Analisis Sentimen Nasabah Pada Layanan Perbankan Menggunakan Metode Regresi Logistik Biner, *Naïve Bayes Classifier* (NBC), dan *Support Vector Machine* (SVM)

Erna Dwi Nurindah Sari dan Irhamah
Departemen Statistika, Fakultas Matematika, Komputasi, dan Sains Data,
Institut Teknologi Sepuluh Nopember (ITS)
e-mail: irhamah@statistika.its.ac.id

Abstrak— ***Di era globalisasi, semua aktivitas yang* dilakukan tidak dapat terlepas dari teknologi. Dengan menggunakan Twitter pendapat masyarakat mengenai layanan dari perbankan dapat diketahui. Pada penelitian ini pendapat nasabah akan dibedakan menjadi sentimen positif atau negatif sehingga hasil analisa sentimen tersebut dapat dijadikan evaluasi sebagai peningkatan layanan pada para nasabahnya. Klasifikasi sentimen positif dan sentimen negatif dilakukan dengan beberapa metode yakni menggunakan Regresi Logistik Biner yang merupakan salah satu metode konvensional, metode Naïve Bayes Classifier yakni metode yang sederhana namun memiliki ketepatan klasifikasi yang baik, dan metode Support Vector Machine (SVM) yang dapat mengklasifikasikan sentimen dengan beberapa jenis kernel. Data yang digunakan pada penelitian ini diklasfikasikan secara manual dan dengan menggunakan kamus lexicon. Karena jumlah data sentimen yang tidak seimbang maka dilakukan SMOTE pada data klasifikasi manual. Hasil dari penelitian menunjukkan bahwa metode terbaik untuk mengklasifikasikan sentimen pada layanan BRI adalah SMOTE-SVM kernel RBF, sedangkan untuk Bank Mandiri adalah SMOTE-NBC karena memiliki nilai AUC paling tinggi.***

***Kata Kunci—Naïve Bayes Classifier*, Perbankan, Regresi Logistik Biner, Sentimen Negatif, Sentimen Positif, *SVM*.**

I. PENDAHULUAN

PADA era globalisasi ini, semua aktivitas yang dilakukan tidak dapat terlepas dari teknologi. Digitalisasi dalam banyak aspek membuat sumber daya manusia terbantu dan lebih mudah dalam menangani berbagai macam pekerjaan sehingga kualitas layanan yang dihasilkan akan meningkat, salah satunya adalah bank. Bank merupakan lembaga yang bergerak dan mengelola keuangan yang berasal dari para nasabahnya. Nasabah merupakan seseorang yang menggunakan jasa dari perbankan. Kepuasan pelayanan nasabah menjadi hal yang penting untuk diperhatikan pada industri perbankan. Oleh karena itu, pelayanan yang baik dapat memberikan kepuasan nasabah dalam bertransaksi di industri perbankan. Ketepatan dan kecepatan layanan yang disuguhkan oleh perbankan akan menjadi hal yang sangat diperlukan pada era sekarang ini.

Di Indonesia terdapat beberapa perusahaan perbankan antara lain adalah Bank Rakyat Indonesia (BRI), Bank Mandiri, Bank Central Asia (BCA), dan lain-lain. Berbagai macam kondisi layanan perbankan tersebut akan menimbulkan berbagai respon dari nasabah. Jika terjadi kesalahan pada dalam melakukan transaksi di perbankan nasabah pada saat ini akan cenderung melakukan protes atau mengajukan keluhan ke pihak yang bersangkutan. Namun jika pelayanan yang dilakukan oleh perbankan cepat dan memuaskan maka masyarakat akan cenderung melakukan apresiasi. Pada saat ini jejaring sosial merupakan sarana yang banyak dipilih masyarakat dalam menceritakan berbagai pendapat dan keluh kesahnya. Salah satu jejaring sosial yang banyak digunakan adalah Twitter. Oleh karena itu, dengan menggunakan Twitter dapat diketahui pendapat nasabah mengenai layanan perbankan. Pendapat masyarakat tersebut nantinya akan dibedakan menjadi sentimen yang bersifat positif atau negatif sehingga berdasarkan sentimen tersebut dapat dijadikan evaluasi dan selanjutnya perusahaan perbankan dapat melakukan peningkatan ataupun perbaikan dalam memberikan layanan kepada nasabahnya.

Klasifikasi sentimen positif dan sentimen negatif dilakukan dengan metode statistika yakni menggunakan *text mining*. *Text mining* merupakan suatu cabang ilmu dari statistika yang mengolah teks menjadi suatu data. Proses *text mining* ini lebih banyak dilakukan pada suatu pengelompokkan data. Salah satu cara untuk pengelompokkan data adalah dengan menggunakan Regresi Logistik Biner. Metode Regresi Logistik Biner merupakan metode klasik yang digunakan untuk mengetahui pola hubungan antara variabel respon yang bersifat biner yakni terdiri dari 0 dan 1 dengan variabel prediktornya [1]. Metode lain yang digunakan adalah *Naïve Bayes Classifier* dan *Support Vector Machine* (SVM). *Naïve Bayes Classifier* adalah suatu klasifikasi yang berdasarkan pada teorema Bayes yang bertujuan dalam menghitung peluang pada tiap kelas serta memiliki asumsi bahwa hubungan antar kelas adalah independen sedangkan *Support Vector Machine* (SVM) memiliki ide dasar bahwa menemukan fungsi pemisah (*hyperlane*) yang nantinya dapat memisahkan dua kelas secara optimal [2].

II. TINJAUAN PUSTAKA

A. *Text Mining*

Text mining adalah suatu proses untuk mengekstraksi suatu pola untuk dapat dieksplorasi yang datanya berasal dari suatu teks. *Text mining* adalah suatu disiplin ilmu yang berdasarkan pada *information retrieval*, *data mining*, *machine learning*, ilmu statistika, dan lingusitik komputasi [3].

B. *Text Preprocessing*

Text Preprocessing adalah suatu tahap pertama dalam mengolah teks untuk digunakan dalam pengubahan dokumen

menjadi data yang terstruktur sesuai dengan kebutuhannya agar dapat diolah lebih lanjut dalam proses *text mining*. Tahapan dalam praproses teks adalah *cleansing, case folding, stemming, stopwords,* dan *tokenizing.*

C. TF-IDF

TF-IDF adalah suatu metode pembobotan yang banyak digunakan dalam membangun model vektor. Dengan menggunakan TF-IDF akan diketahui seberapa penting suatu kata terhadap kumpulan *tweet* berdasarkan frekuensi kemunculannya. TF-IDF bekerja dengan melibatkan perkalian antara *Term Frequency* (TF) dengan *Inverse Document Frequency* (IDF). TF memiliki tujuan untuk menunjukkan jumlah kemunculan sebuah kata pada suatu *tweet*. IDF memiliki tujuan untuk menghitung frekuensi kemunculan suatu kata pada seluruh *tweet* [4].

