JURNAL ILMIAH ELEKTRONIKA DAN KOMPUTER, Vol.13, No.1, Juli 2020, pp. 159 – 175
p-ISSN : 1907-0012 (print)
e-ISSN : 2714-5417 (online)
http://journal.stekom.ac.id/index.php/elkom

EVALUASI EKSTRAKSI FITUR KLASIFIKASI TEKS UNTUK PENINGKATAN AKURASI KLASIFIKASI MENGGUNAKAN NAIVE BAYES

[1]Aji Priyambodo,* [2]Prihati, Prihati

[1]Institut Teknologi dan Bisnis Semarang; priyambodo@itbsemarang.ac.id
[2]Institut Teknologi dan Bisnis Semarang; prihati@itbsemarang.ac.id

ARTICLE INFO

Article history:

Received 30 April 2020
Received in revised form 2 Mei 2020
Accepted 10 Juni 2020
Available online Juli 2020

ABSTRACT

Classification is one of the most widely used techniques in machine learning. Text classification is the process of classifying data according to pre-determined groups or classes. Where in most cases, text classification uses labeled training data to obtain the rules used to classify test data into predefined groups. In this study, it is proposed to use CountVectorizer for Indonesian text classification which will be compared with TF-IDF Term Weighting and its three feature levels, namely Character Level, Word Level and N-gram Level as feature extraction which is implemented together with Naive Bayes classification and the BPPPTIndToEngCorpusHalfM dataset. To compare the classification performance, this study uses 10-Fold Cross Validation and Split Data using a ratio of 90:10, while to evaluate the accuracy of the authors using the F1-Score and AUC with the hope that this study will get good accuracy results so that it can be used as a reference to be developed using another method. The F1-Score accuracy obtained in this study was 0.93 and the AUC score was 0.95.

Keywords: Text classification, feature extraction, Count Vectorizer, Naive Bayes

Abstrak

Klasifikasi adalah salah satu teknik yang paling banyak digunakan dalam pembelajaran mesin. Klasifikasi teks adalah proses pengklasifikasian data menurut kelompok atau kelas yang telah ditentukan sebelumnya. Di mana dalam kebanyakan kasus, klasifikasi teks menggunakan data pelatihan berlabel untuk mendapatkan aturan yang digunakan untuk mengklasifikasikan data uji ke dalam kelompok yang telah ditentukan. Pada penelitian ini diusulkan untuk menggunakan CountVectorizer untuk klasifikasi teks bahasa Indonesia yang akan dibandingkan dengan TF-IDF Term Weighting dan tiga level fiturnya yaitu Character Level, Word Level dan N-gram Level sebagai ekstraksi fitur yang diimplementasikan

Received April 30, 2020; Revised Mei 2, 2020; Accepted Juni 22, 2020

p-ISSN : 1979-0155 e-ISSN : 2614-8870

bersama dengan Naive Klasifikasi Bayes dan set data BPPPTIndToEngCorpusHalfM. Untuk membandingkan performansi klasifikasi, penelitian ini menggunakan 10-Fold Cross Validation dan Split Data menggunakan rasio 90:10, sedangkan untuk mengevaluasi akurasi penulis menggunakan F1-Score dan AUC dengan harapan penelitian ini mendapatkan akurasi yang baik. sehingga dapat dijadikan acuan untuk dikembangkan dengan menggunakan metode lain. Akurasi F1-Score yang diperoleh dalam penelitian ini adalah 0,93 dan skor AUC adalah 0,95.

Kata kunci: Klasifikasi teks, ekstraksi fitur, Count Vectorizer, Naive Bayes

PENDAHULUAN

Klasifikasi adalah salah satu teknik yang paling banyak digunakan dalam machine learning. Klasifikasi bisa menjadi aplikasi mandiri seperti dalam klasifikasi teks atau bagian dari bidang lain seperti dalam penambangan data dan penambangan teks. Klasifikasi teks adalah proses pengklasifikasian data menurut kelompok atau kelas yang telah ditentukan sebelumnya. Di mana, dalam banyak kasus, klasifikasi teks menggunakan data pelatihan berlabel untuk mendapatkan aturan yang digunakan untuk mengklasifikasikan data pengujian ke dalam grup yang telah ditentukan. Menurut Nicolosi klasifikasi terdiri dari dua tahap: tahap pembelajaran yang menganalisis data pelatihan dan menetapkan aturan klasifikasi untuk data tersebut; dan tahap klasifikasi yang mengklasifikasikan data uji menggunakan aturan yang dihasilkan ke dalam kelompok di mana kelompok tersebut didefinisikan berdasarkan nilai atribut data [1].

Klasifikasi teks menurut Purohit dkk didapat dengan mengklasifikasikan dokumen berdasarkan isinya dan atau topiknya ke dalam kategori yang telah ditentukan [2]. Klasifikasi teks sangat penting oleh karena itu banyak metode dan algoritma berbeda yang digunakan untuk menyelesaikannya agar mendapatkan akurasi yang baik seperti yang dipaparkan oleh Scott dan Matwin [3]. Shang dkk dalam papernya [4] telah mengusulkan berbagai metode pemilihan fitur untuk klasifikasi teks, selain itu menurut Somol and Novovičová klasifikasi teks seringkali digunakan untuk mengolah data yang memiliki banyak fitur atau umumnya dikatakan berdimensi tinggi [5], selain itu kumpulan data berdimensi tinggi menimbulkan tiga masalah dalam proses pemilihan fitur. Pertama, pemilihan fitur yang tidak stabil dengan sampel terbatas dan dimensi tinggi menurut Dernoncourt dkk [6]. Kedua, pemilihan fitur menghabiskan lebih banyak waktu untuk sampel dimensi tinggi. Ketiga, kinerja klasifikasi mungkin tidak cukup baik dengan menggunakan metode pemilihan fitur tertentu. Oleh karena itu, beberapa faktor harus dipertimbangkan untuk memilih metode pemilihan fitur yang sesuai untuk mengklasifikasikan teks dengan sampel terbatas. Misalnya, penyaringan spam email, kosakata istilah bisa sangat besar, preferensi pengguna untuk spam dan non-spam sering berbeda, email datang dalam berbagai bahasa dan kualitas. yang mencerminkan berbagai aspek kinerja klasifikasi dan tidak dapat digantikan satu sama lain.

