, tetapi saat ini, masyarakat cenderung lebih memilih media digital yang memberikan akses instan dan *real-time* [1]. Penyebaran berita *online* kini terus meningkat, terutama dengan popularitas berita *online* yang mencakup berbagai topik seperti olahraga, politik, kesehatan, makanan, dan keuangan. Identifikasi kategori berita seringkali dilakukan melalui judul berita, yang tidak hanya mencerminkan esensi berita, tetapi juga menjadi bagian yang paling menarik perhatian pembaca [2].

Namun, muncul masalah akibat penggunaan luas media digital dalam menyampaikan informasi. Jumlah berita digital yang terus meningkat setiap harinya oleh beberapa portal berita menimbulkan tantangan, dimana berita seringkali memiliki keterkaitan dengan lebih dari satu kategori [3]. Masuknya sejumlah besar berita *online* ke dalam portal berita, terkadang terjadi ketidaksesuaian kategori. Kesalahan penempatan kategori ini dapat membuat proses pencarian informasi menjadi tidak efisien dan menurunnya kualitas pengalaman pengguna karena klasifikasi berita yang tidak akurat dapat memberikan penyajian informasi yang kurang sesuai dengan preferensi pengguna. Kesalahan semacam ini sering terjadi akibat kesalahan manusia dalam proses klasifikasi. Untuk mengatasi hal ini, diperlukan penggunaan sistem otomatis dalam melakukan klasifikasi berita dengan baik.

Dalam proses klasifikasi teks, penerapan model pembelajaran mesin dapat menjadi solusi, dan salah satu metodenya yang umum digunakan adalah *Support Vector Machine* (SVM). SVM merupakan algoritma klasifikasi yang melakukan proses pencarian *hyperplane* optimal untuk memisahkan kelas-kelas berdasarkan karakteristiknya. SVM dapat menangani data yang kompleks dan memiliki keunggulan dalam mengolah dataset yang besar bahkan dimensi yang tinggi [4].

Klasifikasi teks pada berita *online* menggunakan SVM telah menjadi objek penelitian sebelumnya dengan tambahan seleksi fitur menggunakan *chi-square* [5]. Data yang dipakai adalah berita bahasa Indonesia yang diambil dari situs berita www.kompas.com. Komposisi data tersebut terdiri dari 6 kategori dengan total 2400 judul berita. Penelitian mencakup pencarian parameter SVM yang optimal dan nilai *threshold chi-square* terbaik. Hasil pengujian menunjukkan performa terbaik dengan akurasi 93,06%, presisi 92,11%, *recall* 93,06%, dan nilai *f1-score* 93,04%. Hasil tersebut diperoleh saat *threshold chi-square* setidaknya ada pada batas 80%, dan parameter SVM menggunakan kernel polinomial derajat 2, C = 1, λ = 1, konstanta γ = 0,01, ε = 10^{-8}, dengan maksimal iterasi sebesar 10.

Penelitian mengenai klasifikasi berita juga telah dilakukan oleh [6] dengan menggunakan algoritma SVM sebagai model utama dan seleksi fitur menggunakan *Mutual Information*. Data yang digunakan dalam penelitian tersebut terdiri dari 360 dokumen berita berbahasa Indonesia, dimana terdapat kategorisasi topik-topik yang masing-masing terdiri dari 30 dokumen berita. Dataset yang digunakan terdiri dari 12 kelas yaitu diantaranya olahraga, ekonomi, hiburan, kesehatan, pendidikan, budaya, gaya hidup, kriminal, otomotif, politik, wisata, dan teknologi. Hasil pengujian menunjukkan bahwa penggunaan *stemming* tanpa penggunaan *stopword removal*, kombinasi antara metode klasifikasi *Support Vector Machine* (SVM) dan fitur seleksi *Mutual Information* (MI) menghasilkan hasil terbaik yaitu 94.24%.

Berdasarkan paparan yang telah dijelaskan, penggunaan SVM dengan pencarian parameter terbaik telah memberikan bukti yang dapat menghasilkan akurasi tinggi dalam klasifikasi berita. Dalam rangka mencapai akurasi yang optimal, penelitian ini memanfaatkan metode SVM dengan seleksi fitur MI untuk mencapai akurasi optimal untuk melakukan klasifikasi judul berita *online*. Hal ini dilakukan sebagai alternatif terhadap seleksi fitur *chi-square*, dengan tujuan untuk mengevaluasi pengaruh metode seleksi fitur yang berbeda. Kombinasi antara SVM dan MI telah dilakukan pada [6], namun pada penelitian tersebut dataset yang digunakan sangat sedikit yaitu hanya 360 dokumen. Selain itu pada penelitian tersebut menggunakan seluruh kalimat dalam berita menjadi fitur sedangkan pada penelitian ini hanya menggunakan judul berita saja. Penggunaan hanya judul berita sebagai fitur satu-satunya memiliki tujuan untuk mengoptimasi sumber daya baik waktu dan *memory*.

2. KAJIAN PUSTAKA

Penelitian difokuskan dalam melakukan klasifikasi terhadap judul berita berbahasa Indonesia. Dalam tinjauan literatur pada penelitian ini, dibedakan ke dalam dua kategori, yaitu literatur yang membahas terkait metode klasifikasi SVM dan literatur yang membahas terkait metode untuk melakukan *reduce* fitur menggunakan MI.

