Jurnal Ilmiah Wahana Pendidikan, Maret 2024, 10 (5), 45-58
DOI: https://doi.org/10.5281/zenodo.10516379
p-ISSN: 2622-8327 e-ISSN: 2089-5364
Accredited by Directorate General of Strengthening for Research and Development
Available online at https://jurnal.peneliti.net/index.php/JIWP

Analisis Sentimen Tweet Penanganan Covid-19 Di Indonesia Menggunakan SVM dan Naïve Bayes dengan Operator Seleksi Fitur Information Gain

Aisyah Nur Hasna[1] , Fajri Rakhmat Umbara[2], Puspita Nurul Sabrina[3]

[1,2,3]Universitas Jenderal Achmad Yani

Received: 15 November 2023
Revised: 13 Desember 2023
Accepted: 15 Januari 2023

Abstract

Opinion that is present from the public is one indicator of sentiment assessment that can be used to assess a matter. In 2020, the world is experiencing a COVID-19 pandemic so that Indonesia is also affected. On Twitter social media at that time there was a lot of discussion about the virus and the state of government policy at that time. Through these tweets, there are those who agree to provide a response to the policy, there are also those who oppose or disagree. Producing these responses is divided into two types of groups, namely positive and negative groups. In this study, tweets were analyzed using two algorithms, namely SVM and Naïve Bayes compared with and without feature selection by the information gain operator so that information is extracted that public opinion tends to be positive or negative. Comparing the algorithms in this study resulted in the highest level of accuracy using the SVM method plus information gain which resulted in an accuracy rate of 66.7% with a precision of 65.5%, a recall value of 66.9% and an f1-score of 66.2%.

Keywords: *Covid-19, SVM; Naïve Bayes; Information Gain; Sentimen Analysis*

(*) Corresponding Author: aisyahnrhasna01@gmail.com

How to Cite: Hasna, A. N., Umbara, F. R., & Sabrina, P. N. (2024). Analisis Sentimen Tweet Penanganan Covid-19 Di Indonesia Menggunakan SVM dan Naïve Bayes dengan Operator Seleksi Fitur Information Gain. https://doi.org/10.5281/zenodo.10516379.

INTRODUCTION

Topik dengan tema politik sangat ramai diperbincangkan di twitter karena disini para pengguna merasa bebas berekspresi serta menyampaikan argumen atau opini (Emeraldien, Sunarsono, and Alit 2019). Sehingga platform twitter ini menjadi salah satu media sosial yang diakses oleh kebanyakan orang ketika ramai masalah yang sedang terjadi serta banyak diperbincangkan(Deller 2011). Di tahun 2020, pemerintahan Indonesia sedang ramai diperbincangkan sehingga banyak opini yang muncul dari para pengguna twitter pada saat itu.

Kebijakan yang menjadi perdebatan di twitter yaitu pemberlakuan lockdown dan juga pada saat pemberlakuan tersebut, DPR RI tetap menyelenggarakan rapat paripurna ke-12 masa persidangan III tahun sidang 2019-2020 yang dipimpin oleh pimpinan DPR(RI 2020). Pemberlakuan lockdown di Indonesia menjadi perdebatan karena ada yang setuju dengan kebijakan tersebut, dan ada pula yang tidak setuju atas kebijakan lockdown di Indonesia.

Penelitian terdahulu lainnya juga yaitu dengan judul 'Analisis Sentimen Pada Ulasan Buku Berbahasa Inggris Menggunakan Information Gain dan Support Vector Machine' telah dilakukan sehingga menghasilkan akurasi 71,04% menggunakan SVM dan 70,69% menggunakan SVM dan information gain(Hilman, Nurjaman, and Mubarok 2017). Penelitian lain dengan judul 'Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes Dan Seleksi Fitur Information Gain' menghasilkan akurasi sebesar 81,00% pada Naïve Bayes dan 86,50% pada penambahan seleksi fitur information gain(Negara, Muhardi, and

Putri 2020). Dan juga terdapat penelitian berjudul 'Penerapan Analisis Sentimen untuk Menilai Suatu Produk pada Twitter Berbahasa Indonesia dengan Metode Naïve Bayes Classifier dan Information Gain' menghasilkan kenaikan 4% tingkat akurasi dengan hasil 70,00% menggunakan Naïve Bayes dan 74,00% dengan melakukan penambahan seleksi fitur information gain(Ahmad Wildan Attabi, Lailil Muflikhah, and Mochammad Ali Fauzi 2018).

Terdapat penelitian terdahulu pada tahun 2020 yang berjudul 'Analisis Sentimen Dewan Perwakilan Rakyat Dengan Algoritma Klasifikasi Berbasis Particle Swarm Optimization' menguji kasus tersebut menggunakan algoritma Naïve Bayes dan SVM serta membandingkannya dengan dan tanpa optimasi PSO sehingga dapat disimpulkan pada kasus ini yaitu penggunaan algoritma SVM dengan optimasi PSO lebih baik jika dibandingkan dengan algoritma lainnya yang diuji. Pada paper tersebut terdapat saran penelitian yaitu melanjutkan kasus tersebut menggunakan operator feature selection by corellation atau feature selection by information gain(Faisal et al. 2020).

