ANALISIS SENTIMEN PEMBELIAN BAHAN BAKAR MINYAK PADA APLIKASI MyPertamina DENGAN METODE *NAIVE BAYES CLASSIFIER* DAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE*

I Komang Damai Armawan[1)], Mochamad Husni[2)], Tubagus M. Akhriza[3)]
STMIK PPKIA Pradnya Paramita Malang
komang.armawan.27@gmail.com[1)], husni@stimata.ac.id[2)], akhriza@stimata.ac.id[3)]

Abstract

The implementation of the MyPertamina application policy for subsidized fuel purchases has received various responses from the public, expressed through social media. These responses can be classified into neutral, positive, and negative feedback. Manual analysis can be time-consuming, so the Naive Bayes Classifier (NBC) method can be used for quick and accurate sentiment analysis of public responses to the implementation of the MyPertamina application for subsidized fuel purchases. The research aims to analyze sentiment using the NBC method and implement the Synthetic Minority Oversampling Technique (SMOTE) on the application of MyPertamina for subsidized fuel purchases in the community. The dataset in this study is divided into three ratios: 30% for testing set, 40%, and 50%. The sentiment analysis results using the NBC and SMOTE classification methods with a 30% training set ratio show the best outcome. Initially, there were 972 data points, which were preprocessed to 712, and then the SMOTE algorithm was implemented to balance the training set. The results showed that 38% were neutral responses, 35% were positive, and 27% were negative, with an accuracy of 84%.

Keywords: *Natural Language Processing, Klasifikasi, Naive Bayes Classifier (NBC), Synthetic Minority Oversampling Technique (SMOTE)*

1. PENDAHULUAN

Kebijakan pengendalian subsidi BBM diterapkan pemerintah melalui PT Pertamina dengan sistem aplikasi MyPertamina. Pengendalian BBM Subsidi terus didorong oleh pemerintah agar pemberian subsidi benar-benar dinikmati oleh orang yang berhak. Aplikasi MyPertamina akan membantu pemerintah untuk mengendalikan penggunaan energi bersubsidi secara lebih baik, termasuk mengetahui secara pasti siapa dan berapa jumlah pengguna BBM subsidi di Indonesia (Sindo, 2022).

MyPertamina adalah aplikasi layanan keuangan digital yang dikembangkan oleh PT Pertamina, memiliki fungsi yang hampir sama dengan platform layanan keuangan lain seperti Ovo, Dana, ShopeePay, dan sebagainya. Pengguna atau konsumen bisa melakukan beberapa pembelian produk buatan Pertamina secara nontunai. Layanan pembayaran nontunai pada aplikasi MyPertamina beroprasi dengan dukungan dari platform LinkAja. Pengguna diharuskan memiliki akun MyPertamina dahulu sebelum bisa menautkannya dengan akun LinkAja. Sesuai kebijakan pemerintah yang baru mengenai pengendalian Subsidi BBM, MyPertamina akan digunakan sebagai syarat untuk bisa membeli BBM bersubsidi. Pengguna diharuskan mendaftarkan diri di aplikasi MyPertamina, kemudian data tersebut akan diverifikasi oleh BPH Migas untuk memastikan kelayakan pengguna menerima subsidi BBM (Safitri, 22).

Penerapan kebijakan aplikasi MyPertamina mendapat respon yang beragam dari publik, yang dituangkan kedalam media sosial. Topik MyPertamia banyak dibicarakan oleh masyarakat melalui media sosial Twitter, dengan 10.500 cuitan (DPR, 2022). Dari jumlah cuitan tersebut, terdapat berbagai tanggapan yang muncul, mulai dari netral, positif, maupun negatif. Tanggapan positif masyarakat ini cenderung tidak seimbang dengan tanggapan negatif, sehingga menyebabkan penerapan MyPertamina untuk pembelian BBM Subsidi ini terkesan sangat buruk. Permasalahan inilah yang menyebabkan perlu dilakukan analisis sentimen publik untuk bisa mengetahui data jelas dari tanggapan tersebut.

2. KAJIAN LITERATUR

2.1. *ARTIFICIAL INTELLIGENCE*

Artificial Intelligence (AI) adalah istilah yang digunakan untuk menyebut yang menyiratkan penggunaan komputer untuk memodelkan perilaku cerdas, yang terintegrasi dengan big data, yang menyerupai kecerdasan manusia (Pabubung, 2021). Istilah AI diciptakan oleh McCarthy pada 1950-an dan mengacu pada cabang ilmu komputer dimana algoritma dikembangkan untuk meniru fungsi kognitif manusia, seperti belajar, penalaran, dan pemecahan masalah. AI juga adalah istilah yang mencakup secara luas, tetapi tidak terbatas pada, Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), dan Computer Vision (CV) (Yin, Ngiam, & Teo, 2021).

AI merupakan sebuah perkembangan dari teknologi informasi yang terus berkembang dalam sepuluh tahun terakhir. Pemanfaatan AI sangat luas dan tidak hanya terbatas di sektor industri telekomunikasi, namun juga pada berbagai sektor seperti perbankan, manufaktur, jasa, analisis, bahkan di sektor pemerintah. Di beberapa negara, implementasi AI sudah mencapai hampir 56%, terutama pada sektor industri (Ririh, Laili, Wicaksono, & Tsurayya, 2020).