Dalam menghadapi tantangan tersebut di atas banyak penelitian yang sudah dilakukan menggunakan pengklasifikasi naive bayes. Menurut Xu [7] pengklasifikasi naive bayes dikenal sebagai grup dari pengklasifikasi probabilistik sederhana di mana semua fitur lepas satu sama lain, menurut variabel kategori. Naive bayes cukup efektif untuk mengklasifikasikan teks, meskipun kurang akurat dibandingkan metode diskriminatif lain seperti SVM menurut Ting dkk [8]. Ada dua model pendekatan untuk naive bayes menurut Lewis [9], dan Vidhya.dan Aghila yaitu sebagai model multivariat bernoulli dan multinomial [10]. Dari model ini model multinomial lebih cocok untuk data berdimensi tinggi, Kim dkk mengusulkan model dokumen estimasi dan normalisasi panjang dokumen mengurangi masalah dalam pendekatan multinomial tradisional untuk klasifikasi teks. Selain itu Myaeng dkk mengusulkan model poisson untuk klasifikasi teks naive bayes dan juga memberikan metode peningkatan bobot untuk meningkatkan kinerja [11]. Naive bayes yang dimodifikasi diusulkan Schneider [12], Jiang dkk [13], Zhang dan Gao [14] untuk meningkatkan kinerja klasifikasi teks, sedangkan Pazzani menyediakan cara untuk meningkatkan klasifikasi naive bayes dengan mencari ketergantungan di antara atribut [15]. Naive bayes mudah untuk

p-ISSN : 1907-0012 (print); e-ISSN : 2714-5417 (online)

implementasi dan komputasi serta digunakan untuk pra-pemrosesan diusulkan Isa dkk yaitu untuk vektorisasi [16].

Pada penelitian ini kami sebagai peneliti memutuskan mengangkat masalah pemrosesan data tentang pemilihan fitur terutama penggunaan ekstraksi fitur klasifikasi teks sebagai fokus utama. Dengan tujuan untuk mengetahui dasar pemilihan klasifikasi naive bayes sebagai klasifikasi yang digunakan pada penelitian ini dan mengevaluasi ekstraksi fitur usulan yaitu CountVectorizer sebagai ekstraksi fitur klasifikasi teks berbahasa Indonesia pada penelitian ini dan 3 level fitur TF-IDF Term Weighting sebagai pembanding, ketiga level tersebut adalah Character Level, Word Level dan N-gram Level yang dikombinasikan dengan Naive Bayes pada klasifikasi teks berbahasa Indonesia sehingga dapat diketahui fitur yang terbaik berdasarkan akurasi klasifikasi. **Keberhasilan penelitian ini dapat dimanfaatkan** sebagai masukan dan informasi bagi peneliti lain yang akan mengadakan penelitian selanjutnya atau tentang masalah yang di teliti untuk menerapkan dalam sistem yang lebih luas dan lebih kompleks. Penelitian ini disusun sebagai berikut pada bagian 2 dijelaskan tinjauan pustaka, metodologi penelitan dijelaskan pada bagian 3, bagian 4 menjelaskan hasil penelitian, kesimpulan dan saran dituliskan pada bagian 5.

TINJAUAN PUSTAKA

1.1. Tahapan SLR

Tinjauan pustaka sistematis atau *systematic literature review* (SLR) secara garis besar dilakukan secara urut dalam empat tahapan. Menurut Gaffar [17] pada saat menyusun *systematic literature review,* tahapan yang dilakukan adalah sebagai berikut :

1. Penentuan perencanaan mencakup mengembangkan ulasan pertanyaan, metode dan protokol rencana.
2. Melakukan penelusuran data mencakup pencarian secara komprehensif terkait judul dan abtsrak, penyaringan dan penilaian artikel yang sesuai dan melakukan ektraksi data.
3. Melakukan analisis mencakup analisis deskriptif atau tematik dan
4. Melakukan sintesis dan diskusi.

1.2. Pertanyaan Penelitian

Untuk menyusun tinjauan pustaka sistematis atau *systematic literature review* selalu diawali dan didasari oleh pertanyaan penelitian atau *research question* (RQ). Menurut Wahono [18] “RQ dibagi menjadi lima tahapan yang dikenal dengan sebutan PICOC. Tahapan PICOC yang digunakan pada penelitian ini termuat dalam tabel 2.1 sebagai berikut ini

Tabel 2.1 Ringkasan PICOC

Population	Ekstraksi fitur
Intervention	Klasifikasi Teks, klasifikasi, metode, atribut, dataset
Comparison	Membandingkan ekstraksi fitur
Outcomes	Pengaruh ekstraksi fitur dalam peningkatan akurasi
Context	Studi klasifikasi teks berbahasa Indonesia, *dataset* kecil dan besar

Pertanyaan penelitian ditunjukan pada Tabel 2.2, pertanyaan penelitian pada bab ini adalah pertanyaan penelitian untuk SLR berbeda dengan pertanyaan penelitian pada penelitian utama dalam makalah ini.

Tabel 2.2 Pertanyaan Penelitian atau *Research Question* (RQ)

ID	Pertanyaan Penelitian	Motivasi

p-ISSN : 1979-0155 e-ISSN : 2614-8870

RQ1	Jurnal Internasional yang mempublikasikan tentang evaluasi ekstraksi fitur pada klasifikasi teks ?	Mengidentifikasi Jurnal Internasional yang sering mempublikasikan penelitian tentang evaluasi ekstraksi fitur pada klasifikasi teks.
RQ2	Tren tahun penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks ?	Mengidentifikasi tren tahun penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks.
RQ3	Apa permasalahan yang muncul dalam penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks ?	Mengidentifikasi permasalahan yang sering muncul dalam penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks.
RQ4	Apa kontribusi yang dihasilkan dari penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks ?	Mengidentifikasi kontribusi-kontribusi yang dihasilkan dari penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks.
RQ5	Metode apa yang digunakan pada penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks ?	Mengidentifikasi metode yang sering digunakan dalam penelitian evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks.
RQ6	Apa saja pengukuran dan hasil penelitian yang digunakan dalam penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks ?	Mengidentifikasi pengukuran yang digunakan dan hasil penelitian dari penelitian evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks.
RQ7	Berapa banyak data yang digunakan dalam penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks?	Mengidentifikasi jumlah data yang sering digunakan dalam penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks.
RQ8	Berapa banyak atribut yang digunakan dalam penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi ?	Mengidentifikasi jumlah atribut yang sering digunakan dalam penelitian tentang evaluasi ekstraksi fitur untuk meningkatkan akurasi klasifikasi pada klasifikasi teks.

1.3. Strategi Pencarian

Proses pencarian sesuai dengan tahap 4 dalam tahapan *systematic literature review* di atas terdiri dari beberapa proses, termasuk pemilihan perpustakaan digital dan pengaturan kata kunci. Sebelum memulai pencarian dilakukan penentuan atau pemilihan *database* yang sesuai untuk menemukan jurnal yang relevan. Berikutini adalah perpustakaan digital dalam penelitian ini:

- IEEE Xplore (ieeexplore.ieee.org)
- ScienceDirect (sciencedirect.com)
- Google Scholar (scholar.google.co.id)
- Springer (link.springer.com)

p-ISSN : 1907-0012 (print); e-ISSN : 2714-5417 (online)

- Elsevier (elsevier.com)
- ACM Digital Library (dl.ac.org)

Kata kunci dikembangkan sesuai dengan langkah-langkah berikut:

1 Indentifikasi istilah pencarian dari PICOC, terutama dari populasi dan intervensi.
2 Identifikasi istilah pencarian dari pertanyaan penelitian.
3 Indentifikasi istilah pencarian dalam judul, abstrak dan kata kunci yang relevan.
4 Indentifikasi sinonim, ejaan alternatif dan antonim dari istilah pencarian.
5 Penentuan kata kunci yang menyeluruh menggunakan identifikasi istilah pencarian Boolean *AND* dan *OR*.

