Journal of Applied Informatics and Computing (JAIC)
Vol.7, No.2, December 2023, pp. 224~230
e-ISSN: 2548-6861

Implementation of Information Gain for Sentiment Analysis of PSE Policy using Naïve Bayes Algorithm

Stevanus Ertito Pramudja [1]*, Yuyun Umaidah, [2], Aries Suharso [3]***
* Informatika, Universitas Singaperbangsa Karawang
** Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang
11910631170139@student.unsika.ac.id[1], yuyun.umaidah@staff.unsika.ac.id[2], aries.suharso@unsika.ac.id[3]

Article Info

Article history:

Received 2023-08-21
Revised 2023-09-11
Accepted 2023-10-03

Keyword:

PSE Kominfo,
Naïve Bayes Classifier,
Information Gain,
Confusion Matrix.

ABSTRACT

The Ministry of Communication and Information Technology of Indonesia (Kominfo) has established the Penyelenggara Sistem Elektronik (PSE) policy as a mandatory registration requirement for both domestic and foreign Electronic Systems (ES). As a result, Kominfo will impose sanctions on all ES by temporarily suspending their access if they fail to register by July 29, 2022, at 23:59 WIB. This policy has sparked both support and opposition among the Indonesian public, and it has become a topic of discussion, including among Twitter users. Therefore, sentiment analysis is employed as a solution to identify public concerns or issues regarding the policy based on negative and positive tweets. The objective of this research is to evaluate the results of feature selection using Information Gain and the Naïve Bayes Classifier algorithm in analyzing Twitter users' sentiment towards the policies of the Information and PSE of the Ministry of Communication and Information Technology. A total of 1153 lines of tweets were collected from the Twitter platform using the keyword "PSE Kominfo," which were then analyzed using the Naïve Bayes Classifier algorithm and Information Gain feature selection with three scenarios: 90:10, 80:20, and 70:30. Based on the evaluation using the confusion matrix, overall, Scenario 1 with a 90:10 ratio and Information Gain feature selection performed the best, achieving an accuracy of 79.7%, recall of 85%, and an F-1 score of 88%. However, the best precision was observed in Scenario 2 with an 80:20 ratio, reaching 92% due to the higher proportion of positive predictions made by the model compared to other scenarios.

I. PENDAHULUAN

Kementerian Komunikasi dan Informatika RI telah menetapkan kebijakan PSE (Penyelenggara Sistem Elektronik) sebagai kewajiban pendaftaran SE (Sistem Elektronik) yang terdiri dari SE domestik ataupun SE asing. Dilansir pada kominfo.go.id, Menurut Peraturan Pemerintah Nomor 71 tahun 2019, dapat diketahui bahwa PSE adalah setiap orang, penyelenggara negara, badan usaha, dan masyarakat yang menyediakan, mengelola, dan mengoperasikan sistem elektronik secara sendiri maupun bersama-sama kepada pengguna sistem elektronik untuk keperluan dirinya atau keperluan pihak lainnya dan berdasarkan peraturan Menteri Komunikasi dan Informatika No 5 tahun 2020 mengenai PSE Lingkup Privat[1]. Kementerian Komunikasi serta Informatika akan memberi sanksi terhadap seluruh SE berbentuk pemutusan akses sementara jika tidak melaksanakan pendaftaran hingga 29 Juli 2022 Jam 23.59 WIB. PSE sendiri bertujuan untuk memberikan, mengurus, dan menjalankan sistem elektronik baik secara individu maupun kolaboratif kepada pengguna sistem elektronik, baik untuk keperluan pribadi maupun keperluan organisasi atau kelompok [2]. Namun kebijakan tersebut telah menuai pro dan kontra terhadap masyarakat Indonesia, terutama kepada netizen Indonesia di media social. Masyarakat yang pro terhadap kebijakan PSE dari berbagai sektor, menilai bahwa hal tersebut sebagai Regulasi

Ekosistem dan perlindungan data, namun masyarakat yang kontra terhadap kebijakan tersebut juga cukup banyak, kritik yang ditujukan kepada kebijakan tersebut dinilai sewenang – wenangnya dalam memblokir platform seperti Paypal, Steam, Yahoo!, dan sebagainya [3] sehingga didapatkan bahwa tujuan dari penelitian ini untuk menyediakan menyediakan wawasan dan pandangan masyarakat terhadap kebijakan PSE Kominfo dari berbagai sudut pandang, baik yang bersifat positif maupun negatif juga untuk mengevaluasi hasil seleksi fitur Information Gain serta algoritma Naïve Bayes Classifier dalam menganilisis sentimen pengguna twitter terhadap kebijakan Penyelenggara Sistem Elektronik Kominfo. Twitter adalah layanan jaringan mikroblogging yang dimulai pada tahun 2006 di mana pengguna dapat berbagi pesan teks, yang disebut sebagai tweet, dan tautan ke konten lain seperti gambar, situs web, dan artikel. Setiap tweet memiliki panjang maksimal 140 karakter, yang menggambarkan suatu acara atau pendapat orang tentang suatu acara atau seseorang, dan dapat memiliki tautan ke artikel berita, video, atau gambar [4].