D. Regresi Logistik Biner

Regresi logistik biner merupakan suatu metode analisis data yang digunakan untuk mencari hubungan antara variabel respon (*y*) yang bersifat biner dengan variabel (*x*) yang bersifat polikotomus [1]. Hasil dari variabel respon yang terdiri dari 2 kategori yaitu sukses dan gagal yang dinotasikan dengan *y*=1 (sukses) dan *y*=0 (gagal). Oleh karena itu, variabel *y* mengikuti distribusi *Bernoulli* untuk setiap observasi tunggal. Fungsi probabilitas untuk setiap observasi dapat dituliskan pada persamaan sebagai berikut.

$$f(y) = \pi^{y}(1-\pi)^{1-y}; y = 0,1 \quad (1)$$

dimana *y* adalah variabel respon jika *y*=0 maka $f(y) = 1-\pi$ dan jika *y* = 1 maka $f(y) = \pi$. Fungsi regresi logistiknya dapat dituliskan pada persamaan (2).

$$f(z) = \frac{1}{1+e^{-z}} \text{ ekuivalen } f(z) = \frac{e^{z}}{1+e^{z}} \quad (2)$$

dimana, $z = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$ dengan *p* adalah banyaknya variabel prediktor

Nilai *z* antara $-\infty$ dan $+\infty$ sehingga nilai $f(z)$ terletak antara 0 dan 1 untuk setiap nilai *z* yang diberikan. Hal tersebut menunjukkan bahwa model logistik menggambarkan probabilitas atau risiko dari suatu objek sehingga model regresi dapat dituliskan pada persamaan.

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x_1)}}{1+e^{(\beta_0+\beta_1 x_1)}} \quad (3)$$

β_p merupakan parameter ke-*p* yang diestmiasi. Untuk mempermudah pendugaan parameter regresi maka model regresi logistik pada persamaan (3) dapat diurutkan dengan menggunakan transformasi logit $\pi(x)$ sehingga diperoleh persamaan (4) berikut

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p . \quad (4)$$

Model tersebut merupakan fungsi linier dari parameter-parameternya. Pada regresi logistik, variabel respon diekspresikan sebagai $y = \pi(x) + \varepsilon$ dimana ε mempunyai salah satu dari kemungkinan dua nilai yaitu $y = \pi(x) + \varepsilon$ dengan peluang $\pi(x)$ jika $y = 1$ dan $\varepsilon = -\pi(x)$ dengan peluang $1-\pi(x)$ jika dan mengikuti distribusi binomial dengan rataan nol dan varians $(\pi(x))(1-\pi(x))$.

1. Uji Serentak

Uji secara simultan pada regresi logistik biner memiliki pada hipotesis sebagai berikut.

H_0: $\beta_1 = \beta_2 = \ldots = \beta_j = 0$ (tidak ada pengaruh antara variabel prediktor terhadap variabel respon)

H_1: minimal ada satu $\beta_j \neq 0$ (ada pengaruh antara variabel prediktor terhadap variabel respon)

Statistik uji yang digunakan adalah *G* yang mengikuti distribusi dengan derajat bebas sama dengan banyaknya parameter, dimana H_0 akan ditolak jika nilai statistik uji *G* akan lebih dari sama dengan nilai $\chi^2_{(p,\alpha)}$ dengan tingkat kepercayaan α. Rumus untuk statistik uji *G* dapat ditulisakan pada persamaan (5) sebagai berikut [1].

$$G = 2\left\{\sum_{i=1}^{n}[y_i \ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n)]\right\} \quad (5)$$

2. Uji Individu

Pengujian secara parsial dilandaskan pada hipotesis adalah sebagai berikut.

H_0: $\beta_j = 0$ (tidak ada pengaruh antara masing-masing variabel prediktor terha-dap variabel respon)

H_1: $\beta_j \neq 0$ (ada pengaruh antara masing-masing variabel prediktor terhadap variabel respon)

H_0 ditolak jika memiliki nilai P *value* = $P(Z>Z_{hitung})$ yang kurang dari taraf signifikasn. Rumus statistik uji *Wald* dapat dituliskan pada persamaan (6) sebagai berikut [1].

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad (6)$$

dimana $\hat{\beta}_j$ merupakan koefien parameter yang ke-*j* serta $SE(\hat{\beta}_j)$ adalah standar error dari koefien parameter yang ke-*j*.

E. Naïve Bayes Classifier

Klasifikasi Bayes merupakan klasifikasi statistik yang digunakan untuk melakukan prediksi pada kelas suatu anggota probabilitas. Ide awal dari *Naive Bayes* ini yaitu dengan *join* probabilitas kata dan kategori yang diberikan oleh sebuah dokumen [5]. Berikut adalah pendekatan yang dipakai dalam *Naïve Bayes Classifier*.

$$V_{MAP} = \arg\max_{V_j \in V} P(V_j) \prod_{k=1}^{n} P\left(x_k \mid V_j\right) \quad (7)$$

dimana, V_{MAP} merupakan nilai *output* hasil klasifikasi *Naïve Bayes*. Nilai $P(V_j)$ dapat dihitung dengan rumus pada persamaan (8).

$$P(V_j) = \frac{\left|doc_j\right|}{\left|training\right|} \quad (8)$$

dimana, $\left|doc_j\right|$ adalah jumlah *tweet* yang memiliki kategori *j* dalam *training*. Untuk $\left|training\right|$ adalah jumlah *tweet* dalam contoh yang digunakan *training*. Untuk setiap probabilitas kata x_k dalam setiap kategori dihitung pada saat *training* dengan persamaan (9).

$$P(x_k \mid V_j) = \frac{n_k + 1}{\left|n + \text{kosakata}\right|}, \quad (9)$$

n_k merupakan jumlah kemunculan pada kata x_k dalam *tweet* berkategori v_j sedangkan *n* adalah banyaknya seluruh kata

dalam *tweet* dengan kategori V_j dan |kosakata| merupakan banyaknya kata dalam *training.*