Dari penelitian-penelitian terdahulu, penggunaan seleksi fitur information gain dalam analisis sentimen diharapkan dapat membantu memperbaiki hasil akurasi pada studi kasus tertentu yang didapatkan dari pemodelan menggunakan metode algortima SVM dan Naïve Bayes. Sehingga pada penelitian ini menerapkan metode SVM dan Naïve Bayes yang hasilnya dibandingkan dengan penggunaan seleksi fitur information gain pada setiap masing-masing algoritma terhadap analisis sentimen penanganan covid-19 di Indonesia, untuk mengetahui apakah terdapat perbaikan akurasi dengan adanya penggunaan seleksi fitur pada kasus ini.

METHODS

Berikut merupakan alur metode penelitian yang dilakukan untuk mengetahui opini masyarakat mengenai penanganan covid-19 di Indonesia yang secara umum digambarkan pada Figure 1.

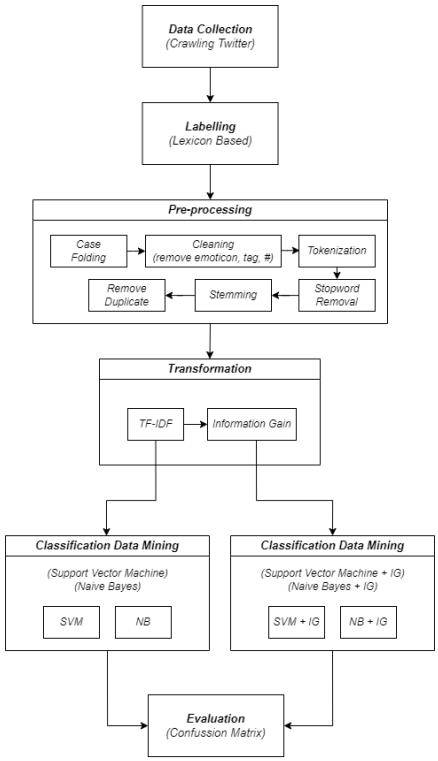

Figure 1. Methods

1. *Data Collection*

Data diperoleh dari *twitter* menggunakan *platform* PhantomBuster.

2. *Labelling*

Pelabelan dilakukan menggunakan *lexicon* dengan menghasilkan data negatif dan positif.

3. *Pre-processing*

Setelah data memiliki label, data tersebut dibersihkan dengan menggunakan tahapan-tahapan *pre-processing* untuk menghasilkan data yang bersih agar dapat diklasifikasi. Tahapan tersebut diantaranya.

a. *Case folding*

Tahap *case folding* merupakan merupakan tahapan dari preprocessing yang digunakan untuk mengubah seluruh teks menjadi huruf kecil. Tahapan ini berfungsi untuk menyamakan bentuk huruf sehingga kata yang berbeda bentuknya akan memiliki makna yang sama. Selain itu, tahap case folding juga memiliki fungsi untuk mengurangi dimensi data karena menghindari duplikasi akibat perbedaan bentuk huruf yang besar dan kecil akibat dari variasi kata.

b. *Cleaning*

Cleaning atau pembersihan karakter merupakan tahap yang digunakan untuk menghapus karakter–karakter yang tidak diperlukan dalam proses pengolahan data. Karakter yang tidak diperlukan yaitu seperti emoticon, link, tag mention dan hashtag, karena data pada twitter kemungkinan mengandung karakter–karakter tersebut.

c. *Tokenization*

Tokenization merupakan tahapan yang digunakan untuk memecah teks menjadi unit–unit kecil yang biasa disebut dengan token. Token tersebut dapat berupa

pemisahan kata, selain pemisahan kata token tersebut juga dapat memisahkan frasa sampai karakter sesuai dengan kebutuhan. Namun token yang digunakan pada pengolahan data penelitian ini yaitu pemisahan kata.

d. *Stopword Removal*

Stopword removal merupakan proses penghapusan kata pada dataset yang terdaftar di dalam kamus stopword. Kumpulan kata yang ada di dalam kamus stopword merupakan kata–kata umum yang dianggap kurang memiliki makna penting atau memiliki nilai informasi yang rendah seperti kata penghubung dan kata ganti sehingga perlu untuk dihapus. Selain itu stopword juga dapat membantu mengurangi dimensi data dan memfokuskan pada kata–kata yang lebih relevan dan bermakna dalam penelitian analisis sentimen ini.

e. *Stemming*

Stemming merupakan tahapan proses mengubah kata–kata di dalam teks menjadi bentuk dasarnya dengan menghapus imbuhan atau akhiran kata yang tidak relevan. Tahapan ini digunakan untuk memperkecil variasi kata karena kata yang kemungkinan memiliki makna sama namun imbuhannya berbeda akan menambah variasi kata sehingga tidak memfokuskan pada kata yang relevan.

f. *Remove Duplicate*

Remove duplicate dilakukan untuk menghindari bobot yang terlalu tinggi pada sentimen yang sama sehingga dapat mencegah bias dan memperoleh representasi yang lebih akurat.