Analisis Sentimen atau Opinion Mining yaitu langkah ekstraksi informasi dari data berbentuk teks, bertujuan untuk memperoleh wawasan tentang kecenderungan penilaian terhadap suatu objek yang sedang diselidiki. Pandangan yang diungkapkan oleh masyarakat umum dalam hal sentimen ini dapat menjadi pedoman dalam memutuskan tentang suatu produk [5]. Analisis sentimen sudah sering sekali dipakai sebagai penelitian memakai macam-macam algoritma machine learning semisal penelitian analisis sentimen pengguna e-wallet memakai Naïve Bayes Classifier serta seleksi fitur Information Gain, pada penelitian itu dihasilkan evaluasi serta performa tahap data mining memakai Naïve Bayes Classifier memiliki akurasi 84% serta akurasi meningkatkan menggunakan seleksi fitur Information Gain yaitu mencapai 92%[6] dan seperti penelitian Bijaksana A, 2020[7] dihasilkan bahwa evaluasi serta performa tahap data mining memakai Naïve Bayes Classifier juga meningkat setelah melakukan seleksi fitur Information Gain yaitu dari 77% menjadi 83%.

Berdasarkan paparan diatas, tujuan dari penelitian ini adalah untuk mengevaluasi hasil seleksi fitur menggunakan Information Gain dan algoritma Klasifikasi Naïve Bayes dalam menganalisis sentimen pengguna Twitter terhadap kebijakan Penyelenggara Sistem Elektronik dari Kementerian Komunikasi dan Informatika (Kominfo).

II. METODE PENELITIAN

Metodologi penelitian yang diterapkan pada penelitian ini adalah Knowledge Discovery in Database (KDD) dengan proses tahapan yaitu Data Selection, Preprocessing, Transformation, Data Mining, dan Evaluation. KDD adalah bidang yang terus berkembang yang menggabungkan teknik database, statistika, dan kecerdasan buatan untuk mengekstraksi pengetahuan (data tingkat tinggi) dari informasi besar (data tingkat rendah)[8], dapat dilihat pada gambar 1 alur penelitian dengan Knowledge Discovery in Database (KDD).

Gambar 1. Alur Penelitian dengan Metode Knowledge Discovery in Database (KDD)

A. Data Selection

Proses pengumpulan data menggunakan teknik crawling pada media sosial twitter merupakan tahap awal dari metode yang digunakan. Dataset yang diperoleh memanfaatkan API Twitter dengan opini publik berbahasa Indonesia terkait kebijakan "PSE Kominfo" dengan rentang waktu 29 Juli 2022 hingga 30 Juli 2022 dengan jumlah 1153 rows yang berupa teks kemudian diklasifikasikan menjadi label positif, negatif, dan netral dengan bantuan ahli atau pakar bahasa Indonesia sebagai validasi pelabelan data kemudian data berubah menjadi 683 rows setelah data validasi label netral di drop.

B. Preprocessing Data

Tahap selanjutnya adalah preprocessing yang merupakan pengolahan data. Pada tahap ini, peneliti memproses dataset mentah atau kotor hasil crawling untuk menghilangkan emoji/emoticon, url, hastag ataupun simbol-simbol yang mengganggu dalam klasifikasi maupun teks yang tidak relevan dengan penelitian. Berikut tahapan preprocessing data[9] :

1) Cleansing

Cleansing adalah tahapan pertama yaitu dengan membersihkan dataset yang digunakan seperti atribut atau teks yang berpengaruh pada hasil klasifikasi antaranya penggunaan emoji/emoticon, url, hastag ataupun simbol-simbol yang tidak diperlukan.

2) Case Folding

Case Folding adalah proses perubahan teks atau kalimat dari huruf kapital menjadi huruf kecil (lowercase) yang bertujuan untuk mengurangi redudansi pada dataset.

3) Tokenizing

Tokenizing merupakan langkah pemecahan kata per kata pada suatu teks atau kalimat. Kata tersebut digunakan untuk keperluan pembobotan dan dipisah dalam satu spasi.

4) Normalization

Normalization adalah tahapan untuk merubah atau memperbaiki kata-kata yang tidak baku menjadi kata baku.

5) Stopword Removal

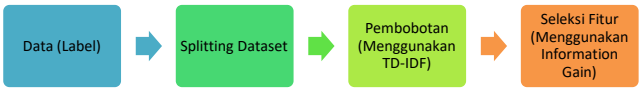
Penggunaan Stopward Removal biasanya untuk menghilangkan kata pada teks yang tidak relevan dengan penelitian Dalam tahapan ini, kata-kata yang tidak memiliki kepentingan dalam pemrosesan data akan dihapus.

