Article - M. Diki Armanda.docx

by 2 Perpustakaan UMSIDA

Submission date: 12-Feb-2024 10:56AM (UTC+0700)
Submission ID: 2292390330
File name: Article - M. Diki Armanda.docx (595.86K)
Word count: 4680
Character count: 30386

Social Network Analysis of BSI Data Leakage Using Naive Bayes, Support Vector Machine (SVM), and Random Forest Classification Algorithms

[Social Network Analysis Terhadap Kebocoran Data BSI Menggunakan Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), Dan Random Forest]

Mokhamad Diki Armanda[1)], Rohman Dijaya[*,2)], Cindy Taurusta[3)]

[1)] *Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
[2)] *Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
[3)] *Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia*
*Email Penulis Korespondensi: rohman.dijaya@umsida.ac.id

Abstract. *This research aims to design a sentiment analysis detection system in the context of the BSI bank digital data hacking case. The main focus of the research is to compare the performance of three different classification algorithms, namely Random Forest, Naive Bayes, and Support Vector Machine (SVM), which were used to analyze sentiment in the BSI bank digital data hacking case. Positive sentiment expresses joy and appreciation for something, reflecting feelings of pleasure and satisfaction with the expected results. In contrast, negative sentiment reflects disappointment, and disapproval, perhaps caused by an unsatisfactory experience or results that do not meet expectations. There is also a neutral sentiment that reflects impartiality in presenting an opinion, which tends to indicate an inability to take sides, perhaps due to a lack of emotional involvement or confusion in responding to a given situation. The data collected was 24,000 tweets, with neutral sentiment reaching 83.86%, negative sentiment 8.43%, and positive sentiment 7.71%. The evaluation results show that the Support Vector Machine classification algorithm has the highest accuracy in identifying sentiment in tweet data related to BSI, namely 95.72%. Meanwhile, Naive Bayes has an accuracy of 89.73%, and Random Forest reaches 95.7%. This research provides a deep understanding of sentiment analysis, especially in the context of data leaks in the digital era.*

Keywords - *Sentiment Analysis, BSI Data Leak, Random Forest, Support Vector Machine, Naive Bayes*

Abstrak. *Penelitian ini bertujuan untuk merancang sistem deteksi sentimen analisis dalam konteks kasus peretasan data digital bank BSI. Fokus utama penelitian adalah membandingkan kinerja tiga algoritma klasifikasi yang berbeda, yaitu Random Forest, Naive Bayes, dan Support Vector Machine (SVM), yang digunakan untuk menganalisis sentimen pada kasus peretasan data digital bank BSI. Sentimen positif mengekspresikan kegembiraan, dan penghargaan terhadap suatu hal, mencerminkan perasaan senang dan kepuasan atas hasil yang diharapkan. Sebaliknya, sentimen negatif mencerminkan kekecewaan, dan ketidaksetujuan, mungkin disebabkan oleh pengalaman yang tidak memuaskan atau hasil yang tidak sesuai dengan harapan. Ada juga sentimen netral yang mencerminkan ketidakberpihakan dalam menyampaikan suatu pendapat, yang cenderung menunjukkan ketidakmampuan untuk memihak, mungkin karena kurangnya keterlibatan emosional atau kebingungan dalam merespons situasi yang diberikan. Data yang dikumpulkan sebanyak 24.000 tweet, dengan hasil sentimen netral mencapai 83,86%, sentimen negatif 8,43%, dan sentimen positif 7,71%. Hasil evaluasi menunjukkan bahwa algoritma klasifikasi Support Vector Machine memiliki akurasi tertinggi dalam mengidentifikasi sentimen pada data tweet terkait BSI, yaitu sebesar 95,72%. Sementara itu, Naive Bayes memiliki akurasi sebesar 89,73%, dan Random Forest mencapai 95,7%. Penelitian ini memberikan pemahaman yang mendalam tentang analisis sentimen, khususnya dalam konteks kebocoran data di era digital.*

Kata Kunci - *Analisis Sentimen, Kebocoran Data BSI, Random Forest, Support Vector Machine, Naive Bayes*

I. PENDAHULUAN

Pada masa era digital yang semakin berkembang, pertukaran informasi melalui platform media sosial telah menjadi elemen tak terpisahkan dalam kehidupan sehari-hari. Berbagai topik dan isu diperdebatkan, berbagi gagasan, dan mengungkapkan pendapat melalui platform tersebut[1]. Namun, dengan pertumbuhan pesatnya penggunaan media sosial, terbuka pula potensi risiko terkait kerahasiaan dan privasi informasi. Oleh karena itu, diperlukan adanya perlindungan data pribadi yang kuat untuk melindungi pelanggan dari risiko kebocoran data[2].

