Analisis Sentimen Masyarakat terhadap Kebijakan Iuran Tabungan Perumahan Rakyat (Tapera) pada *Platform* X Menggunakan Algoritma *Naïve Bayes Classifier* dan *Support Vector Machine*

Anis Maulidatur Rizqiyah[1], I Kadek Dwi Nuryana[2]

[1,2] Sistem Informasi, Fakultas Teknik, Universitas Negeri Surabaya
[1]anis.18026@mhs.unesa.ac.id
[2]dwinuryana@unesa.ac.id

***Abstrak*—Pemerintah Indonesia menetapkan perubahan terhadap PP Nomor 25 Tahun 2020 tentang Penyelenggaraan Tabungan Perumahan Rakyat (Tapera) melalui PP Nomor 21 Tahun 2024. Dalam perubahan tersebut gaji pekerja Indonesia akan dipotong 3% untuk Tapera. Hal tersebut menimbulkan perdebatan dikalangan masyarakat, terutama pengguna *platform* X. Pada *platform* tersebut, masyarakat berbagi opini dan pandangan mereka terhadap kebijakan Tapera. Penelitian ini bertujuan untuk mengklasifikasikan *tweet* terkait kebijakan Tabungan Perumahan Rakyat (Tapera) menggunakan algoritma Naïve Bayes dan *Support Vector Machine* (SVM), sehingga didapatkan informasi mengenai sentimen masyarakat terhadap kebijakan tersebut. Data sejumlah 1280 *tweet* didapatkan dari hasil *crawling* web X. Data tersebut diproses menggunakan *library* sklearn dan diberikan label menggunakan *InSet Lexicon*. Data juga diproses menggunakan SMOTE. Klasifikasi dilakukan dengan membagi data ke dalam rasio 80:20, 70:30 dan 60:40. Hasil klasifikasi menggunakan algoritma Naïve Bayes dan SVM kemudian dievaluasi menggunakan *confusion matrix* dan *k-fold cross validation*. Dari hasil klasifikasi didapatkan bahwa sentimen masyarakat cenderung kearah negatif terhadap kebijakan Tapera. Didapatkan juga bahwa algoritma SVM memiliki akurasi yang lebih baik dibandingkan dengan algoritma Naïve Bayes. Sebelum SMOTE, SVM memiliki akurasi 84% pada rasio 80:20 dengan kernel linear dan C=2, sedangkan Naïve Bayes memiliki akurasi 81% pada rasio 80:20 dengan model Complement dan alpha 0.01. Setelah SMOTE, SVM memiliki akurasi 93% pada rasio 80:20 dengan kernel rbf dan C=3, sedangkan Naïve Bayes memiliki akurasi 89% pada rasio 60:40 dengan model Complement dan alpha 0.1.**

***Kata Kunci*—Tapera, Naïve Bayes, SVM, *InSet Lexicon*, analisis sentimen**

I. PENDAHULUAN

Menurut Peraturan Pemerintah (PP) Nomor 25 tahun 2020, Tabungan Perumahan Rakyat (Tapera) merupakan tabungan yang bersifat berkala untuk pembiayaan perumahan dan akan dikembalikan kepada peserta pada akhir periode beserta hasil pemupukannya. Peserta Tapera adalah Warga Negara Indonesia (WNI) dan Warga Negara Asing (WNA) yang telah bekerja di Indonesia kurang lebih enam bulan dan telah membayar uang jaminan. Peserta Tapera adalah pegawai negeri, swasta, dan *freelance* yang sudah menikah dan berusia minimal 20 tahun. Keikutsertaan dalam Tapera bersifat wajib bagi pekerja yang memenuhi syarat-syarat tersebut. Pada tahun 2024, pemerintah menetapkan perubahan terhadap PP Nomor 25 Tahun 2020 tentang Penyelenggaraan Tabungan Perumahan Rakyat (Tapera) melalui PP Nomor 21 tahun 2024. Dalam PP 21 Tahun 2024, Tapera akan dipotong sebesar 3% dari gaji pekerja Indonesia. Untuk pegawai negeri dan swasta, pemotongan akan didistribusikan antara 2,5% pekerja dan 0,5% pemberi kerja. Bagi *freelance*, pemotongan akan ditanggung sepenuhnya oleh pekerja.

Kebijakan Tapera menyebabkan perdebatan dikalangan masyarakat, terutama pengguna *platform* X. Masyarakat berbagi opini dan pandangan mereka mengenai kebijakan tersebut. X atau yang sebelumnya dikenal dengan Twitter merupakan media sosial dimana penggunanya dapat mengungkapkan dan berbagi opini atau pandangan secara singkat dan *real-time*. Opini atau pandangan tersebut diungkapkan pengguna melalui *tweet.* Berdasarkan Databoks, pada tahun 2023 Indonesia merupakan pengguna X keempat terbesar di dunia setelah USA, Jepang dan India dengan 25,25 juta pengguna. Hal tersebut membuktikan bahwa saat ini X telah menjadi salah satu media sosial yang utama bagi masyarakat Indonesia. Perbincangan maupun perdebatan yang terjadi di X, mempengaruhi arah pandang masyarakat yang mengonsumsi informasi dari *platform* tersebut. Berdasarkan hal tersebut, diperlukan klasifikasi *tweet* mengenai kebijakan Tapera untuk mengetahui sentimen masyarakat mengenai hal tersebut.

Dalam mencari sentimen dari opini masyarakat diperlukan adanya analisis sentimen. Analisis sentimen merupakan suatu teknik atau metode yang digunakan untuk menentukan bagaimana emosi diungkapkan dalam tulisan dan bagaimana mengklasifikasikan perasaan tersebut sebagai perasaan positif atau negatif [1]. Hasil analisis sentimen berupa kelas yang berisi sekumpulan teks yang diklasifikasikan ke dalam positif, negatif dan netral. Analisis ini juga memberikan hasil lain berupa visualisasi frekuensi kata pada suatu data, grafik, atau tabel [2].

Analisis sentimen dapat dilakukan dengan menggunakan metode klasifikasi *text mining*, seperti *Support Vector Machine*, Naïve Bayes dan *K-Nearest Neighbor* [3]. Dalam penelitian ini akan membandingkan dua algoritma untuk mencari algoritma terbaik dalam melakukan analisis sentimen terhadap kebijakan Tabungan Perumahan Rakyat (Tapera). Algoritma yang akan dibandingkan yakni naïve bayes dan *support vector machine*, kedua algoritma tersebut akan dibandingkan berdasarkan tingkat akurasinya.