4. *Transformation*

a. TF–IDF

TF–IDF (Term Frequency–Invers Document Frequency) merupakan jenis fitur ekstraksi berbasis kata yang digunakan untuk mengevaluasi pentingnya kata per kata di dalam teks berdasarkan frekuensi kemunculam kata tersebut dalam suatu data (term frequency) dan jumlah data yang mengandung kata tersebut di dalam dataset (invers document frequency). Menghitung bobot nilai TF-IDF adalah menggunakan persamaan (1.1) (Akbari, Novianty, and Casi 2017):

$$W_{t,d} = W_{tf_{t,d}} \times idf_t \quad \text{.....(1.1)}$$

Keterangan :

$W_{t,d}$ = bobot TF-IDF

$W_{tf_{t,d}}$ = bobot frekuensi term

idf_t = frekuensi term

b. *Information Gain*

Seleksi fitur *information gain* merupakan metode seleksi fitur yang paling sederhana melalui klasifikasi atribut dan banyak digunakan dalam aplikasi klasifikasi teks, analisis data microarray, dan analisis data citra(Syafitri Hidayatul AA, Yuita Arum S 2018).

$$Entropy(S) = \sum_{j=1}^{k} -p_j log_2 p_j \quad \text{.....(1.2)}$$

$$Entropy(S, A) = \sum_{i=1}^{v} (\frac{|S_v|}{|S|} \times Entropy(S_v) \quad \text{.....(1.3)}$$

$$Gain(S, A) = Entropy(S) - \sum values(S, A) \quad \text{.....(1.4)}$$

Keterangan :

S = dataset

k = jumlah *S*

p_j = jumlah sampel untuk kelas *j*
Sv = himpunan bagian dari *S* dimana atribut *A* bernilai sebesar *v*
|*Sv*| = jumlah sampel untuk nilai v
|*S*| = jumlah seluruh sampel data
Entropy(S) = nilai *entropy* dilakukan sebelum pemisahan
Entropy(S,A) = nilai entropi yang dilakukan setelah pemisahan.

5. *Classification*

a. *Support Vector Machine*

Support Vector Machine (SVM) merupakan suatu metode yang termasuk ke dalam supervised learning yang umumnya digunakan untuk klasifikasi. SVM dipopulerkan oleh Vapnik, Boser, dan Guyon pada tahun 1992 berdasarkan prinsip Structural Risk Minimization, bertujuan untuk menemukan hyperplane terbaik yang memisahkan dua buah kelas pada sebuah tempat sehingga terbentuk classifier(Faisal et al. 2020). Persamaan SVM dengan kernel linear dapat dilihat pada persamaan (1.5)(Patle and Chouhan 2013):

$k(X, X_i) = X.X^T$ …….(1.5)

b. *Naïve Bayes*

Naïve Bayes merupakan suatu metode atau mekanisme untuk klasifikasi probabilistik dari teorema Bayesian, dimana teorema tersebut merupakan teori anumerta dari Thomas Bayes. Penggunaan naive bayes classifier memiliki kelebihan yaitu metode algoritma ini hanya membutuhkan sedikit data training untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi data(Kawani 2019). Pada naïve bayes, klasifikasi dihitung dengan P(H|X). Dimana peluang bahwa hipotesa benar untuk data *sample* X dapat diamati dan diterapkan pada persamaan (1.6).

$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$ ……...(1.6)

Keterangan:
X = data sample dengan *class* (label) yang belum diketahui
H = hipotesa bahwa X adalah data pada *class* (label) spesifik
$P(H|X)$ = probabilitas (peluang) hipotesa H berdasarkan kondisi X
$P(H)$ = probabilitas (peluang) dari hipotesa H
$P(X|H)$ = probabilitas X berdasarkan kondisi hipotesis H
$P(X)$ = probabilitas dari data *sample* yang diamati (data X)

6. *Evaluation*

Untuk mengukur performa dari model dapat digunakan 4 metode evaluasi pada conflusion matrix. Metode tersebut yaitu:

a. Accuracy, menggambarkan persentase jumlah record data yang terklasifikasi dengan benar oleh sistem dengan persamaan (1.7).

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ ……(1.7)

b. Precision, menggambarkan presentase akurasi antara data yang diminta dengan hasil klasifikasi dengan persamaan (1.8).

$Precision = \frac{TP}{TP+FP}$ ……(1.8)

c. Recall, menggambarkan keberhasilan model klasifikasi dalam menentukan kembali informasi atau hasil nilai dengan menggunakan persamaan (1.9).

$Recall = \frac{TP}{TP+FN}$ …….(1.9)

d. F1-Score, menggambarkan suatu perbandingan nilai rata-rata dari precision dan recall dengan menggunakan persamaan (1.10).

$F1 - Score = \frac{2\ x\ Recall\ x\ Precision}{Recall+Precision}$ ……...(1.10)

RESULTS & DISCUSSION

Results

Data diperoleh dari *twitter* menggunakan *platform* PhantomBuster dengan mengambil *tweet* pada tanggal 30 Maret 2020 memakai kata kunci 'pemerintah', 'lockdown', 'karantina' dan 'covid'. Sehingga didapatkan sebanyak 3228 data, namun setelah dilakukan pengecekan manual terdapat beberapa data yang tidak sesuai tema dan ada data yang serupa, maka total data yang digunakan pada tahap ini terdapat sebanyak 1560 data. *Sample data* dapat dilihat pada tabel 1.