6) Stemming

Stemming merupakan proses mengubah kata yang memiliki imbuhan menjadi bentuk kata dasarnya.

C. Transformation

Setelah melakukan preprocessing data masih berupa data nominal sehingga kita perlu melakukan transformation untuk mengubah kedalam data numerikal. Selanjutnya data tersebut dilakukan splitting dataset dengan 3 skenario yaitu 70:30, 80:20, dan 90:10, setelah itu dilakukan pembobotan menggunakan TF-IDF, pentingnya suatu fitur kemudian ditentukan bersama oleh hasil perkalian dari nilai TF dan IDF-nya pada setiap katanya[10] setelah itu melakukan seleksi fitur menggunakan information gain guna mengidentifikasi fitur yang relevan atau penting. Setelah itu dilakukan pembobotan menggunakan TF-IDF pada setiap katanya setelah itu melakukan seleksi fitur menggunakan information gain guna mengidentifikasi fitur yang relevan atau penting. Alur transformasi data dapat dilihat pada gambar 2.



Gambar 2. Alur Transformation Data

D. Data Mining

Data mining merupakan proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi[11]. Pada penelitian ini proses data mining menggunakan algoritma Naïve Bayes Classifier.

E. Evaluation

Evaluation adalah evaluasi dari performa hasil pengolahan data mining, dari hasil yang diperoleh dapat disimpulkan valid atau tidaknya dengan hipotesa yang dibuat. Pengukuran validitas hasil akan menggunakan confusion matrix dengan nilai accuracy, precision, dan recall.

III. HASIL DAN PEMBAHASAN

Hasil penelitian yang dilakukan adalah dengan menerapkan metodologi dari *Knowledge Discovery in Database (KDD)* dalam proses implementasi *information gain* untuk analisis sentimen kebijakan PSE dengan algoritma *Naïve Bayes*.

A. Data Selection

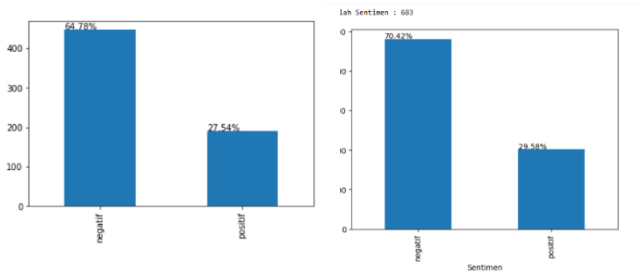
Rentang waktu 29 Juli 2022 hingga 30 Juli 2022 menggunakan sebanyak 1153 *rows* hasil *crawling* yang berupa teks kemudian diklasifikasikan menjadi label positif, negatif, dan netral secara manual kemudian divalidasi oleh ahli atau pakar bahasa Indonesia kemudian data berubah menjadi 683 *rows* setelah data validasi label netral di *drop*. Data tersebut diambil menggunakan *Twitter API Key V2* yang sebelumnya wajib registrasi pada website *https://dev.twitter.com*[12], setelah *API Key* didapatkan menggunakan *Bearer Token* maka *database twitter* dapat diakses menggunakan bahasa pemrograman *python* serta *package tweepy*. Gambar 3 merupakan hasil *data crawling* dan pelabelan data secara manual termasuk validasi oleh ahli bahasa.

	Username	Text	Validasi	Sentiments
0	tsetiady	kominfo ini bisa jadi lembaga paling hipokrit ...	Negatif	Negatif
1	pepo_infp	@tilehopper Kalau sistem registrasi PSE tdk di...	Negatif	Negatif
2	21BeritaTerkini	Situs dan Aplikasi yang Diblokir Kominfo karen...	Netral	Negatif
3	aagunawan_	UU ITE ⁄ kominfo\nPSE Ranah Privat ⁄ kominfo\n...	Negatif	Netral
4	anggreannie	@pancasyah Salfok sama yg di iklan, situs tida...	Negatif	Positif
...	...	...	...	...
1148	kelas_teknisi	Lewat Aturan PSE, Kominfo Bisa Intip Percakapa...	Netral	Netral
1149	radaraktual	Duh! Lewat Aturan PSE, Kominfo Bisa Intip Perc...	Netral	Netral
1150	CISSReC	Aturan PSE Wajib Daftar, Kominfo Bisa Lihat Pe...	Netral	Netral
1151	Gorajuara	Begini Cara Kominfo Bisa Intip Perbincangan di...	Netral	Netral
1152	Acehzonedotcom	Kominfo Bisa Intip Pesan Whatsapp-Gmail Pakai ...	Netral	Netral

1153 rows × 4 columns

Gambar 3. Hasil data crawling dan pelabelan data secara manual serta validasi data

Berikut hasil diagram batang berdasarkan jumlah sentimen setelah label "netral" dihapus dapat dilihat gambar 4. (Gambar kiri merupakan sentimen secara manual dan sebelah kanan merupakan sentimen yang sudah divalidasi).



