Analisis Sentimen Nasabah terhadap Pelayanan PT. Bank Mandiri (TBK) dengan Metode Regresi Logistik Biner

Carica Deffa Yullinda
Dept. Computer Science & Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
carica.deffa@mail.ugm.ac.id

Devito Karunia Susilo
Dept. Computer Science & Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
devito.280202@mail.ugm.ac.id

Irbah Asfarina
Dept. Computer Science & Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
irbahasfarina@mail.ugm.ac.id

Maulana Akbar Ibrahim
Dept. Computer Science & Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
maulanaakbar02@mail.ugm.ac.id

Petra Bayu Pangestu
Dept. Computer Science & Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
petrabayu19@mail.ugm.ac.id

Rama Destrian Hartadi
Dept. Computer Science & Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
ramadestrianhartadi@mail.ugm.ac.id

Setyo Aji Pratomo
Dept. Computer Science & Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
setyoaji02@mail.ugm.ac.id

Abstrak*—Kualitas pelayanan bank dapat terlihat dari penilaian para nasabah. Untuk mendapatkan penilaian masyarakat tentang pelayanan bank yang pada akhirnya dapat membantu bank melakukan perbaikan pelayanan, maka dilakukan analisis sentimen dengan data yang diperoleh dari jejaring sosial Twitter. Bank yang dianalisis yaitu Bank Mandiri dengan metode Regresi Logistik Biner. Analisis sentimen diklasifikasikan ke dalam dua kelas, yaitu sentimen positif dan negatif. Dilakukan serangkaian tahapan yaitu pengambilan data twitter, *text prepocessing,* ekstraksi fitur, representasi fitur, hingga membuat machine learning model dengan Regresi Logistik Biner. Jumlah tweet yang digunakan adalah 100 tweet yang dibagi ke dalam 90 data pelatihan dan 10 data tes. Hasil uji dengan 10-fold cross validation menunjukkan nilai rata-rata akurasi 71%, precision 70%, dan recall 58%. Sentimen yang didapatkan cenderung negatif dengan presentase sebesar 57% sentimen negatif dan 43% sentimen positif pada data *training* serta 71,8% sentimen negatif dan 28,2% sentimen positif pada data *testing.

Keywords—analisis sentimen, twitter, klasifikasi, Regresi Logistik Biner, bank

I. PENDAHULUAN

Bank merupakan suatu badan usaha yang melaksanakan kegiatan berupa pengelolaan keuangan. Dalam dunia perbankan, dikenal istilah nasabah sebagai pengguna jasa atau fasilitas bank. Nasabah mengambil peran penting bagi eksistensi bank karena pelayanan suatu bank akan mempengaruhi kepuasan nasabah. Kualitas pelayanan bank dapat dilihat dari respon atau penilaian para nasabahnya. Pelayanan yang memuaskan akan memberikan penilaian positif dari nasabah dan sebaliknya, pelayanan yang kurang baik akan memberikan penilaian negatif. Tentunya, nasabah akan memilih bank dengan kualitas pelayanan yang baik sehingga menghasilkan suatu kepercayaan dan akhirnya menggunakan jasa dari bank tersebut.

Dengan perkembangan teknologi saat ini, setiap orang dapat lebih mudah menyampaikan opini atau pendapatnya terhadap suatu topik tertentu di ruang publik melalui media sosial. Dari berbagai media sosial yang ada, Twitter merupakan platform yang populer dan banyak digunakan masyarakat untuk mengemukakan pendapat atau memberikan komentar. Menurut laporan statistik per Januari 2022, pengguna Twitter di Indonesia menempati posisi tertinggi kelima dengan jumlah pengguna sebanyak 18, 45 juta. Melalui Twitter, dapat diketahui pendapat atau pandangan nasabah berupa respon positif maupun keluhan-keluhan terhadap pihak bank yang bersangkutan. Pendapat atau pandangan terhadap sesuatu tersebut dikenal dengan sentimen. Sentimen dari nasabah akan dianalisis sehingga dapat diklasifikasikan menjadi sentimen positif dan negatif untuk mengetahui pandangan nasabah tentang kualitas dari bank yang digunakan. Dengan demikian, bank dapat melakukan perbaikan pelayanan berdasarkan hasil analisis sentimen.

Twitter memiliki *Application Programming Interface (API)* yang memungkinkan dalam mengakses dan memperoleh data. Dalam melakukan analisis sentimen, diperlukan pengolahan data teks atau *text preprocessing* dengan *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 temu kembali informasi, penambangan data, pembelajaran mesin, ilmu statistika, dan linguistik komputasi [1]. Tahapan dalam *text preprocessing* meliputi *Parsing, Lexical Analysis, Stop-word Removal, Phrase Detection, Stemming,* dan *Lemmatization*. Terdapat berbagai metode klasifikasi dalam ilmu statistik, contohnya Regresi Logistik Biner. Regresi logistik biner adalah metode analisis data yang digunakan untuk mencari hubungan antara variabel respon (y) yang bersifat biner

dengan variabel (x) yang bersifat polikotomus [1]. Pada project ini akan dilakukan analisis sentimen nasabah terhadap pelayanan PT. Bank Mandiri (TBK) dengan metode Regresi Logistik Biner. Hasil analisis dengan metode tersebut diharapkan menghasilkan sentimen positif dan negatif dari nasabah PT. Bank Mandiri.