A. State of The Art

Klasifikasi merupakan proses pengelompokkan data ke dalam kategori atau kelas tertentu. Proses ini dilakukan dengan memeriksa hubungan diantara data dan menentukkan atribut atau label kelas untuk setiap sampel yang akan diklasifikasikan [7]. Klasifikasi umumnya digunakan dalam konteks teks untuk menyusun informasi ke dalam kategori atau topik tertentu. Proses klasifikasi ini membantu memahami struktur dan konten teks, sehingga memudahkan analisis berdasarkan kesamaan karakteristik atau subjek dalam teks tersebut. Klasifikasi memiliki beragam penerapan yang luas, mencakup berbagai konteks, seperti klasifikasi dokumen, analisis sentimen, klasifikasi berita, dan beragam penggunaan lainnya. Sebagai contoh, penelitian yang dilakukan oleh [8] mengenai klasifikasi judul berita *online* Radar Banjarmasin menggunakan metode *Naïve Bayes* dengan 40 data dan 4 kategori. Hasil penelitian menunjukkan bahwa metode *Naïve Bayes* memberikan hasil akurasi akhir sebesar 78.75%. Selain itu, hasil *recall* mencapai 80.56% dan *precision* mencapai 78.75%. Penelitian

terbaru mengenai judul berita *online* dilakukan oleh [5] berkaitan dengan klasifikasi judul berita *online* di kompas.com. Penelitian ini melibatkan 2400 judul berita yang terbagi dalam 6 kategori. Penelitian tersebut mencakup identifikasi parameter SVM optimal dan penemuan nilai *threshold chi-square* terbaik. Hasil uji coba menunjukkan kinerja terbaik dengan akurasi mencapai 93,06%, presisi sebesar 92,11%, *recall* mencapai 93,06%, dan f1-score mencapai 93,04%. Keberhasilan ini dicapai dengan menggunakan *threshold chi-square* setidaknya 80%, serta parameter SVM dengan kernel polinomial derajat 2, C = 1, λ = 1, dan konstanta lainnya.

B. Berita

Berita adalah informasi penting dan menarik yang melibatkan fakta dan pendapat. Tujuannya adalah untuk memberikan informasi yang cepat kepada sebagian besar orang [9]. Dalam suatu berita, terdapat komponen utama seperti judul dan isi. Judul berita adalah komponen penting karena dapat membantu pembaca memahami dengan cepat pokok bahasan yang akan dibahas dalam berita tersebut [10].

C. Pre-processing

Preprocessing merupakan proses awal untuk menyiapkan dokumen atau data mentah agar siap untuk proses selanjutnya. Tujuannya adalah mengubah data yang masih mentah dan belum terstruktur menjadi format yang lebih sesuai untuk analisis [11]. Terdapat beberapa tahap *preprocessing text* yaitu *case folding, cleaning, tokenizing, stopword removal,* dan *stemming*.

D. Term Weighting

Term weighting adalah proses memberikan bobot atau nilai ke berbagai kata *(term)* dalam suatu dokumen [12]. Metode umum untuk memberikan bobot pada kata yang umum digunakan yaitu *Term Frequency-Inverse Document Frequency* (TF-IDF). *Term frequency (TF)* adalah konsep yang mencerminkan seberapa sering suatu kata muncul dalam suatu dokumen, sementara konsep *Inverse Document Frequency (DF)* merupakan suatu konsep probabilitas kemunculan suatu kata dalam pada suatu dokumen.

1. Nilai *TF* dihitung berdasarkan jumlah kemunculan kata pada suatu dokumen.
2. Nilai *IDF* dihitung dengan dengan persamaan (1).
   $$\boldsymbol{IDF}_{(\boldsymbol{a},\boldsymbol{b})} = \log \boldsymbol{N}/\boldsymbol{DF} \quad (1)$$
3. Selanjutnya nilai *TF-IDF* dihitung dengan persamaan (2).
   $$\boldsymbol{TF} - \boldsymbol{IDF}_{(\boldsymbol{a},\boldsymbol{b})} = \boldsymbol{TF}_{(\boldsymbol{a},\boldsymbol{b})} \times \boldsymbol{IDF}_{(\boldsymbol{a})} \quad (2)$$

Keterangan:

$\boldsymbol{TF}_{(\boldsymbol{a},\boldsymbol{b})}$ = jumah suatu kata *a* dalam dokumen *b*.

$\boldsymbol{IDF}_{(\boldsymbol{a})}$ = nilai proporsi suatu kata *a* terhadap keseluruhan dokumen.