2.2. *DATA MINING*

Data mining adalah bagaimana mencari data yang tersedia untuk menciptakan sebuah model, lalu memanfaatkan model tersebut untuk mengenali pola data lain yang tidak tersedia di dalam basis data yang tersimpan. Dalam data mining pengelompokan data dapat juga dilakukan untuk mengetahui pola secara universal dari data yang tersedia agar dilakukan langkah tindak lanjut lainnya yang berguna sebagai pendukung kegiatan dan tujuan akhir tertentu (Utomo & Purba, 2019). Penjelasan lain mengatakan Data mining merupakan proses iteratif dan interaktif untuk menemukan pola atau model baru yang sempurna, bermanfaat dan dapat dimengerti dalam suatu basis data yang sangat besar (massive database) (Sikumbang, 2018). Dapat disimpulkan bahwa data mining adalah sebuah metode yang digunakan untuk menemukan pola dari kumpulan data, kemudian pola tersebut dimanfaatkan untuk proses analisa dari kumpulan data yang bersifat besar yang bertujuan mendapatkan suatu pengetahuan.

Tahapan proses dalam penggunaan data mining merupakan salah satu dari rangkaian Knowledge Discovery in Database (KDD). KDD adalah sebuah kegiatan untuk mengekstrak suatu pengetahuan dari sekumpulan data, hasil data mining kemudian diubah menjadi informasi yang mudah dipahami (Afdal & Rosadi, 2019).

2.3. *MACHINE LEARNING*

Machine Learning (ML) adalah aplikasi atau bagian dari AI yang membuat sistem memiliki kemampuan belajar secara otomatis dan meningkatkan kemampuannya berdasarkan pengalaman tanpa diprogram secara eksplisit. Dalam hal ini, program komputer tidak ditulis secara statis. Fokus ML terdapat pada pengembangan program komputer yang dapat mengakses data dan belajar dari data tersebut. ML juga dapat didefinisikan sebagai algoritma yang bertujuan menemukan dan mengaplikasikan pola-pola di dalam data. Algoritma pada ML menggunakan teknik-teknik statistik untuk menemukan pola-pola tersebut (Kusuma, 2020). ML bisa diartikan bahwa adalah sebuah cara untuk mesin atau komputer untuk bisa bertindak pintar dan belajar dari data yang dimasukkan dan dengan demikian mesin tersebut bisa bertindak dengan “pintar” untuk menyelesaikan suatu tugas yang diberikan (Chinnamgari, 2019).

2.4. *WEB SCRAPING*

Web scraping adalah suatu proses yang melibatkan pengambilan dokumen semi-terstruktur dari internet, umumnya halaman web berupa halaman dalam bahasa markup seperti HyperText Markup Language (HTML) atau Extensible HyperText Markup Language (XHTML), dan kemudian dokumen tersebut dianalisis untuk mengekstraksi data tertentu yang dapat digunakan dalam konteks lain. Web scraping juga dikenal sebagai screen scraping (Mufidah & Siahaan, 2021).

Pada dasarnya teknik web scraping adalah suatu proses copy-paste dari suatu web namun dilakukan secara otomatis dan terorganisir. Teknik web scraping hanya mengambil data yang sudah tersaji dalam suatu web, bukan secara ilegal dengan langsung masuk kedalam database pemilik

web. Dengan demikian, baik pemilik web masih tetap mendapatkan traffics yang dapat meningkatkan valuasinya, sementara pihak scraper mendapatkan data yang dibutuhkan (Saurkar, Pathare, & Gode, 2018).

2.5. *TWITTER*

Twitter adalah layanan jejaring sosial yang memungkinkan penggunanya untuk memposting teks, gambar dan video yang dikenal dengan sebutan kicauan (tweet). Twitter didirikan oleh Jack Dorsey pada bulan maret 2006, dan situs jejaring sosialnya diluncurkan pada bulan juli. Sejak diluncurkan, Twitter telah menjadi salah satu dari sepuluh situs yang paling sering dikunjungi di Internet (D'Monte, 2009). Pengguna Twitter di tahun 2023 berdasarkan laporan dari We Are Social dan Hootsuite, terdapat 556 juta pengguna Twitter di seluruh dunia. Jumlah tersebut meningkat 27,4% dibandingkan pada periode yang sama tahun sebelumnya. Indonesia menempati urutan ke-5 untuk pengguna Twitter terbanyak, dengan jumlah pengguna aktif sekitar 24 juta (Santika, 2023).

2.6. *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)*

TF-IDF merupakan algoritma yang digunakan untuk melakukan pembobotan sebuah kata dalam dataset. Penggunaan algoritma ini bertujuan untuk membuat sebuah vektor dengan banyak term sehingga tiap kata yang ada dalam dataset bisa dikenali dan dihitung sebagai satu fitur. Perhitungan TF-IDF terbagi menjadi dua bagian, yaitu Term Frequency (TF) dan Inverse Document Frequency (IDF). TF digunakan untuk menghitung jumlah kemunculan kata dalam sebuah dataset, sedangkan IDF merupakan jumlah data yang berisi term yang dicari (Darwis, Pratiwi, & Pasaribu, 2020).