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 temu kembali informasi, penambangan data, pembelajaran mesin, ilmu statistika, dan linguistik komputasi. [1]

B. *Text Preprocessing*

Text preprocessing adalah tahap pertama dalam memproses data. Hal ini dilakukan untuk mengubah bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan. Tahapan dalam *text preprocessing* adalah *Parsing, Lexical Analysis, Stop-word Removal, Phrase Detection, Stemming*, dan *Lemmatization*. *Parsing* adalah proses pemecahan struktur dokumen menjadi komponen-komponen terpisah yang terdiri dari beberapa unit dokumen. *Lexical Analysis* atau Tokenisasi adalah proses penghasilan token dari tiap unit dokumen dengan melakukan penghilangan angka dan tanda baca, *case folding*, serta *cleaning*. *Stop-word Removal* atau *Filtering* adalah proses penyaringan kata-kata yang dianggap dapat mewakili dokumen dari token yang dihasilkan. *Phrase Detection* adalah proses deteksi 2 kata atau lebih yang dapat menjadi frasa. *Stemming* adalah proses pengubahan pengubahan setiap kata menjadi bentuk kata dasarnya dengan menghilangkan imbuhan pada setiap kata yang bukan merupakan kata dasar. Hasil dari proses *Stemming* disebut sebagai *term index* yang dapat digunakan untuk *text mining*.

C. *TF-IDF*

TF-IDF adalah suatu metode yang dimaksudkan untuk memberikan pembobotan terhadap pentingnya sebuah kata dalam koleksi atau korpus. Dengan menggunakan *TF-IDF* akan diketahui seberapa pentingnya suatu kata dalam kumpulan tweet berdasarkan frekuensi kemunculannya. Dengan melibatkan antara *Term Frequency* dan *Inverse Document Frequency* maka akan didapatkan nilai *TF-IDF*. *TF* bertujuan untuk menunjukkan jumlah kemunculan sebuah kata dalam suatu tweet, sedangkan *IDF* bertujuan untuk menghitung frekuensi kemunculan kata pada seluruh tweet. [2]

D. Regresi Logistik Biner

Regresi logistik biner adalah 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) = π. Fungsi logistiknya dapat dituliskan pada persamaan (2).

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

E. *K-Fold Cross Validation*

K-fold Cross Validation adalah salah satu metode statistik yang digunakan untuk mengevaluasi kinerja model. Metode ini digunakan untuk mengetahui rata-rata keberhasilan dari suatu model. Model dilatih oleh subset data pembelajaran dan divalidasi oleh subset validasi.

K-fold Cross Validation diawali dengan membagi data sejumlah *n-fold* yang diinginkan. Dalam *n-fold* CV, data dibagi menjadi *n-fold* berukuran kira kira sama. Selanjutnya proses uji dan latih dilakukan sebanyak n kali.

III. METODOLOGI PENELITIAN

A. Sumber Data

Sumber data yang digunakan berasal dari kumpulan tweet yang diambil dengan menggunakan Twitter API. Kata kunci yang digunakan untuk mengambil tweet adalah '@bankmandiri'. Rentang waktu tweet yang diambil untuk data pelatihan adalah 17 Mei 2022 hingga 24 Mei 2022 sejumlah 100 tweet. Waktu tweet yang diambil untuk data tes adalah 25 Mei 2022 sejumlah 39 data.

B. Variabel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri dari variabel bebas X yang berupa hasil *TF-IDF* dari setiap tweet dan variabel terikat Y yang berupa sentimen dari tweet tersebut, dilambangkan dengan 0 untuk tweet dengan sentiment negatif dan 1 untuk tweet dengan sentimen positif.

C. Tahap Analisis

Tahapan analisis yang dilakukan dalam penelitian ini adalah sebagai berikut.

1. Mengambil data tweet menggunakan Twitter API.
2. Melabeli sentimen setiap tweet pada data pelatihan, 0 untuk sentimen negatif dan 1 untuk sentimen positif.
3. Melakukan *text preprocessing*, ekstraksi fitur, dan representasi fitur menggunakan *TF-IDF.*
4. Membuat model Pembelajaran Mesin menggunakan Regresi Logistik Biner
5. Membagi data menjadi 10-*fold cross validation* dan menguji skor akurasi dan presisi tiap *fold*
6. Melakukan analisis menggunakan model Regresi Logistik Biner
7. Membuat kesimpulan dan saran.

IV. ANALISIS DAN PEMBAHASAN

A. *Text Preprocessing* dan Pelabelan

Data tweet yang sudah dikumpulkan menggunakan Twitter API kemudian disimpan dalam bentuk file .csv dan diberi label untuk setiap tweet.

Tabel 1.