Tabel 1 Data Collection

Username	***Date***	***Tweet***
aiziz_m	2020-03-30 15:19:03	Bukti pemerintah tidak mampu mensejahterakan warganya #TolakDaruratSipil
garislurus0	2020-03-30 20:15:53	@onekey57467761 @saincis darurat sipil saat itu pasca dom karena adanya separatisme ya berbeda bos berbeda saat darurat sipil ketika ada bencana dan darurat sipil tetap dipimpin menteri dan gubernur @onekey57467761 tumben nih eks cebong pd nentang pemerintah
partner_pnk	2020-03-30 17:47:56	Mari Kita Dukung Langkah Pemerintah dalam mengatasi wabah Corona, hilangkan perbedaan politik, agama, suku dan sebagainya, mari kita bersatu bahu membahu saling mendukung agar wabah ini bisa cepat hilang dari negara Indonesia. Selamatkan Indonesia Your Are Not Alone!!!

Pada tahap pelabelan menggunakan lexicon based dengan Vader Sentiment data mulai diklasifikasikan, apabila data mendapat compound score < 0 maka data tersebut bernilai 'negatif' dan apabila menghasilkan nilai > 0 maka data tersebut bernilai 'positif'. Sehingga didapatkan 1493 data dengan label negatif dan positif. Pelabelan dapat dilihat pada tabel 2.

Tabel 2 *Labelling*

Content	***Score***	***Label***
Bukti pemerintah tidak mampu mensejahterakan warganya #TolakDaruratSipil	-0,765	Negatif
@onekey57467761 @saincis darurat sipil saat itu pasca dom karena adanya separatisme ya berbeda bos berbeda saat darurat sipil ketika ada bencana dan darurat sipil tetap dipimpin menteri dan gubernur @onekey57467761 tumben nih eks cebong pd nentang pemerintah	-0,7717	Negatif

Mari Kita Dukung Langkah Pemerintah dalam mengatasi wabah Corona, hilangkan perbedaan politik, agama, suku dan sebagainya, mari kita bersatu bahu membahu saling mendukung agar wabah ini bisa cepat hilang dari negara Indonesia. Selamatkan Indonesia Your Are Not Alone!!!	0,9405	Positif

Selanjutnya dilakukan tahap *pre-processing* dengan menggunakan *case folding, cleaning, tokenization, stopword removal, stemming dan remove duplicate*. Hasil dari tahap preprocessing dapat dilihat pada tabel 3.

Tabel 3 *Pre-Processing*

Content	***Case folding***	***Cleaning***	***Tokenization***	***Stopword Removal***	***Stemming***
Bukti pemerintah tidak mampu mensejahterakan warganya #TolakDaruratSipil	bukti pemerintah tidak mampu mensejahterakan warganya #tolakdaruratsipil	bukti pemerintah tidak mampu mensejahterakan warganya tolakdaruratsipil	['bukti', 'pemerintah', 'tidak', 'mampu', 'mensejahterakan', 'warganya', 'tolakdaruratsipil ']	['bukti', 'pemerintah', 'tidak', 'mensejahterakan', 'warganya', 'tolakdaruratsipil ']	bukti pemerintah tidak sejahtera warga tolak darurat sipil

Setelah *pre-processing* selesai dilakukan, tahap berikutnya yaitu dengan mentransformasikan teks yang telah didapat menggunakan ekstraksi fitur TF-IDF agar menjadi representasi numerik. Data yang digunakan terdapat pada tabel 4.

Tabel 4 Data yang digunakan

Tweet	**Ket**
bagus pemerintah pusat realokasi apbn	Train positif 1
nampaknya pemerintah bantu masyarakat korban	Train positif 2
setuju percaya pemerintah alias negara	Train positif 3
bukti pemerintah tidak sejahtera warga tolak darurat sipil	Train negatif 1
pemerintah darurat sipil tolak darurat sipil	Train negatif 2
pemerintah pelit rakyat	Train negatif 3
ayo bantu pemerintah atas covid selesai	Test positif
pemerintah egois	Test negatif

Selanjutnya data yang sudah bersih dapat melalui tahap berikutnya yaitu TF-IDF. Perhitungan TF-IDF dilakukan dengan mencari nilai TF (*term frequency*) lalu IDF yang selanjutnya nilai tersebut dikalikan, perhitungan nilai TF dapat ditunjukkan pada Tabel 5.

Tabel 5 Perhitungan Term Frequency

Term	TF								DF
	Train P 1	Train P 2	Train P 3	Train N 1	Train N 2	Train N 3	Test P 1	Test N 1	
bagus	1	0	0	0	0	0	0	0	1
pemerintah	1	1	1	1	1	1	1	1	8
pusat	1	0	0	0	0	0	0	0	1
realokasi	1	0	0	0	0	0	0	0	1

apbn	1	0	0	0	0	0	0	0	1
nampaknya	0	1	0	0	0	0	0	0	1
bantu	0	1	0	0	0	0	1	0	2
masyarakat	0	1	0	0	0	0	0	0	1
korban	0	1	0	0	0	0	0	0	1
setuju	0	0	1	0	0	0	0	0	1
percaya	0	0	1	0	0	0	0	0	1
alias	0	0	1	0	0	0	0	0	1
negara	0	0	1	0	0	0	0	0	1
bukti	0	0	0	1	0	0	0	0	1
tidak	0	0	0	1	0	0	0	0	1
sejahtera	0	0	0	1	0	0	0	0	1
warga	0	0	0	1	0	0	0	0	1
tolak	0	0	0	1	1	0	0	0	2
darurat	0	0	0	1	2	0	0	0	2
sipil	0	0	0	1	2	0	0	0	2
pelit	0	0	0	0	0	1	0	0	1
rakyat	0	0	0	0	0	1	0	0	1
ayo	0	0	0	0	0	0	1	0	1
atas	0	0	0	0	0	0	1	0	1
covid	0	0	0	0	0	0	1	0	1
selesai	0	0	0	0	0	0	1	0	1
egois	0	0	0	0	0	0	0	1	1