E. Mutual Information

Mutual Information (MI) telah menjadi salah satu metode yang umum digunakan untuk melakukan seleksi fitur [6]. MI mengukur sejauh mana informasi terkandung dalam suatu fitur, memungkinkan penilaian terhadap pengaruh fitur tersebut dalam pembuatan keputusan klasifikasi yang akurat. Dalam konteks ini, MI menjadi alat yang sangat berguna dalam mengevaluasi relevansi dan signifikansi setiap fitur terhadap tugas klasifikasi. Dengan mengukur berapa banyak informasi yang terkandung dalam fitur-fitur tersebut, penelitian dapat mengidentifikasi pengaruhnya dalam mencapai hasil klasifikasi yang optimal. Rumus perhitungan nilai MI secara formal dapat dilihat pada persamaan (3).

$$I(U,C)\sum_{et\in(1,0)}\sum_{ec\in(1,0)}P(U=et,C=ec)\log_2\frac{P(U=et,C=ec)}{P(U=et)P(C=ec)} \quad (3)$$

Sesuai dengan persamaan (3), variabel *U* adalah variabel acak dengan nilai *et* = 1 (yaitu suatu dokumen mengandung kata *t*) atau *et* = 0 (yaitu suatu dokumen yang tidak mengandung *kata t*). Variabel C adalah variabel acak dengan nilai *ec* = 1 (jika suatu dokumen merupakan anggota suatu kelas *c*) dan *ec* = 0 (jika suatu dokumen bukan merupakan anggota suatu kelas *c*). Persamaan diatas (3) dapat dijabarkan menjadi seperti persamaan (4).

$$I(U,C) = \frac{N_{11}}{N}\log 2\frac{N.N_{11}}{N_1N_1} + \frac{N_{01}}{N}\log 2\frac{N.N_{01}}{N_0N_1} + \frac{N_{10}}{N}\log 2\frac{N.N_{10}}{N_1N_0} + \frac{N_{00}}{N}\log 2\frac{N.N_{00}}{N_0N_0} \quad (4)$$

Sebagai contoh, N_{10} merupakan jumlah dokumen yang mengandung *t* (*et* = 1) dan tidak dalam kelas *c* (*ec* = 0).
Keterangan:

N = banyaknya dokumen yang mengandung et dan ec atau ($N = N_{00} + N_{01} + N_{10} + N_{11}$)
$N_{1.}$ = banyaknya dokumen yang mengandung et atau ($N_{1.} = N_{10} + N_{11}$)
N_1 = banyaknya dokumen yang mengandung ec atau ($N_1 = N_{01} + N_{11}$)
$N_{0.}$ = banyaknya dokumen yang tidak mengandung et atau ($N_{0.} = N_{01} + N_{00}$)
N_0 = banyaknya dokumen yang tidak mengandung ec atau ($N_0 = N_{10} + N_{00}$)

F. Support Vector Machine

SVM merupakan salah satu algoritma klasifikasi yang digunakan dalam pembelajaran mesin. Prinsip kerja SVM melibatkan pencarian *hyperplane* atau garis pemisah yang mampu memisahkan dua kelas data, sambil juga memaksimalkan nilai margin atau jarak antara dua kelas data yang berbeda [5]. SVM awalnya berfungsi sebagai *linear classifier* dengan pemisahan menggunakan garis lurus. Untuk menangani masalah *non-linear*, SVM menggunakan fungsi transformasi $\phi(x)$ ke ruang dimensi yang lebih tinggi. Namun, untuk menghindari perhitungan nilai transformasi, SVM mengadopsi kernel *trick* berdasarkan teori Mercer. Berbagai fungsi kernel yang digunakan dapat ditemukan dalam Tabel 1.

Tabel 1. Fungsi Kernel

Nama Kernel	Fungsi Kernel
Linear	$K(x_i x_j) = (x_i x_j)$
Polynomial	$K(x_i x_j) = (x_i x_j)^d$
Gaussian RBF	$K(x_i x_j) = \exp\frac{-\|\|x_i - x_j\|\|^2}{2\sigma^2}$
Sigmoid	$K(x_i x_j) = tanh(\sigma(x_i x_j) + c)$

Proses klasifikasi SVM dilakukan dengan mencari nilai $f(x)$ dengan x adalah data yang ingin diklasifikasikan. Rumus dari mencari nilai $f(x)$ dapat dilihat pada Persamaan (5).

$$f(x) = \sum_i^n = 1\, a_i y_i K(x_i, x) + b \quad (5)$$

keterangan:
$f(x)$ = Fungsi klasifikasi untuk data x
n = Banyak data latih
α_i = Lagrange multiplier untuk data ke-i
y_i = Label data latih ke-i
$K(x_i, x)$ = Fungsi kernel untuk data latih $x_i\ dan\ x$
b = Nilai bias

Pada persamaan di atas terdapat nilai alpha (α) dan juga bias (b) yang mana nilai ini diperoleh dari proses pelatihan SVM. Salah satu metode pelatihan SVM adalah *sequential training*.

Dalam klasifikasi biner, output $f(x)$ menggunakan fungsi sign menghasilkan +1 untuk positif dan -1 untuk negatif. Pada klasifikasi multi-class, seperti metode "*one-against-all*", diperlukan model SVM terpisah untuk setiap kelas. Setiap model memberikan keputusan untuk kelasnya, dan hasilnya digabungkan untuk prediksi akhir.

G. Sequential Training

Sequential training adalah metode pelatihan di mana model atau algoritma pembelajaran mesin dilatih secara berurutan, satu *instance* atau *batch* data pada satu waktu. Sebaliknya dengan pendekatan *batch training* di mana seluruh dataset digunakan sekaligus, sequential *training* memperbarui model secara bertahap seiring masuknya data baru. Pendekatan ini berguna ketika data masuk secara berurutan atau ketika sumber daya komputasi terbatas [5].