Naïve Bayes adalah metode pengklasifikasian kelompok objek berdasarkan karakteristiknya dengan menggunakan basis probabilitas [4]. Naïve bayes sangat baik untuk klasifikasi teks karena bekerja secara efektif, memiliki waktu *training* yang cepat dengan data *training* yang relatif kecil, dan mampu menangani data dalam ukuran yang besar dengan baik, dimana asumsi independen data tersebut valid [5]. Algoritma naïve bayes dapat melakukan klasifikasi sentimen pada ulasan aplikasi PLN *mobile* lebih baik daripada algoritma *K-Nearest Neighbor* (KNN). Hasil akurasi algoritma tersebut lebih tinggi 1,29% daripada algoritma KNN dengan persentase 77,69%. Naïve bayes memiliki komputasi yang cukup efisien dan pengujian yang lebih cepat dibanding KNN [6]. Naïve bayes juga memiliki akurasi yang lebih baik daripada algoritma *decision tree* dalam menganalisis sentimen masyarakat mengenai vaksinasi COVID-19 di Indonesia. Dimana naïve bayes memiliki akurasi 100%, sedangkan *decision tree* hanya 50,39% [7].

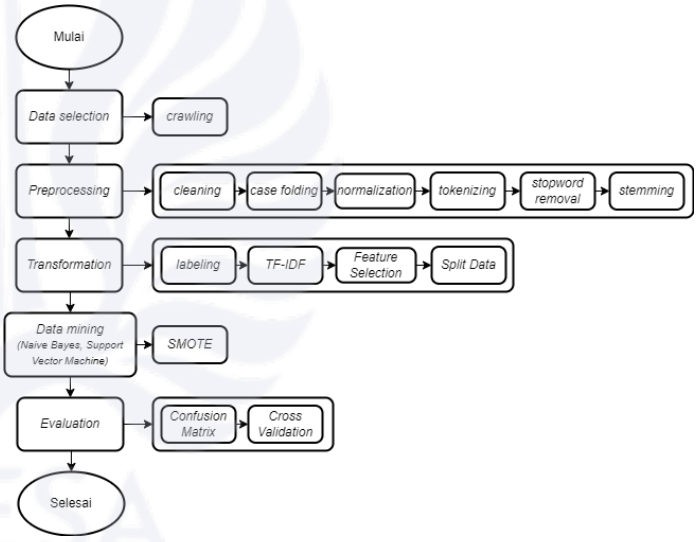
Support Vector Machine atau SVM adalah metode klasifikasi yang memisahkan titik dari dua kelas berbeda dalam ruang fitur untuk mencari *hyperplane* [8]. Dalam melakukan klasifikasi teks, *support vector machine* memiliki performa yang bagus dan dapat menangani data yang sangat besar [9]. *Support vector machine* juga dapat meminimalkan *noise* dan kuat terhadap *overfitting* [10]. Pada analisis sentimen mengenai ulasan aplikasi *home credit*, algoritma *support vector machine* memiliki akurasi yang lebih baik daripada algoritma KNN dengan perbedaan 9%, algoritma *support vector machine* memiliki akurasi 88% [11]. Algoritma *support vector machine* juga memiliki performa lebih baik daripada algoritma *decision tree* dalam menganalisis sentimen top 10 *traveler ranked* hotel di kota Makassar dengan nilai akurasi 98,98%, sedangkan *decision tree* hanya 93,60% [12].

Penelitian mengenai analisis sentimen menggunakan naïve bayes dan SVM sudah ada sebelumnya, yakni penelitian mengenai komparasi metode naïve bayes dan SVM pada sentimen twitter mengenai persoalan perppu cipta kerja. Penelitian tersebut bertujuan untuk mengetahui sentimen masyarakat mengenai pengesahan perppu cipta kerja dan hasilnya dapat membuat masyarakat lebih teredukasi mengenai kebijakan tersebut. Dari penelitian ini didapatkan bahwa dari 622 data yang terkumpul, sebanyak 332 data mencerminkan sentimen positif, 224 data mencerminkan sentimen negatif dan 66 data mencerminkan sentimen netral. Dari hasil komparasi algortima SVM memiliki akurasi yang lebih tinggi daripada *naïve bayes* dengan 78%, sedangkan *naïve bayes* hanya 73% [13]. Terdapat pula penelitian lain mengenai *E-Wallet Sentiment Analysis Using Naïve Bayes and Support Vector Machine Algorithm*. Penelitian tersebut bertujuan untuk mencari metode terbaik dalam analisis sentimen *e-wallet*. Dari penelitian ini didapatkan bahwa *naïve bayes* memiliki akurasi yang lebih baik daripada SVM dengan 94,90%, sedangkan SVM hanya 91% [14].

Berdasarkan hal tersebut, pada penelitian ini akan dilakukan analisis sentimen masyarakat terhadap kebijakan Tabungan Perumahan Rakyat (Tapera) pada platform X menggunakan algoritma naïve bayes dan SVM. Sehingga, dapat diketahui metode terbaik dalam melakukan analisis tersebut.

II. METODE PENELITIAN

Penelitian ini menggunakan metode *Knowledge Discovery in Database (KDD)* yang mengidentifikasi pola pada dataset secara runtut sehingga data mudah dipahami. KDD terdiri dari 5 tahapan, yakni *data selection, preprocessing, transformation, data mining* dan *evaluation*. Gbr 1. merupakan alur dari penelitian ini.



Gbr 1. Diagram Alir Metode Penelitian

A. *Data Selection*

Data dikumpulkan dengan teknik *crawling*, sebuah teknik mengumpulkan data dengan berpindah-pindah antar halaman *web* [15]. Proses *crawling* dilakukan dengan menggunakan tweet-harvest, sebuah *command-line tool* untuk mengambil data dari X berdasarkan pencarian spesifik *keywords* dan rentang tanggal [16]. Proses ini dilakukan dengan bahasa pemrograman *python* melalui Google colab. Data yang diambil adalah data dari *platform* X berupa *tweet* dengan *keyword* tapera yang menggunakan bahasa Indonesia.