Langkah kedua yaitu dengan mencari nilai IDF (*Invers Document Frequency*) lalu mengalikan nilai TF dengan IDF. Perhitungan TF-IDF dapat ditunjukkan pada Tabel 6.

Tabel 6 Perhitungan Term Frequency - Invers Document Frequency

DF	IDF	TF-IDF							
		Train P 1	Train P 2	Train P 3	Train N 1	Train N 2	Train N 3	Test P 1	Test N 1
1	0,90309	0,90309	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
1	0,90309	0,90309	0	0	0	0	0	0	0
1	0,90309	0,90309	0	0	0	0	0	0	0
1	0,90309	0,90309	0	0	0	0	0	0	0
1	0,90309	0	0,90309	0	0	0	0	0	0
2	0,60206	0	0,60206	0	0	0	0	0,60206	0
1	0,90309	0	0,90309	0	0	0	0	0	0
1	0,90309	0	0,90309	0	0	0	0	0	0
1	0,90309	0	0	0,90309	0	0	0	0	0
1	0,90309	0	0	0,90309	0	0	0	0	0
1	0,90309	0	0	0,90309	0	0	0	0	0
1	0,90309	0	0	0,90309	0	0	0	0	0
1	0,90309	0	0	0	0,90309	0	0	0	0

1	0,90309	0	0	0	0,90309	0	0	0	0
1	0,90309	0	0	0	0,90309	0	0	0	0
1	0,90309	0	0	0	0,90309	0	0	0	0
2	0,60206	0	0	0	0,60206	0,60206	0	0	0
2	0,60206	0	0	0	0,60206	1,20412	0	0	0
2	0,60206	0	0	0	0,60206	1,20412	0	0	0
1	0,90309	0	0	0	0	0	0,90309	0	0
1	0,90309	0	0	0	0	0	0,90309	0	0
1	0,90309	0	0	0	0	0	0	0,90309	0
1	0,90309	0	0	0	0	0	0	0,90309	0
1	0,90309	0	0	0	0	0	0	0,90309	0
1	0,90309	0	0	0	0	0	0	0,90309	0
1	0,90309	0	0	0	0	0	0	0	0,90309

Setelah dilakukan pembobotan kata, tahap berikutnya yaitu dengan memasukkan data tersebut pada ekstraksi fitur information gain. Misal contoh data *sample* yang digunakan memiliki total data sebanyak 15 buah dimana sebanyak 10 dokumen dengan label positif dan 5 dokumen dengan label negatif terdapat kata 'dukung' sebagai fitur yang akan diproses. Setelah dilihat, fitur 'dukung' pada dataset terdapat 12 dokumen dimana sebanyak 9 dokumen berlabel positif dan 3 memiliki label negatif. Dari data di atas, maka perhitungan selanjutnya yaitu:

a. Total dokumen dan sentimen probabilitas:

- Jumlah total dokumen dalam dataset = 15
- Probabilitas sentimen positif = 10/15 = 0,67
- Probabilitas sentimen negatif = 5/15) = 0,33

b. *Conditional Probabilities*:

- Probabilitas fitur "dukung" muncul dalam dokumen dengan sentimen positif = 9/12 = 0,75
- Probabilitas fitur "dukung" tidak muncul dalam dokumen dengan sentimen positif = 1 – 0,75 = 0,25
- Probabilitas fitur "dukung" muncul dalam dokumen dengan sentimen negatif = 3/12 = 0,25
- Probabilitas fitur "dukung" tidak muncul dalam dokumen dengan sentimen negatif = 1 – 0,25 = 0,75

c. Entropy Sentimen Awal (Sebelum Pemisahan)

- Entropy sentimen awal = – (probabilitas sentimen positif * log2(probabilitas sentimen positif) + probabilitas sentimen negatif * log2(probabilitas sentimen negatif))
- Entropy sentimen awal = – (0,67 * log2(0,67) + 0,33 * log2(0,33))
  = – (0,67 * –0.58 + 0,33 * –1,59)
  = – (–0,3886 + –0,5247) = 0,9133

d. Entropy Sentimen Setelah Pemisahan Fitur "dukung"

- Entropy Sentimen Setelah Pemisahan Fitur "dukung" = –(probabilitas fitur "dukung" muncul dalam dokumen positif * log2(probabilitas fitur "baik" muncul dalam dokumen positif) + probabilitas fitur "dukung" tidak muncul dalam dokumen positif * log2(probabilitas fitur "dukung" tidak muncul dalam

dokumen positif) + probabilitas fitur "dukung" muncul dalam dokumen negatif * log2(probabilitas fitur "dukung" muncul dalam dokumen negatif) + probabilitas fitur "dukung" tidak muncul dalam dokumen negatif * log2(probabilitas fitur "dukung" tidak muncul dalam dokumen negatif))