F. Support Vector Meachine

Support Vector Machine (SVM) pertama kali diperkenalkan pada akhir 1979 oleh Vapnik yang selanjutnya mendapatkan perhatian yang lebih pada tahun 1992 bersama dengan Boser [6]. *Support Vector Machine* adalah salah satu metode dalam *machine learning* yang digunakan untuk melakukan suatu prediksi dan pengklasifikasian. Ide dasar dalam *Support Vector Machine* merupakan gabungan dari beberapa konsep yang pernah ada sebelumnya dalam mengatasi permasalah terutama pada kasus klasifikasi dan prediksi. Konsep klasifikasi dengan menggunakan metode *Support Vector Machine* memiliki ide dasar bahwa menemukan fungsi pemisah (*hyperlane*) yang nantinya dapat memisahkan dua kelas secara optimal [2]. Persamaan *hyperplane* dikatakan baik jika memiliki margin terbesar. Margin adalah dua kali jarak antara *hyperplane* dan *support vector,* dimana *support vector* adalah titik yang berada paling dekat dengan *hyperplane.* Pada dasarnya SVM dikembangkan dengan prinsip *linier classifier* yang dapat dibedakan menjadi dua yaitu *linear separable* dan non *separable.*

G. K-fold Cross Validation

Metode *k-fold cross validation* adalah suatu metode yang diketahui untuk memprediksi tingkat kesalahan dalam teknik klasifikasi. Metode ini banyak digunakan untuk mengurangi bias yang berhubungan dengan data random. *K-fold cross validation* membagi data kedalam beberapa bagian (*subset)* yang mana disebut *fold* [7]. Data secara random dipartisi menjadi *k* himpunan *fold* sehingga dapat ditulis menjadi $D_1, D_2, \ldots, D_k$, dimana setiap bagian memiliki ukuran yang sama [8].

H. Synthetic Minority Oversampling Technique (SMOTE*)*

Pada jumlah data yang tidak seimbang atau *imbalance* terdapat perbedaan jumlah antara kategori yang sangat signifikan, yang mana terdapat kategori yang dominan dan disebut mayoritas serta adanya kategori yang lebih sedikit dan disebut minoritas [9]. Pada metode ini diterapkan salah satu metode *oversampling* dimana dilakukan dengan menerapkan metode sampling. Untuk replikasi datanya, SMOTE melakukan dengan pemilihan *k nearest neighbor* sebagai penentuannya, sehingga nantinya jumlah data yang minoritas akan menjadi seimbang dengan data mayoritas. Pada metode SMOTE ini dicari *k nearest neighbour* untuk setiap data pada kategori minoritas dan akan dibuat data sintesis sebanyak persentase replikasi data minoritas.

I. Ketepatan Klasifikasi

Pengukuran ketepatan klasifikasi dilakukan untuk melihat performa klasifikasi yang telah dilakukan. Dalam mengukur ketepatan klasifikasi, perlu diketahui jumlah pada setiap kelas prediksi dan kelas aktual yang terdiri dari *TP* (*True Positif*) yaitu jumlah *tweet* bersentimen positif yang tepat diprediksi dalam kelas positif, *TN* (*True Negatif*) yaitu *tweet* yang tepat terprediksi dalam kelas negatif, *FP* (*False Positif*) yaitu *tweet* bersentimen negatif yang terprediksi dalam kelas positif, dan *FN* (*False Negatif*) yaitu *tweet* bersentimen positif yang terprediksi dalam kelas negative [10].

Tabel 1.
Confusion Matrix

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	*TP*	*FN*
Negatif	*FP*	*TN*

Tabel 1 menunjukkan bahwa *confusion matrix* digunakan untuk menghitung ketepatan klasifikasi. Nilai yang berada pada diagonal utama menunjukkan hasil keputusan yang benar. AUC dapat mengestimasi dengan berbagai teknik, salah satunya adalah metode trapezoidal yang merupakan metode geomterik yang berdasarkan pada interpolasi linier antara nilai pada ROC.

J. Word Cloud

Word Cloud merupakan suatu metode untuk membuat visualisasi data yang berasal dari dokumen atau teks. *Word cloud* adalah perwakilan grafis dari sebuah teks atau dokumen dengan cara melakukan *plotting* pada kata-kata yang mempunyai frekuensi kemunculan paling banyak dan digambarkan pada ruang berdimensi dua.

III. METODOLOGI PENELITIAN

A. Sumber Data

Sumber data yang digunakan adalah data sekunder. Data sekunder diambil langsung dari kumpulan *tweet* para pengguna Twitter di Indonesia. Akun Twitter yang digunakan pada penelitian ini adalah @kontakBRI dan akun Twitter @mandiricare. Data *tweet* diambil mulai tanggal 7 Februari 2019 hingga 8 April 2019 dengan menggunakan Twitter API.

B. Variabel Penelitian

Variabel penelitian yang digunakan dalam penelitian ini terdiri dari variabel respon (*Y*) dan variabel prediktor (*X*) yang disajikan pada Tabel 2 berikut.

Tabel 2.
Variabel Penelitian

Variabel	Nama Variabel	Skala Data
Y	Sentimen 0 = Sentimen positif 1 = Sentimen negatif	Nominal
X_k	Bobot kata kunci dari setiap *tweet* yang diperoleh dari hasil *tf-idf*	Rasio

C. Langkah Analisis

Langkah analisis yang dilakukan dalam penelitian ini adalah sebagai berikut :

1. Mengambil *tweet* menggunakan Twitter API
2. Melakukan *text preprocssing*
3. Membagi data menjadi 4-*fold cross validation*
4. Melakukan analisis menggunakan Regresi Logistik Biner
5. Melakukan analisis menggunakan *Naïve Bayes Classifier*
6. Melakukan analisis menggunakan *Support Vector Machine* dengan kernel linier dan RBF
7. Membandingkan hasil klasifikasi
8. Menarik kesimpulan dan saran.

IV. ANALISIS DAN PEMBAHASAN

A. Praproses Teks

Tweet yang telah dikumpulkan kemudian dilakukan praproses teks dengan urutan seperi Tabel 3.

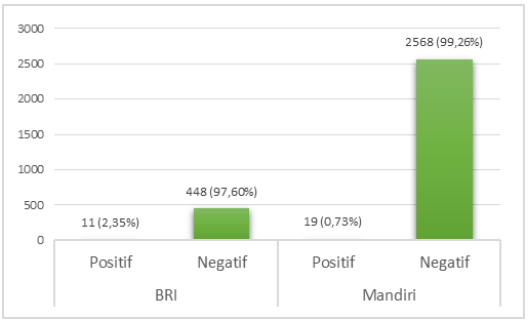
Tabel 3.
Contoh Praproses Teks

Praproses	Hasil
Tweet	@kontakBRI halo min, maaf ini BRI kenapa gangguan terus ya, susah mau transfer, sekalinya bisa malah keambil dua kali dananya.. Kan rugi saya,, mohon bantuannya..
Menghapus *username*	halo min, maaf ini BRI kenapa gangguan terus ya, susah mau transfer, sekalinya bisa malah keambil dua kali dananya.. Kan rugi saya,, mohon bantuannya..
Menghapus baris kosong	halo min, maaf ini BRI kenapa gangguan terus ya, susah mau transfer, sekalinya bisa malah keambil dua kali dananya.. Kan rugi saya,, mohon bantuannya..
Menghapus *punctuation*	halo min maaf ini BRI kenapa gangguan terus ya susah mau transfer sekalinya bisa malah keambil dua kali dananya Kan rugi saya mohon bantuannya
Menghapus spasi berlebih	halo min maaf ini BRI kenapa gangguan terus ya susah mau transfer sekalinya bisa malah keambil dua kali dananya Kan rugi saya mohon bantuannya
Case folding	halo min maaf ini bri kenapa gangguan terus ya susah mau transfer sekalinya bisa malah keambil dua kali dananya kan rugi saya mohon bantuannya
Stemming	'halo min maaf ini bri kenapa ganggu terus ya susah mau transfer sekali bisa malah ambil dua kali dana kan rugi saya mohon bantu'
Stopwords	['halo', 'maaf', 'bri', 'ganggu', 'susah', 'transfer', 'dana', 'rugi', 'mohon', 'bantu']