Gambar 4. Perbandingan grafik sentiment positif dan negatif sebelum dan sesudah validasi

B. Preprocessing Data

Setelah data telah diberikan label dan di validasi, langkah selanjutnya adalah menghilangkan *noise* atau kata – kata yang tidak diperlukan dengan tahap *preprocessing* dalam mengoptimalkan proses klasifikasi.

1) Cleansing dan Case folding

Pada tahapan ini dilakukan proses dalam menghilangkan emoji,angka,URL maupun simbol – simbol yang tidak diperlukan dan mengubah seluruh huruf teks pada tweet dari

bentuk kapital menjadi huruf kecil atau *lowercase* untuk mengoptimalkan proses klasifikasi.

	Username	Text	Sentimen	Text_clean
0	tsetiady	kominfo ini bisa jadi lembaga paling hipokrit ...	negatif	kominfo ini bisa jadi lembaga paling hipokrit ...
1	pepo_infp	@tilehopper Kalau sistem registrasi PSE tdk di...	negatif	kalau sistem registrasi pse tdk dihentikan da...
3	aagunawan_	UU ITE ◆ kominfo PSE Ranah Privat ◆ kominfo p...	negatif	uu ite kominfo pse ranah privat kominfo paham ...
4	anggreannie	@pancasyah Salfok sama yg di iklan, situs tida...	negatif	salfok sama yg di iklan situs tidak bisa di a...
5	San_Sunsh	Ini yang di blokir kominfo bisa ga mereka daft...	negatif	ini yang di blokir kominfo bisa ga mereka daft...

Gambar 5. Hasil Tahapan Cleaning dan Case Folding data

2) *Tokenizing*

Tahapan berikutnya dilakukan pemecahan menjadi kata per kata pada suatu teks atau kalimat, proses ini dilakukan untuk proses transformasi data atau pembobotan.

Text	Sentimen	Text_clean	tokenisasi
kominfo ini bisa jadi lembaga paling hipokrit ...	negatif	kominfo ini bisa jadi lembaga paling hipokrit ...	[kominfo, ini, bisa, jadi, lembaga, paling, hi...
@tilehopper Kalau sistem registrasi PSE tdk di...	negatif	kalau sistem registrasi pse tdk dihentikan da...	[kalau, sistem, registrasi, pse, tdk, dihentik...
UU ITE ◆ kominfo PSE Ranah Privat ◆ kominfo p...	negatif	uu ite kominfo pse ranah privat kominfo paham ...	[uu, ite, kominfo, pse, ranah, privat, kominfo...
@pancasyah Salfok sama yg di iklan, situs tida...	negatif	salfok sama yg di iklan situs tidak bisa di a...	[salfok, sama, yg, di, iklan, situs, tidak, bi...
Ini yang di blokir kominfo bisa ga mereka daft...	negatif	ini yang di blokir kominfo bisa ga mereka daft...	[ini, yang, di, blokir, kominfo, bisa, ga, mer...

Gambar 6. Hasil Tahapan Tokenizing

3) *Normalization*

Tahapan selanjutnya dilakukan proses normalisasi dimana mengubah kata tidak baku atau mengandung unsur slang, redudansi yang tidak sesuai dengan kaidah Bahasa Indonesia.

Username	Text	Sentimen	Text_clean	tokenisasi	normalisasi
tsetiady	kominfo ini bisa jadi lembaga paling hipokrit ...	negatif	kominfo ini bisa jadi lembaga paling hipokrit ...	[kominfo, ini, bisa, jadi, lembaga, paling, hi...	[komunikasi, dan, informatika, ini, bisa, jadi...
pepo_infp	@tilehopper Kalau sistem registrasi PSE tdk di...	negatif	kalau sistem registrasi pse tdk dihentikan da...	[kalau, sistem, registrasi, pse, tdk, dihentik...	[kalau, sistem, pendaftaran, penyelenggara, si...
aagunawan_	UU ITE ◆ kominfo PSE Ranah Privat ◆ kominfo p...	negatif	uu ite kominfo pse ranah privat kominfo paham ...	[uu, ite, kominfo, pse, ranah, privat, kominfo...	[undang, undang, informasi, dan, transaksi, el...
anggreannie	@pancasyah Salfok sama yg di iklan, situs tida...	negatif	salfok sama yg di iklan situs tidak bisa di a...	[salfok, sama, yg, di, iklan, situs, tidak, bi...	[salah, fokus, sama, yang, di, iklan, situs, t...
San_Sunsh	Ini yang di blokir kominfo bisa ga mereka daft...	negatif	ini yang di blokir kominfo bisa ga mereka daft...	[ini, yang, di, blokir, kominfo, bisa, ga, mer...	[ini, yang, di, blokir, komunikasi, dan, infor...