Entropy sentimen setelah pemisahan dengan fitur "dukung"

= –(0,75* log2(0,75) + 0,25* log2(0,25) + 0,25* log2(0,25) + 0,75* log2(0,75))

= – (0,75* –0,42 + 0,25* –2 + 0,25* –2 + 0,75 * –0,42)

= – (–0,315 – 0,5 – 0,5 – 0,315) = – (–1,63)

= 1,63

e. Information Gain

- Information Gain untuk fitur "dukung" = Entropy sentimen awal – Entropy sentimen setelah pemisahan dengan fitur "dukung"
- Information Gain untuk fitur "dukung" = 0,9133 – 1,63 = –0,7167

Karena pada contoh hasil dari perhitungan *information gain* menghasilkan angka yang negatif, maka dapat disimpulkan bahwa fitur "dukung" tidak memberikan kontribusi yang signifikan dalam memisahkan dokumen berdasarkan sentimen. Sehingga dalam kasus ini, mungkin bukan fitur yang baik untuk memprediksi sentimen dalam analisis sentimen ini.

Hasil dari data TF-IDF serta perhitungan TF-IDF dan *information gain* dimasukkan ke dalam klasifikasi svm dan naïve bayes. Pada tahap ini, dengan mengasumsikan terdapat 3 kata unik dalam masing-masing data dengan nilai numerik merupakan nilai acak yang bukan sesungguhnya. Dilakukan pengklasifikasisan teks menggunakan algoritma SVM menggunakan kernel linear. Untuk contoh data yang digunakan yaitu dapat dilihat pada Tabel 7 dan Tabel 8.

Tabel 7 Tabel Sampel Data Latih Perhitungan SVM

Data Latih	Representasi Numerik
setuju percaya pemerintah alias negara	[0,234; 0,091; 0,423]
pemerintah pelit rakyat	[0,132; 0,211; 0,122]

Tabel 8 Tabel Sampel Data Uji Perhitungan SVM

Data Uji	Representasi Numerik
mari bantu pemerintah atas covid telah beres	[0,322; 0,111; 0,122]
pemerintah egois	[0,212; 0,123; 0,201]

Dimana persamaan rumus yang digunakan untuk kernel linear yaitu menggunakan persamaan (1.5). Berikut contoh perhitungannya.

$k(data_{latih_1}, data_{uji_1}) = [0{,}234;\ 0{,}091;\ 0{,}423].\,[0{,}322;\ 0{,}111;\ 0{,}122]$

$[(0{,}234 \times 0{,}322) + (\ 0{,}091 \times 0{,}111) + (0{,}423 \times 0{,}122)]$

$0{,}075 + 0{,}0101 + 0{,}051 = 0{,}1361$

$k(data_{latih_1}, data_{uji_2}) = [0{,}234;\ 0{,}091;\ 0{,}423].\,[0{,}212;\ 0{,}123;\ 0{,}201]$

$[(0{,}234 \times 0{,}212) + (\ 0{,}091 \times 0{,}123) + (0{,}423 \times 0{,}201)]$

$0{,}049 + 0{,}011 + 0{,}085 = 0{,}145$

$k(data_{latih_2}, data_{uji_1}) = [0{,}132;\ 0{,}211;\ 0{,}122].\,[0{,}322;\ 0{,}111;\ 0{,}122]$

$[(0{,}132 \times 0{,}322) + (\ 0{,}211 \times 0{,}111) + (0{,}122 \times 0{,}122)]$

$0{,}043 + 0{,}023 + 0{,}015 = 0{,}081$

$k(data_{latih_2}, data_{uji_2}) = [0{,}132;\ 0{,}211;\ 0{,}122].\,[0{,}212;\ 0{,}123;\ 0{,}201]]$

$[(0{,}132 \times 0{,}212) + (\ 0{,}211 \times 0{,}123) + (0{,}122 \times 0{,}201)]$

$0{,}028 + 0{,}026 + 0{,}025 = 0{,}079$

Hasil dari perhitungan manual algoritma SVM dengan data *sample* 2 data latih dan 2 data uji menghasilkan nilai 0,1361 untuk kombinasi data latih 1 dan data uji, nilai 0,145 dari kombinasi data latih 1 dan data uji 2, nilai 0,081 dari kombinasi data latih 2 dan data uji 1, dan menghasilkan nilai 0,079 dari kombinasi perhitungan data latih 2 dan data uji 2.

Setelah itu dilakukan perhitungan Naïve Bayes sebagai berikut. Misalkan terdapat data latih dengan 100 dokumen, di mana 60 dokumen memiliki sentimen positif dan 40 dokumen memiliki sentimen negatif.

a. Probabilitas Kelas:

- P (Positif) = 60 / 100 = 0,6
- N (Negatif) = 40 / 100 = 0,4

b. Probabilitas Kata dalam Kelas

Misalkan kita memiliki dua kata unik dalam semua dokumen, yaitu "bantu" dan "pelit". Berikut adalah contoh frekuensi kemunculan kata-kata dalam dokumen dengan sentimen positif dan negatif:

- freqPositif("bantu") = 45
- freqNegatif("bantu ") = 15
- freqPositif("pelit ") = 10
- freqNegatif("pelit") = 30

Jumlah kata unik dalam semua dokumen adalah 2.