Risiko kebocoran data (*data leaks*) pada platform media sosial, di mana informasi pribadi atau data sensitif dapat jatuh ke tangan yang salah dan disalahgunakan. Di Indonesia Data Pribadi diatur dalam Undang – Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi, Data Pribadi adalah data tentang orang perseorangan yang teridentifikasi atau dapat diidentifikasi secara tersendiri atau dikombinasi dengan informasi lainnya baik secara

langsung maupun tidak langsung melalui sistem elektronik atau non-elektronik[3]. Dalam konteks ini, penelitian dan analisis kinerja algoritma dalam menghadapi tantangan seperti ini menjadi sangat penting. Algoritma-algoritma ini dapat mengolah, menganalisis, dan mengklasifikasikan teks berdasarkan berbagai parameter seperti sentimen, emosi, atau konteks tertentu. Dengan penerapan algoritma yang tepat, informasi dalam bentuk teks di platform media sosial seperti Twitter dapat diidentifikasi dengan akurasi tinggi[4].

Analisis sentimen, sebagai alat yang kuat dalam bidang ilmu data, memeriksa sentimen publik untuk mengungkap reaksi terhadap peristiwa atau kejadian tertentu. Dengan mengkaji pola bahasa dalam unggahan media sosial atau data teks, analisis sentimen mengurai pendapat publik, memahami emosi, dan sikap yang dominan dalam percakapan daring[5]. Analisis sentimen berperan penting dalam ranah finansial dengan menyelidiki dan mengevaluasi pandangan dan perasaan pelaku pasar yang tercermin dalam data teks, memberikan wawasan berharga tentang potensi perubahan pasar dan investor. Sentimen analisis finansial dapat memberikan gambaran tentang persepsi dan opini masyarakat terhadap bank, memungkinkan bank untuk merespons secara cepat terhadap perubahan sentimen pasar, meningkatkan kepercayaan pelanggan, dan merancang strategi bisnis yang lebih tepat dalam mendukung pertumbuhan dan keberlanjutan operasional[6].

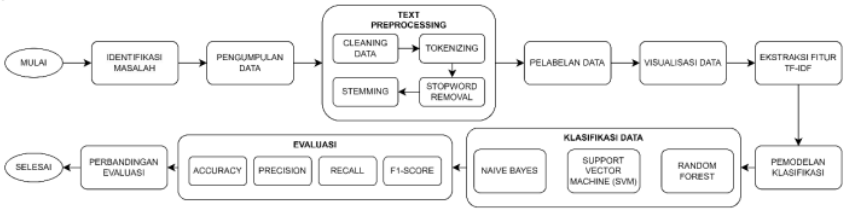
Sentimen yang muncul dalam berbagai konteks mencerminkan beragam emosi, mulai dari yang positif, netral, hingga negatif. Ekspresi positif sering kali menunjukkan kepuasan, kegembiraan, dan penghargaan terhadap suatu hal, yang mencerminkan perasaan senang dan kepuasan atas hasil yang diharapkan[7]. Sebaliknya, ekspresi negatif sering menunjukkan ketidakpuasan, kekecewaan, dan ketidaksetujuan, yang mungkin disebabkan oleh pengalaman yang tidak memuaskan atau hasil yang tidak sesuai dengan harapan. Di antara kedua ekstrem ini, terdapat juga ekspresi netral yang mencerminkan ketidakberpihakan atau ketidakberdayaan dalam menyampaikan suatu pendapat, ekspresi netral ini cenderung menunjukkan ketidakmampuan untuk memihak, mungkin karena kurangnya keterlibatan emosional atau kebingungan dalam merespons situasi yang diberikan[8].