B. *Preprocessing*

Preprocessing dilakukan untuk membuat data yang digunakan terjamin kualitasnya dan benar-benar mewakili data

yang sebenarnya [17]. *Preprocessing* terdiri dari 6 tahapan, yaitu:

1) Cleaning: Pada data *tweet* yang telah dikumpulkan terdapat URL, *hastag*, *emoticon* maupun *mention*. Atribut tersebut tidak diperlukan dan perlu dihilangkan melalui proses *cleaning*.

2) Case Folding: Data *tweet* yang telah dikumpulkan akan disamakan seluruhnya menjadi huruf kecil agar tidak mempengaruhi proses *data mining*.

3) Normalization: Dalam data yang dikumpulkan terdapat kata baku dan tidak baku, hal ini perlu diperbaiki atau diolah agar kata tersebut sesuai dengan Kamus Besar Bahasa Indonesia (KBBI) dan lebih mudah diproses pada tahap selanjutnya.

4) Tokenizing: Data teks dipecah per kata atau token berdasarkan kamus data agar frekuensi data yang terdapat dalam *corpus* lebih gampang ditemukan.

5) Stopword Removal: Kata yang tidak penting pada data dibersihkan berdasarkan daftar *stopwords* yang terdapat pada *library sastrawi*.

6) Stemming: Data kata yang ada, dihapus imbuhan dan akhirannya hingga membentuk kata dasar. Sehingga kata yang sama tidak memiliki variasi dan lebih mudah diproses pada tahap selanjutnya.

C. Transformation

Transformation merupakan proses mengubah data ke dalam format yang dapat digunakan dalam tahap *data mining*. Sebelum dilakukan *transformation*, data diberikan label positif, negatif dan netral untuk mendeteksi sentimen dari data tersebut. Pelabelan dilakukan dengan menggunakan kamus *InSet Lexicon* (*Indonesia Sentiment Lexicon*). Pada kamus tersebut terdapat daftar kata positif dan negatif beserta skor dari tiap kata. Untuk kata positif dan negatif yang terdapat dalam *tweet* akan diberikan skor sesuai dengan skor kata tersebut dalam kamus *InSet Lexicon*. Skor dari setiap kata akan dijumlahkan dan menjadi skor akhir. Jika skor akhir > 0 maka *tweet* akan diklasifikasikan ke dalam kelas positif. Jika skor akhir < 0 maka *tweet* akan diklasifikasikan ke dalam kelas negatif. Dan jika skor akhir = 0, maka akan diklasifikasikan ke dalam kelas netral.

Transformation dilakukan dengan menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) pada *library* sklearn. Pada tahap ini data yang telah diberikan label akan dilakukan pembobotan untuk menghitung seberapa banyak suatu kata muncul dalam setiap kalimat. TF-IDF memiliki kemampuan yang baik dalam meningkatkan akurasi klasifikasi data [18].

Setelah *transformation*, dilakukan *feature selection.* Dimana data diseleksi dengan menggunakan *chi-square* untuk mencari fitur yang dibutuhkan. *Feature selection* dapat meningkatkan performa dan menghindari *overfitting* [19]. Penelitian ini menggunakan *chi-square* pada sklearn yang dibantu oleh SelectKBest dengan memfilter data ke dalam jumlah yang diinginkan. Pada penelitian ini, dari fitur yang terbentuk, dipilih sebanyak 2000 fitur teratas untuk digunakan pada perhitungan chi^2 antara setiap fitur dengan target.

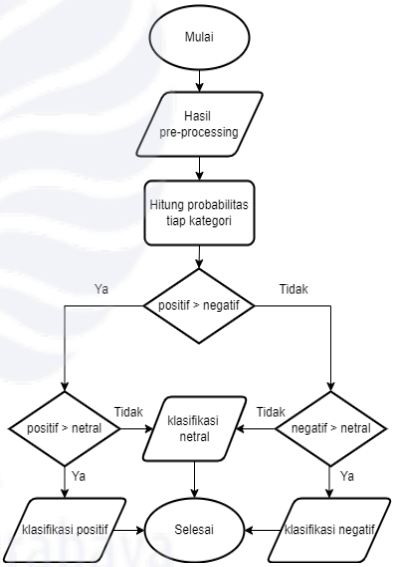
D. Data Mining

Data mining memproses data untuk menghasilkan pola atau informasi yang menarik [20]. Data yang telah diolah pada tahap sebelumnya, diproses menggunakan naïve bayes dan SVM.

Penelitian ini membandingkan data asli dengan data yang telah diolah menggunakan *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE menyeimbangkan data yang tidak seimbang dan dapat meningkatkan performa klasifikasi sehingga hasil yang didapatkan memiliki akurasi yang tinggi [21].

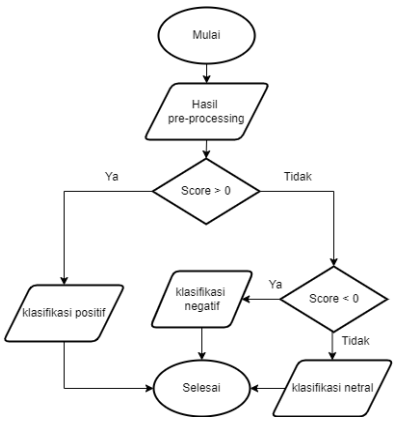
Penelitian ini membagi data menjadi data *training* dan data *testing* dengan 3 rasio yaitu, 80:20, 70:30 dan 60:40. Interpretasinya adalah pada rasio 80:20, 80% data digunakan sebagai data *training* dan 20% sisanya sebagai data *testing*.

1) Naïve Bayes: Naïve Bayes merupakan metode klasifikasi yang mengelompokkan objek berdasarkan ciri-cirinya menggunakan dasar probabilitas [4]. Naïve bayes memiliki perhitungan yang mudah dipahami, cepat dan efisien dan tidak memerlukan jumlah data yang banyak. Gbr 2. merupakan alur dari algoritma naïve bayes.



Gbr 2. Diagram Alir Naive Bayes

2) Support Vector Machine: Support Vector Machine atau SVM merupakan metode klasifikasi yang memisahkan titik dari dua kelas yang berbeda dalam ruang fitur untuk mencari *hyperplane* [8]. SVM dapat membuat model yang baik meskipun dengan data yang sedikit dan memiliki tingkat generalisasi data yang tinggi [1]. Gbr 3. merupakan alur dari algoritma SVM.