2.7. *SYNTHETICS MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)*

SMOTE adalah sebuah metode yang pertama kali diusulkan oleh Chawla pada tahun 2002. Metode ini bertujuan untuk melakukan oversamping pada kelas minoritas dan membuat data training sintetik. Imbalance dataset adalah sebuah kondisi dimana perbandingan masing-masing kelas di dataset yang tidak seimbang. Secara umum ada dua cara untuk mengatasi imbalance dataset, yaitu bisa dengan Random Undersampling (RUC) dan Random Oversampling (ROC). ROS akan menduplikasi secara acak data dari kelas yang minoritas, secara umum oversampling lebih baik dari undersampling. Oversampling sendiri adalah sebuah cara untuk menaikan jumlah kelas minoritas sampai setara dengan kelas mayoritas. ROC bisa menjadi salah satu cara yang cukup baik untuk membantu proses penyeimbangan dataset, namun kemungkinan hasil dari klasifikasi nantinya mengalami overfitting karena metode ini membuat salinan/duplikat yang sama persis dari data yang berasal dari kelas minoritas (Amalia & Wahyuni, 2020).

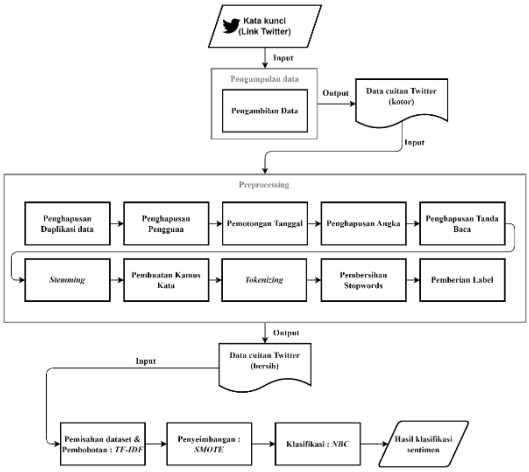
SMOTE sendiri adalah sebuah metode yang mirip dengan ROC namun data sintetik yang dibuat tidak hanya asal duplikat, melainkan dibuat dengan menggunakan konsep nearest neighbour. Data sintetik yang dibuat dengan metode SMOTE dihitung dari jarak kedekatan fiturnya. Jumlah tetangga terdekat dari k yang dipilih menyesuaikan dengan jumlah data sintetik yang dibutuhkan. Data sintetik dibuat dengan mengambil perbedaan antara vektor fitur yang dipertimbangkan dan tetangga terdekatnya dan mengalikan perbedaan ini dengan angka acak antara 0 dan 1. Data sintetik yang dihasilkan menyebabkan wilayah keputusan kelas minoritas menjadi lebih umum, yang mana mengarah ke generalisasi yang lebih baik untuk proses klasifikasi nantinya (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

2.8. *NAÏVE BAYES CLASSIFIER (NBC)*

Algoritma Naïve Bayes Classifier (NBC) adalah algoritma klasifikasi berdasarkan probabilitas dalam statistik yang diusulkan oleh Thomas Bayes yang digunakan untuk memprediksi peluang masa depan berdasarkan peluang masa lalu (Teorema Bayes). Algoritma tersebut kemudian dikombinasikan dengan “naiv” dimana kondisi antara atribut satu sama lain tidak terikat satu sama lain (Pratmanto, Rousyati, Wati, Widodo, & Suleman, 2020).

3. METODE PENELITIAN

3.1. KERANGKA KERJA PENELITIAN



Gambar 1 *Framework*

3.2. KATA KUNCI

Kata kunci yang digunakan untuk pencarian data di fitur pencarian *Twitter* terdiri dari dua belas kata kunci. Tabel 1 berikut ini menunjukan kata kunci yang digunakan:

Tabel 1 Kata Kunci Dan Link

No	Kata Kunci/Tagar	Link
1	Beli BBM pakai MyPertamina	https://*Twitter*.com/search?q=Beli%20BBM%20pakai%20MyPertamina&src=typed_query
2	MyPertamina	https://*Twitter*.com/search?q=MyPertamina&src=typed_query&f=top
3	BBM MyPertamina	https://*Twitter*.com/search?q=BBM%20MyPertamina&src=typed_query&f=top
4	Pertalite tepat sasaran MyPertamina	https://*Twitter*.com/search?q=Pertalite%20tepat%20sasaran%20MyPertamina&src=typed_query&f=top
5	BBM tepat sasaran MyPertamina	https://*Twitter*.com/search?q=BBM%20tepat%20sasaran%20MyPertamina&src=typed_query&f=top
6	Subsidi BBM tepat sasaran MyPertamina	https://*Twitter*.com/search?q=Subsidi%20BBM%20tepat%20sasaran%20MyPertamina&src=typed_query&f=top
7	Subsidi BBM tepat sasaran	https://*Twitter*.com/search?q=Subsidi%20BBM%20tepat%20sasaran&src=typed_query&f=top
8	Subsidi MyPertamina	https://*Twitter*.com/search?q=Subsidi%20MyPertamina&src=typed_query&f=top
9	Beli pertalite pakai MyPertamina	https://*Twitter*.com/search?q=Beli%20pertalite%20pakai%20MyPertamina&src=typed_query&f=top
10	Beli bensin MyPertamina	https://*Twitter*.com/search?q=Beli%20bensin%20MyPertamina&src=typed_query&f=top
11	#MyPertaminaUnFaedah	https://*Twitter*.com/search?q=%23MyPertaminaUnFaedah&src=typed_query&f=top
12	Pertalite MyPertamina	https://*Twitter*.com/search?q=Pertalite%20MyPertamina&src=typed_query&f=top

3.3. PENGAMBILAN DATA

Pengambilan data dalam penelitian ini akan memiliki satu proses yaitu *Scraping*. Tahap ini menggunakan teknik *web scraping* yang dibantu dengan tool *Apify*. Link yang telah dikumpulkan kemudian ditempelkan satu persatu ke *form* Link *Twitter* di *Console Apify*.