BSI (Bank Syariah Indonesia) adalah salah satu bank BUMN yang ada di negara Indonesia dan merupakan salah satu perusahaan perbankan terkemuka dalam menyediakan solusi perangkat lunak bisnis. Terlepas dari keberhasilannya dalam menghadirkan solusi inovatif, informasi yang diungkapkan terkait pelanggan, proyek, atau layanan bisnis, apabila jatuh ke tangan yang salah, dapat menimbulkan dampak serius, termasuk kerugian finansial dan reputasi[9]. Pada tanggal 8 Mei 2023 awalnya PT. Bank Syariah Indonesia mengalami gangguan pelayanan digital, layanan seperti BSI Mobile, mesin ATM, dan teller di kantor cabang tidak dapat digunakan oleh nasabah. Pada Sabtu, 13 Mei 2023 media sosial Twitter dihebohkan oleh akun @darktracer_int yang menulis *tweet* bahwa LockBit 3.0 kelompok peretas asal Rusia mengaku sebagai pihak yang telah melakukan serangan ke sistem layanan Bank Syariah Indonesia[10].

Berdasarkan permasalahan di atas, maka penelitian ini akan mengembangkan sistem "*Social Network Analysis* Terhadap Kebocoran Data BSI Menggunakan Algoritma Klasifikasi *Naive Bayes, Support Vector Machine* (SVM), dan *Random Forest*". Penelitian dilakukan dengan membandingkan performa sentimen analisis melalui tiga algoritma klasifikasi utama yang sering digunakan dalam sentimen analisis, yaitu *Random Forest, Naive Bayes,* dan *Support Vector Machine* (SVM) menggunakan bahasa pemrograman python dalam konteks kasus BSI *Data Leak*. Penelitian dimulai dengan mengumpulkan data berupa tweet dari media sosial Twitter yang selanjutnya dilakukan *text preprocessing* untuk mempersiapkan data set sebelum diolah, kemudian dilakukan pelabelan dan pembobotan kata sebelum masuk ke proses klasifikasi model. Dan terakhir dievaluasi untuk mengetahui nilai *F1-score, recall*, presisi, dan akurasi.

II. METODE

Penelitian dilakukan berdasarkan data berupa *tweet* yang berkaitan dengan kasus peretasan data digital Bank Syariah Indonesia (BSI) berdasarkan media sosial Twitter. Diagram alir penelitian sebagaimana dapat dilihat pada Gambar 1.



Gambar 1. Diagram Alir Penelitian

Berdasarkan diagram alir pada Gambar 1 dapat dilihat proses dalam tahapan penelitian yang akan menjadi acuan dalam pengerjaan penelitian ini.

A. Identifikasi Masalah

Seperti yang sudah dijelaskan sebelumnya, penelitian ini akan berfokus pada:

1. Bagaimana merancang sistem deteksi sentimen analisis pada kasus peretasan data digital bank BSI?
2. Bagaimana membandingkan algoritma klasifikasi Random Forest, Naive Bayes, dan Support Vector Machine (SVM) yang digunakan untuk sistem social network analysis pada kasus peretasan data digital bank BSI?

B. Pengumpulan Data

Data yang diperoleh untuk penelitian ini terdiri dari serangkaian tweet yang dihimpun melalui media sosial Twitter menggunakan library Tweepy. Pengumpulan data dilakukan dalam rentang waktu satu minggu, mulai dari tanggal 9 Mei 2023 hingga 17 Mei 2023. Setelah berhasil mengumpulkan data, informasi tersebut disimpan dalam format CSV.

C. Text Processing

Text processing adalah langkah awal yang sangat penting dalam pengolahan data, bertujuan untuk menyederhanakan data dan mempersiapkannya agar dapat diolah dengan lebih efisien pada tahap selanjutnya. Proses preprocessing melibatkan pemilihan data yang relevan dan mengubahnya menjadi format yang lebih terstruktur. Dalam tahap ini, data yang berlimpah dan tidak relevan akan dieliminasi terlebih dahulu, memastikan bahwa hanya data yang dibutuhkan yang diproses dalam analisis selanjutnya. Dengan melakukan preprocessing, dataset menjadi lebih bersih dan siap untuk dijalani melalui proses analisis data lebih lanjut[11].

D. Pelabelan Data

Dalam konteks penelitian di bidang text mining, pelabelan data memainkan peran krusial dalam analisis sentimen teks. Penerapan teknik ini memungkinkan peneliti untuk mengklasifikasikan data teks ke dalam kategori sentimen yang sesuai, seperti positif, negatif, atau netral. Salah satu pendekatan yang umum digunakan adalah menggunakan library VADER Sentiment Analysis, sebuah alat analisis sentimen yang dirancang khusus untuk teks sosial media. Dengan memanfaatkan VADER, peneliti dapat secara otomatis menilai sentimen teks dalam data set dengan akurasi yang tinggi.