H. One-Againts-All

One-Against-All (OAA) atau disebut juga *One-Versus-All* (OVA) adalah metode yang umum digunakan dalam *Support Vector Machine* untuk menangani masalah klasifikasi *multi-class* [5]. SVM pada dasarnya dirancang untuk klasifikasi biner, tetapi dengan menggunakan pendekatan OAA, dapat diadaptasi untuk menangani lebih dari dua kelas. Dengan menggunakan pendekatan ini, kita dapat

memperluas SVM untuk menangani *multi-class* dengan memanfaatkan keunggulan SVM dalam menangani klasifikasi biner. Setiap model SVM dalam OAA bertanggung jawab untuk memisahkan satu kelas dari kelas-kelas lainnya, membuatnya menjadi pendekatan yang sederhana dan intuitif.

3. METODE

A. Alur Proses Sistem

Penelitian yang diusulkan adalah untuk mengklasifikan berita ke dalam 6 kelas, berdasarkan fitur judul berita. Alur dari sistem klasifikasi judul berita yang dilakukan terdiri dari beberapa proses yaitu 1) Preprocessing, 2) Data Splitting, 3) TF-IDF, 4) Pengujian pada model SVM tanpa MI, 5) transformasi *matrix* dengan MI, 6) Pengujian pada model SVM dengan MI, 7) Evaluasi dengan *confussion matrix*. Adapun alur system yang dilakukan dapat dilihat melalui diagram alir pada Gambar 1 yang dijelaskan pada sub-poin 3C-3G.

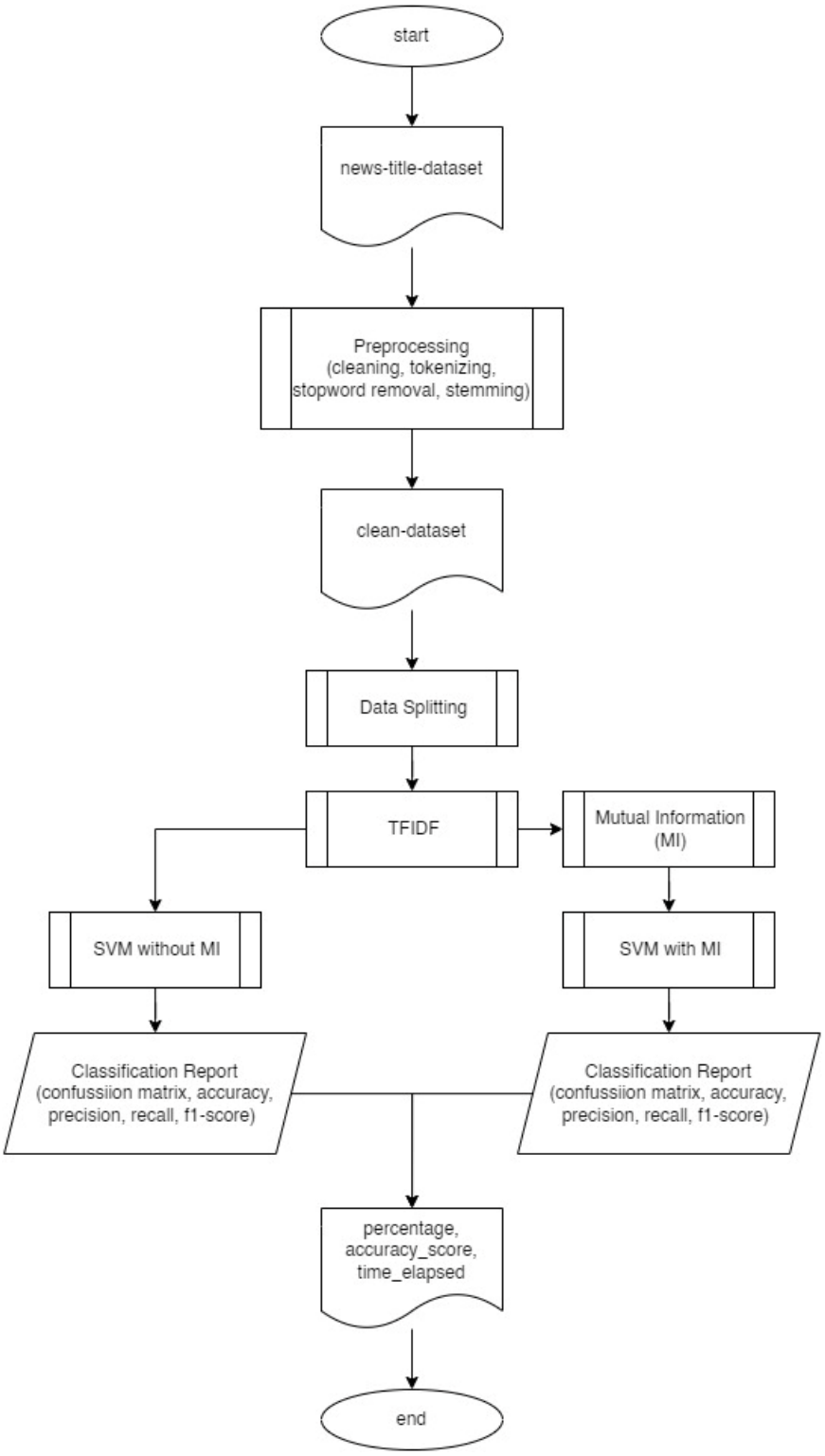
B. Deskripsi *Dataset*

Dataset yang digunakan merupakan dataset sekunder yang tersedia pada platform *open source kaggle*. Dataset merupakan data judul berita berbahasa Indonesia yang terdiri dari 91.017 judul berita yang bersumber dari salah satu portal berita terbesar di Indonesia, detik.com. Dataset ini terdiri dari 9 kategori yang dapat dilihat pada Tabel 2.