Gambar 7. Hasil Tahapan Normalization

4) *Stopword Removal*

Selanjutnya pada tahapan ini untuk melakukan penyaringan dengan menghilangkan kata-kata yang tidak diperlukan atau tidak memiliki kepentingan dalam proses klasifikasi. Pada proses ini bahasa pemrograman *python* telah menyediakan *library* khusus yatu *nltk stopwords* dan sudah bisa bahasa Indonesia , namun untuk memaksimalkan tahapan ini peneliti membuat dokumen khusus yang berisi beberapa kata yang tidak terdeteksi oleh *library nltk*.

	Username	Text	Sentimen	Text_clean	tokenisasi	normalisasi	stopwords_removal
0	tsetiady	kominfo ini bisa jadi lembaga paling hipokrit ...	negatif	kominfo ini bisa jadi lembaga paling hipokrit ...	[kominfo, ini, bisa, jadi, lembaga, paling, hi...	[komunikasi, dan, informatika, ini, bisa, jadi...	[komunikasi, informatika, lembaga, kalih, muna...
1	pepo_infp	@tilehopper Kalau sistem registrasi PSE tdk di...	negatif	kalau sistem registrasi pse tdk dihentikan da...	[kalau, sistem, registrasi, pse, tdk, dihentik...	[kalau, sistem, pendaftaran, penyelenggara, si...	[sistem, pendaftaran, penyelenggara, sistem, e...
3	aagunawan_	UU ITE ◆ kominfo PSE Ranah Privat ◆ kominfo p...	negatif	uu ite kominfo pse ranah privat kominfo paham ...	[uu, ite, kominfo, pse, ranah, privat, kominfo...	[undang, undang, informasi, dan, transaksi, el...	[undang, undang, informasi, transaksi, elektro...
4	anggreannie	@pancasyah Salfok sama yg di iklan, situs tida...	negatif	salfok sama yg di iklan situs tidak bisa di a...	[salfok, sama, yg, di, iklan, situs, tidak, bi...	[salah, fokus, sama, yang, di, iklan, situs, t...	[salah, fokus, iklan, situs, akses, terindikas...
5	San_Sunsh	Ini yang di blokir kominfo bisa ga mereka daft...	negatif	ini yang di blokir kominfo bisa ga mereka daft...	[ini, yang, di, blokir, kominfo, bisa, ga, mer...	[ini, yang, di, blokir, komunikasi, dan, infor...	[blokir, komunikasi, informatika, daftar, peny...

Gambar 8. Hasil Tahapan Stopword Removal

5) *Stemming*

Tahapan terakhir adalah *stemming* yaitu proses mengubah kata yang memiliki imbuhan menjadi ke bentuk dasarnya, pada proses ini bahasa pemrograman *python* juga menyediakan *library* yaitu *StemmerFactory* dari Sastrawi.

Username	Text	Sentimen	Text_clean	tokenisasi	normalisasi	stopwords_removal	stemming
tsetiady	kominfo ini bisa jadi lembaga paling hipokrit ...	negatif	kominfo ini bisa jadi lembaga paling hipokrit ...	[kominfo, ini, bisa, jadi, lembaga, paling, hi...	[komunikasi, dan, informatika, ini, bisa, jadi...	[komunikasi, informatika, lembaga, kalih, muna...	[komunikasi, informatika, lembaga, kalih, muna...
pepo_infp	@tilehopper Kalau sistem registrasi PSE tdk di...	negatif	kalau sistem registrasi pse tdk dihentikan da...	[kalau, sistem, registrasi, pse, tdk, dihentik...	[kalau, sistem, pendaftaran, penyelenggara, si...	[sistem, pendaftaran, penyelenggara, sistem, e...	[sistem, daftar, selenggara, sistem, elektroni...
aagunawan_	UU ITE ◆ kominfo PSE Ranah Privat ◆ kominfo p...	negatif	uu ite kominfo pse ranah privat kominfo paham ...	[uu, ite, kominfo, pse, ranah, privat, kominfo...	[undang, undang, informasi, dan, transaksi, el...	[undang, undang, informasi, transaksi, elektro...	[undang, undang, informasi, transaksi, elektro...
anggreannie	@pancasyah Salfok sama yg di iklan, situs tida...	negatif	salfok sama yg di iklan situs tidak bisa di a...	[salfok, sama, yg, di, iklan, situs, tidak, bi...	[salah, fokus, sama, yang, di, iklan, situs, t...	[salah, fokus, iklan, situs, akses, terindikas...	[salah, fokus, iklan, situs, akses, indikasi, ...
San_Sunsh	Ini yang di blokir kominfo bisa ga mereka daft...	negatif	ini yang di blokir kominfo bisa ga mereka daft...	[ini, yang, di, blokir, kominfo, bisa, ga, mer...	[ini, yang, di, blokir, komunikasi, dan, infor...	[blokir, komunikasi, informatika, daftar, sele...	[blokir, komunikasi, informatika, daftar, sele...