Gbr 3. Diagram Alir SVM

E. Evaluation

Evaluation mengukur akurasi model dalam melakukan prediksi sentimen. *Evaluation* dilakukan dengan menggunakan *confusion matrix* dan menghitung nilai *accuracy*, *precision*, *recall* dan *f1-score* dari hasil *k-fold cross validation* untuk menguji performa model yang terbentuk. Pada penelitian ini menggunakan *k* dengan rentang 2 hingga 10.

III. HASIL DAN PEMBAHASAN

A. Data Selection

Data selection dilakukan dengan teknik *crawling* menggunakan *library* tweet-harvest yang memanfaatkan *auth_token* pengguna X. Dari *crawling* tersebut didapatkan 1280 data *tweet* terbaru yang menggunakan *keyword* 'tapera' berbahasa Indonesia dan disimpan dengan format csv. Penelitian ini hanya menggunakan variabel full_text karena hanya berfokus pada analisis isi *tweet*. Tabel I menunjukkan data hasil *crawling*.

TABEL I
DATA HASIL CRAWLING

id_str	username	created_at	full_text
1802652766147670000	share_addict	Mon Jun 17 10:41:25 +0000 2024	@quarterdedie @snnatchy Takut di suruh bayar tapera
1802650902974820000	Dewacorner	Mon Jun 17 10:34:01 +0000 2024	@quarterdedie @snnatchy Mungkin dia ga mau ikut bpjs dan tapera juga
1802650762067290000	LeeRona16	Mon Jun 17 10:33:27 +0000 2024	@5teV3n_Pe9eL duit gaji ditilep apalagi duit tapera

B. Preprocessing

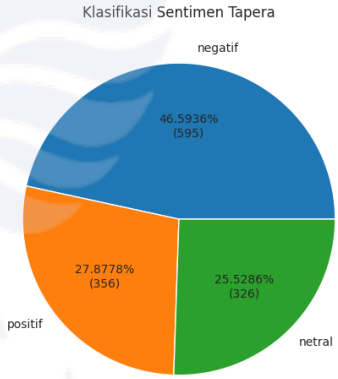
Preprocessing membuat data yang diperoleh siap digunakan pada proses selanjutnya. Namun sebelum dilakukan *preprocessing*, dilakukan penghapusan data duplikasi, sehingga data yang didapatkan bersifat *unique*. Dari hasil penghapusan tersebut dari 1280 data tereduksi menjadi 1277 data. Tabel II menunjukkan contoh hasil *preprocessing*.

TABEL II
PROSES PREPROCESSING

***Tweet* asli**	@quarterdedie @snnatchy Mungkin dia ga mau ikut bpjs dan tapera juga
Cleaning	Mungkin dia ga mau ikut bpjs dan tapera juga
Case Folding	mungkin dia ga mau ikut bpjs dan tapera juga
Normalization	mungkin dia tidak mau ikut bpjs dan tapera juga
Tokenizing	[mungkin, dia, tidak, mau, ikut, bpjs, dan, tapera, juga]
Stopword Removal	[mungkin, mau, ikut, bpjs, tapera]
Stemming	mungkin mau ikut bpjs tapera

C. Transformation

Sebelum dilakukan *transformation* data diberikan label positif, negatif dan netral dengan bantuan kamus *InSet Lexicon*. Dari hasil pelabelan didapatkan bahwa 46.59% teks atau 595 teks diklasifikasikan sebagai data negatif. 27,88% teks atau 356 teks diklasifikasikan sebagai data positif dan 35,53% teks atau 326 teks diklasifikasikan sebagai data netral. Gbr 4. menunjukkan hasil pelabelan.



Gbr 4. Hasil Pelabelan

Setelah diberikan label, dilakukan *transformation* dengan memberikan bobot kata menggunakan TF-IDF. Tabel III menunjukkan contoh hasil TF-IDF.

TABEL III
CONTOH HASIL TF-IDF

Kata	Skor
tapera	0.133857
bayar	0.472877
kerja	0.136266

Setelah diberikan bobot, dilakukan *feature selection* dengan menggunakan *chi-square*. Dari 3490 fitur yang ada, dipilih sebanyak 2000 fitur untuk digunakan pada perhitungan chi^2 antara setiap fitur dengan target. Tabel IV menunjukkan contoh fitur terbaik dari hasil *feature selection*.

TABEL IV
CONTOH HASIL *FEATURE SELECTION*

Kata	Skor
rumah	28.127681
ikut	13.738677
tapera	13.210902

D. Data Mining

Data mining dilakukan dengan mengklasifikasikan data ke dalam algoritma naïve bayes dan SVM untuk menghasilkan model yang diinginkan. Data yang telah dilakukan *transformation* dibagi ke dalam 3 rasio yakni 80:20, 70:30 dan 60:40. Data juga diolah menggunakan SMOTE untuk meningkatkan performa model. Tabel V menunjukkan jumlah kelas sentimen pada data *training* sebelum dan setelah dilakukan SMOTE.

TABEL V
KELAS SENTIMEN SEBELUM DAN SETELAH SMOTE

Rasio	Sebelum SMOTE			Setelah SMOTE		
	+	-	=	+	-	=
80:20	128	65	63	467	467	467
70:30	186	101	97	409	409	409
60:40	253	135	123	342	342	342

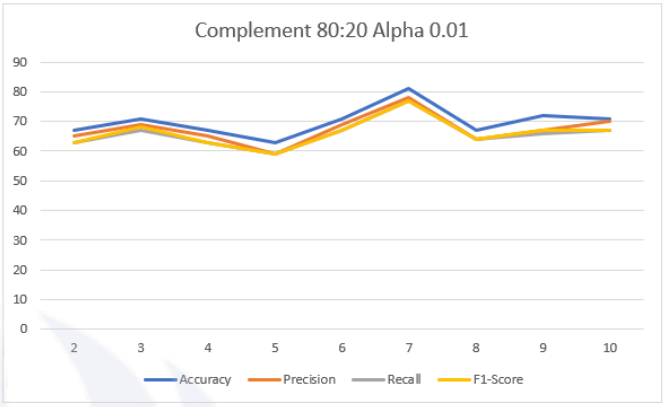
1) Naïve Bayes: Pada naïve bayes, pemodelan dilakukan dengan menggunakan 3 model, yakni Multinomial, Complement dan Gaussian. Pada model Multinomial dan Complement menggunakan parameter alpha 0.01, 0.1 dan 1. Sedangkan pada model Gaussian menggunakan parameter var_smoothing 0.01, 0.1 dan 1.