Kata kunci yang digunakan dalam pencarian guna penelitian ini adalah:

(Classification OR Classify OR Prediction OR Predict) AND Text Classification AND (Performance OR Evaluation) AND Feature Extraction

1.4. Seleksi Studi dan Ekstraksi Data

Proses pencarian dan seleksi studi utama dilakukan dalam dua langkah yaitu pengecualian studi utama berdasarkan judul dan abstrak dan pengecualian studi utama berdasarkan teks lengkap. Seleksi studi yang digunakan hanya jurnal, sedangkan buku dan prosiding tidak digunakan pada seleksi studi. Hasil akhir dari seleksi terdapat 42 (empat puluh dua) jurnal pada studi utama. Sedangkan ekstraksi data dirancang untuk mengumpulkan data dari studi utama yang dibutuhkan untuk menjawab pertanyaan penelitian seperti tabel 2.3 berikut.

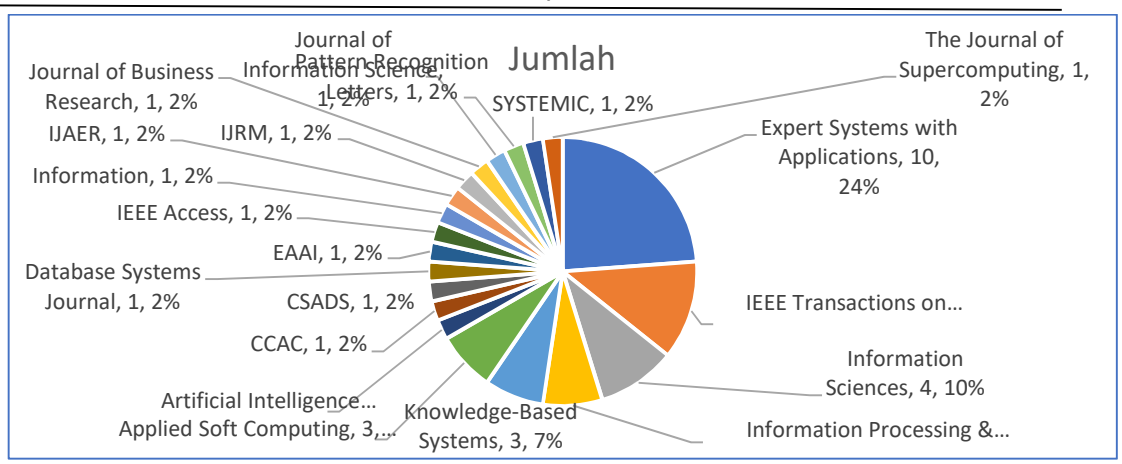
Tabel 2.3 Properti Ekstraksi Data untuk Pertanyaan Penelitian

Properti	Pertanyaan Penelitian
Peneliti dan tahun publikasi	RQ1, RQ2
Judul dan Abstrak	RQ3, RQ4
Metode yang sering digunakan	RQ5, RQ6
Dataset yang digunakan	RQ7, RQ8

1.5. Publikasi Jurnal Penting

Dari proses pemilihan studi, didapatkan 42 jurnal terkait klasifikasi teks yang berkontribusi dalam analisis *review* untuk menjawab pertanyaan penelitian yang telah dibuat sebelumnya. Kemudian dari jurnal yang terpilih, jurnal internasional yang berkontribusi di bidang evaluasi fitur untuk klasifikasi teks diidentifikasi. Gambar diagram 2.1 menunjukkan jurnal yang mempublikasikan topik tentang evaluasi fitur untuk klasifikasi teks.

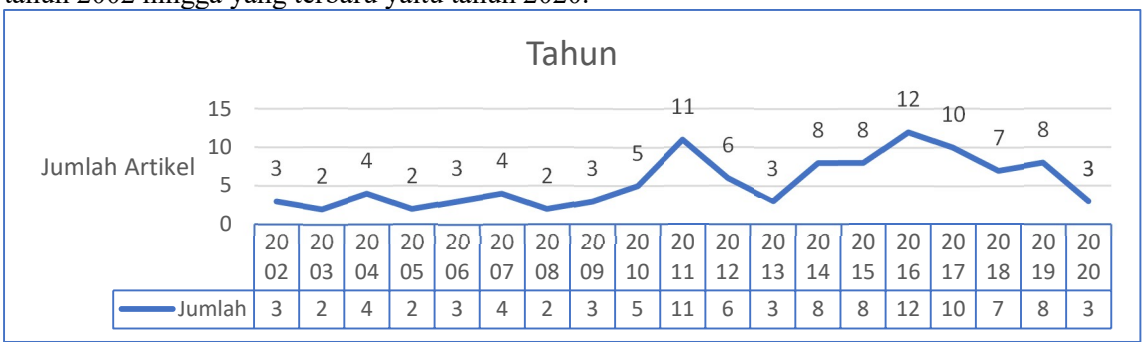
p-ISSN : 1979-0155 e-ISSN : 2614-8870



Gambar Diagram 2.1 Publikasi Jurnal yang Digunakan

Dari gambar diagram 2.1 dapat diketahui 3 (tiga) publikasi jurnal tertinggi yang sering mempublikasikan jurnal dengan topik tentang evaluasi fitur untuk klasifikasi teks yaitu *Expert Systems with Applications*, *IEEE Transaction on Knowledge and Data Engineering* dan *Information Sciences*. Ketiga publikasi jurnal tersebut memiliki persentase berturut-turut 24%, 12%, dan 10% dari 42 (empat puluh dua) publikasi jurnal SLR yang telah diidentifikasi.

Sedangkan pada gambar diagram 2.2 menunjukan jumlah penelitian tentang klasifikasi teks pertahun. Pada *review* ini tahun penelitian yang diambil adalah mulai dari tahun 2002 hingga yang terbaru yaitu tahun 2020.