Gambar 9. Hasil Tahapan Stemming

C. *Transformation Data*

Setelah data sudah clean atau bersih, maka proses selanjutnya adalah transformasi data atau pembobotan menggunakan *TF-IDF* pada setiap katanya, langkah awal dari transformasi data adalah melakukan *splitting data* terlebih dahulu, dikarenakan pada penelitian ini menggunakan 3 Skenario maka dapat dilihat pembagian dataset pada tabel 1.

TABEL 1
SKENARIO SPLITTING DATASET

Skenario	Presentase Data		Presentase Data	
	Training	Testing	Training	Testing
Skenario 1	90%	10%	614	69
Skenario 2	80%	20%	546	137
Skenario 3	70%	30%	478	205
Total			683	

Setelah melakukan *splitting data*, langkah berikutnya melakukan perhitungan *term* atau pembobotan pada setiap kata menggunakan TF-IDF dikarenakan presentase data pada skenario 1 lebih banyak daripada skenario 2 dan skenario 3 maka dalam implementasi pembobotan menggunakan skenario 1 sebagai acuan. Dalam analisis teks, setiap kata memiliki skor TF-IDF yang menunjukkan seberapa sering kata tersebut muncul dalam teks tersebut. Namun, kata-kata yang jarang muncul dianggap kurang berpengaruh dalam proses klasifikasi, sehingga perlu dilakukan seleksi fitur menggunakan algoritma tertentu untuk memperbaiki performa model yaitu dengan menggunakan *information gain*[13], *information gain* adalah salah satu algoritma penyaringan klasik, di mana rasio *information gain* mewakili perbandingan antara keuntungan informasi dan informasi intrinsik[14]. Tahapan awal pada seleksi fitur *information gain* adalah menghitung nilai *mutual information* terlebih dahulu kemudian mencari nilai rata – rata atau *mean* dari *mutual information* tersebut dapat menggunakan library *numpy* pada *python*.

```
: avg_train_mi = np.mean(train_mi)

  print(f"Average Mutual Information: {avg_train_mi}")

  Average Mutual Information: 0.0056918050115630755
```

Setelah itu selanjutnya adalah menentukan nilai *threshold* yaitu **MI < *threshold***, sehingga berdasarkan hasil eksperimen *threshold* yang terbaik untuk performa model yaitu 0,008 hasil bisa dilihat pada **tabel 2** berdasarkan skenario 1 hingga skenario 3 dengan berdasarkan **MI < *threshold*** yaitu 0,005 < 0,008.

TABEL 2
HASIL SELEKSI FITUR INFORMATION GAIN

Skenario	*Threshold 0,008*	
	Jumlah fitur sebelum *Information Gain*	Jumlah fitur sesudah *Information Gain*
Skenario 1 (90:10)	1732	1023
Skenario 2 (80:20)	1597	926
Skenario 3 (70:30)	1477	854

D. Data Mining

Setelah melakukan seleksi fitur, langkah berikutnya adalah proses *data mining*. Pada penelitian ini proses pengolahan *data mining* atau membuat model menggunakan algoritma *Naïve Bayes Classifier*, *a*lgoritma ini menginferensikan probabilitas bahwa sebuah contoh baru termasuk ke dalam suatu kelas berdasarkan asumsi bahwa semua atribut saling independen satu sama lain jika diberikan kelasnya asumsi tersebut didorong oleh kebutuhan untuk mengestimasi probabilitas multivariat dari data latihan[15] kemudian hasil proses modelling akan mendeteksi kalimat dengan bilangan 1(*true*) dan 0(*false*) berarti dalam kasus ini 1 akan dinyatakan positif dan 0 dinyatakan negatif kemudian hasil prediksi dari model tersebut akan di *export* ke dalam *excel*(.csv).

E. Evaluation

Evalution merupakan tahapan terakhir dari hasil penelitian ini berdasarkan metodologi *Knowledge Discovery in Database* (KDD) dimana dilakukan pengujian dan evaluasi performa model yang digunakan juga membandingkan performa antara impelemtasi menggunakan dan tidak menggunakan fitur seleksi *Information Gain*. Tabel 3 menunjukkan hasil dari performa dari *confusion matrix* Naïve *Bayes Classifier* dan *Naïve Bayes Classifier + Information Gain*.