2) Support Vector Machine: Pada SVM, pemodelan dilakukan dengan menggunakan parameter C dan kernel. Dimana C merupakan parameter regularisasi dan kernel merupakan tipe kernel yang akan digunakan. Nilai C yang digunakan yakni 1, 2 dan 3. Sedangkan kernel yang digunakan yakni linear, rbf dan poly.

E. Evaluation

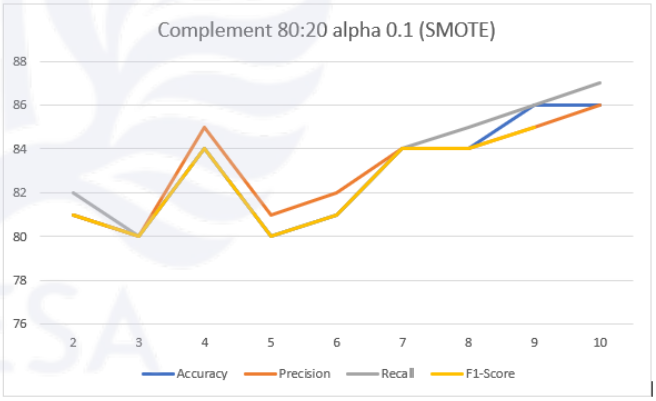
Evaluation dilakukan dengan menggunakan *confusion matrix* dan menghitung nilai *accuracy, precision, recall* dan *f1-score* dari hasil hasil *k-fold cross validation*.

1) Naïve Bayes: Dari hasil evaluasi algoritma naïve bayes menggunakan *k-fold cross validation* didapatkan bahwa model Complement dengan rasio 80:20 dan parameter alpha 0.01 merupakan model dengan nilai *accuracy* paling baik diantara model yang lain sebelum SMOTE. Nilai *accuracy* model tersebut yakni 81% dengan *precision* 78%, *recall* 77% dan *f1-score* 77%. Nilai tersebut didapatkan dari nilai *k* terbaik yakni *k*=7. Gbr 5. menunjukkan hasil *k-fold cross validation* pada algoritma naïve bayes sebelum SMOTE.



Gbr 5. *k-Fold Cross Validation* Naive Bayes Sebelum SMOTE

Sedangkan setelah SMOTE, model Complement dengan rasio 80:20 dan parameter alpha 0.1 merupakan model dengan nilai *accuracy* terbaik dengan nilai *accuracy* 89%, *precision*, *recall* dan *f1-score* 88%. Nilai tersebut didapatkan dari nilai *k* terbaik yakni *k*=10. Gbr 6. menunjukkan hasil *k-fold cross validation* pada algoritma naïve bayes setelah SMOTE.



Gbr 6. *k-Fold Cross Validation* Naive Bayes Setelah SMOTE

Untuk model dengan nilai *accuracy* paling rendah sebelum dilakukan SMOTE yakni model Gaussian dengan rasio 80:20 dan alpha 1 dengan nilai *accuracy* 54%, *precision* 84%, *recall* 38% dan *f1-score* 32%. Sedangkan, untuk model dengan nilai *accuracy* paling rendah setelah dilakukan SMOTE yakni model Gaussian dengan rasio 60:40 dan alpha 1 dengan nilai *accuracy* 42%, *precision* 80%, *recall* 36% dan *f1-score* 25%. Tabel VI dan VII menunjukkan hasil evaluasi naïve bayes sebelum dan setelah dilakukan SMOTE.

TABEL VI
EVALUATION NAÏVE BAYES SEBELUM SMOTE

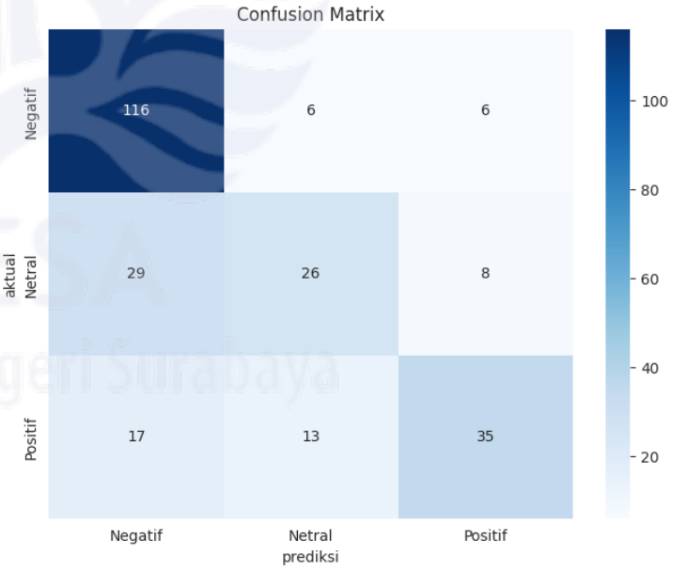
Rasio	Model	*Accuracy*	*Precision*	*Recall*	*F1_score*

80:20	Multinomial	0.01	75%	71%	67%	67%
		0.1	75%	72%	65%	67%
		1	63%	81%	50%	46%
	Complement	0.01	81%	78%	77%	77%
		0.1	79%	76%	74%	75%
		1	79%	76%	72%	73%
	Gaussian	0.01	69%	67%	65%	66%
		0.1	68%	71%	60%	61%
		1	54%	84%	38%	32%
70:30	Multinomial	0.01	71%	67%	63%	64%
		0.1	70%	72%	61%	63%
		1	62%	66%	49%	47%
	Complement	0.01	75%	69%	68%	69%
		0.1	75%	70%	69%	69%
		1	76%	73%	67%	69%
	Gaussian	0.01	69%	67%	62%	62%
		0.1	68%	70%	58%	59%
		1	56%	51%	42%	37%
60:40	Multinomial	0.01	72%	71%	69%	69%
		0.1	70%	80%	61%	64%
		1	60%	73%	51%	49%
	Complement	0.01	74%	75%	68%	69%
		0.1	75%	75%	69%	70%
		1	72%	73%	69%	70%
	Gaussian	0.01	68%	66%	64%	65%
		0.1	64%	60%	54%	54%
		1	60%	79%	45%	44%