Tabel 2. Dataset Judul Berita

Category	Jumlah Data
news	32360
hot	16330
finance	14168
travel	6466
inet	5640
health	4919
oto	4383
food	4315
sport	2436

Pada dataset ini diasumsikan bahwa pada kategori news dan hot tidak dapat dikatakan sebagai sebuah kategori yang spesifik sehingga kategori yang akan diambil berjumlah 6 yaitu *finance, travel, health, oto, food*, dan *sport* dengan jumlah data yang diambil akan diseimbangkan pada masing-masing kategori sejumlah 2000 data sehingga total data yang digunakan adalah 12000 data. Komposisi data *training* adalah 80% dan 20% sisanya digunakan untuk data *test*.



Gambar 1. Alur Proses Sistem Klasifikasi Judul Berita

C. *Preprocessing*

Preprocessing data merupakan suatu metode yang digunakan untuk menyiapkan kumpulan data guna keperluan pemodelan. Dalam ranah *text mining*, proses ini melibatkan serangkaian langkah untuk mentransformasikan data teks asli menjadi suatu struktur data yang membedakan fitur-fitur tekstual antar kategori teks. Fase *preprocessing* memiliki signifikansi yang besar dalam analisis sentimen karena tujuannya adalah untuk mengatasi nilai yang hilang (data kosong), mengidentifikasi dan mengeliminasi

data duplikat, serta menangani format data yang tidak sesuai. Beberapa langkah yang tercakup dalam tahap preprocessing mencakup:

- Pembersihan Data *(Data Cleaning):* Pada tahap ini, juga dilakukan proses *casefolding* yang merujuk pada transformasi semua huruf dalam teks atau opini dalam dataset menjadi huruf kecil. Kemudian dilakukan proses eliminasi karakter dan simbol tertentu, kecuali huruf, dari kumpulan data.
- Tokenisasi *(Tokenization):* Tokenisasi adalah proses pemisahan kata-kata pada dataset berdasarkan spasi. Hal ini diperlukan untuk membentuk token atau unit dasar dalam pemodelan data, memfasilitasi analisis lebih lanjut.
- Penghapusan *Stopword (Stopword Removal):* Penghapusan *stopword* melibatkan eliminasi kata-kata dalam teks yang memiliki arti yang minim, seperti kata penghubung, dan sejenisnya.
- *Stemming: Stemming* merupakan proses mengubah kata-kata dalam dataset menjadi bentuk dasarnya dengan menghilangkan imbuhan seperti awalan, sisipan, akhiran, dan kombinasi.

D. Ekstraksi Fitur dengan TF-IDF

Langkah berikutnya setelah adalah memberikan bobot pada data yang sudah dibersihkan sebelumnya. Dalam tahap ini, setiap *term unigram* yang dihasilkan dari prapemrosesan akan diberikan bobot menggunakan perhitungan nilai TF-IDF sesuai dengan rumus TF-IDF. Bobot ini digunakan untuk mengukur relevansi suatu kata dalam kelompok dokumen, sehingga dapat menentukan apakah kata tersebut cocok sebagai fitur pada tahap seleksi fitur. Selain itu, tahap ini juga mencakup vektorisasi, yang mengubah data teks menjadi vektor fitur yang diperlukan oleh model klasifikasi yang akan digunakan.

E. *Mutual Information*

Dalam menghitung *Mutual Information (MI)* antara dua variabel, langkah awal melibatkan perhitungan seberapa sering kombinasi nilai keduanya muncul bersama (probabilitas bersama) dan seberapa sering masing-masing variabel muncul sendiri (probabilitas marginal). Entropi dari setiap variabel dihitung untuk mengukur tingkat ketidakpastian. *MI* kemudian dihitung dengan membandingkan entropi variabel tersebut. Proses pemilihan fitur dilakukan dengan mengambil sejumlah persentase terbaik dari fitur-fitur tersebut, yang disebut sebagai threshold. Jumlah persentase fitur yang diambil dapat diuji sebagai nilai *threshold* untuk menentukan fitur terbaik.