Contoh data sebelum *Text Preprocessing* dan Pelabelan

Sentimen	Teks
	LIVIN TOLONG LAAHHH MAU BELANJA NIH YAKALI EROR TERUS @livinpoin @bankmandiri
	Mantap @bankmandiri sebagai bank dengan asset terbesar di Indonesia

Karena metode yang digunakan adalah algoritma Regresi Logistik Biner, maka hanya ada 2 jenis label, yaitu 0 untuk negatif dan 1 untuk positif. Selanjutnya dilakukan *text preprocessing* untuk setiap data tweet.

Tabel 2.

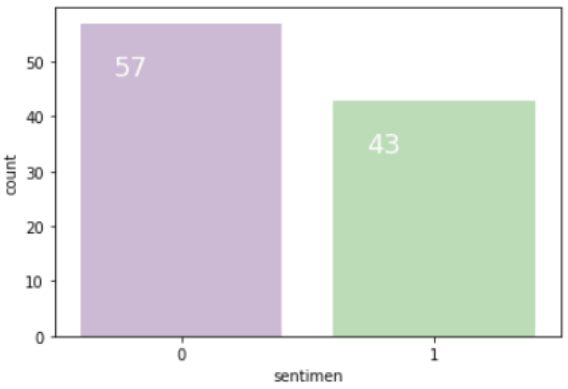
Contoh data setelah *Text Preprocessing* dan Pelabelan

Sentimen	Teks
0	[‘tolong’,‘mau’, ‘belanja’, ‘eror’, ‘terus’]
1	[‘mantap’, ‘bank’, ‘aset’, ‘besar’]

Hasil dari tahap *preprocessing* dan ekstraksi fitur didapatkan kamus dengan 236 kata yang berbeda. Kamus tersebut selanjutnya digunakan untuk menghitung nilai *TF-IDF* yang menjadi representasi fitur dari tiap *tweet*.

B. Karakteristik Data

Data tweet yang diambil menggunakan Twitter API berjumlah lebih dari 100. Saat pelabelan, data tersebut disaring sehingga hanya berjumlah 100 data untuk melatih model. Penyaringan yang dilakukan bertujuan agar rasio sentimen negatif dan positif tidak terlalu timpang. Setelah disaring, didapat 100 data tweet dengan rasio sentimen seperti yang ditunjukkan pada Gambar 1.



Gambar 1. Rasio Sentimen Negatif dan Positif dari Data Tweet untuk Pelatihan Model.

Dari 100 data tweet yang berhasil dikumpulkan, 57% data memiliki sentimen negatif, sementara untuk data dengan sentimen positif terdapat 43% data.

C. Klasifikasi Data dengan Regresi Logistik Biner

Data *TF-IDF* yang telah didapat selanjutnya digunakan untuk membuat model klasifikasi dengan Regresi Logistik Biner. Untuk validasi, digunakan 10-fold cross validation sehingga pada masing-masing fold akan terdapat 90 data untuk pelatihan dan 10 data untuk tes. Berikut adalah nilai performa yang didapatkan.

TABEL 4.

PERFORMA KLASIFIKASI DENGAN REGRESI LOGISTIK BINER

Fold	**Skor**		
	Accuracy	***Precision***	***Recall***
1	0.8	0.8	0.8
2	0.9	1.0	0.666
3	0.7	1.0	0.5
4	0.6	0.0	0.0
5	0.7	0.25	1.0
6	0.7	0.7142	0.833
7	0.9	0.75	1.0
8	0.3	1.0	0.125
9	0.7	1.0	0.4

Fold	Skor		
	Accuracy	*Precision*	*Recall*
10	0.8	0.5	0.5
Rata-rata	0.71	0.7014	0.5825

Performa terbaik didapat dari fold ke-7 dengan nilai akurasi 0.9, presisi 0.75, dan recall 1.0.

Setelah diperoleh model, dilakukan uji coba klasifikasi terhadap data tes sejumlah 39 *tweet* yang diperoleh dari tanggal 25 Mei 2022. Hasilnya adalah 28 *tweet* diklasifikasi sebagai *tweet* dengan sentimen negatif dan 11 *tweet* diklasifikasi sebagai *tweet* dengan sentimen positif.

V. KESIMPULAN DAN SARAN

Dari hasil uji, didapatkan rata-rata akurasi klasifikasi data tweet dengan kunci '@bankmandiri' melalui metode Regresi Logistik Biner sebesar 71%. Dengan menggunakan 10-fold cross validation, maka nilai terbaik didapatkan pada fold 7 dengan akurasi 90%, presisi 75%, dan recall 100%. Kemudian untuk analisis sentimen, pada data *training* didapatkan hasil sentimen negatif sebesar 57% dan sentimen positif sebesar 43% dengan pelabelan manual, dan pada data *testing* didapatkan hasil sentimen negatif sebesar 71.8% dan sentimen positif sebesar 28.2%. Hal tersebut mengindikasikan bahwa sentimen nasabah Bank Mandiri di akun twitter Bank Mandiri cenderung negatif.

DAFTAR PUSTAKA

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

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