Berdasarkan contoh praproses teks pada Tabel 3 dapat diketahui hasil akhir dari *stopwords* adalah berupa kata dasar yang selanjutnya kata dasar tersebut akan menjadi kata kunci-kata kunci.

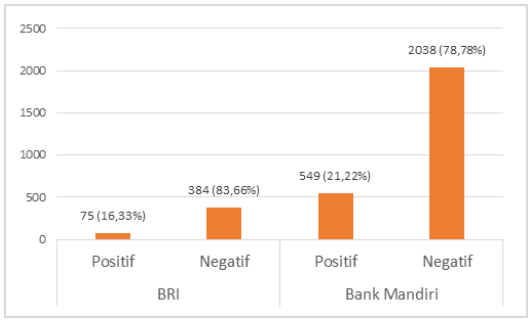
B. Karakteristik Data

Pada penelitian ini data yang digunakan merupakan data *tweet* BRI dan Bank Mandiri. Berikut adalah perbandingan data sentimen positif dan negatif pada masing-masing bank.



Gambar 2. Frekuensi Data *Tweet* Setiap Sentimen Untuk Data dengan Klasifikasi Manual.

Gambar 2 untuk data dengan klasifikasi *tweet* pelayanan nasabah bank secara manual atau untuk selanjutnya disebut data awal. Frekuensi data *tweet* setiap sentimen pada Gambar 2 menunjukkan bahwa *tweet* dengan sentimen negatif memiliki jumlah lebih banyak daripada *tweet* dengan sentimen positif. Jumlah *tweet* yang mengandung sentimen negatif ada sebanyak 448 data dari 459 data *tweet* untuk BRI dan 2568 *tweet* untuk Bank Mandiri. Untuk *tweet* yang mengandung setimen positif hanya ada sebanyak 11 data *tweet* untuk BRI dan 19 data *tweet* Bank Mandiri. Hal ini menunjukkan bahwa jumlah *tweet* sentimen positif dan negatif memiliki jumlah yang tidak seimbang. Selanjutnya akan digambarkan karakteristik data *tweet* untuk data dengan klasifikasi menggunakan *lexicon*. Gambar 3 dibawah ini menunjukkan bahwa *tweet* pada data klasifikasi dengan kamus *lexicon* memiliki sentimen negatif yang jumlahnya lebih banyak daripada *tweet* yang memiliki sentimen positif baik antara pada *tweet* BRI atau Bank Mandiri. Jumlah *tweet* yang mengandung sentimen negatif ada sebanyak 384 data dari 459 data *tweet* untuk BRI dan 2038 *tweet* untuk Bank Mandiri.



Gambar 3. Frekuensi Data *Tweet* Setiap Sentimen Untuk Data dengan Klasifikasi *Lexicon*

Untuk *tweet* yang mengandung setimen positif hanya ada sebanyak 75 data *tweet* untuk BRI dan 549 data *tweet* untuk Bank Mandiri. Hal ini menunjukkan bahwa data dengan klasfikasi kamus *lexicon* masih memiliki data yang tidak seimbang.

C. Klasifikasi pada data BRI

Dalam penelitian ini digunakan 4-*fold cross validation* yaitu membagi data keseluruhan menjadi 4 bagian dimana masing-masing bagian akan menjadi *training* dan *testing*. Pada penelitian ini terdapat tiga jenis data yang digunakan yakni data dengan klasifikasi sentimen secara manual disebut dengan data awal, data dengan klasifikasi sentimen dengan menggunakan *lexicon* disebut data *lexicon*, dan data dengan klasifikasi manual yang telah dilakukan SMOTE disebut dengan data SMOTE

1. Klasifikasi Data BRI dengan Regresi Logistik Biner

Langkah pertama yang dilakukan dalam analisis dengan metode Regresi Logistik Biner adalah dengan mendefinisikan data Y (variabel respon) dan X (variabel prediktor) yang digunakan. Berikut akan dilakukan uji serentak dan parsial pada data pelayanan nasabah BRI dan variabel prediktornya dengan data yang digunakan sebagai contoh pengujian merupakan data yang memiliki nilai performa klasifikasi paling tinggi. Berdasarkan hasil yang diperoleh dapat diketahui bahwa nilai G^2 adalah sebesar 570,7 yang berarti dapat diputuskan tolak H_0 karena nilai G^2 kurang dari nilai $\chi^2_{(0.05,8)}$ sebesar 15,507 sehingga dapat disimpulkan bahwa terdapat pengaruh yang signifikan antara variabel kata kunci terhadap variabel klasifikasi sentimen. Selanjutnya, berdasarkan pada hasil uji signifikansi parameter didapatkan keputusan bahwa variabel bantu, *call*, kredit hingga sukses signifikan terhadap model sehingga dapat ditulis model regresi logistik biner sebagai berikut.

$$\pi(x) = \frac{\exp(1,79 - 3,89X_{12} - 6,45X_{28} + \ldots - 4,24X_{120})}{1 + \exp(1,79 - 3,89X_{12} - 6,45X_{28} + \ldots - 4,24X_{120})}$$

Berikut adalah performa klasifikasi yang didapatkan dengan Regresi Logistik Biner untuk data *tweet* pada pelayanan BRI.