F. *SVM Multi Class*

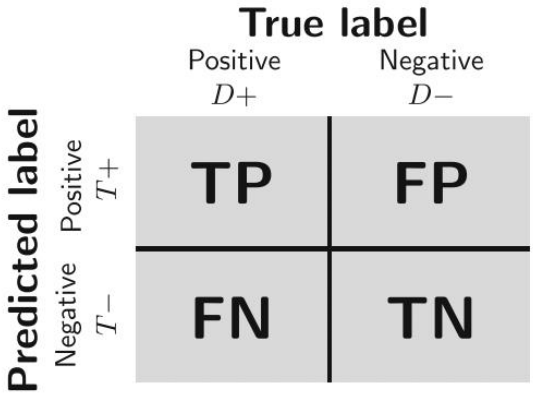
Proses SVM pada *multi class* dengan menggunakan metode *one-against-all* melibatkan iterasi untuk setiap kelas yang ada. Iterasi dimulai dengan mengubah label, di mana label diatur menjadi 1 jika sama dengan kelas yang sedang diiterasi dan -1 jika tidak. Setelah itu, dilakukan pelatihan SVM menggunakan *sequential training*, diikuti dengan pengujian SVM untuk mencari nilai $f(x)$. Selama pengujian, suatu data uji akan diklasifikasikan ke dalam kelas yang memiliki nilai $f(x)$ tertinggi. Proses ini memastikan bahwa setiap kelas dievaluasi secara terpisah. Proses akhir dari SVM adalah evaluasi, di mana performa model dievaluasi berdasarkan metrik yang relevan, seperti akurasi, presisi, *recall*, atau *F1-score*. Evaluasi ini memberikan gambaran tentang sejauh mana model dapat mengklasifikasikan data dengan benar dan efektif untuk setiap kelas yang terlibat dalam masalah *multi-class classification*.

G. Model Evaluation

Matriks memungkinkan digunakan sebagai metode dasar untuk mengukur kinerja model *machine learning*. Dalam masalah klasifikasi, kinerja model dapat digambarkan dalam sebuah matriks yang disebut *confusion matrix*. Pada klasifikasi biner (dengan dua kelas), *confusion matrix* membagi hasil sampel data uji menjadi empat kategori berdasarkan label yang sebenarnya (*true label*) dan label yang diprediksi (*predicted label*) seperti yang ditunjukkan pada Gambar 2, diantaranya [13]:

- *True Positive* (TP): sampel data bernilai positif dan model memprediksi dengan benar bernilai positif. Contoh: suatu berita berlabel ‘finance’ (1) dan model mengklasifikasi sampel sebagai berita finance (1).
- *True Negative* (TN): sampel data bernilai negatif dan model memprediksi dengan benar bernilai negatif. Contoh: suatu berita berlabel bukan ‘finance’ (0) dan model mengklasifikasi sampel sebagai berita bukan finance (0).

- *False Positive* (FP) : sampel data bernilai negatif dan model memprediksi dengan salah bernilai positif. Contoh: suatu berita berlabel bukan 'finance' (0) dan model mengklasifikasi sampel sebagai berita berlabel 'finance' (1).
- *False Negative* (FN) : sampel data bernilai positif dan model memprediksi dengan salah bernilai negatif. Contoh: suatu berita berlabel 'finance' (1) dan model mengklasifikasi sampel sebagai berita berlabel bukan 'finance' (0).



Gambar 2. confusion matrix

Berdasarkan *confusion matrix*, nilai evaluasi model *machine learning* dapat direpresentasikan sebagai berikut sesuai dengan persamaan (6)-(9).

- Akurasi : persentase sampel yang diklasifikasikan dengan benar.
$$Accuracy = \frac{TP + TN}{TP+FP +TN +FN} \quad (6)$$
- Presisi atau *Positive predictive value* (PPV): persentase sampel yang diklasifikasikan positif yang sebenarnya memang berlabel positif.
$$Precision = \frac{TP}{TP+FP} \quad (7)$$
- Recall atau *sensitivity* : persentase sampel positif yang benar-benar berhasil diidentifikasi.
$$Recall = \frac{TP}{TP +FN} \quad (8)$$
- F1-score : rata-rata harmonik PPV (presisi) dan sensitivitas (recall).
$$F1 - score = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = \frac{2TP}{2TP+FP +FN} \quad (9)$$

Penelitian ini menggunakan model *multi-class* maka formula yang digunakan sesuai dengan penelitian [14], dimana nilai presisi, recall dan F1-score akan dicari berdasarkan tiap kategori dan akan dihitung untuk mencari rata-rata sebagai nilai akhir.

4. HASIL DAN PEMBAHASAN

A. Pengujian Kernel SVM

Pengujian kernel SVM menggunakan 3 jenis kernel yaitu *linear, polynomial, dan RBF*. Pengujian dilakukan dengan mencoba 4 kemungkinan kernel SVM dengan berbagai nilai parameter di dalamnya. Parameter SVM lainnya yang digunakan adalah C = 0.1, 1, 10, dan 100. Hasil dari pengujian kernel SVM dapat dilihat melalui Tabel 3. Nilai C 10 dan kernel RBF menjadi komposisi yang terbaik untuk model SVM. Nilai rata-rata akurasi yang diraih adalah 0,863.

Tabel 3. Pencarian Parameter Terbaik SVM

C	Kernel	Rata-rata akurasi	Standar Deviasi
0,1	Linear	0,7614	0,0109
1	Linear	0,8576	0,0045
10	Linear	0,8351	0,0070
100	Linear	0,8259	0,0051
0,1	Poly	0,3290	0,0410
1	Poly	0,8359	0,0032
10	Poly	0,8382	0,0039