E. Pembobotan Kata

Pembobotan kata menggunakan TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode yang umum digunakan dalam analisis teks ilmiah. TF-IDF memberikan bobot pada kata-kata dalam suatu dokumen dengan mempertimbangkan frekuensi kemunculan kata tersebut dalam dokumen tersebut (TF) dan invers dari frekuensi kemunculan kata tersebut dalam seluruh koleksi dokumen (IDF). Dengan menggunakan metode ini, kata-kata yang sering muncul dalam suatu dokumen tetapi jarang muncul dalam koleksi dokumen secara keseluruhan akan memiliki bobot yang lebih tinggi. TF-IDF digunakan untuk mengekstraksi dan memberikan nilai penting pada kata-kata yang dapat mencirikan konten suatu dokumen secara lebih akurat, yang dapat digunakan dalam analisis teks ilmiah dan pengambilan informasi[12].

F. Naive Bayes

Algoritma klasifikasi Naive Bayes, umum digunakan dalam text mining, memisahkan dokumen ke dalam kategori berdasarkan teorema Bayes. Meskipun asumsi independensi antar fitur mungkin terlalu sederhana, prosesnya melibatkan pelatihan model dengan data terkategori dan perhitungan probabilitas saat menghadapi dokumen baru. Langkah pertama dalam metode naive bayes adalah menetapkan data set dan menentukan atribut serta class yang akan digunakan. Selanjutnya, probabilitas class dan probabilitas atribut dihitung, dan akhirnya, kesimpulan dibuat berdasarkan hasil perhitungan prediksi[13]. Secara matematis, jika C adalah kategori dan W adalah himpunan kata dalam dokumen, probabilitas posterior dapat dihitung dengan persamaan (1) berikut:

$$P(C|W) = \frac{P(C) \cdot P(W|C)}{P(W)} \quad (1)$$

G. Support Vector Machine

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang digunakan untuk tugas klasifikasi dan regresi. SVM berfokus pada pencarian hyperplane terbaik yang dapat memisahkan dua kelas dalam ruang fitur dengan margin maksimum. Margin ini diukur sebagai jarak antara hyperplane dan titik terdekat dari setiap kelas. SVM juga dapat menangani pemisahan non-linear dengan menggunakan fungsi kernel, yang memetakan data ke dimensi yang lebih tinggi. Keunggulan SVM melibatkan kemampuannya menangani data yang kompleks dan memiliki performa yang baik, terutama ketika dimensi fitur tinggi atau ketika terdapat pemisahan yang jelas antara kelas-kelas dalam data. Data set yang akan diteliti dipersiapkan terlebih dahulu dalam cara kerja algoritma Support Vector Machine. Atribut yang berbeda dan telah diketahui dipastikan dalam proses ini[14]. Selanjutnya, garis khayal dibuat untuk memisahkan dua kelompok data atau atribut, dan kesimpulan dihasilkan dari data set yang diolah. Untuk kernel linear menggunakan persamaan (2) berikut:

$$K(x_1, x_2) = x_1^T \cdot x_2 \quad (2)$$

H. Random Forest

Random Forest adalah algoritma pembelajaran mesin yang beroperasi berdasarkan prinsip *ensemble learning*. Cara kerja algoritma ini ialah membangun sejumlah besar pohon keputusan secara acak dan menggabungkan hasil prediksi mereka untuk membuat keputusan akhir. Setiap pohon dibangun dengan menggunakan sub set acak dari data dan fitur,

mengurangi risiko overfitting dan meningkatkan keakuratan prediksi. Keputusan akhir diambil berdasarkan mayoritas suara (klasifikasi) atau rata-rata (regresi) dari hasil pohon-pohon tersebut. Keunggulan utama dari *Random Forest* melibatkan kemampuannya menangani data set yang kompleks, mengatasi *overfitting*, dan memberikan estimasi pentingnya masing-masing fitur dalam proses pengambilan keputusan[15].