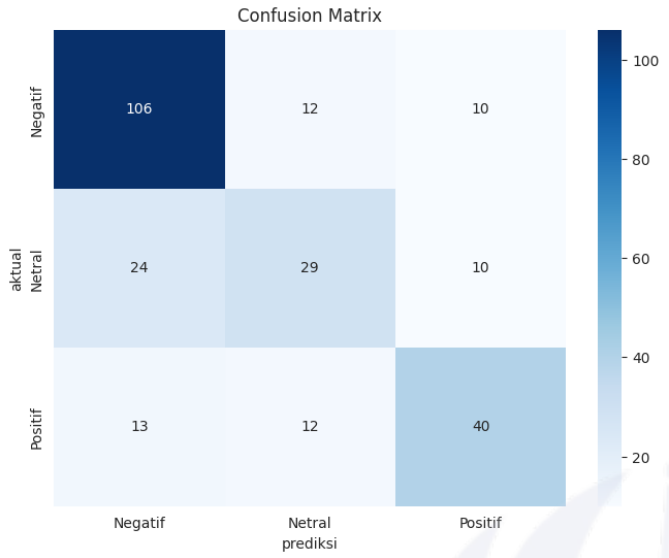
TABEL VII
EVALUATION NAÏVE BAYES SETELAH SMOTE

Rasio	**Model**		***Accuracy***	***Precision***	***Recall***	***F1_score***
80:20	Multinomial	0.01	88%	88%	88%	88%
		0.1	87%	88%	88%	87%
		1	85%	85%	85%	85%
	Complement	0.01	86%	86%	87%	86%
		0.1	86%	86%	87%	86%
		1	87%	88%	88%	87%
	Gaussian	0.01	84%	84%	84%	84%
		0.1	67%	76%	64%	65%
		1	46%	80%	41%	34%
70:30	Multinomial	0.01	87%	87%	87%	87%
		0.1	86%	86%	86%	86%
		1	86%	87%	86%	86%
	Complement	0.01	86%	86%	86%	86%
		0.1	86%	86%	85%	85%
		1	86%	87%	86%	85%
	Gaussian	0.01	83%	85%	85%	84%
		0.1	72%	77%	75%	73%
		1	47%	48%	36%	26%
60:40	Multinomial	0.01	85%	86%	86%	86%
		0.1	87%	87%	87%	87%
		1	86%	87%	86%	86%
	Complement	0.01	87%	87%	87%	87%
		0.1	89%	88%	88%	88%
		1	87%	88%	87%	87%
	Gaussian	0.01	85%	86%	84%	85%
		0.1	69%	78%	71%	69%
		1	42%	80%	36%	25%

Pada Gbr 7. Dan Gbr 8. menunjukkan *confusion matrix* dari model terbaik pada algoritma naïve bayes sebelum dan setelah dilakukan SMOTE. Dari gambar tersebut dapat diketahui bahwa nilai *True Negative* baik sebelum dan setelah dilakukan SMOTE merupakan yang paling tinggi dibanding nilai *True Positive* dan *True Neutral.*

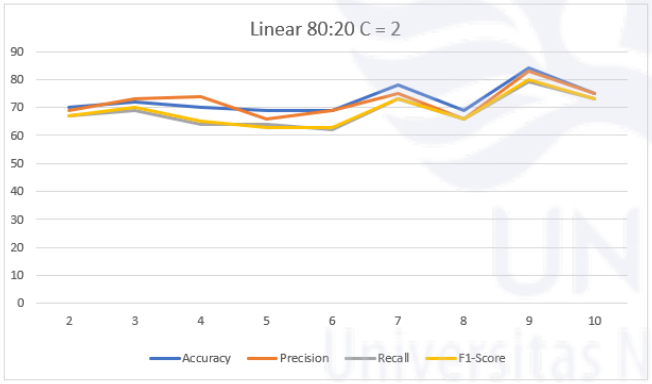


Gbr 7. *Confusion Matrix* Naive Bayes Sebelum SMOTE



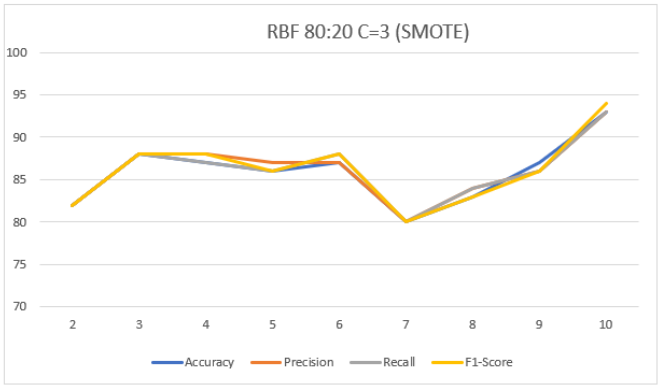
Gbr 8. *Confusion Matrix* Naive Bayes Setelah SMOTE

2) Support Vector Machine: Dari hasil evaluasi algoritma SVM menggunakan *k-fold cross validation* didapatkan bahwa rasio 80:20 dan parameter kernel linear, C=2 merupakan model dengan nilai *accuracy* paling baik diantara model yang lain sebelum SMOTE. Nilai *accuracy* model tersebut yakni 84% dengan *precision* 83%, *recall* 79% dan *f1-score* 80%. Nilai tersebut didapatkan dari nilai *k* terbaik yakni *k*=9. Gbr 9. menunjukkan hasil *k-fold cross validation* pada algoritma SVM sebelum SMOTE.



Gbr 9. *k-Fold Cross Validation* SVM Sebelum SMOTE

Sedangkan setelah SMOTE, model dengan rasio 80:20 dan parameter kernel rbf dan C=3 merupakan model dengan nilai *accuracy* terbaik dengan nilai *accuracy* 93%, *precision* 93%, *recall* 93% dan *f1-score* 94%. Nilai tersebut didapatkan dari nilai *k* terbaik yakni *k*=10. Gbr 10. menunjukkan hasil *k-fold cross validation* pada algoritma SVM setelah SMOTE.