Tabel 6.
Performa Klasifikasi dengan Regresi Logistik Biner pada Data BRI

	Folds	*Accuracy*	*Precision*	*Recall*	*AUC*
Data Awal	1	0,973	0,00	0,00	0,50
	2	0,973	0,00	0,00	0,50
	3	0,973	0,00	0,00	0,50
	4	0,982	0,00	0,00	0,50
	Rata-Rata	0,975	0,00	0,00	0,50

Tabel 6.
Performa Klasifikasi dengan Regresi Logistik Biner pada Data BRI (Lanjutan)

	Folds	*Accuracy*	*Precision*	*Recall*	*AUC*
Data SMOTE	1	0,982	0,97	1,00	0,982
	2	0,995	0,99	1,00	0,995
	3	0,986	0,97	1,00	0,986
	4	0,982	0,97	1,00	0,982
	Rata-Rata	0,986	0,975	1,00	0,986
Data dengan *Lexicon*	1	0,834	0,00	0,00	0,50
	2	0,834	0,00	0,00	0,50
	3	0,834	0,00	0,00	0,50
	4	0,842	0,00	0,00	0,50
	Rata-Rata	0,836	0,00	0,00	0,50

Berdasarkan pada Tabel 6 tersebut data yang memiliki rata-rata performa klasifikasi paling tinggi yakni pada data yang telah dilakukan SMOTE dimana *subset* ke-2 memiliki nilai performa klasifikasi yang terbaik yaitu dengan nilai akurasi sebesar 99,5%, nilai presisi sebesar 99%, nilai *recall* sebesar 100%, dan nilai AUC sebesar 99,5%. Sehingga, dapat diketahui bahwa *subset* ke-4 pada data SMOTE adalah *subset* yang memiliki nilai performa klasifikasi terbaik.

Tabel 7.
Confusion Matrix untuk Regresi Logistik Biner pada Data SMOTE BRI

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	113	0
Negatif	1	111

Tabel 7 menunjukkan untuk data negatif yang diklasifikasikan benar negatif adalah 111 data, sedangkan ada 113 data positif diklasifikasikan positif. Terdapat 1 data *tweet* negatif yang salah diklasifikasikan ke dalam *tweet* positif dan tidak ada data *tweet* positif diklasifikasikan negatif.

2. Klasifikasi Data BRI dengan Naïve bayes Classifier

Berikut adalah performa klasifikasi yang didapatkan dengan *Naïve Bayes Classifier* pada data *tweet* pelayanan di BRI berdasarkan pada hasil klasifikasi dengan perhitungan nilai probabilitas tertinggi untuk setiap *tweet*.

Tabel 8.
Performa Klasifikasi dengan Naïve bayes Classifier pada Data BRI

Data	*Fold*	Akurasi	Presisi	*Recall*	*AUC*
Data Awal	1	0,956	0,00	0,00	0,491
	2	0,973	0,00	0,00	0,50
	3	0,965	0,00	0,00	0,495
	4	0,973	0,00	0,00	0,495
	Rata-Rata	0,966	0,00	0,00	0,495
Data SMOTE	1	0,991	0,99	0,99	0,991
	2	0,986	0,97	1,00	0,986
	3	0,991	0,98	1,00	0,991
	4	0,982	0,99	0,97	0,982
	Rata-Rata	0,988	0,983	0,990	0,988
Data dengan *Lexicon*	1	0,791	0,35	0,32	0,600
	2	0,782	0,12	0,05	0,489
	3	0,800	0,30	0,16	0,542
	4	0,833	0,00	0,00	0,494
	Rata-Rata	0,801	0,192	0,132	0,531

Data yang memiliki hasil paling baik adalah data dengan klasifikasi menggunakan SMOTE dengan *fold* pada *subset* ke-1 dengan nilai akurasi sebesar 99,1%, nilai presisi sebesar 99%, nilai *recall* sebesar 99%, dan nilai AUC sebesar 99,1%. Sehingga, dapat diketahui bahwa *subset* ke-1 pada data SMOTE adalah *subset* yang memiliki nilai performa klasifikasi terbaik. Berikut adalah *confusion matrix* untuk *subset* ke-1.

Tabel 9.
Confusion Matrix untuk NBC pada Data SMOTE BRI

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	112	1
Negatif	1	112

Tabel 9 menunjukkan hasil bahwa *confusion matrix u*ntuk data negatif yang diklasifikasikan benar negatif adalah 112 data, sedangkan terdapat 112 data positif diklasifikasikan positif.

3. Klasifikasi Data BRI dengan Suport Vector Machine

Untuk klasifikasi dengan SVM kernel liner diketahui hasil nilai akurasi, presisi, *recall*, dan AUC dengan nilai parameter C yang optimum yakni 0,1. *Fold subset* terbaik pada data SMOTE dengan nalai C sebesar 0,1 terdapat pada *subset* ke-1 yang memiliki nilai performa klasifikasi yang terbaik pula yaitu dengan nilai akurasi sebesar 98,2%, nilai presisi sebesar 97%, nilai *recall* sebesar 100%, dan AUC sebesar 98,2%. Berdasarkan *subset* ke-1 tersebut dapat dikatakan bahwa *subset* tersebut adalah yang terbaik sehingga model yang didapatkan untuk metode SVM kernel linier dengan hasil yang paling optimum dituliskan pada persamaan berikut.

$$K(\mathrm{x}_i, \mathrm{x}_j) = (\mathrm{x}^T\ \mathrm{x}) + 0{,}1$$

Selanjutnya dari model SVM kernel linier dengan menggunakan C yang paling optimum yakni sebesar 0,1 dapat dibuat *confusion matrix* untuk melihat ketepatan klasifikasi pada data pelayan di Bank BRI. Berikut adalah *confusion matrix* untuk *subset* ke-1 pada data SMOTE.

Tabel 10.
Confusion Matrix SVM Kernel Linear pada Data SMOTE BRI

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	113	0
Negatif	4	108

Tabel 10 menunjukkan bahwa nilai yang berada di diagonal utama adalah hasil keputusan yang benar sehingga dapat diketahui bahwa ketepatan klasifikasi data negatif yang diklasifikasikan benar negatif adalah sebanyak 108 data *tweet*, sedangkan untuk data positif yang benar diklasifikasikan positif adalah sebanyak 113 data *tweet*. Pada hasil klasifikasi pelayanan Bank BRI ini terdapat 0 data *tweet* positif yang diklasifikasikan negatif dan terdapat 4 data *tweet* negatif yang diklasifikasikan positif.

Selanjutnya untuk SVM dengan kernel RBF, parameter paling optimum pada kernel RBF terdapat pada pada C sebesar 100 serta γ sebesar 0,01 dengan *fold subset* terbaik yakni pada *subset* ke-3 yang memiliki nilai performa klasifikasi yaitu nilai akurasi sebesar 100%, nilai presisi sebesar 100%, nilai *recall* sebesar 100%, dan nilai AUC sebesar 100%. Nilai γ yang digunakan yakni sebesar 0,01 kemudian disubstitusikan pada persamaan kernel RBF fungsi kernel RBF dapat dituliskan pada persamaan berikut.

$$K(\mathrm{x}_1, \mathrm{x}_2) = \exp(-0{,}01\|\mathrm{x}_1, \mathrm{x}_2\|^2)$$

Selanjutnya akan dibuat *confusion matrix* untuk melihat ketepatan klasifikasi pada data pelayan di Bank BRI. Berikut adalah *confusion matrix* untuk *subset* ke-3 pada data awal.