I. Evaluasi

Proses validasi ini bertujuan untuk memperoleh hasil prediksi dari model yang telah dikembangkan, yang kemudian dibandingkan dengan model lainnya. Langkah validasi ini memegang peranan kunci dalam rangkaian pembangunan model, sebab memungkinkan penilaian objektif terhadap performa model yang telah dibuat dengan membandingkannya dengan model alternatif[16]. Evaluasi kinerja algoritma klasifikasi menghasilkan nilai akurasi, presisi, *recall*, dan *f1-score*. Akurasi (*Accuracy*) mengukur seberapa baik model dapat mengklasifikasikan seluruh kelas dengan benar. Presisi (*Precision*) memberikan informasi tentang seberapa baik model dalam mengidentifikasi positif dari semua yang diprediksi sebagai positif. Recall (*Sensitivity*) memberikan informasi tentang seberapa baik model dapat menemukan semua *instance* yang positif. F1-Score memberikan keseimbangan antara presisi dan *recall*, dan sering digunakan jika kelas target tidak seimbang (*imbalance*). Rumus perhitungan untuk evaluasi algoritma klasifikasi dapat dilihat pada persamaan (3), (4), (5), dan (6).

$$Accuracy = \frac{TP + TN}{(TP + FP + FN + TN)} \times 100\% \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (4)$$

$$Recall = \frac{TP}{TP + FP} \times 100\% \quad (5)$$

$$Precision = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

Dalam mengukur kinerja model klasifikasi, terdapat beberapa metrik evaluasi yang digunakan. *True Positive* (TP) mengindikasikan hasil klasifikasi yang tepat, sementara *True Negative* (TN) menunjukkan hasil klasifikasi yang tidak tepat. Di sisi lain, *False Positive* (FP) merujuk pada hasil klasifikasi yang sebenarnya tidak tepat, meskipun model mengklasifikasikannya dengan benar, dan *False Negative* (FN) menggambarkan hasil klasifikasi yang seharusnya benar tetapi dinyatakan salah oleh model. Metrik-metrik ini memainkan peran penting dalam mengevaluasi sejauh mana model dapat mengenali dan membedakan antara kelas-kelas yang berbeda, memberikan pemahaman yang mendalam terkait ketepatan dan ketelitian model klasifikasi yang telah dikembangkan[17].

III. HASIL DAN PEMBAHASAN

Hasil analisis sentimen terhadap kasus kebocoran data digital Bank BSI dibahas secara menyeluruh dalam bab pembahasan penelitian ini. Dalam bab ini, berbagai aspek yang berkaitan dengan bagaimana masyarakat dan pemangku kepentingan menanggapi kebocoran data tersebut.

A. Data set

Dari 9 Mei 2023 hingga 17 Mei 2023, data dari media sosial Twitter yang berhasil dikumpulkan sebanyak 24.401 *tweet*. Memiliki 15 atribut antara lain URL, Tanggal, *Tweet*, ID, *Username, Likes, Quotes* dan sebagainya yang disimpan dalam format .csv. Data selanjutnya diolah ke tahapan *preprocessing* untuk meningkatkan struktur dari data. Tabel 1 menunjukkan hasil *crawling tweet*.

Tabel 1. Sampel *tweet* hasil *Crawling*

No	Tweet
1	@bankbsi_id Layanan BAYAR ke @TakafulKeluarga sudah bisa blm ya ,kok jadi repot begini ðŸ«
2	@NurmansyahAff @berlianidris @Paltiwest @bankbsi_id Apa mungkin dah ganti vendor?
	...
24401	@adym311286 @bankbsi_id @FikiHari lewat atm mandiri bisa kak

B. Pre-processing

Tahapan *preprocessing* data dilakukan untuk meningkatkan performa masing-masing algoritma klasifikasi dalam melakukan prediksi sehingga didapatkan data yang lebih presisi. Tahapan ini meliputi *cleaning* data, Setelah dilakukan proses *preprocessing* data jumlah data yang dihasilkan sejumlah 24.400 *tweet*. Hasil dari proses *preprocessing* dapat dilihat pada Tabel 2.