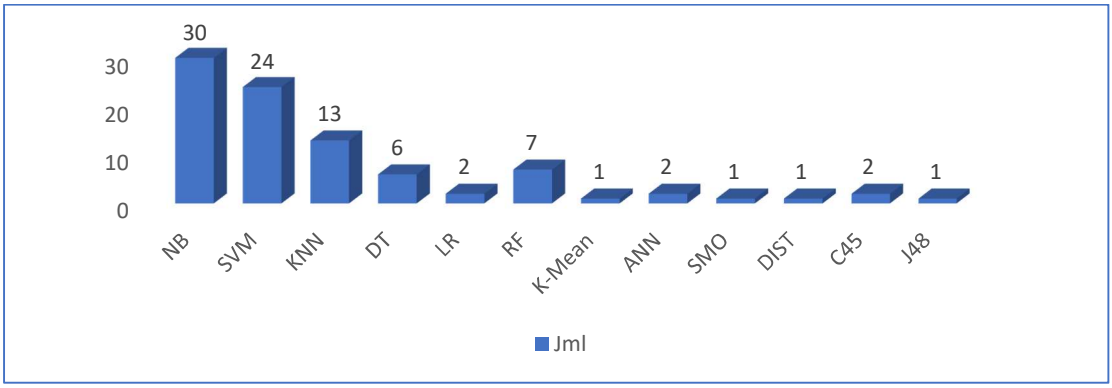


Gambar Diagram 2.2 Tren Tahun Penelitian

Dari gambar diagram 2.2 menunjukkan bahwa pada setiap tahunnya terdapat beberapa penelitian tentang evaluasi fitur untuk klasifikasi teks. Bisa dilihat dari gambar diagram 2.2, tren penelitian tentang evaluasi fitur untuk klasifikasi teks masih menjadi penelitian yang relevan saat ini karena setiap tahun masih terdapat penelitian tentang evaluasi fitur untuk klasifikasi teks.

Pada penelitian ini peneliti mencoba untuk *mereview* metode klasifikasi teks yang digunakan pada jurnal SLR tentang evaluasi fitur untuk klasifikasi teks dan dari hasil penelitian jurnal SLR peneliti mengidentifikasi beberapa metode klasifikasi teks yang sering digunakan seperti nampak pada gambar diagram 2.3

p-ISSN : 1907-0012 (print); e-ISSN : 2714-5417 (online)



Gambar Diagram 2.3 Metode yang Digunakan

Dari berbagai metode klasifikasi yang digunakan untuk klasifikasi teks seperti nampak pada gambar diagram 2.3 di atas bisa dilihat bahwa *Naive Bayes* selanjutnya disebut NB merupakan metode yang paling populer dipakai untuk klasifikasi teks. NB dipakai sebanyak 30 jurnal penelitian, metode *Support Vector Machine* selanjutnya disebut SVM digunakan sebanyak 24 jurnal penelitian, selanjutnya ada *Random Forest* (RF), *K-Nearest Neighbour* (KNN), dan *Decision Tree* (DT). Kelima metode klasifikasi tersebut sering digunakan oleh peneliti karena merupakan metode klasifikasi sederhana menurut para peneliti dan cocok diterapkan dalam proses klasifikasi teks.

Selanjutnya dari 42 (empat puluh dua) jurnal SLR hanya diambil 5 (lima) jurnal yang dijadikan sebagai jurnal SLR studi utama atau patokan dalam pelaksanaan penelitian penelitian ini dan hanya jurnal terpilih saja. Jurnal terpilih digunakan karena relevansinya terhadap penelitian yang akan dilakukan saat ini seperti tampak pada tabel 2.4 berikut:

Tabel 2.4 Jurnal Studi Utama

No	Penelitian	Masalah	Pendekatan	Hasil
1.	Penelitian [19]	Ketidakseimbangan kelas (class imbalance) pada dataset .	*Naive Bayes* + IG + IGFSS dan SVM + IG+IGFSS (Komparasi)	*Naive Bayes* menghasilkan akurasi yang signifikan hingga mencapai 98% sedangkan SVM dengan akurasi sebesar 97%.
2.	Penelitian [20]	*Reduction* fitur pada klasifikasi teks dalam hal ini tingkat redundansi fitur dan dimensi ruang fitur.	*Naive Bayes* + *Information Gain* + *Maximal Marginal Relevance for Feature Selection* (MMR-FS)	*Naive Bayes* menghasilkan akurasi yang meningkat signifikan hingga mencapai 86% .
3.	Penelitian [21]	*Reduction* fitur pada klasifikasi	*Naive Bayes* +	Peningkatan kinerja SABigram lebih signifikan

p-ISSN : 1979-0155 e-ISSN : 2614-8870

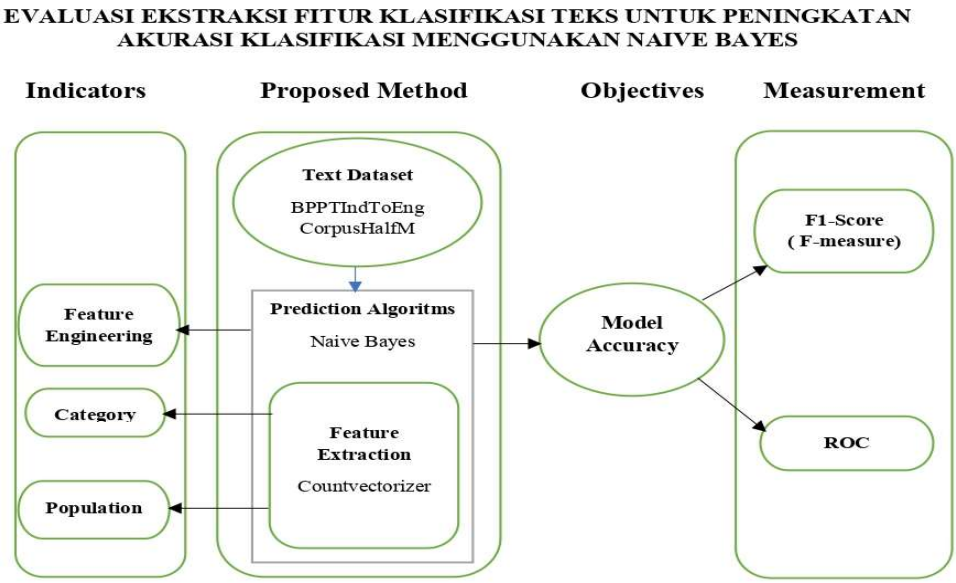
		teks dalam hal ini tingkat redundansi fitur dan dimensi ruang fitur.	SABigram, SVM + SABigram, (Komparasi)	yaitu sebesar 75% saat menggunakan pengklasifikasi *Naive Bayes* daripada SVM. Akurasi dengan NB juga menunjukkan hasil yang bagus yaitu sebesar 78%
4.	Penelitian [22]	*Reduction* fitur pada klasifikasi teks dalam hal ini tingkat redundansi fitur dan dimensi ruang fitur	Integrasi MRMR dan MCC pada *Naive Bayes*	*Naive Bayes* mampu mengklasifikasikan ulasan hotel di Las Vegas dengan *precision* dan *recall* yang tinggi (skor F1> 0,84).
5.	Penelitian [23]	Ketidakseimbangan kelas (*class imbalance*) pada dataset .	*Naive Bayes* + *SmartBT*	*Naive Bayes* + *SmartBT* adalah metode yang efisien dalam dataset berdimensi rendah dengan akurasi sebesar 73%.