TABEL 3.
HASIL KLASIFIKASI

Hasil Klasifikasi		**Actual**			
		Naïve Bayes Classifier		*Naïve Bayes Classifier + IG*	
Skenario 1 (90:10)		Positif	Negatif	Positif	Negatif
Prediction	Positif	6	10	3	5
	Negatif	6	47	9	52
Skenario 2 (80:20)		Positif	Negatif	Positif	Negatif
Prediction	Positif	17	20	5	8
	Negatif	15	85	27	97
Skenario 3 (70:30)		Positif	Negatif	Positif	Negatif
Prediction	Positif	27	38	20	28
	Negatif	26	114	33	124

Confusion matrix [16] dipakai untuk mengukur kinerja model klasifikasi dengan menghitung *accuracy, precision, recall,* serta *F1- score*. *Accuracy* yaitu rasio diantara jumlah data yang benar diprediksi dengan jumlah total data berikut rumus dari *Accuracy* :

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

Precision adalah rasio antara jumlah data positif yang benar diprediksi dengan jumlah total data positif yang diprediksi, dapat diketahui rumus dari *precision* adalah sebagai berikut :

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

Recall adalah rasio antara jumlah data positif yang benar diprediksi dengan jumlah total data positif yang sebenarnya, rumus dari *recall* :

$$Recall = \frac{TP}{(TP+TN)}$$

F1-score adalah rasio atau rata – rata harmonis antara precision dan recall, rumus dari *f-1 score* adalah sebagi berikut :

$$\text{F-1} = \frac{2(Recall\ x\ Precision)}{Recall+Precision}$$

Kemudian saat *confusion matrix* didapatkan evaluasi kembali dengan menggunakan *accuracy, precision, recall,* dan *f-1 score* seperti tabel 4 yang merupakan tabel *summary* hasil klasifikasi.

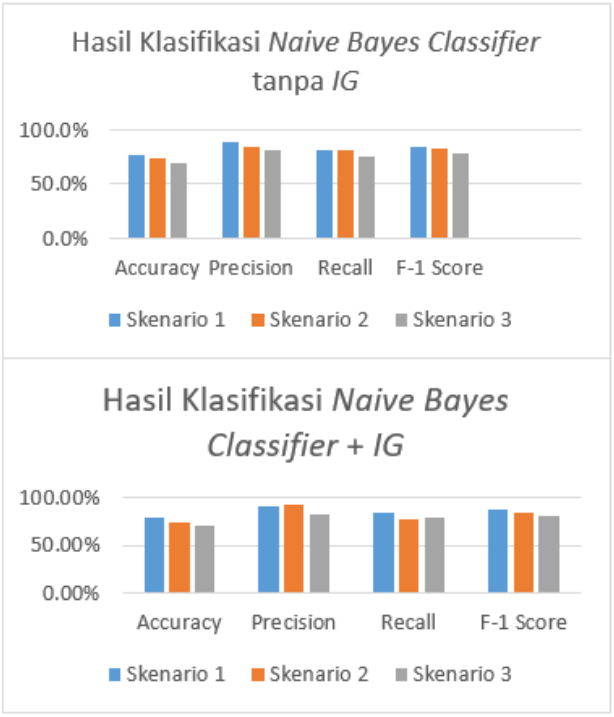
TABEL 4.
RINGKASAN HASIL KLASIFIKASI

Skenario	*Naïve Bayes Classifier*			
	Accuracy	*Precision*	*Recall*	*F-1 Score*
90:10	76,8%	89%	82%	85%
80:20	74,5%	85%	81%	83%
70:30	68,8%	81%	75%	78%
Skenario	*Naïve Bayes Classifier + IG*			
	Accuracy	*Precision*	*Recall*	*F-1 Score*
90:10	79,7%	91%	85%	88%
80:20	74,5%	92%	78%	85%
70:30	70,2%	82%	79%	80%

Hasil evaluasi performa model dengan algoritma *Naïve Bayes Classifier* terbagi menjadi 3 skenario dan skenario 1 dengan rasio 90:10 dengan menggunakan seleksi fitur *Information Gain* menjadi performa terbaik yaitu *Accuracy* 79,7%, *Recall* 85%, dan *F-1 Score* sebesar 88% namun pada *Precision* terbaik ada pada skenario 2 dengan rasio 80:20 yaitu sebesar 92% dikarenakan proporsi prediksi positif yang dilakukan oleh model lebih besar dibandingkan skenario lainnya.

Frekuensi data *tweet* dapat diketahui dengan visualiasi menggunakan *wordcloud* untuk memunculkan kata yang sering muncul atau memiliki kemunculan tertinggi, Dengan menggunakan *wordcloud*, kita dapat dengan mudah mengetahui kata-kata yang lebih dominan pada suatu teks, sehingga dapat membantu kita menjawab pertanyaan penelitian dengan cepat[17]. Lihat gambar 11 untuk melihat

wordcloud kata yang sering muncul pada kebijakan PSE Kominfo.



Gambar 10. Hasil Klasifikasi

Gambar 11. Wordcloud pada kebijakan PSE Kominfo

Adapun visualisasi dengan top 10 kata yang paling sering muncul, dapat dilihat pada diagram batang pada gambar 11. untuk memudahkan informasi yang telah didapatkan.