3.3. *PREPROCESSING*

Data yang telah didapatkan kemudian akan melalui rangkaian proses *preprocessing* dimulai dengan proses penghapusan duplikasi data. kemudian dilanjutkan dengan proses penghapusan pengguna yang terdiri dari adalah akun “MyPertamina”, “LinkAja Syariah”, “LinkAja”, “Pertamina Papua Maluku”, “Pertamax Series ID”, “CNN Indonesia”, “kumparan”, “KOMPAS TV”, “Kompas.com”, “LIPUTAN6”, “detikcom”, “DetikFinance”, “Detik jabar”, “detik jatim”, “detikoto”, “detikinet”, “Jawa Pos”, “tempo.co”, dan “brightgas”. Proses selanjutnya adalah pemotongan data berdasarkan tanggal yang dimulai dari “2022-5-25” dan berakhir pada “2022-12-31”. Setelah rangkaian proses tersebut, data kemudian melalui proses penghapusan angka, tanda baca, *stemming*, pembuatan kamus kata, *tokenizing*, pembersihan *stopwords*, dan *labeling*.

3.4. *SPLITTING* DAN TF-IDF

Dataset yang sudah bersih kemudian dipisahkan menjadi *training set* dan *testing set*. Jumlah *testing set* akan dibuat menjadi 30%, 40%, dan 50%. Setiap *testing set* tersebut kemudian akan diuji untuk mengetahui rasio terbaik untuk model dibuat. Setelah proses tersebut *training set* dan *testing set* kemudian dibobotkan dengan menggunakan TF-IDF.

3.5. NBC

NBC digunakan untuk memprediksi kelas pada *training set* dengan perhitungan yang telah ditentukan. Contoh perhitungan akan dilakukan dengan lima data, yang dimana satu data berperan sebagai *testing set* dan empat data berperan sebagain *training set*. Empat data dalam *training set* memiliki label masing-masing satu dari data tersebut memiliki label netral, satu positif, dan dua negatif serta satu data yang akan diprediksi seperti pada Tabel 2 berikut:

Tabel 2 *Training Set* Dan *Testing Set* Untuk Perhitungan NBC

Dokumen ke-	Isi Dokumen	Label
1	kenapa bijak negeri lucu lucu beli bensin wajib pakai mypertamina masuk wilayah spbu petugas larang hidup hp bijak lucu tidak bijak cerdas	Negatif
2	pernah beli bensin pakai mypertamina meledak	Negatif
3	spbu solo bisa bayar pakai mypertamina mana	Netral
4	bagus bijak distribusi bensin subsidi mypertamina dukung bayar cashless psimis antri lama semoga mekanisme	Positif
5	bisa pakai banyak ewallet beli bensin pakai mypertamina	?

Data tersebut kemudian diproses menggunakan algoritma NBC dengan beberapa rangkaian tahap. Tahap pertama adalah menghitung *prior probability* sebagai berikut:

Prior probability untuk kelas positif :

$$P(positif) = 1/4 = 0{,}25$$

Prior probability untuk kelas negatif :

$$P(negatif) = 2/4 = 0{,}50$$

Prior probability untuk kelas netral :

$$P(netral) = 1/4 = 0{,}25$$

Tahap selanjutnya adalah melakukan perhitungan untuk probabilitas kemunculan kata dalam data di tiap kelas seperti berikut ini:

$P(bisa|positif) = (0+1)/(14+49) = 1/63 = 0{,}015873$
$P(pakai|positif) = (0+1)/(14+49) = 1/63 = 0{,}015873$
$P(banyak|positif) = (0+1)/(14+49) = 1/63 = 0{,}015873$
$P(ewallet|positif) = (0+1)/(14+49) = 1/63 = 0{,}015873$
$P(beli|positif) = (0+1)/(14+49) = 1/63 = 0{,}015873$
$P(bensin|positif) = (1+1)/(14+49) = 2/63 = 0{,}031746$
$P(pakai|positif) = (0+1)/(14+49) = 1/63 = 0{,}015873$
$P(mypertamina|positif) = (1+1)/(14+49) = 2/63 = 0{,}031746$

$P(bisa|negatif) = (0+1)/(28+49) = 1/77 = 0{,}012987$
$P(pakai|negatif) = (2+1)/(28+49) = 3/77 = 0{,}038961$
$P(banyak|negatif) = (0+1)/(28+49) = 1/77 = 0{,}012987$
$P(ewallet|negatif) = (0+1)/(28+49) = 1/77 = 0{,}012987$
$P(beli|negatif) = (2+1)/(28+49) = 3/77 = 0{,}038961$
$P(bensin|negatif) = (2+1)/(28+49) = 3/77 = 0{,}038961$
$P(pakai|negatif) = (2+1)/(28+49) = 3/77 = 0{,}038961$
$P(mypertamina|negatif) = (2+1)/(28+49) = 3/77 = 0{,}038961$