Dari kelima penelitian terkait tersebut pada tabel 2.4 dapat disimpulkan bahwa metode *Naive Bayes Classifier* mampu dalam penyelesaian masalah klasifikasi teks dengan akurasi tertinggi yaitu sebesar 98% ketika *Naive Bayes* dikombinasikan dengan *Information Gain* dan IGFSS [19], dan pada penelitian [20], [22], [23] akurasi *Naive Bayes* yang dikombinasikan dengan beberapa fitur hanya mampu memberi hasil sebesar 73% s.d 86%, sedangkan pada penelitian [21] tidak secara signifikan membandingkan akurasi *Naive Bayes* dengan *Support Vector Machine*, karena pada penelitian [21] ini titik berat penelitian hanya pada pengukuran kinerja fitur ekstraksi SABigram yang naik sebesar 75% ketika dikombinasikan dengan *naive bayes* dimana akurasi yang didapat sebesar 78%. Dengan demikian pada penelitian ini kami memutuskan untuk menggunakan klasifikasi *Naive Bayes* sebagai metode klasifikasi berdasarkan kesimpulan yang didapat dari kelima jurnal penelitan pada tabel 2.4.

1.6. Kerangka Berpikir

Agar penelitian ini terarah dan mencapai tujuan penelitian yang telah ditetapkan maka peneliti membuat kerangka berpikir seperti gambar bagan di bawah ini :

p-ISSN : 1907-0012 (print); e-ISSN : 2714-5417 (online)



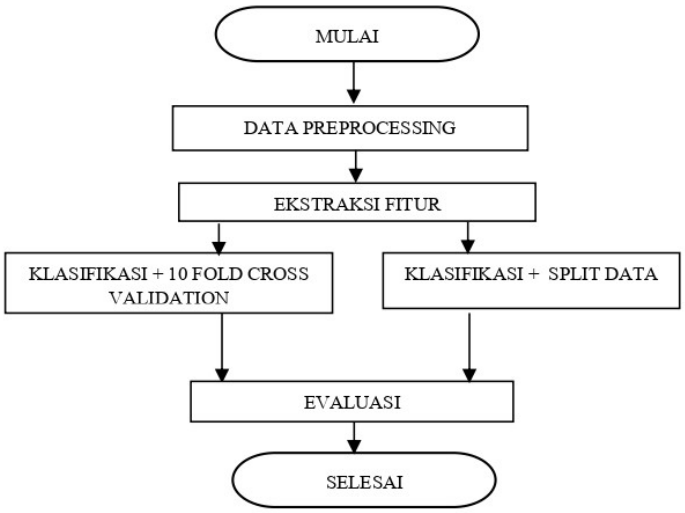
Gambar 2.4 Kerangka Berpikir

METODE PENELITIAN

Penelitian ini menggunakan sampel penelitian dataset BPPTIndToEngCorpusHalfM [24] yaitu dataset yang berisi 500.463 kata dari berbagai sumber *online*, dataset ini merupakan kumpulan data *Word Bahasa Indonesia Corpus* dan *Parallel English Translation* yang dibuat oleh BPPT (Badan Pengkajian dan Penerapan Teknologi Indonesia), dimana dataset ini memiliki 4 kelas yakni : (0 : *Ekonomic*, 1 : *International*, 2 : *Sciene*, 3: *Sport*).

1.7. Metode Usulan

Tahapan metode usulan yang peneliti lakukan seperti pada bagan gambar 3.1 sebagai berikut :



Gambar 3.1 Metode Usulan

Berdasarkan gambar kegiatan pertama yang peneliti lakukan sebelum memulai penelitian yaitu menyiapkan peralatan yang digunakan. Peralatan yang peneliti gunakan yaitu sebuah laptop dengan spesifikasi sistem operasi windows 10 *home edition*, *processor* AMD A9-Radeon R5 3.0 GHz dan RAM berkapasitas 4 GB. selanjutnya peneliti

p-ISSN : 1979-0155 e-ISSN : 2614-8870

melakukan proses *data preprocessing* dengan langkah sebagai berikut pertama mengimport dataset dari file untuk dibagi menjadi 2 (dua) jenis data yaitu data independen dan data dependen, kedua menangani nilai data kosong, ketiga menggunakan encode untuk mengangani data, keempat membagi dataset menjadi data latih, data valid, dan data uji, serta yang kelima melakukan *feature scaling*

Bedasarkan hasil *data preprocessing* maka dataset yang digunakan sebanyak 24024 record yang dibagi menjadi 21621 data latih dan 2403 data uji. Dengan dataset perkelas yaitu *Economy* 6544 data, *International* 6642, *Science* 6355 data, dan *Sport* 4483 data

1.8. Feature Engineering

Feature engineering adalah proses menggunakan pengetahuan data untuk membuat fitur atau variabel yang membuat algoritma pembelajaran mesin bekerja lebih efisien. Ini adalah tugas mendasar untuk meningkatkan performa model *machine learning* dan akurasi prediksi., pada langkah ini, data teks mentah akan diubah menjadi fitur vektor dan pada penelitian ini peneliti menggunakan fitur-fitur sebagai berikut:

Tabel 3. 1. Contoh Dokumen Data Awal

d1	tinggi harga minyak dunia sentuh dolar as barel boediono amat tren naik harga minyak turun

Dengan menggunakan contoh data pada tabel 3.1 maka hasil fitur vektor yang dihasilkan dari masing-masing fitur sebagai berikut

Tabel 3.2. CountVectorizer

Index	Kata	tf
		d1
0	tinggi	1
1	harga	2
2	minyak	2
3	dunia	1
4	sentuh	1
5	dolar	1
6	as	1
7	barel	1
8	boediono	1
9	amat	1
10	tren	1
11	naik	1
12	turun	1

Tabel 3.3. Character Level

Char	df	D/df + 1	log (D/df +1)	tf x (log (D/df+1)
a	13	0,076923077	0,032184683	0,0322
b	2	0,5	0,176091259	0,1761
d	3	0,333333333	0,124938737	0,1249
e	4	0,25	0,096910013	0,0969
g	4	0,25	0,096910013	0,0969
h	3	0,333333333	0,124938737	0,1249
i	7	0,142857143	0,057991947	0,0580
k	3	0,333333333	0,124938737	0,1249
l	2	0,5	0,176091259	0,1761
m	3	0,333333333	0,124938737	0,1249
n	9	0,111111111	0,045757491	0,0458
o	4	0,25	0,096910013	0,0969
r	6	0,166666667	0,06694679	0,0669
s	2	0,5	0,176091259	0,1761
t	5	0,2	0,079181246	0,0792
u	4	0,25	0,096910013	0,0969
y	2	0,5	0,176091259	0,1761

Tabel 3.4. Ngram Level

Tabel 3.5. Word Level

p-ISSN : 1907-0012 (print); e-ISSN : 2714-5417 (online)

NO	Bigram	df	D/df + 1	log (D/df +1)	tf x (log (D/df+1)
1	tinggi harga	1	1	0,301029996	0,30103
2	harga minyak	2	0,5	0,176091259	0,176091
3	minyak dunia	1	1	0,301029996	0,30103
4	dunia sentuh	1	1	0,301029996	0,30103
5	sentuh dolar	1	1	0,301029996	0,30103
6	dolar as	1	1	0,301029996	0,30103
7	as barel	1	1	0,301029996	0,30103
8	barel boediono	1	1	0,301029996	0,30103
9	boediono amat	1	1	0,301029996	0,30103
10	amat tren	1	1	0,301029996	0,30103
11	tren naik	1	1	0,301029996	0,30103
12	naik harga	1	1	0,301029996	0,30103
13	harga minyak	2	0,5	0,176091259	0,176091
14	minyak turun	1	1	0,301029996	0,30103