Gbr 10. *k-Fold Cross Validation* SVM Sebelum SMOTE

Untuk model dengan nilai *accuracy* paling rendah sebelum dilakukan SMOTE yakni model dengan rasio 60:40 dan parameter kernel poly, C=1 dengan nilai *accuracy* 55%, *precision* 79%, *recall* 43% dan *f1-score* 39%. Sedangkan, untuk model dengan nilai *accuracy* paling rendah setelah dilakukan SMOTE yakni model dengan rasio 80:20 dan parameter kernel poly, C=1 dengan nilai *accuracy* 75%, *precision* 83%, *recall* 74% dan *f1-score* 75%. Tabel VIII dan IX menunjukkan hasil evaluasi SVM sebelum dan setelah dilakukan SMOTE.

TABEL VIII
EVALUATION SVM SEBELUM SMOTE

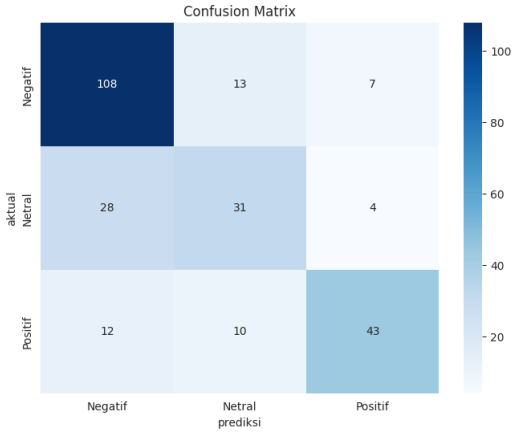
Rasio	Model		*Accuracy*	*Precision*	*Recall*	*F1_score*
80:20	Linear	1	80%	76%	74%	75%
		2	84%	83%	79%	80%
		3	80%	83%	73%	75%
	Rbf	1	82%	79%	76%	77%
		2	78%	74%	74%	74%
		3	78%	75%	73%	73%
	Poly	1	57%	84%	41%	38%
		2	58%	85%	42%	40%
		3	58%	79%	42%	40%
70:30	Linear	1	72%	73%	64%	66%
		2	77%	76%	71%	72%
		3	72%	73%	64%	66%
	Rbf	1	76%	75%	67%	68%
		2	76%	77%	75%	73%
		3	77%	785	76%	75%
	Poly	1	56%	84%	41%	37%
		2	56%	78%	41%	36%
		3	57%	85%	41%	38%
60:40	Linear	1	72%	70%	71%	69%
		2	76%	78%	73%	73%
		3	72%	70%	71%	69%
	Rbf	1	75%	77%	65%	64%
		2	79%	80%	72%	73%

		3	81%	79%	77%	77%
	Poly	1	55%	79%	43%	39%
		2	60%	80%	46%	43%
		3	60%	80%	46%	43%

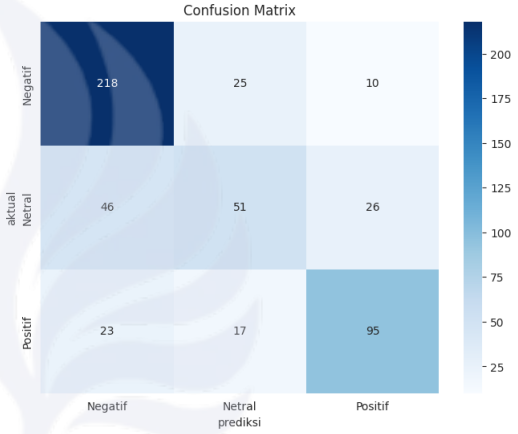
TABEL IX
EVALUATION SVM SETELAH SMOTE

Rasio	**Model**		***Accuracy***	***Precision***	***Recall***	***F1_score***
80:20	Linear	1	91%	91%	91%	91%
		2	91%	90%	90%	91%
		3	91%	91%	91%	91%
	Rbf	1	91%	91%	92%	91%
		2	92%	92%	93%	92%
		3	93%	93%	93%	94%
	Poly	1	75%	83%	74%	75%
		2	79%	84%	78%	79%
		3	81%	84%	81%	82%
70:30	Linear	1	87%	87%	88%	87%
		2	89%	88%	89%	88%
		3	87%	87%	88%	87%
	Rbf	1	89%	90%	89%	89%
		2	91%	91%	91%	91%
		3	90%	90%	90%	90%
	Poly	1	76%	87%	76%	77%
		2	78%	85%	79%	79%
		3	81%	84%	82%	82%
60:40	Linear	1	87%	88%	87%	87%
		2	89%	89%	89%	89%
		3	87%	88%	87%	87%
	Rbf	1	90%	90%	89%	90%
		2	89%	89%	89%	89%
		3	89%	89%	88%	88%
	Poly	1	77%	82%	76%	76%
		2	78%	82%	77%	77%
		3	80%	82%	80%	81%

Pada Gbr 11. Dan Gbr 12. menunjukkan *confusion matrix* dari model terbaik pada algoritma SVM sebelum dan setelah dilakukan SMOTE. Dari gambar tersebut dapat diketahui bahwa nilai *True Negative* baik sebelum dan setelah dilakukan SMOTE merupakan yang paling tinggi dibanding nilai *True Positive* dan *True Neutral.*



Gbr 11. *Confusion Matrix* SVM Sebelum SMOTE



Gbr 12. *Confusion Matrix* SVM Setelah SMOTE

IV. KESIMPULAN

Dari hasil pelabelan yang dilakukan dengan menggunakan *InSet Lexicon* diperoleh bahwa sebanyak 46.59% teks atau 595 teks diklasifikasikan sebagai data negatif. 27,88% teks atau 356 teks diklasifikasikan sebagai data positif dan 35,53% teks atau 326 teks diklasifikasikan sebagai data netral. Dari hasil klasifikasi menggunakan algoritma naïve bayes dan SVM diperoleh bahwa nilai *True Negative* pada *confusion matrix* merupakan yang paling besar pada kedua algoritma tersebut. Hal ini menunjukkan bahwa sentimen masyarakat terhadap kebijakan Tapera cenderung negatif.