$P(bisa|netral) = (1+1)/(7+49) = 2/56 = 0{,}035714$
$P(pakai|netral) = (1+1)/(7+49) = 2/56 = 0{,}035714$
$P(banyak|netral) = (0+1)/(7+49) = 1/56 = 0{,}017857$
$P(ewallet|netral) = (0+1)/(7+49) = 1/56 = 0{,}017857$
$P(beli|netral) = (0+1)/(7+49) = 1/56 = 0{,}017857$
$P(bensin|netral) = (0+1)/(7+49) = 1/56 = 0{,}017857$

$P(pakai \mid netral) = (1 + 1)/(7 + 49) = 2/56 = 0{,}035714$
$P(mypertamina|netral) = (1 + 1)/(7 + 49) = 2/56 = 0{,}035714$

$P(positif|d5) = 0{,}25 \times 0{,}015873 \times 0{,}015873 \times 0{,}015873 \times 0{,}015873 \times 0{,}031746 \times 0{,}015873 \times 0{,}031746 = 0{,}001512$
$P(negatif|d5) = 0{,}50 \times 0{,}012987 \times 0{,}038961 \times 0{,}012987 \times 0{,}012987 \times 0{,}038961 \times 0{,}038961 \times 0{,}038961 \times 0{,}038961 = 0{,}003795$
$P(netral|d5) = 0{,}25 \times 0{,}035714 \times 0{,}035714 \times 0{,}017857 \times 0{,}017857 \times 0{,}017857 \times 0{,}017857 \times 0{,}035714 \times 0{,}035714 = 0{,}005102$

Dari perhitungan tersebut data 5 masuk kedalam kategori kelas berlabel netral karena dari hasil perhitungan yang dilakukan kelas netral menunjukan nilai tertinggi.

4. HASIL DAN PEMBAHASAN

4.1. PEMBACAAN *DATASET*

Proses pembacaan *dataset* oleh *Jupyter Notebook* menggunakan *library* pandas pada data sebelumnya yang masih dalam bentuk excel. Total data yang terbaca terdiri dari 972 *records.*

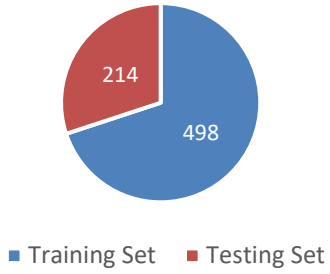
4.2. *PREPROCESSING*

Data yang telah melalui rangkaian *preprocessing* mengalami beberapa perubahan. Pada proses pertama penghapusan duplikasi, data tidak mengalammi perubahan jumlah tetap di 972 *records*. Proses selanjutnya penghapusan pengguna, data mengalami pengurangan menjadi 747 *records*. Pada proses pemotongan tanggal, data kembali mengalami pengurangan menjadi 712 *records*.

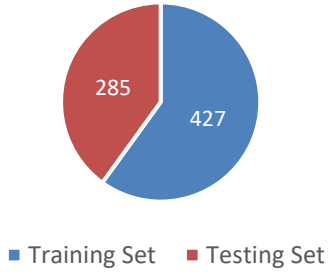
Data sebelumnya kemudian melalui beberapa proses *text-processing* seperti penghapusan angka, penghapusan tanda baca, *stemming*, pembuatan kamus kata, *tokenizing*, pembersihan *stopwords*, dan *labeling*

4.3. *SPLITTING DATASET* DAN PEMBOBOTAN TF-IDF

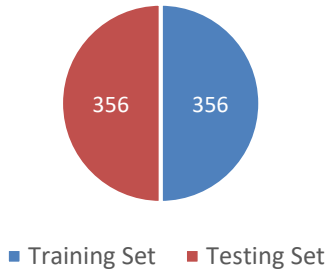
Splitting atau pemisahan dataset dilakukan dengan beberapa rasio yaitu 30%, 40% dan 50% untuk *testing set* dari total data. Gambar 2 sampai dengan 4 berikut ini menunjukan proses pemisahan dan perbadingan jumlah *training set* serta *testing set*:



Gambar 2 *Chart* Perbandingan *Training Set* Dan *Testing Set* Rasio 30%



Gambar 3 *Chart* Perbandingan *Training Set* Dan *Testing Set* Rasio 40%



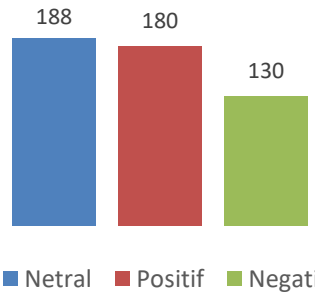
Gambar 4 *Chart* Perbandingan *Training Set* Dan *Testing Set* Rasio 50%

Dataset yang telah di pisah kemudian akan di lakukan pembobotan kata menggunakan TF-IDF pada data latih dan data yang akan di uji. Proses pembobotan ini menggunakan fungsi TfidfVectorizer() dari library sklearn.