Index	Kata	tf	df	D/df + 1	log (D/df +1)	tf x (log (D/df+1)
0	tinggi	1	1	2	0,30103	0,30103
1	harga	2	2	1,5	0,17609	0,35218
2	minyak	2	2	1,5	0,17609	0,35218
3	dunia	1	1	2	0,30103	0,30103
4	sentuh	1	1	2	0,30103	0,30103
5	dolar	1	1	2	0,30103	0,30103
6	as	1	1	2	0,30103	0,30103
7	barel	1	1	2	0,30103	0,30103
8	boediono	1	1	2	0,30103	0,30103
9	amat	1	1	2	0,30103	0,30103
10	tren	1	1	2	0,30103	0,30103
11	naik	1	1	2	0,30103	0,30103
12	turun	1	1	2	0,30103	0,30103

HASIL PENELITIAN

Kinerja Klasifikasi

Untuk membandingkan kinerja klasifikasi peneliti memakai 10-Fold Cross Validation dan Split Data dengan memakai rasio 90:10, pemakaian 10-Fold Cross Validation direkomendasikan untuk pemilihan model terbaik karena cenderung memberikan estimasi akurasi yang bagus dibandingkan dengan yang lainnya, selain itu banyak hasil eksperimen menunjukan bahwa 10-Fold Cross Validation adalah pilihan terbaik untuk mendapatkan estimasi yang akurat. Kinerja klasifikasi menggunakan 10-Fold Cross Validation seperti berikut

Precision, Recall, F1-Score

	precision	recall	f1-score	support
0	0.90	0.90	0.90	640
1	0.87	0.90	0.88	669
2	0.85	0.85	0.85	670
3	0.96	0.91	0.93	424
accuracy			0.89	2403
macro avg	0.89	0.89	0.89	2403
weighted avg	0.89	0.89	0.89	2403

Gambar 4.1 F1-Score CountVectorizer + CV

Precision, Recall, F1-Score

	precision	recall	f1-score	support
0	0.90	0.74	0.81	640
1	0.64	0.91	0.75	669
2	0.75	0.79	0.77	670
3	0.99	0.50	0.66	424
accuracy			0.76	2403
macro avg	0.82	0.73	0.75	2403
weighted avg	0.80	0.76	0.76	2403

Gambar 4.2 F1-Score Character Level + CV

Precision, Recall, F1-Score

	precision	recall	f1-score	support
0	0.81	0.77	0.79	640
1	0.72	0.78	0.75	669
2	0.66	0.76	0.71	670
3	0.98	0.70	0.81	424
accuracy			0.76	2403
macro avg	0.79	0.75	0.76	2403
weighted avg	0.77	0.76	0.76	2403

Gambar 4.3 F1-Score Ngram Level + CV

Precision, Recall, F1-Score

	precision	recall	f1-score	support
0	0.90	0.90	0.90	640
1	0.84	0.90	0.87	669
2	0.84	0.85	0.85	670
3	0.99	0.88	0.93	424
accuracy			0.88	2403
macro avg	0.89	0.88	0.89	2403
weighted avg	0.88	0.88	0.88	2403

Gambar 4.4 F1-Score Word Level + CV

Sedangkan kinerja klasifikasi mengunakan Split Data dengan rasio 90:10 seperti berikut

p-ISSN : 1979-0155 e-ISSN : 2614-8870

	precision	recall	f1-score	suppor
0	0.94	0.93	0.94	64
1	0.92	0.92	0.92	66
2	0.90	0.92	0.91	67
3	0.97	0.95	0.96	42
accuracy			0.93	240
macro avg	0.93	0.93	0.93	240
weighted avg	0.93	0.93	0.93	240

Gambar 4.5 F1-Score CountVectorizer

	precision	recall	f1-score	support
0	0.89	0.81	0.85	640
1	0.81	0.87	0.84	669
2	0.78	0.86	0.82	670
3	0.96	0.81	0.87	424
accuracy			0.84	2403
macro avg	0.86	0.84	0.85	2403
weighted avg	0.85	0.84	0.84	2403

Gambar 4.6 F1-Score Character Level

	precision	recall	f1-score	support
0	0.87	0.79	0.83	640
1	0.71	0.84	0.77	669
2	0.74	0.71	0.72	670
3	0.93	0.84	0.88	424
accuracy			0.79	2403
macro avg	0.81	0.80	0.80	2403
weighted avg	0.80	0.79	0.79	2403

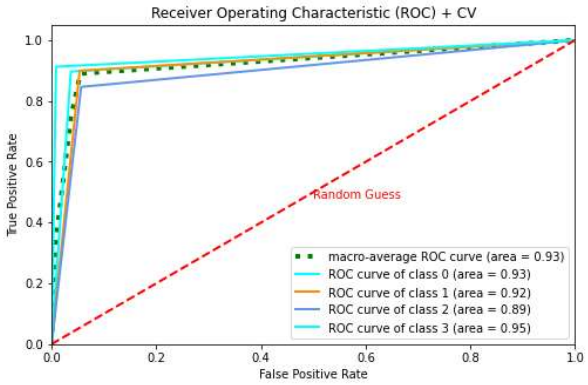
Gambar 4.7 F1-Score Ngram Level

	precision	recall	f1-score	support
0	0.93	0.94	0.93	640
1	0.91	0.90	0.91	669
2	0.89	0.91	0.90	670
3	0.98	0.95	0.96	424
accuracy			0.92	2403
macro avg	0.93	0.92	0.93	2403
weighted avg	0.92	0.92	0.92	2403

Gambar 4.8 F1-Score Word Level

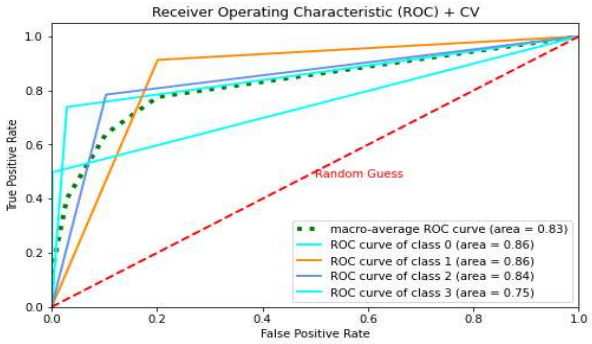
1.9. Keakuratan Akurasi

Untuk mengetahui keakuratan akurasi peneliti menghitung luas daerah dibawah kurva ROC yang disebut AUC, dengan harapan penelitian ini mendapatkan hasil akurasi yang baik sehingga dapat dijadikan acuan untuk dikembangkan menggunakan metode yang lain. Keakuratan klasifikasi menggunakan 10-Fold Cross Validation seperti berikut