Tabel 11.
Confusion Matrix untuk SVM Kernel RBF pada Data Awal BRI

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	113	0
Negatif	0	112

Tabel 11 dapat diketahui bahwa ketepatan klasifikasi data *tweet* bersentimen negatif yang diklasifikasikan benar memiliki sentimen negatif adalah sebanyak 112 data *tweet*, sedangkan untuk data bersentimen positif yang diklasifikasikan benar memiliki sentimen positif ada sebanyak 113 data *tweet*.

D. Klasifikasi Pada Data Bank Mandiri

Sebelum dilakukan klasifikasi data *tweet* pelayanan BRI setelah dilakukan praproses akan dihitung nilai TF-IDF untuk setiap kata kunci yang telah didapatkan.

1. Klasifikasi Data Mandiri dengan Regresi Logistik Biner

Berikut adalah hasil klasifikasi dengan menggunakan metode Regresi Logistik Biner. Namun terlebih dahulu dilakukan uji signifikansi parameter untuk mengetahui variabel yang signfikan terhadap model. Berdasarkan hasil yang diperoleh dapat diketahui bahwa nilai G^2 adalah sebesar 717,81 yang berarti dapat diputuskan tolak H_0 karena nilai G^2 lebih dari nilai $\chi^2_{(0.05,340)}$ sebesar 527,862 sehingga dapat disimpulkan bahwa terdapat pengaruh yang signifikan antara variabel kata kunci terhadap variabel klasifikasi sentimen. Untuk hasil pengujian signifikansi parameter menghasilkan bahwa variabel kata kunci dari variabel besok, duit, *edc*, *emoney*, hingga *user* signfikan terhadap model dan model regresi logistik biner dapat ditulis sebagai berikut.

$$\pi(x) = \frac{\exp(-2,95\times10+1,27\times10X_{5}+2,45\times10X_{53}+\ldots+1,24\times10X_{463})}{1+\exp(-2,95\times10+1,27\times10X_{5}+2,45\times10X_{53}+\ldots+1,24\times10X_{463})}$$

Berikut adalah performa klasifikasi yang didapatkan dengan Regresi Logistik Biner untuk data *tweet* pada pelayanan Bank Mandiri.

Tabel 15.
Performa Klasifikasi dengan Regresi Logistik Biner pada Data Bank Mandiri

	Folds	*Accuracy*	*Precision*	*Recall*	*AUC*
Data Awal	1	0,996	0,00	0,00	0,50
	2	0,990	0,00	0,00	0,50
	3	0,990	0,00	0,00	0,50
	4	0,992	0,00	0,00	0,50
	Rata-Rata	0,992	0,00	0,00	0,50
Data SMOTE	1	0,989	0,98	1,00	0,989
	2	0,991	0,98	1,00	0,991
	3	0,987	0,98	1,00	0,987
	4	0,995	0,99	1,00	0,995
	Rata-Rata	0,991	0,983	1,00	0,991
Data dengan *Lexicon*	1	0,809	0,63	0,18	0,576
	2	0,822	0,85	0,24	0,615
	3	0,820	0,70	0,26	0,613
	4	0,814	0,69	0,21	0,593
	Rata-Rata	0,816	0,717	0,222	0,599

Berdasarkan pada Tabel 15 tersebut data yang memiliki rata-rata performa klasifikasi paling tinggi yakni pada data dengan klasifikasi menggunakan SMOTE dimana *subset* ke-4 memiliki nilai performa klasifikasi yang terbaik Berikut adalah *confusion matrix* untuk *subset* ke-4.

Tabel 16.
Confusion Matrix untuk Regresi Logistik Biner pada Data SMOTE Bank Mandiri

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	641	0
Negatif	6	636

Tabel 16 menunjukkan bahwa ketepatan klasifikasi untuk data negatif yang diklasifikasikan benar negatif adalah 636 data, sedangkan untuk data positif diklasifikasikan positif adalah sebanyak 641 data *tweet*. Tidak terdapat data *tweet* positif yang salah diklasifikasikan ke dalam *tweet* negatif dan ada 6 data *tweet* negatif diklasifikasikan positif.

2. Klasifikasi Data Bank Mandiri dengan Naïve bayes Classifier

Selanjutnya akan dilakukan klasifikasi *tweet* pelayanan pada Bank Mandiri sama seperti pada data BRI. Berikut adalah performa klasifikasi yang didapatkan dengan *Naïve Bayes Classifier* pada data *tweet* pelayanan di Bank Mandiri.

Tabel 17.
Performa Klasifikasi dengan NBC pada Data Bank Mandiri

	Folds	*Accuracy*	*Precision*	*Recall*	*AUC*
Data Awal	1	0,996	0,00	0,00	0,50
	2	0,990	0,00	0,00	0,50
	3	0,990	0,00	0,00	0,50
	4	0,992	0,00	0,00	0,50
	Rata-Rata	0,992	0,00	0,00	0,50
Data SMOTE	1	0,995	0,99	1,00	0,995
	2	0,997	1,00	1,00	0,997
	3	0,994	0,99	1,00	0,994
	4	0,996	0,99	1,00	0,996
	Rata-Rata	0,996	0,993	1,00	0,996
Data dengan *Lexicon*	1	0,771	0,43	0,35	0,616
	2	0,777	0,50	0,40	0,641
	3	0,803	0,54	0,42	0,662
	4	0,764	0,43	0,36	0,616
	Rata-Rata	0,778	0,475	0,382	0,633

Data yang memiliki hasil paling baik adalah pada data SMOTE dengan *fold* terbaik yakni pada *subset* ke-2 yang memiliki nilai akurasi sebesar 99,7%, nilai presisi sebesar 100%, nilai *recall* sebesar 100%, dan nilai AUC sebesar 99,7%. Berikut adalah *confusion matrix* untuk *subset* ke-2.

Tabel 18.
Confusion Matrix untuk NBC pada Data *Lexicon* Bank Mandiri

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	641	0
Negatif	3	639

Tabel 18 menunjukkan hasil bahwa untuk data negatif yang diklasifikasikan benar negatif adalah sebanyak 639 data *tweet*, sedangkan untuk data positif diklasifikasikan benar positif ada sebanyak 641 data *tweet*. Tidak data positif yang salah diklasifikasikan negatif dan untuk data *tweet* negatif yang salah diklasifikasikan positif terdapat sebanyak 3 data *tweet*.