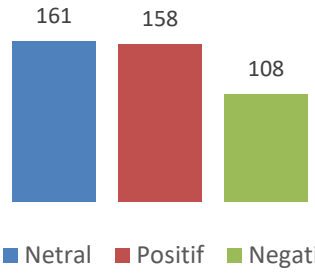
4.4. SMOTE

Hasil pemisahan dataset yang sebelumnya sudah di lakukan, di temukan sebuah permasalahan yang di mana perbandingan kelas di *training set* yang

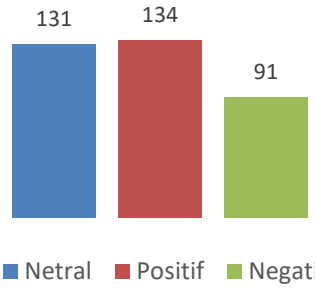
tidak seimbang. Gambar 5 sampai dengan 7 berikut ini menunjukan perbandingan kelas di *training set* untuk semua rasio sebelum proses SMOTE:



Gambar 5 *Chart* Perbandingan Kelas Di *Training Set* Sebelum SMOTE Rasio 30%

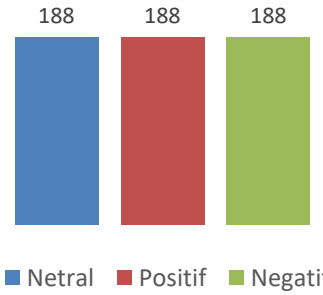


Gambar 6 *Chart* Perbandingan Kelas Di *Training Set* Sebelum SMOTE Rasio 40%

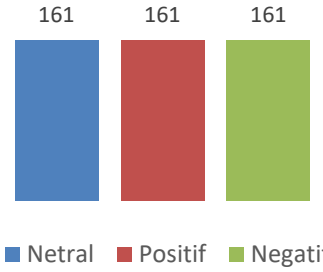


Gambar 7 *Chart* Perbandingan Kelas Di *Training Set* Sebelum SMOTE Rasio 50%

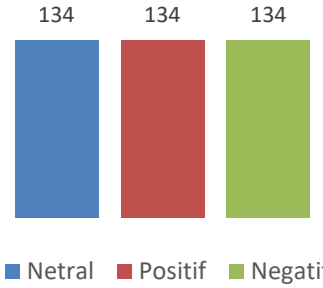
SMOTE digunakan untuk mengatasi permasalahan tersebut, dilakukan dengan bahasa *python* dan *library* imblearn. Gambar 8 sampai dengan 10 berikut ini menunjukan perbandingan kelas di *training set* untuk semua rasio setelah proses SMOTE:



Gambar 8 *Chart* Perbandingan Kelas Di *Training Set* Setelah SMOTE Rasio 30%



Gambar 9 *Chart* Perbandingan Kelas Di *Training Set* Setelah SMOTE Rasio 40%



Gambar 10 *Chart* Perbandingan Kelas Di *Training Set* Setelah SMOTE Rasio 50%

4.5. NBC

Proses klasifikasi ini menggunakan metode NBC MultinomialNB oleh library sklearn. MultinomialNB adalah salah satu model dari pengklasifikasian NBC yang di khususkan untuk mengklasifikasikan data yang berupa sebuah data. Model yang akan dibuat pada penelitian ini ada dua, yaitu model untuk klasifikasi tanpa SMOTE dan dengan SMOTE. Gambar 11 berikut ini menunjukan pembuatan model klasifikasi:

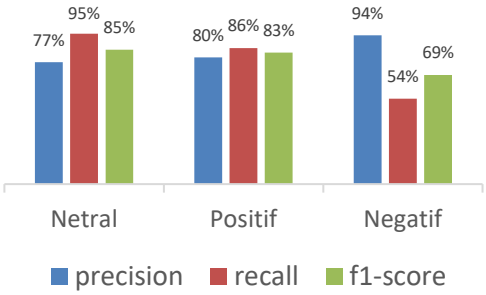
```
# model dengan SMOTE
clfsm = MultinomialNB()
clfsm.fit(x_sm, y_sm)

# model tanpa SMOTE
clf = MultinomialNB()
clf.fit(train_tfidf, y_train)
```

Gambar 11 *Pseudocode* Pembuatan Model Klasifikasi dengan SMOTE

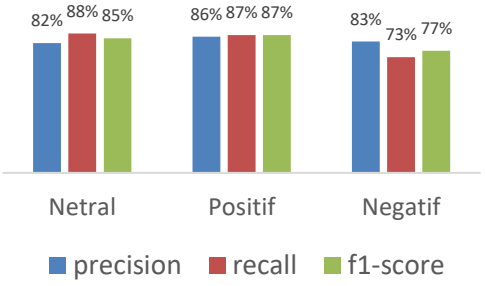
4.5.1. KLASIFIKASI PADA *TESTING SET* 30%

Klasifikasi pada kasus pertama yaitu *testing set* 30% dari model pertama yaitu adalah klasifikasi tanpa mengimplementasikan SMOTE mendapatkan hasil prediksi dengan akurasi 81%, *recall* untuk kelas netral 95%, positif 86%, dan negatif 54%. Data testing dengan jumlah 214 data yang telah diklasifikasikan oleh model ini menghasilkan 84 data masuk sebagai kelas netral, 71 positif, dan 59 negatif. Gambar 12 berikut ini menunujukan hasil klasifikasi oleh model yang tidak mengimplementasikan SMOTE:



Gambar 12 *Chart* Hasil Klasifikasi NBC Tanpa Implementasi SMOTE Rasio 30%

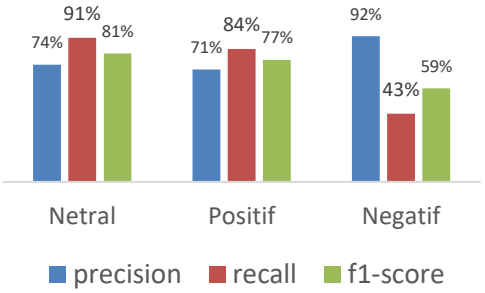
Hasil prediksi oleh model yang mengimplementasikan SMOTE mendapatkan tingkat akurasi 84%, *recall* untuk kelas netral bernilai 88%, positif 87%, dan negatif 77%. Data testing yang diklasifikasikan mendapatkan hasil 84 data masuk kedalam kelas netral, 71 positif, dan 59 negatif. Gambar 13 berikut ini menunujukan hasil klasifikasi oleh model yang mengimplementasikan SMOTE:



Gambar 13 *Chart* Hasil Klasifikasi NBC Dengan Implementasi SMOTE Rasio 30%

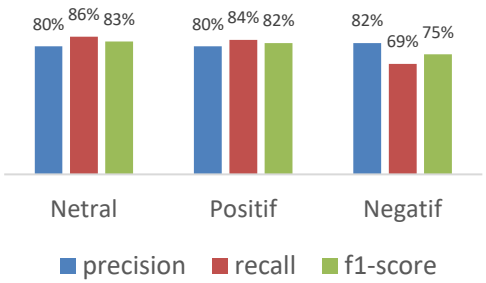
4.5.2. KLASIFIKASI PADA *TESTING SET* 40%

Hasil klasifikasi pada rasio *testing set* 40% juga dilakukan dengan dua model yang telah dibuat sebelumnya. Model pertama adalah model yang tidak mengimplementasikan SMOTE mendapatkan bahwa dari total 285 data testing yang diklasifikasikan, terdapat 111 data netral, 93 positif, dan 81 negatif. Tingkat akurasi yang didapatkan oleh model tersebut adalah 75% dengan *recall* kelas netral sebesar 91%, positif 84%, dan negatif 43%. Gambar 14 berikut ini menunujukan hasil klasifikasi oleh model tersebut:



Gambar 14 *Chart* Hasil Klasifikasi NBC Tanpa Implementasi SMOTE Rasio 40%

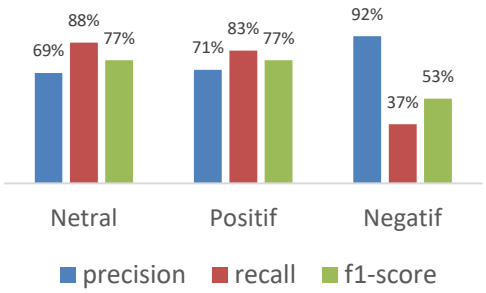
Model kedua adalah model yang mengimplementasikan SMOTE sebelum diklasifikasikan dengan algoritma NBC. Model kedua dengan rasio *testing set* 40% menghasilkan 111 data netral, 93 positif, dan 81 negatif. Tingkat akurasi yang didapatkan adalah 81% dengan *recall* netral 86%, positif 84% dan negatif 69%. Gambar 15 berikut ini menunujukan hasil klasifikasi oleh model tersebut:



Gambar 15 Chart Hasil Klasifikasi NBC Dengan Implementasi SMOTE Rasio 40%

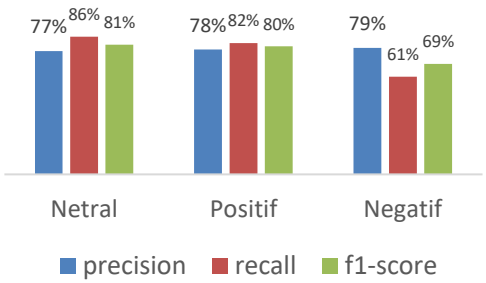
4.5.3. KLASIFIKASI PADA *TESTING SET* 50%

Rasio terakhir yang diujikan adalah *testing set* berjumlah 50% dari total data. Pada rasio ini data testing juga akan diklasifikasikan menggunakan dua model yang telah dibuat. Model pertama yaitu model yang tidak mengimplementasikan SMOTE dengan 365 data menghasilkan 141 data masuk kedalam kelas netral, 117 positif, dan 98 negatif. Tingkat akurasi yang didapatkan oleh model tersebut adalah 72% dengan *recall* kelas netral 88%, positif 83%, dan negatif 37%. Gambar 16 berikut ini menunujukan hasil klasifikasi oleh model tersebut:



Gambar 16 *Chart* Hasil Klasifikasi NBC Tanpa Implementasi SMOTE Rasio 50%

Model kedua yang mengimplementasikan SMOTE untuk menyeimbangkan data di *training set* mendapatkan hasil klasifikasi 141 data masuk kedalam kelas netral, 117 positif, dan 98 negatif. Tingkat akurasi yang didapatkan oleh model tersebut adalah 78% dengan *recall* 86% untuk kelas netral, 82% positif, dan 61% netral. Gambar 17 berikut ini menunujukan hasil klasifikasi oleh model tersebut:



Gambar 17 *Chart* Hasil Klasifikasi NBC Dengan Implementasi SMOTE Rasio 50%

Tiga rasio yang telah diujikan dengan kedua model yang telah dibuat mendapatkan hasil yang berbeda-beda. Rasio dengan 30% *testing set* mendapatkan hasil klasifikasi terbaik oleh model yang telah dibuat. Algoritma SMOTE juga berpengaruh pada hasil klasifikasi untuk penyeimbangan *precision*, *recall* dan *f1-score*. Akurasi pada setiap rasio juga meningkat berkat implementasi dari algoritma SMOTE pada *training set*.