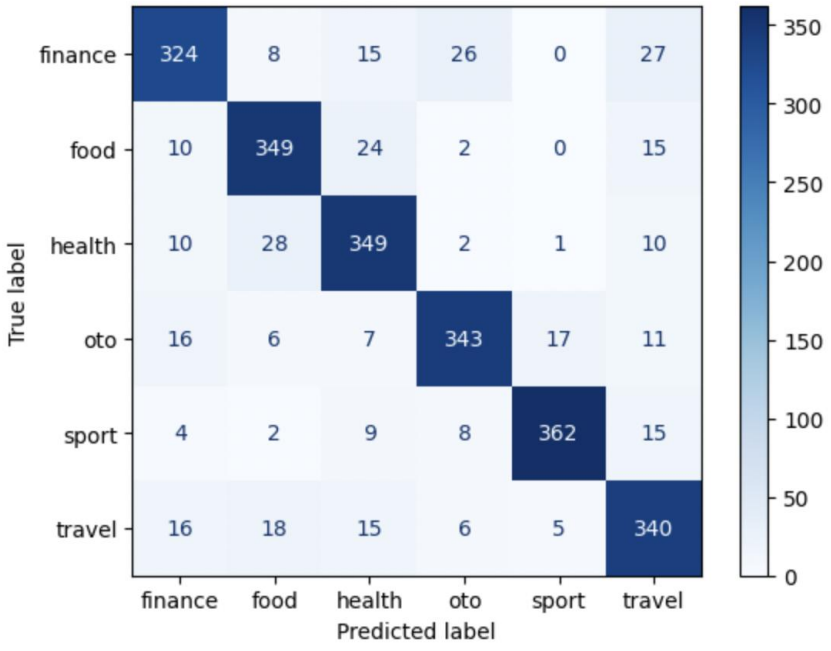
100	Poly	0,5590	0,0205
0,1	RBF	0,6096	0,0267
1	RBF	0,8607	0,0064
10	**RBF**	**0,8630**	**0,0052**
100	RBF	0,8630	0,0052

B. Pengujian Threshold MI

Threshold MI adalah parameter kunci dalam seleksi fitur. Nilai *threshold* ini menentukan berapa banyak fitur yang akan digunakan dalam klasifikasi menggunakan SVM. Seleksi fitur MI bertujuan mencari akurasi terbaik dengan mempertahankan hanya fitur-fitur penting. Jika nilai *threshold* terlalu kecil akan membuat banyak kehilangan informasi dari data sehingga hasil yang didapat menjadi menurun. Hasil dari pengujian *threshold* MI diperlihatkan melalui Tabel 4. Persentase fitur yang terbaik yang ditemukan adalah 85%. Pengurangan fitur sebanyak 15% tersebut terbukti mampu memberikan hasil yang identik (F1 = 0,8615) dengan penggunaan tanpa seleksi fitur MI (F1 = 0,8625). Terjadi penurunan waktu proses dari 14,25 detik menjadi 13,70 detik ini berarti terdapat penurunan waktu proses sekitar 3,86%.

Tabel 4. Hasil Pengujian SVM + MI

Persentase Fitur MI	Waktu Proses (detik)	Akurasi	Presisi	Recall	F1
50	12,1554	0,8350	0,8387	0,8350	0,8359
55	11,8732	0,8408	0,8431	0,8408	0,8413
60	12,6989	0,8433	0,8453	0,8433	0,8437
65	13,5343	0,8467	0,8483	0,8467	0,8470
70	13,3343	0,8475	0,8491	0,8475	0,8478
75	13,5006	0,8529	0,8544	0,8529	0,8531
80	13,6465	0,8563	0,8575	0,8563	0,8565
85	**13,7044**	**0,8613**	**0,8624**	**0,8613**	**0,8615**
90	13,7638	0,8600	0,8613	0,8600	0,8603
95	13,9062	0,8604	0,8618	0,8604	0,8606
100	**14,2536**	**0,8625**	**0,8635**	**0,8625**	**0.8625**



Gambar 3. *Confusion Matrix* saat *threshold* MI = 85%

Gambar 3. menunjukkan bahwa masih terdapat *miss-classification error* dari beberapa kategori. Salah satu yang paling menonjol adalah *food* dan *health*. Kedua terminologi *food* dan *health* memang secara nyata tampak berhubungan satu dengan yang lainnya dimana makanan adalah satu bahasan yang beririsan dengan kesehatan. Hal ini memberikan suatu pandangan terkait perlunya pembangkitan jenis representasi text lain seperti ujicoba berbagai jenis *word embedding* yang berbeda dengan TF-IDF seperti *BERT* [15], *Glove* [16], *Word2Vec* [17] atau *Fastext* [18].

5. SIMPULAN DAN SARAN

Berdasarkan hasil dari pengujian-pengujian yang dilakukan dalam penelitian ini, dapat disimpulkan bahwa penggunaan metode SVM dengan seleksi fitur *Mutual Information* (MI) mampu menghasilkan akurasi yang sama dengan metode SVM tanpa MI, namun dengan waktu eksekusi yang lebih singkat dan penggunaan dimensi yang lebih sedikit sehingga secara tidak langsung akan mengurangi penggunaan *resource* komputasi dalam klasifikasi judul berita dalam bahasa Indonesia. Hasil penelitian menunjukkan bahwa pengujian berbagai kernel SVM dan MI dengan *threshold* 85% memberikan tingkat akurasi tertinggi pada mesin SVM dengan kernel *RBF* dan nilai C=10 yaitu sebesar 86,15%. Selain itu, penelitian ini juga memberikan pemahaman yang lebih baik tentang proses klasifikasi *multi-class* menggunakan metode SVM dengan pendekatan satu lawan semua (*one-against-all*). Sebagai langkah lanjutan, penelitian masa depan dapat memperluas cakupan dataset, mempertimbangkan teknik preprocessing teks yang lebih mutahir, dan mengeksplorasi penggunaan metode *word embedding* seperti *BERT*, *Glove*, *Word2Vec* dan *Fastext* untuk klasifikasi judul berita dalam bahasa Indonesia.