3. Klasifikasi Data Bank Mandiri dengan Suport Vector Machine

Selanjutnya akan dilakukan klasifikasi dengan menggunakan SVM kernel linier dan RBF. Untuk SVM kernel linier *fold subset* terbaik untuk data SMOTE dengan nalai C sebesar 0,01 memiliki nilai performa klasifikasi yang terbaik pula yaitu dengan nilai akurasi sebesar 94,3%, nilai presisi sebesar 98%, nilai *recall* sebesar 91%, dan nilai AUC sebesar 94,3%. Sehingga model yang didapatkan untuk metode SVM kernel linier dengan hasil yang paling optimum dituliskan pada persamaan berikut.

$$K(\mathrm{x}_i, \mathrm{x}_j) = (\mathrm{x}^T\ \mathrm{x}) + 0,01$$

Selanjutnya dari model SVM kernel linier dengan menggunakan C yang paling optimum yakni sebesar 0,01 dapat dibuat *confusion matrix* untuk melihat ketepatan klasifikasi pada data pelayan di Bank Mandiri. Berikut adalah *confusion matrix* untuk *subset* ke-3.

Tabel 19.
Confusion Matrix untuk SVM Kernel Linear pada Data SMOTE Bank Mandiri

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	582	59
Negatif	14	628

Tabel 19 menunjukkan hasil bahwa data negatif yang diklasifikasikan benar negatif sebanyak 628 data *tweet*, sedangkan untuk data positif yang benar diklasifikasikan positif ada sebanyak 582 data *tweet*. Untuk data positif yang salah diklasifikasikan negatif ada sebanyak 59 data *tweet* dan untuk data *tweet* negatif yang diklasifikasikan positif ada sebanyak 14 data *tweet*.

Sedangkan untuk SVM kernel RBF parameter paling optimum terdapat pada pada C sebesar 10 serta γ sebesar 0,01 pada data klasifikasi menggunakan SMOTE. *Fold subset* terbaik yakni pada *subset* ke-3 yang memiliki nilai performa klasifikasi yaitu nilai akurasi sebesar 99,6%, nilai presisi sebesar 99%, nilai *recall* sebesar 100%, dan nilai AUC sebesar 99,6% sehingga fungsi kernel RBF dapat dituliskan pada persamaan berikut.

$$K(\mathrm{x}_1, \mathrm{x}_2) = \exp(-0,01\|\mathrm{x}_1, \mathrm{x}_2\|^2)$$

Selanjutnya akan dibuat *confusion matrix* untuk melihat ketepatan klasifikasi pada data pelayan di Bank Mandiri.

Tabel 20.
Confusion Matrix untuk SVM Kernel RBF pada Data Awal Bank Mandiri

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	641	0
Negatif	5	637

Untuk data negatif yang diklasifikasikan benar negatif pada Tabel 20 adalah sebanyak 637 data *tweet*, sedangkan untuk data positif diklasifikasikan benar positif ada sebanyak 641 data *tweet* juga. Tidak ada data *tweet* positif yang salah diklasifikasikan negatif dan terdapat 5 data *tweet* negatif yang salah diklasifikasikan positif.

E. *Pemilihan* Metode Terbaik

Tabel 21.
Pemilihan Metode Terbaik pada Data Pelayanan BRI

Data	Metode	Jenis	Akurasi	Presisi	*AUC*
Data Awal	RBL	*Training*	0,975	0,00	0,50
		Testing	0,975	0,00	0,50
	NBC	*Training*	0,971	0,00	0,498
		Testing	0,966	0,00	0,495
	SVM Linier	*Training*	0,975	0,00	0,50
		Testing	0,975	0,00	0,50
	SVM RBF	*Training*	0,982	1,00	0,654
		Testing	0,973	0,00	0,498
Data SMOTE	RBL	*Training*	0,991	0,985	0,991
		Testing	0,985	0,973	0,985
	NBC	*Training*	0,987	0,987	0,987
		Testing	0,987	0,981	0,987
	SVM Linier	*Training*	0,983	0,970	0,983
		Testing	0,978	0,960	0,977
	SVM RBF	*Training*	0,997	0,993	0,995
		Testing	0,992	0,984	0,992
Data dengan *Lexicon*	RBL	*Training*	0,836	0,00	0,50
		Testing	0,836	0,00	0,50
	NBC	*Training*	0,865	0,70	0,647
		Testing	0,818	0,355	0,537
	SVM Linier	*Training*	0,836	0,00	0,50
		Testing	0,836	0,00	0,50
	SVM RBF	*Training*	0,908	0,94	0,732
		Testing	0,829	0,442	0,538

Tabel 21 merupakan ringkasan nilai akurasi, presisi, *recall*, dan AUC dengan masing-masing metode untuk data *tweet* pelayanan di BRI.

Secara keseluruhan, metode yang paling cocok digunakan untuk melakukan klasifikasi data pelayanan nasabah BRI adalah dengan metode SVM kernel RBF karena nilai AUC yang dimiliki selalu paling tinggi dengan menggunakan data SMOTE. Karena data awal dan data *lexicon* memiliki proporsi yang tidak seimbang sehingga data SMOTE lebih baik digunakan. Selanjutnya, akan dibandingkan pula performa klasifikasi data *tweet* pelayanan di Bank Mandiri.

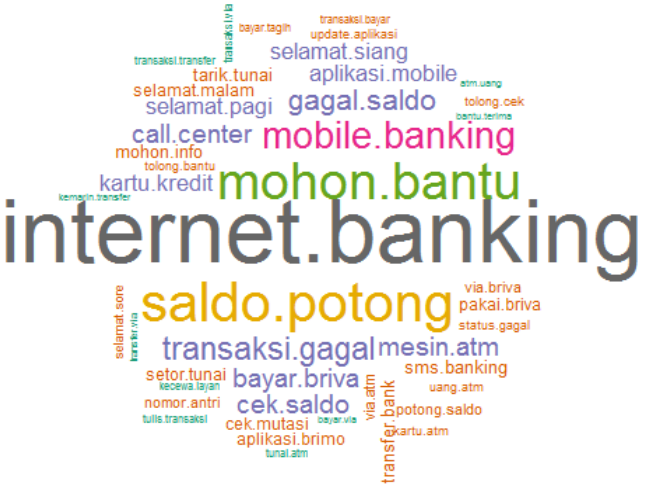
Tabel 22.
Pemilihan Metode Terbaik Pada Data Pelayanan Bank Mandiri