IV. KESIMPULAN

Penelitian ini bertujuan untuk mengimplementasikan seleksi fitur *information gain* untuk analisis sentimen pada opini pengguna *twitter* terhadap kebijakan Penyelenggara Sistem Elektronik Kominfo menggunakan algoritma *Naïve Bayes Classifier* dengan beberapa tahapan yaitu pengumpulan data *tweet* pada *twitter* menggunakan *API Key Twitter V2* dengan, kemudian data labeli positif dan negatif. Tahapan selanjutnya dilakukan *preprocessing*, *splitting data* dengan 2 skenario, pembobotan dengan TF-IDF, seleksi fitur *Information Gain*, *data mining*, dan evaluasi performa model.

Adapun hasil evaluasi performa model dengan algoritma *Naïve Bayes Classifier* terbagi menjadi 3 skenario dan skenario 1 dengan rasio 90:10 dengan menggunakan seleksi fitur *Information Gain* menjadi performa terbaik yaitu *Accuracy* 79,7%, *Recall* 85%, dan *F-1 Score* sebesar 88% namun pada *Precision* terbaik ada pada skenario 2 dengan rasio 80:20 yaitu sebesar 92% dikarenakan proporsi prediksi positif yang dilakukan oleh model lebih besar dibandingkan skenario lainnya

DAFTAR PUSTAKA

[1] Kominfo, "Peraturan Menteri Komunikasi dan Informatika Nomor 5 Tahun 2020 tentang Penyelenggara Sistem Elektronik Lingkup Privat," *https://pse.kominfo.go.id/*, Jul. 22, 2022. .

[2] C. Rahmawati and P. Sukmasetya, "Sentimen Analisis Opini Masyarakat Terhadap Kebijakan Kominfo atas Pemblokiran Situs non-PSE pada Media Sosial Twitter," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 5, p. 1393, Oct. 2022, doi: 10.30865/jurikom.v9i5.4950.

[3] CNN Indonesia, "Analisis PSE Kominfo Buat Kepentingan Siapa?," *https://www.cnnindonesia.com/teknologi/20220804012642-192-830038/pse-kominfo-buat-kepentingan-siapa*, Aug. 04, 2022. .

[4] O. Y. Adwan, M. Al-Tawil, A. M. Huneiti, R. A. Shahin, A. A. Abu Zayed, and R. H. Al-Dibsi, "Twitter sentiment analysis approaches: A survey," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 15, pp. 79–93, 2020, doi: 10.3991/ijet.v15i15.14467.

[5] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, A. Mulyani, and J. I. Komputer, "Comparison Of Classification Algorithm And Feature Selection In Bitcoin Sentiment Analysis," *J. Tek. Inform.*, vol. 3, no. 3, 2022, doi: 10.20884/1.jutif.2022.3.3.343.

[6] A. Isnanda, Y. Umaidah, and J. H. Jaman, "Implementasi Naïve Bayes Classifier Dan Information Gain Pada Analisis Sentimen Penggunaan E-Wallet Saat Pandemi," *J. Teknol. Inform. dan Komput.*, vol. 7, no. 2, pp. 144–153, Sep. 2021, doi: 10.37012/jtik.v7i2.648.

[7] A. Bijaksana, P. Negara, H. Muhardi, and I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes Dan Seleksi Fitur Information Gain Sentiment Analysis On Airlines Using Naïve Bayes Method And Feature Selection Information Gain," vol. 7, no. 3, pp. 599–606, 2020, doi: 10.25126/jtiik.202071947.

[8] F. Nasser K, "A Review of Data Mining and Knowledge Discovery Approaches for Bioinformatics," *Iraqi J. Sci.*, vol. 63, 2022, Accessed: Aug. 21, 2023. [Online]. Available: 10.24996/ijs.2022.63.7.37.

[9] I. A. Ahmad Sabri, M. Man, W. A. W. Abu Bakar, and A. N. Mohd Rose, "Web Data Extraction Approach for Deep Web using WEIDJ," in *Procedia Computer Science*, 2019, vol. 163, pp. 417–426, doi: 10.1016/j.procs.2019.12.124.

[10] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.01054.

[11] B. Al Hadits, "Analisis Sentimen Kepuasan Pelanggan Indihome Pada Media Sosial Twitter Menggunakan Metode Naïve Bayes Classifiers Proposal Skripsi," 2020.

[12] C. Barrie and J. Ho, "academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint," *J. Open Source Softw.*, vol. 6, no. 62, p. 3272, Jun. 2021, doi: 10.21105/joss.03272.

[13] Jurafsky, Daniel, & Martin, James H. 2019. *Speech and Language Processing. Pearson*.

[14] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, "Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier," Apr. 2019, doi: 10.1016/j.comnet.2020.107247

[15] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," Knowl Based Syst, vol. 192, Mar. 2020, doi: 10.1016/j.knosys.2019.105361.

[16] Provost, F., & Fawcett, T. (2013). Data Science and Its Relationship to Big Data and Data-Driven Decision Making. Big Data, 1, 51-59

[17] A. S. Ramadhani, "Analisis Sentimen Netizen Terhadap Trailer Film di YouTube SKRIPSI," 2020.