DAFTAR PUSTAKA

[1] F. A. Ramadhan, S. H. Sitorus, and T. Rismawan, ‘Penerapan Metode Multinomial Naïve Bayes untuk Klasifikasi Judul Berita Clickbait dengan Term Frequency - Inverse Document Frequency’, *Jurnal Sistem dan Teknologi Informasi (JustIN)*, vol. 11, no. 1, p. 70, 2023, doi: 10.26418/justin.v11i1.57452.

[2] A. Alfando and R. Hayami, ‘Klasifikasi Teks Berita Berbahasa Indonesia Menggunakan Machine Learning Dan Deep Learning: Studi Literatur’, *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 681–686, 2023, doi: 10.36040/jati.v7i1.6486.

[3] B. H. Mahendra, Adiwijaya, and U. N. Wisesty, ‘Kategorisasi Berita Multi-Label Berbahasa Indonesia Menggunakan Algoritma Random Forest’, *e-Proceeding of Engineering*, vol. 6, no. 2, pp. 9030–9041, 2019.

[4] R. Yuranda, T. Sutabri, and D. Wahyuningsih, ‘Pendekatan Macine Learning dalam Evaluasi Label Berita Berdasarkan Judul : Studi Kasus Media Online’, vol. 12, pp. 434–439, 2023.

[5] P. Rama, B. Putra, and R. S. Perdana, ‘Klasifikasi Judul Berita Online menggunakan Metode Support Vector Machine (SVM) dengan Seleksi Fitur Chi-square’, vol. 7, no. 5, pp. 2132–2141, 2023.

[6] L. G. Irham, A. Adiwijaya, and U. N. Wisesty, ‘Klasifikasi Berita Bahasa Indonesia Menggunakan Mutual Information dan Support Vector Machine’, *Jurnal Media Informatika Budidarma*, vol. 3, no. 4, p. 284, 2019, doi: 10.30865/mib.v3i4.1410.

[7] N. Ajijah, A. Kurniawan, and Susilawati, ‘Klasifikasi Teks Mining Terhadap Analisa Isu Kegiatan Tenaga Lapangan Menggunakan Algoritma K-Nearest Neighbor (KNN)’, *Jurnal Sains Komputer & Informatika (J-SAKTI*, vol. 7, no. 1, pp. 254–262, 2023.

[8] M. Sholih ’afif, M. Muzakir, M. I. Al, and G. Al Awalaien, ‘Text Mining Untuk Mengklasifikasi Judul Berita Online Studi Kasus Radar Banjarmasin Menggunakan Metode Naïve Bayes’, *Kumpulan jurnaL Ilmu Komputer (KLIK)*, vol. 08, no. 2, pp. 199–208, 2021.

[9] S. Waljinah, H. J. Prayitno, E. Purnomo, A. Rufiah, and E. W. Kustanti, ‘Tindak Tutur Direktif Wacana Berita Online: Kajian Media Pembelajaran Berbasis Teknologi Digital’, *SeBaSa*, vol. 2, no. 2, p. 118, 2019, doi: 10.29408/sbs.v2i2.1590.

[10] W. Afandi, S. N. Saputro, A. M. Kusumaningrum, H. Adriansyah, M. H. Kafabi, and S. Sudianto, ‘Klasifikasi Judul Berita Clickbait menggunakan RNN-LSTM’, *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 7, no. 2, pp. 85–89, 2022, doi: 10.30591/jpit.v7i2.3401.

[11] M. U. Albab, Y. Karuniawati, and M. N. Fawaiq, ‘Optimization of the Stemming Technique on Text preprocessing President 3 Periods Topic’, *Jurnal TRANSFORMATIKA*, vol. 20, no. 2, pp. 1–10, 2023.

[12] M. D. Hendriyanto and B. N. Sari, 'Penerapan Algoritma K-Nearest Neighbor Dalam Klasifikasi Judul Berita Hoax', *Jurnal Ilmiah Informatika*, vol. 10, no. 02, pp. 80–84, 2022, doi: 10.33884/jif.v10i02.5477.

[13] C. O. Varoquaux G, 'Evaluating Machine Learning Models and Their Diagnostic Value', *Machine Learning for Brain Disorders [Internet]. New York, NY: Humana; 2023. Chapter 20.*, vol. Chapter 20, pp. 601–629, 2023, doi: 10.1515/9780823295258-024.

[14] M. Grandini, E. Bagli, and G. Visani, 'Metrics for Multi-Class Classification: an Overview', pp. 1–17, 2020, [Online]. Available: http://arxiv.org/abs/2008.05756

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding', *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[16] J. Pennington, R. Socher, and C. Manning, 'GloVe: Global Vectors for Word Representation', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP})*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, 'Efficient estimation of word representations in vector space', *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pp. 1–12, 2013.

[18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, 'Enriching Word Vectors with Subword Information', *Trans Assoc Comput Linguist*, vol. 5, pp. 135–146, 2017, doi: 10.1162/tacl_a_00051.