Tabel 2. Proses Preprocessing Data

Proses	Tweet
Cleaning data	@bankbsi_id Layanan BAYAR ke @TakafulKeluarga sudah bisa blm ya ,kok jadi repot begini ðŸ™
Tokenization	['layanan', 'bayar', 'ke', 'sudah', 'bisa', 'blm', 'ya', 'kok', 'jadi', 'repot', 'begini']
Stopword Removal	['layanan', 'bayar', 'ke', 'sudah', 'bisa', 'blm', 'ya', 'kok', 'jadi', 'repot', 'begini']
Stemming	layanan bayar ke sudah bisa blm ya kok jadi repot begini

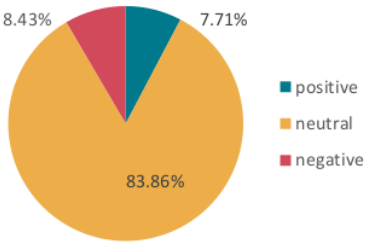
C. Pelabelan Data

Tahap pelabelan data dilakukan secara otomatis menggunakan VADER Sentimen Analisis dengan memberikan skor polaritas pada masing-masing data *tweet*. Jika skor > 0.05 maka menghasilkan sentimen positif, skor 0.05 hingga -0.05 menghasilkan sentimen netral dan skor < -0.05 menghasilkan sentimen negatif.

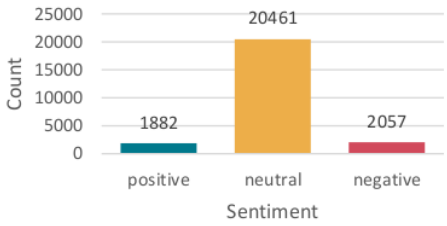
Tabel 3. Sampel Hasil Pelabelan Data

Tweet	sentiment	overall_sentiment
layanan bayar ke sudah bisa blm ya kok jadi repot begini	0	*neutral*
setuju mas jarang banget ciso yang mikirin strategi back up dan recovery end point protection diremehin padahal attack surfacenya jauuuhhh lebih luas dibanding data center	-0.4767	*negative*
dengan kasusnya bank bsi ini next bootcam security ada dimana mana dan jadi incaran banyak orang gimana klo datamatrix buka kelas cyber security	0.5859	*positive*

Dari proses pelabelan data menggunakan VADER sentimen analisis didapatkan hasil dari 24.400 data sejumlah 20.461 data memiliki sentimen netral dan 2.057 data memiliki sentimen negatif, sedangkan sentimen positif hanya sejumlah 1.882 data. Setelah didapatkan hasil dari pelabelan data selanjutnya akan divisualisasikan menggunakan diagram batang yang dapat dilihat pada Gambar 2 dan diagram lingkaran pada Gambar 3.



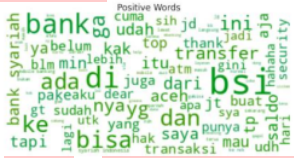
Gambar 2. Diagram Lingkaran Sentimen



Gambar 3. Diagram Batang Sentimen

D. Visualisasi Data

Visualisasi data digunakan untuk mendapatkan frekuensi kata yang paling banyak muncul dalam data set yang telah didapatkan. Visualisasi data disajikan dalam bentuk *word cloud*. Visualisasi data ini dilakukan pada keseluruhan data positif, negatif maupun netral secara terpisah. Dapat dilihat pada Gambar 4 menunjukkan *word cloud* label sentimen positif. Gambar 5 menunjukkan word cloud label sentimen negatif. Gambar 6 menunjukkan *word cloud* label sentimen netral.



Gambar 4. *Word Cloud* Sentimen Positif

Gambar 5. *Word Cloud* Sentimen Negatif

Gambar 6. *Word Cloud* Sentimen Netral

Visualisasi *word cloud* tersebut menunjukkan secara keseluruhan kata yang paling banyak pada data set yang menggandung tentang kasus kebocoran data digital bank BSI. Hasil *word cloud* negatif terhadap persepsi masyarakat

tentang kesulitan melakukan transaksi dan kekhawatiran nasabah tentang diretasnya data pribadi mereka. *Word cloud* netral terhadap opini masyarakat menunjukkan bahwa orang lebih banyak bingung dan tidak tahu apa yang sebenarnya terjadi, mengingat juga tingkat literasi digital di Indonesia khususnya dalam keamanan digital masih mendapat skor rendah yaitu 3.12 (skala 1-5) pada tahun 2022[18]. Hasil *word cloud* positif menunjukkan bahwa respons masyarakat saat ini menyadari pentingnya pemahaman tentang keamanan digital sederhana seperti data diri sampai ke hal teknis seperti mengadakan seminar *cyber security*.