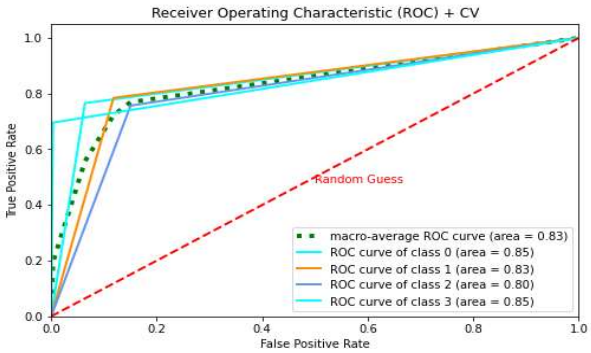


Gambar 4.9 ROC CountVectorizer + CV

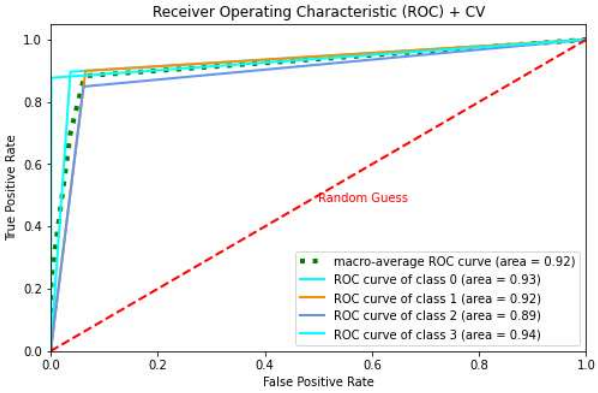
p-ISSN : 1907-0012 (print); e-ISSN : 2714-5417 (online)



Gambar 4.10 ROC Character Level + CV



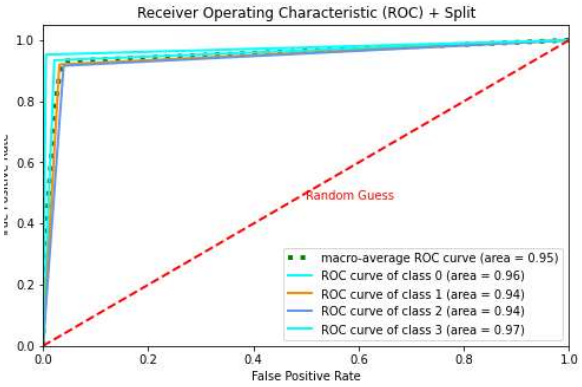
Gambar 4.11 ROC Ngram Level + CV



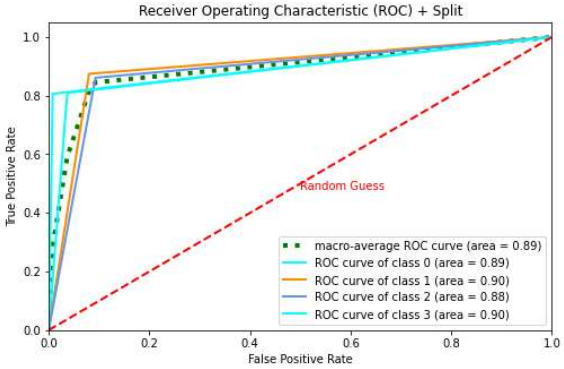
Gambar 4.12 ROC Word Level + CV

Sedangkan keakuratan klasifikasi mengunakan Split Data dengan rasio 90:10 seperti berikut

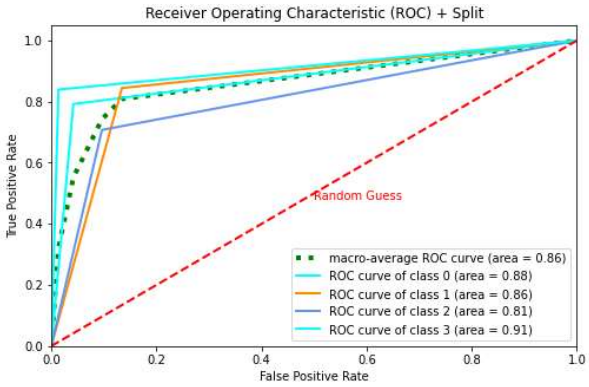
p-ISSN : 1979-0155 e-ISSN : 2614-8870



Gambar 4.13 ROC CountVectorizer

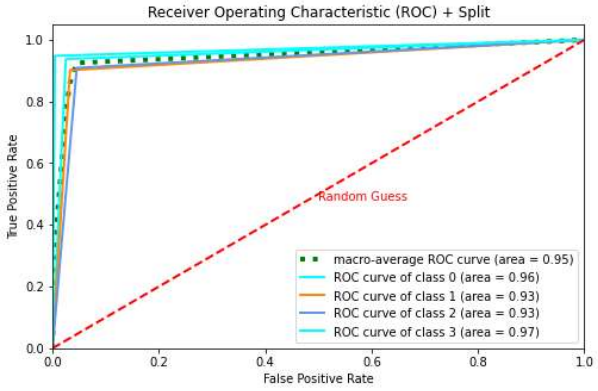


Gambar 4.14 ROC Character Level



Gambar 4.15 ROC Ngram Level

p-ISSN : 1907-0012 (print); e-ISSN : 2714-5417 (online)



Gambar 4.12 ROC Word Level

2. KESIMPULAN DAN SARAN

2.1. Kesimpulan

Berdasarkan hasil penelitian yang sudah peneliti lakukan dapat ditarik beberapa kesimpulan sebagai berikut :

1. Sesuai dengan hasil *systematic literatur review* pada penelitian tesis ini peneliti memutuskan untuk menggunakan klasifikasi Naive Bayes sebagai metode klasifikasi berdasarkan kesimpulan yang didapat dari kelima jurnal penelitan pada tabel 2.4.

2. Hasil dari evaluasi terhadap ekstraksi fitur yang diusulkan pada klasifikasi teks berbahasa Indonesia dan diterapkan pada dataset BPPTIndToEngCorpusHalfM yaitu *counvectorizer* untuk dibandingkan dengan *word level tf-idf*, *n-gram level tf-idf*, dan *character level tf-idf* menunjukan bahwa *countvectorizer* mempunyai akurasi yang bagus dibandingkan dengan fitur yang lainnya untuk semua metode klasifikasi baik menggunakan *10-fold cross validation* maupun split data. Hasil perhitungan *confusion matrik* dan evaluasi menggunakan nilai akurasi *F-1 score* terlihat pada tabel berikut :

Tabel 5.1 Akurasi *F1-Score* Klasifikasi

Fitur	Akurasi *F1-Score*	
	Split	*10-fold*
CountVectorizer	0,93	0,89
Word Level	0,92	0,88
N-Gram Level	0,79	0,75
Character Level	0.84	0.76

Selain menggunakan *confusion matrik* kinerja klasifikasi juga diukur menggunakan ROC, dan hasil pengukuran keakuratan klasifikasi ROC menghasilkan AUC seperti terlihat pada tabel berikut :

Tabel 5.2 AUC Keakuratan Klasifikasi

Fitur	*ROC Split*		*ROC 10-fold*	
	AUC	Keakuratan	AUC	Keakuratan
CountVectorizer	0,95	Sangat Baik	0,93	Sangat Baik
Word Level	0,95	Sangat Baik	0,92	Sangat Baik

p-ISSN : 1979-0155 e-ISSN : 2614-8870

N-Gram Level	0,85	Baik	0,83	Baik
Character Level	0,89	Baik	0,83	Baik

3. Berdasarkan tabel 5.1. dan tabel 5.2. dapat disimpulkan bahwa pemilihan ekstraksi fitur yang tepat dapat meningkatkan hasil akurasi klasifikasi dengan bukti bahwa pemakaian fitur usulan *countvectorizer* menghasilkan akurasi sebesar 0,93 dan AUC sebesar 0,95.