- totalKataPositif = 2 (karena hanya ada dua kata unik)
- totalKataNegatif = 2 (karena hanya ada dua kata unik)
- totalKata = 2 (karena hanya ada dua kata unik)

Berdasarkan rumus, kita dapat menghitung probabilitas kata dalam kelas:

- P("bantu "|Positif) = (45 + 1) / (2 + 2) = 46 / 4 = 11,5
- P("bantu "|Negatif) = (15 + 1) / (2 + 2) = 16 / 4 = 4
- P("pelit "|Positif) = (10 + 1) / (2 + 2) = 11 / 4 = 2,75
- P("pelit "|Negatif) = (30 + 1) / (2 + 2) = 31 / 4 = 7,75

c. Probabilitas Posterior:

Misalkan kita memiliki sebuah dokumen uji yang berisi kata "bantu" dan "pelit". Kita dapat menghitung probabilitas posterior untuk sentimen positif (Positif|Dokumen) dan sentimen negatif (Negatif|Dokumen).

- P(Positif|Dokumen) = P(Positif) * P("bantu"|Positif) * P("pelit"|Positif)
  = 0,6 * 11,5 * 2,75 = 18,975
- P(Negatif|Dokumen) = P(Negatif) * P("bantu"|Negatif) * P("pelit"|Negatif)
  = 0,4 * 4 * 7,75 = 12,4

Untuk mengetahui perhitungan keseluruhan, digunakan pemodelan dalam bentuk *python* yang menghasilkan nilai akurasi yang dapat ditunjukkan pada Tabel 9.

Tabel 9 Pengujian 70% data latih dan 30% data uji

70:30			
SVM	**NB**	**SVM+IG**	**NB+IG**

Accuracy: 64,7% *Precision*: 63,5% *Recall*: 64,7% *F1-Score*: 64,0%	*Accuracy*: 64,5% *Precision*: 64,5% *Recall*: 60,0% *F1-Score*: 62,2%	K = 1814 *Accuracy*: 66,7% *Precision*: 65,5% *Recall*: 66,9% *F1-Score*: 66,2%	K = 1814 *Accuracy*: 65,2% *Precision*: 67,2% *Recall*: 55,5% *F1-Score*: 60,8%

Perhitungan *accuracy*, *precision*, *recall* dan *f1-score* dapat dilihat sebagai berikut.

a. *Accuracy*

$$Accuracy = \frac{(146+153)}{(146+153+77+72)} \times 100\% = 0{,}66741 \times 100\%$$

$$Accuracy = 66{,}7\%$$

b. *Precision*

$$Precision = \frac{146}{146+77} \times 100\% = 0{,}65470 \times 100\%$$

$$Precision = 65{,}5\%$$

c. *Recall*

$$Recall = \frac{146}{146+72} \times 100\% = 0{,}66972 \times 100\%$$

$$Recall = 66{,}9\%$$

d. *F1-Score (F-Measurement)*

$$F1 - Score = \frac{2(0{,}65470 \times 0{,}66972)}{0{,}65470+0{,}66972} \times 100\% = 0{,}6621 \times 100\%$$

$$F1 - Score = 66{,}2\%$$

Sedangkan hasil pengujian menggunakan dataset lain diapatkan hasil seperti pada Tabel 10.

Tabel 10 Pengujian Dataset Lain

70:30				
Dataset	**SVM**	**NB**	**SVM+IG**	**NB+IG**
A (995 data) 1714 fitur	*Accuracy*: 78,6% *Precision*: 81,0% *Recall*: 74,5% *F1-Score*: 77,6%	*Accuracy*: 78,6% *Precision*: 77,7% *Recall*: 79,9% *F1-Score*: 78,8%	*Accuracy*: 75,9% *Precision*: 77,7% *Recall*: 72,5% *F1-Score*: 75,0%	*Accuracy*: 78,9% *Precision*: 77,9% *Recall*: 80,5% *F1-Score*: 79,2%
B (923 data) 2340 fitur	*Accuracy*: 89,2% *Precision*: 80,0% *Recall*: 59,3% *F1-Score*: 68,0%	*Accuracy*: 83,0% *Precision*: 54,9% *Recall*: 72,2% *F1-Score*: 62,4%	*Accuracy*: 89,2% *Precision*: 80,0% *Recall*: 59,3% *F1-Score*: 68,0%	*Accuracy*: 83,8% *Precision*: 56,7% *Recall*: 70,4% *F1-Score*: 62,8%

Hasil menunjukkan bahwa penggunaan seleksi fitur *information gain* pada dataset yang digunakan pada penelitian ini meningkatkan sedikit akurasi, untuk dataset lainnya mendapatkan hasil yang berbeda-beda. Ini dapat disebabkan oleh data yang digunakan memiliki jumlah dan fitur data yang bervariasi, sehingga penggunaan fitur mempengaruhi hasil. Selain itu juga dapat disebabkan oleh fitur yang masih memiliki makna sama dengan penulisan yang berbeda atau fitur yang tidak memiliki makna masih ada di dalam data tersebut.