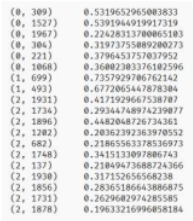
E. Ekstraksi Fitur

Dokumen teks yang telah dilakukan proses *preprocessing* sebelumnya kini melalui tahapan yang lebih lanjut, yaitu pembobotan kata menggunakan algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF). Proses ini memiliki tujuan untuk memberikan penekanan yang lebih tinggi pada kata-kata yang dianggap lebih signifikan dalam konteks tertentu.

Algoritma TF-IDF secara khusus menghitung seberapa sering sebuah kata muncul dalam sebuah dokumen (Frekuensi Kata) dan seberapa unik kata tersebut dibandingkan dengan seluruh koleksi dokumen (Frekuensi Dokumen Kebalikan). Jika nilai dari tahap pembobotan kata ini lebih besar, kata tersebut memiliki bobot yang lebih besar. Ini menunjukkan bahwa kata tersebut memiliki signifikansi yang lebih besar dan dapat dianggap sebagai kata kunci yang lebih penting untuk dokumen atau kalimat.

Penting untuk diingat bahwa nilai TF-IDF yang tinggi menunjukkan bahwa kata itu penting dalam konteks tertentu. Oleh karena itu, proses ini membantu menentukan makna dan relevansi kata-kata yang ada dalam dokumen, dan memudahkan menemukan kata-kata kunci yang dapat memberikan pemahaman yang lebih baik tentang apa yang ada di dalam dokumen[19].

Setelah dilakukan proses TF-IDF maka akan didapat hasil pembobotan kata seperti pada Gambar 7. Pada sampel data pertama, 0 merepresentasikan indeks dari sampel dalam hal ini adalah indeks dari data['Tweet']. 309 merepresentasikan indeks dari fitur pada *stopwords* dan 0.5319652965003833 merepresentasikan nilai atau bobot dari kata tersebut.

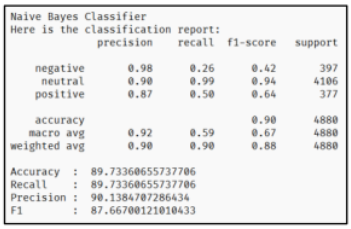


```
(0, 309)     0.5319652965003833
(0, 1527)    0.5391944919917319
(0, 1967)    0.22428313700065103
(0, 304)     0.31973755089200273
(0, 221)     0.3796453757037952
(0, 1068)    0.3600230337610259 6
(1, 699)     0.7357929706762142
(1, 493)     0.6772065447878304
(2, 1931)    0.4171929667538707
(2, 1734)    0.2934474897423907 7
(2, 1896)    0.4482048726734361
(2, 1202)    0.20362392363970552
(2, 682)     0.21865563378536973
(2, 1748)    0.34151330978067 43
(2, 137)     0.21049473688724366
(2, 1930)    0.3171526565682 38
(2, 1856)    0.28365186643886875
(2, 1731)    0.26296029742855 85
(2, 1878)    0.19633216996058184
```

Gambar 7. Sampel Hasil algoritma TF-IDF

F. Klasifikasi *Naive Bayes*

Setelah melalui tahapan proses TF-IDF, langkah berikutnya adalah memasukkan data ke dalam fase pemodelan klasifikasi. Sebelumnya, data akan dibagi menjadi dua bagian, yakni data latih dan data uji, dengan rasio 80:20, di mana 80% digunakan untuk data latih dan 20% untuk data uji dalam eksperimen ini. Data uji berperan sebagai alat evaluasi model, sementara data latih berfungsi untuk pembangunan model dan identifikasi pola. Setelah pembagian data diselesaikan, langkah selanjutnya melibatkan proses klasifikasi model. Algoritma pertama yang akan dilakukan percobaan adalah *Naive Bayes*. Dapat dilihat pada Gambar 8 bahwa pada algoritma *Naive Bayes* nilai akurasi sebesar 90%, dimana nilai presisi negatif 98%, nilai presisi netral 90%, nilai presisi positif 87%, nilai *recall* negatif 26%, nilai *recall* netral 99%, nilai *recall* positif 50%, dan nilai *f1-score* negatif 42%, nilai *f1-score* netral 94%, nilai *f1-score* positif 64%.