Data	Metode	Jenis	Akurasi	Presisi	*AUC*
Data Awal	RBL	*Training*	0,992	0,00	0,50
		Testing	0,992	0,00	0,50
	NBC	*Training*	0,992	0,00	0,507
		Testing	0,992	0,00	0,50
	SVM Linier	*Training*	0,992	0,00	0,50
		Testing	0,992	0,00	0,50
	SVM RBF	*Training*	0,993	1,00	0,585
		Testing	0,992	0,00	0,50
Data SMOTE	RBL	*Training*	0,993	0,990	0,993
		Testing	0,991	0,983	0,991
	NBC	*Training*	0,996	0,992	0,996
		Testing	0,996	0,993	0,996
	SVM Linier	*Training*	0,937	0,978	0,937
		Testing	0,936	0,976	0,936
	SVM RBF	*Training*	0,992	0,986	0,992
		Testing	0,990	0,983	0,990
Data dengan *Lexicon*	RBL	*Training*	0,836	0,84	0,635
		Testing	0,811	0,725	0,589
	NBC	*Training*	0,820	0,602	0,691
		Testing	0,775	0,475	0,624
	SVM Linier	*Training*	0,790	0,732	0,507
		Testing	0,788	0,45	0,504
	SVM RBF	*Training*	0,814	0,898	0,568
		Testing	0,811	0,878	0,559

Secara keseluruhan, berdasarkan Tabel 22 maka metode *Naïve Bayes Classifier* patut dipilih atau layak digunakan dalam melakukan klasifikasi untuk pelayanan nasabah Bank Mandiri dengan menggunakan data yang telah di SMOTE.

F. *Visualisasi Word Cloud*

Word cloud disini akan menampilkan kata-kata yang sering muncul sehingga berdasarkan kata-kata yang sering muncul tersebut dapat diketahui saran atau peningkatan yang harus dilakukan oleh masing-masing bank. Berikut adalah hasil dari visualisasi *owrd cloud* untuk BRI.



Gambar 4. *Word Cloud* Bigram Sentimen Negatif BRI

Berdasarkan pada Gambar 4 dapat diketahui pula bahwa kata "terima kasih" merupakan kata yang paling banyak muncul. Lalu ada kata "saldo potong", "transaksi gagal", "gagal saldo", dan "mohon bantu" yang juga sering muncul. Hal tersebut berarti bahwa pada Gambar 4 tersebut terdapat kata-kata negatif yang mengindikasikan bahwa terdapat banyak *tweet* dengan sentimen negatif didalamnya.

Gambar 5. *Word Cloud* Bigram Sentimen Negatif Bank Mandiri.

Berdasarkan pada Gambar 5 tersebut dapat diketahui bahwa kata "saldo potong" serta "internet banking" merupakan kata yang paling sering muncul dalam data *tweet* pelayanan nasabah Bank Mandiri. Pengaduan tentang internet banking kebanyakan terjadi karena saldo yang terpotong namun saldo tidak masuk kedalam transaksi yang dituju sehingga banyak nasabah yang melakukan komplainnya. Jika dilihat lebih lanjut pada data *tweet* pelayanan nasabah Bank Mandiri dengan sentimen negatif kebanyakan merupakan komplain tentang adanya masalah pada aplikasi mandiri *online.*

V. KESIMPULAN DAN SARAN

Berdasarkan analisis dan pembahasan diatas dapat diperoleh kesimpulan sebagai berikut.

1. Hasil klasifikasi data awal, data SMOTE, dan data *lexicon* memiliki hasil terbaik pada data SMOTE untuk data layanan BRI dan Bank Mandiri.
2. Hasil klasifikasi dengan Regresi Logistik Biner diketahui memiliki akurasi terbaik untuk BRI adalah sebesar 08,5% dengan nilai AUC sebesar 98,5% dengan data SMOTE. Untuk Bank Mandiri nilai akurasinya adalah sebesar 99,3% dengan nilai AUC 99,3% dengan data SMOTE.
3. Hasil klasifikasi menggunakan metode *Naïve Bayes Classifier* mempunyai nilai AUC sebesar 98,7% untuk BRI dengan data SMOTE dan 99,6% untuk Bank Mandiri dengan data SMOTE.
4. Untuk hasil klasifikasi menggunakan SVM diperoleh kernel yang menghasilkan akurasi dan AUC paling tinggi adalah dengan kernel RBF jika menggunakan SMOTE untuk BRI dan Bank Mandiri. Nilai AUC untuk BRI adalah sebesar 99,2% dan untuk Bank Mandiri adalah sebesar 99%.
5. Metode terbaik untuk mengklasifikasikan sentimen layanan pada BRI adalah SMOTE-SVM dengan kernel RBF, sedangkan untuk Bank Mandiri adalah dengan SMOTE-NBC karena memiliki nilai AUC yang paling tinggi.

Saran untuk penelitian selanjutnya adalah jika menggunakan klasifikasi berdasarkan pada kamus *lexicon*, perlu diperhatikan daftar kata positif dan daftar kata negatif yang dimasukkan agar hasil klasfikasi dapat memberikan hasil yang tepat dan akurat, karena pada penelitian masih banyak *tweet* yang kurang tepat pengklasifikasiannya dengan menggunakan kamus *lexicon.* Hal lain yang perlu diperhatikan yakni, jika data tidak seimbang maka untuk Regresi Logistik Biner sebaiknya digunakan *threshold* yang sesuai agar hasil klasifikasi dapat sesuai dan menghasilkan performa klasifikasi yang tinggi. Selanjutnya untuk data yang tidak seimbang jika tidak dilakukan SMOTE akan menghasilkan performa klasifikasi yang sangat kecil sehingga sangat dianjurkan untuk melakukan SMOTE pada data yang tidak seimbang.

DAFTAR PUSTAKA

[1] D. Hosmer and J. Lemeshow, *Applied Logistic Regression*, 2nd ed. USA: John Wiley & Sons Inc, 2000.

[2] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston: Addison-Wesley Longman Publishing Co, 2005.

[3] N. Kumar, "World towards Advance Web Mining: A Review," *Am. J. Syst. Softw.*, vol. 3, pp. 44–61, 2015.

[4] J. Brownlee, "How to Prepare Text Data for Machine Learning with Scikit-Learn," 2019. [Online]. Available: https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/.

[5] C. Durajati and A. B. Gumelar, "Pemanfaatan Teknik Supervised untuk Klasifikasi Teks Bahasia Indonesia," *J. Link*, vol. 16, pp. 1–8, 2012.

[6] K. Hardle, D. Prastyo, and M. Hafner, "Support Vector Machines with Evolutionary Model Selection for Default Prediction." 2013.

[7] E. Gokgoz and A. Subasi, "Comparison of Decision Tree Algorithms for EMG Signal Classification using DWT," *Biomed. Signal Process. Control*, vol. 18, pp. 138–144, 2015.

[8] J. Han and M. Kamber, *Data mining: concepts and techniques*, 3rd ed. Burlington, MA: Elsevier, 2011.

[9] V. Chawla, K. W. Bowyer, L. O. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[10] M. Bekkar, D. Djemaa, and D. Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Set," *J. Inf. Eng. Appl.*, vol. 3, pp. 27–38, 2013.