CONCLUSION

Hasil pengujian menunjukkan nilai akurasi yang paling tinggi didapat dari pembagian data sebesar 70:30, akurasi yang didapatkan sebesar 66.7%, presisi 65.5%, *recall* 66.9% dan *f1-score* 66,2%. Hasil tersebut masih perlu ditingkatkan lagi karena algoritma svm dan naïve bayes terkenal merupakan mesin pembelajaran yang terkenal baik dalam mengklasifikasikan sentiment. Ini dapat disebabkan oleh pelabelan data oleh *lexicon based* yang kebanyakan tidak sesuai antara label positif dan negatif. Bisa dilihat pada percobaan menggunakan dataset lain mendapatkan hasil yang lebih besar sehingga dapat dipastikan bahwa pelabelan menggunakan *lexicon based* untuk dataset penanganan covid-19 di Indonesia hasilnya kurang sesuai. Sehingga pada penelitian berikutnya dapat dilanjutkan dengan mencari cara pelabelan terbaik menggunakan kamus bahasa indonesia dan membersihkan kalimat yang mengandung bahasa daerah untuk diubah terlebih dahulu agar memiliki makna yang sama.

ACKNOWLEDGEMENT

Peneliti sangat bersyukur kepada Allah SWT karena berkat rahmat, hidayah dan karunia-Nya yang sudah diberikan. Serta peneliti mengucapkan terima kasih kepada dosen pembimbing, penguji dan juga seluruh dosen dan staff di jurusan Informatika Universitas Jenderal Achmad Yani atas seluruh ilmu yang telah diberikan kepada peneliti. Tidak lupa juga peneliti mengucapkan terima kasih yang sebesar-besarnya kepada keluarga, sahabat, hingga kerabat yang sudah mendoakan dan selalu memberi semangat dalam bentuk apapun selama penulisan penelitian ini. Semua bentuk bantuan dan dukungan yang sudah diberikan kepada peneliti semoga mendapat balasan yang insyaallah baik dari Allah SWT. Aamiin.

REFERENCES

Ahmad Wildan Attabi, Lailil Muflikhah, and Mochammad Ali Fauzi. 2018. "Penerapan Analisis Sentimen Untuk Menilai Suatu Produk Pada Twitter Berbahasa Indonesia Dengan Metode Naïve Bayes Classifier Dan Information Gain." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 2(11): 4548–54.

Akbari, M Indra Halim Arsya Dwi, Astri Novianty, and Setianingsih Casi. 2017. "Analisis Sentimen Menggunakan Metode Learning Vector Quantization Sentiment Analysis Using Learning Vector Quantization Method." *e-Proceeding of Engineering* 4(2): 2283–92. https://openlibrary.telkomuniversity.ac.id/pustaka/files/135356/jurnal_eproc/analisis-sentimen-menggunakan-metode-learning-vector-quantization.pdf.

Deller, Ruth. 2011. "Twittering on: Audience Research and Participation Using Twitter." *Participations: Journal of Audience & Reception Studies* 8(1): 216–45. http://www.participations.org/Volume 8/Issue 1/deller.htm.

Emeraldien, Fikry Zahria, Rifan Jefri Sunarsono, and Ronggo Alit. 2019. "TWITTER SEBAGAI PLATFORM KOMUNIKASI POLITIK DI." XIV.

Faisal, Anas, Yuris Alkhalifi, Achmad Rifai, and Windu Gata. 2020. "Analisis Sentimen Dewan Perwakilan Rakyat Dengan Algoritma Klasifikasi Berbasis Particle Swarm Optimization." *JOINTECS (Journal of*

Information Technology and Computer Science) 5(2): 61.

Hilman, Muhammad, Aprilian Nurjaman, and Mohamad Syahrul Mubarok. 2017. "Analisis Sentimen Pada Ulasan Buku Berbahasa Inggris Menggunakan Information Gain Dan Support Vector Machine." *e-Proceeding of Engineering : Vol.4, No.3 Desember 2017* 4(3): 4900–4906.

Kawani, Gigih Putra. 2019. "Implementasi Naive Bayes." *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)* 1(2): 73–81.

Negara, Arif Bijaksana Putra, Hafiz Muhardi, and Indira Melinda Putri. 2020. "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes Dan Seleksi Fitur Information Gain." *Jurnal Teknologi Informasi dan Ilmu Komputer* 7(3): 599.

Patle, Arti, and Deepak Singh Chouhan. 2013. "SVM Kernel Functions for Classification." *2013 International Conference on Advances in Technology and Engineering, ICATE 2013*.

RI, Sekretariat Jenderal DPR. 2020. "BAGIAN PERSIDANGAN PARIPURNA." *dpr.go.id*. https://www.dpr.go.id/setjen/index/id/Agenda-BAGIAN-PERSIDANGAN-PARIPURNA (February 10, 2023).

Syafitri Hidayatul AA, Yuita Arum S, Achmad Arwan. 2018. "Seleksi Fitur Information Gain Untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor Dan Naïve Bayes." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 2(9): 2546–54. http://j-ptiik.ub.ac.id.