Dari hasil klasifikasi juga menunjukkan bahwa algoritma SVM memiliki akurasi yang lebih baik daripada algoritma naïve bayes. Sebelum SMOTE, SVM memiliki akurasi 84% pada rasio 80:20 dengan kernel linear dan C=2, sedangkan Naïve Bayes memiliki akurasi 81% pada rasio 80:20 dengan model Complement dan alpha 0.01. Setelah SMOTE, SVM memiliki akurasi 93% pada rasio 80:20 dengan kernel rbf dan C=3, sedangkan Naïve Bayes memiliki akurasi 89% pada rasio 60:40 dengan model Complement dan alpha 0.1.

V. SARAN

Berdasarkan hasil penelitian ini, saran yang dapat disampaikan yaitu:

1) Penelitian selanjutnya dapat menggunakan algoritma lain seperti *K-Nearest Neighbor, Random Forest* atau lainnya untuk mendapatkan performa yang lebih baik.

2) Pemerintah dapat mengevaluasi kebijakan yang dibuat dengan melihat sentimen yang diberikan masyarakat terhadap kebijakan tersebut, terutama kebijakan mengenai Tapera.

REFERENSI

[1] W. Ningsih, B. Alfianda, R. Rahmaddeni, and D. Wulandari, "Perbandingan Algoritma SVM dan Naïve Bayes dalam Analisis Sentimen Twitter pada Penggunaan Mobil Listrik di Indonesia," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 556–562, Feb. 2024, doi: 10.57152/malcom.v4i2.1253.

[2] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 4, no. 2, p. 113, Dec. 2018, doi: 10.26418/jp.v4i2.27526.

[3] F. S. Pamungkas and I. Kharisudin, "Analisis Sentimen dengan SVM," *PRISMA : Prosiding Seminar Nasional Matematika*, vol. 4, pp. 628–634, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[4] N. Dewi and R. E. Putra, "Pengembangan Sistem Analisis Sentimen Masyarakat Terhadap Chatgpt Pada Twitter Dengan Perbandingan Metode Naive Bayes Classifier Dan K-Nearest Neighbors," *JINACS : Journal of Informatics and Computer Science*, vol. 05, no. 03, 2023.

[5] A. Felicia Watratan, A. B. Puspita, D. Moeis, S. Informasi, and S. Profesional Makassar, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," 2020. [Online]. Available: http://journal.isas.or.id/index.php/JACOST

[6] S. Syafrizal, M. Afdal, and R. Novita, "Analisis Sentimen Ulasan Aplikasi PLN Mobile Menggunakan Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 10–19, Dec. 2023, doi: 10.57152/malcom.v4i1.983.

[7] A. Harun and D. P. Ananda, "Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve Bayes dan Decission Tree," *MALCOM: Indonesian Journal of Machine Learning and Computer Science* , vol. 1, pp. 58–63, 2021.

[8] A. R. Isnain, A. I. Sakti, D. Alita, and N. S. Marga, "Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma SVM," *Jurnal Data Mining dan Sistem Informasi*, vol. 2, no. 1, p. 31, Feb. 2021, doi: 10.33365/jdmsi.v2i1.1021.

[9] Oryza Habibie Rahman, Gunawan Abdillah, and Agus Komarudin, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 17–23, Feb. 2021, doi: 10.29207/resti.v5i1.2700.

[10] F. Nurrachmat Hidayat and Sugiyono, "Analisis Sentimen Masyarakat Terhadap Perekrutan PPPK Pada Twitter Dengan Metode Naive Bayes Dan Support Vector Machine," *Jurnal Sains dan Teknologi*, vol. 5, no. 2, pp. 665–672, 2023, doi: 10.55338/saintek.v5i1.1359.

[11] A. Ardiansyah and Wahyudin, "Analisis Sentimen Pada Ulasan Aplikasi Home Credit dengan Metode SVM dan KNN," *Jurnal Komputer Antartika*, vol. 1, no. 4, 2023, [Online]. Available: https://ejournal.mediaantartika.id/index.php/jka

[12] Y. A. Singgalen, "Analisis Sentimen Top 10 Traveler Ranked Hotel di Kota Makassar Menggunakan Algoritma Decision Tree dan Support Vector Machine," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 1, pp. 323–332, 2023, doi: 10.30865/klik.v4i1.1153.

[13] N. M. Farhan and B. Setiaji, "Komparasi Metode Naive Bayes dan SVM pada Sentimen Twitter Mengenai," *Indonesian Journal of Computer Science Attribution*, vol. 12, no. 5, 2023.

[14] D. A. Kristiyanti, D. A. Putri, E. Indrayuni, A. Nurhadi, and A. H. Umam, "E-Wallet Sentiment Analysis Using Naïve Bayes and Support Vector Machine Algorithm," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1742-6596/1641/1/012079.

[15] P. Ayar and C. Sandip, "Efficient Focused Web Crawling Approach for Search Engine," *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 5, pp. 545–551, 2015.

[16] H. Satria, "Tweet-harvest," https://github.com/helmisatria/tweet-harvest.

[17] A. Dwi Indriyanti and P. N. Fadhilah, "Analisis Sentimen terhadap Opini Publik Mengenai Childfree dalam Pernikahan pada Twitter Menggunakan K-Nearest Neighbor (K-NN)," *Journal of Informatics and Computer Science*, vol. 05, 2023.

[18] G. Paltoglou and M. Thelwall, "A study of Information Retrieval weighting schemes for sentiment analysis," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 11–16.

[19] A. P. P. Wardani, A. Adiwijaya, and M. D. Purbolaksono, "Sentiment Analysis on Beauty Product Review Using Modified Balanced Random Forest Method and Chi-Square," *Journal of Information System Research (JOSH)*, vol. 4, no. 1, pp. 1–7, Oct. 2022, doi: 10.47065/josh.v4i1.2047.

[20] Y. Irawan, "Penerapan Data Mining Untuk Evaluasi Data Penjualan Menggunakan Metode Clustering dan Algoritma Hierarki Divisive," *JTIULM*, vol. 04, no. 1, 2019.

[21] R. A. Nurdian, Mujib Ridwan, and Ahmad Yusuf, "Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 1, Apr. 2022, doi: 10.28932/jutisi.v8i1.4004.