2.2. Saran

Berdasarkan kesimpulan di atas maka untuk penelitian yang akan datang, saran yang dapat diajukan peneliti sebagai berikut :

1. Perlu dilakukan studi penambahan *stemming* untuk melihat pengaruh terhadap hasil akurasi klasifikasi.

2. Perlu dilakukan studi perbandingan menggunakan dataset yang berbeda untuk mengetahui seberapa besar pengaruh fitur *countvectorizer* jika dibandingkan dengan *word level tf-idf*, *n-gram level tf-idf*, dan *character level tf-idf,* masih memberi hasil yang bagus atau akan memberikan hasil yang kurang bagus.

3. Perlu dilakukan studi perbandingan dengan fitur yang lain seperti *word2vec* atau *bert*.

4. Perlu dilakukan studi perbandingan dengan metode klasifikasi yang lain seperti SVM*, Logisctic Regresion*, *Random Forest* atau *Neural Network*.

DAFTAR PUSTAKA

[1] N. Nicolosi, “Feature Selection Methods for Text Classification,” *Res. Pap. Present. Rochester Inst. Technol.*, pp. 1–11, 2008.

[2] A. Purohit, D. Atre, and P. Jaswani, “Text Classification in Data Mining,” *Int. J. Sci. Res. Publ.*, vol. 5, no. 6, pp. 1–6, 2015.

[3] S. Scott and S. Matwin, “FEATURE ENGINEERING FOR TEXT CLASSIFICATION,” *Mach. Learn. Work.*, vol. 6, pp. 1–13, Oct. 1999.

[4] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, “A novel feature selection algorithm for text categorization,” *Expert Syst. Appl.*, vol. 33, no. 1, pp. 1–5, 2007.

[5] P. Somol and J. Novovičová, “Evaluating Stability and Comparing Output of Feature Selectors that Optimize Feature Subset Cardinality,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1921–1939, Nov. 2010.

[6] D. Dernoncourt, B. Hanczar, and J.-D. Zucker, “Analysis of feature selection stability on high dimension and small sample data,” *Comput. Stat. Data Anal.*, vol. 71, pp. 681–693, Mar. 2014.

[7] S. Xu, “Bayesian Naïve Bayes classifiers to text classification,” *J. Inf. Sci.*, vol. 44, no. 1, pp. 48–59, Feb. 2018.

[8] S. L. Ting, W. H. Ip, and A. H. C. Tsang, “Is Naïve bayes a good classifier for document classification?,” *Int. J. Softw. Eng. its Appl.*, vol. 5, no. 3, pp. 37–46, 2011.

[9] D. D. Lewis, “A Comparison of Two Learning Algorithms for Text Categorization 1 Introduction 2 Text Categorization : Nature and Approaches,” *Proceeding Third Annu. Symp. Doc. Anal. Inf. Retr.*, pp. 1–14, 1994.

[10] Vidhya. K. A and G. Aghila, “A Survey of Na\"ive Bayes Machine Learning approach in Text Document Classification,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 7, no. 2, pp. 206–211,

p-ISSN : 1907-0012 (print); e-ISSN : 2714-5417 (online)

2010.

[11] S. H. Myaeng, K. S. Han, and H. C. Rim, "Some effective techniques for naive bayes text classification," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 11, pp. 1457–1466, 2006.

[12] K.-M. Schneider, "Techniques for Improving the Performance of Naive Bayes for Text Classification," in *Lecture Notes in Computer Science*, vol. 3406, 2005, pp. 682–693.

[13] Y. Jiang, H. Lin, X. Wang, and D. Lu, "A Technique for Improving the Performance of Naive Bayes Text Classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6988 LNCS, no. PART 2, 2011, pp. 196–203.

[14] W. Zhang and F. Gao, "Performance analysis and improvement of naïve Bayes in text classification application," *2013 IEEE Conf. Anthol. Anthol. 2013*, pp. 1–4, 2013.

[15] M. J. Pazzani, "Searching for Dependencies in Bayesian Classifiers," in *Learning from Data. Lecture Notes in Statistics*, Springer, New York, NY, 1996, pp. 239–248.

[16] D. Isa, L. H. Lee, V. P. Kallimani, and R. Rajkumar, "Text document preprocessing with the bayes formula for classification using the support vector machine," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1264–1272, 2008.

[17] V. Gaffar, "Kupas Tuntas Systematic Literature Review," *UPI*, 2020. [Online]. Available: https://berita.upi.edu/kupas-tuntas-systematic-literature-review/. [Accessed: 08-Feb-2021].

[18] R. S. WAHONO, "SYSTEMATIC LITERATURE REVIEW: PENGANTAR, TAHAPAN DAN STUDI KASUS," 2016. [Online]. Available: https://romisatriawahono.net/2016/05/15/systematic-literature-review-pengantar-tahapan-dan-studi-kasus/#:~:text=Systematic literature review atau sering,pertanyaan penelitian (research question) yang. [Accessed: 08-Feb-2021].

[19] A. K. Uysal, "An improved global feature selection scheme for text classification," *Expert Syst. Appl.*, vol. 43, pp. 82–92, 2016.

[20] M. A. Fauzi, S. Gosario, A. Z. Arifin, and I. S. Prabowo, "Klasifikasi Berita Berbahasa Indonesia Menggunakan Seleksi Fitur Dua Tahap dan Naive Bayes," *SYSTEMIC*, vol. 03, no. 02, pp. 7–12, 2017.

[21] C. Wan, Y. Wang, Y. Liu, J. Ji, and G. Feng, "Composite Feature Extraction and Selection for Text Classification," *IEEE Access*, vol. 7, pp. 35208–35219, 2019.

[22] M. J. Sánchez-Franco, A. Navarro-García, and F. J. Rondán-Cataluña, "A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services," *J. Bus. Res.*, vol. 101, no. December, pp. 499–506, 2019.

[23] F. Asdaghi and A. Soleimani, "An effective feature selection method for web spam detection," *Knowledge-Based Syst.*, vol. 166, pp. 198–206, 2019.

[24] C. Wirawan, "Indonesian Language Models," 2018. [Online]. Available: https://github.com/cahya-wirawan/indonesian-language-models/tree/master/data. [Accessed: 21-Jul-2021].