5. KESIMPULAN

Hasil analisis sentimen dengan mengimplementasikan metode klasifikasi NBC menghasilkan bahwa rasio *testing set* 30% adalah rasio terbaik untuk penelitian ini karena memiliki akurasi tertinggi. Dari total 972 data mentah yang kemudian melalui proses *preprocessing* hingga jumlah data tersisa 712 data, kemudian untuk rasio 30% data dibagi menjadi 498 data *training set* dan 214 *testing set*. Perbandingan anggota dalam *training set* terdiri dari 188 data netral, 180 positif, dan 130 negatif yang kemudian algoritma SMOTE di implementasikan untuk menyeimbangkan anggota dari *training set*. Hasil analisis sentimen dengan implementasi SMOTE menunjukan bahwa 38% data masuk kedalam kelas netral, 35% data positif, dan 27% data negatif dengan tingkat akurasi 84%. Hasil klasifikasi pada rasio yang sama namun tanpa implementasi algoritma SMOTE mendapatkan hasil akurasi yang lebih rendah yaitu 81%. Algoritma SMOTE dalam model yang telah dibuat menunjukan hasil dalam peningkatan akurasi dan penyeimbangan nilai *precision*, *recall*, dan *f1-score* untuk setiap kelas.

6. REFERENSI

Afdal, M., & Rosadi, M. (2019, Februari). Penerapan Association Rule Mining Untuk Analisis Penempatan Tata Letak Buku Di Perpustakaan Menggunakan Algoritma Apriori. *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informas, 5*, 100-101. Retrieved from https://ejournal.uin-suska.ac.id/index.php/RMSI/article/view/7379

Amalia, A. A., & Wahyuni, E. D. (2020, February). SMOTE: Metode Penyeimbang Kelas Pada Klasifikasi Data Mining. *Scan* - Jurnal Teknologi Informasi dan Komunikasi, 35. doi:10.33005/scan.v15i1.1850

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: *Synthetic Minority Over-sampling Technique*. *Journal of Artificial Intelligence Research*, 328. doi:10.1613/jair.953

Chinnamgari, D. S. (2019). *Machine Learning Projects: Implement supervised, unsupervised, and reinforcement learning techniques using R 3.5*. Mumbai: packt .

Darwis, D., Pratiwi, E. S., & Pasaribu, A. F. (2020). Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia. *Jurnal Ilmiah Edutic*, 4-5.

D'Monte, L. (2009, April 29). Swine Flu's Tweet Tweet Causes Online Flutter.

DPR, R. (2022, August). *MyPertamina: Antara Pengawasan dan Kesulitan*. Retrieved from EMedia DPR RI: https://emedia.dpr.go.id/article/mypertamina-antara-pengawasan-dan-kesulitan/

Kusuma, P. D. (2020). *Machine Learning Teori, Program, Dan Studi Kasus.* Purba Daru Kusuma.

Mufidah, U., & Siahaan, M. (2021). http://pusdansi.org/index.php/pusdansi/article/view/12. *Pusdansi.org, 1*. Retrieved from http://pusdansi.org/index.php/pusdansi/article/view/12

Pabubung, M. R. (2021). Epistemologi Kecerdasan Buatan (AI) dan Pentingnya Ilmu Etika dalam Pendidikan Interdisipliner. *Jurnal Filsafat Indonesia*, 152-158.

Pratmanto, D., Rousyati, R., Wati, F. F., Widodo, A. E., & Suleman, S. (2020). *App Review Sentiment Analysis Shopee Application In Google Play Store Using Naive Bayes Algorithm. Journal of Physics: Conference Series*, 1-6.

Ririh, K. R., Laili, N., Wicaksono, A., & Tsurayya, S. (2020). Studi Komparasi Dan Analisis Swot Pada Implementasi Kecerdasan Buatan *(Artificial Intelligence)* Di Indonesia. *Garuda*, 122-130.

Safitri, K. (22, July 21). *Pertamina Perluas Cakupan Uji Coba MyPertamina hingga 50 Kota, Termasuk DKI Jakarta*. Retrieved from Kompas: https://money.kompas.com/read/2022/07/21/115000626/pertamina-perluas-cakupan-uji-coba-mypertamina-hingga-50-kota-termasuk-dki?page=all

Santika, E. F. (2023, February 27). Pengguna Twitter di Indonesia Capai 24 Juta hingga Awal 2023, Peringkat Berapa di Dunia?

Saurkar, A. V., Pathare, K. G., & Gode, S. A. (2018). *An overview on web scraping techniques and tools. International Journal on Future Revolution in Computer Science & Communication Engineering*.

Sikumbang, E. D. (2018, Februari 1). Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori. *Jurnal Teknik Komputer, 4*, 1. Retrieved from https://ejournal.bsi.ac.id/ejurnal/index.php/jtk/article/view/2560/1918

Sindo, K. (2022, July 01). *Pembatasan BBM Subsidi lewat MyPertamina*. Retrieved from SindoNews: https://nasional.sindonews.com/read/

814471/16/pembatasan-bbm-subsidi-lewat-mypertamina-1656662870

Utomo, D. P., & Purba, B. (2019, September). Penerapan Datamining pada Data Gempa Bumi Terhadap Potensi Tsunami di Indonesia. *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, 3.

Yin, J., Ngiam, K. Y., & Teo, H. H. (2021). *Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review*. *JMIR.*