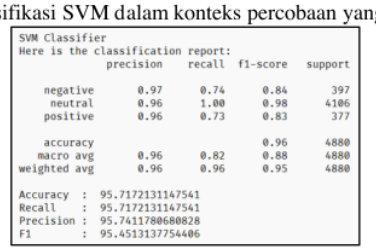


```
Naive Bayes Classifier
Here is the classification report:
              precision    recall  f1-score   support

    negative       0.98      0.26      0.42       397
     neutral       0.90      0.99      0.94      4106
    positive       0.87      0.50      0.64       377

    accuracy                           0.90      4880
   macro avg       0.92      0.59      0.67      4880
weighted avg       0.90      0.90      0.88      4880

Accuracy  :  89.73360655737706
Recall    :  89.73360655737706
Precision :  90.13847072864 34
F1        :  87.66700121010433
```

Gambar 8. Hasil Klasifikasi *Naive Bayes*

G. Klasifikasi *Support Vector Machine*

Percobaan berikutnya melibatkan implementasi model klasifikasi *Support Vector Machine* (SVM). Hasil dari klasifikasi SVM dapat di perhatikan pada Gambar 9, yang menunjukkan bahwa nilai akurasi model ini mencapai 96%. Mencakup nilai presisi negatif sebesar 97%, nilai presisi netral 96%, dan nilai presisi positif 96%. Selain itu, nilai *recall* negatif mencapai 74%, nilai *recall* netral 100%, dan nilai *recall* positif 73%. Secara keseluruhan, nilai *f1-score* negatif mencapai 84%, nilai *f1-score* netral 98%, dan nilai *f1-score* positif 83%. Hasil ini memberikan pemahaman mendalam tentang kinerja model klasifikasi SVM dalam konteks percobaan yang dilakukan.

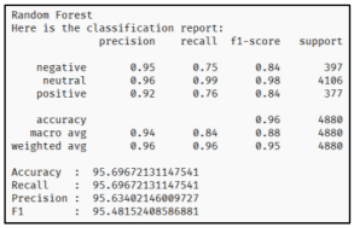


```
SVM Classifier
Here is the classification report:
              precision    recall  f1-score   support

    negative       0.97      0.74      0.84       397
     neutral       0.96      1.00      0.98      4106
    positive       0.96      0.73      0.83       377

    accuracy                           0.96      4880
   macro avg       0.96      0.82      0.88      4880
weighted avg       0.96      0.96      0.95      4880

Accuracy  :  95.7172131147541
Recall    :  95.7172131147541
Precision :  95.7411780680828
F1        :  95.4513137754406
```

Gambar 9. Hasil Klasifikasi SVM

H. Klasifikasi Random Forest

Percobaan terakhir melibatkan penerapan algoritma klasifikasi *Random Forest*. Hasil dari model klasifikasi ini menunjukkan tingkat akurasi sebesar 96%. Detail evaluasi melibatkan nilai presisi negatif 95%, nilai presisi netral 96%, nilai presisi positif 92%, nilai *recall* negatif 75%, nilai *recall* netral 99%, nilai *recall* positif 76%, dan nilai *f1-score* negatif 84%, nilai *f1-score* netral 98%, nilai *f1-score* positif 84%. Hasil klasifikasi *Random Forest* dapat dilihat pada Gambar 10.



```
Random Forest
Here is the classification report:
              precision    recall  f1-score   support

    negative       0.95      0.75      0.84       397
     neutral       0.96      0.99      0.98      4106
    positive       0.92      0.76      0.84       377

    accuracy                           0.96      4880
   macro avg       0.94      0.84      0.88      4880
weighted avg       0.96      0.96      0.95      4880

Accuracy  :  95.69672131147541
Recall    :  95.69672131147541
Precision :  95.63402146009727
F1        :  95.48152408586881
```

Gambar 10. Hasil Klasifikasi *Random Forest*

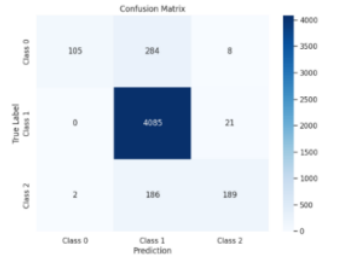
I. Evaluasi

Penilaian performa algoritma klasifikasi *Naive Bayes*, *Support Vector Machine* (SVM), dan *Random Forest* dilakukan melalui penerapan *confusion matrix*. *Confusion matrix*, sebagai alat analisis, digunakan untuk mengevaluasi sejauh mana kualitas klasifikasi yang dihasilkan. Pada Tabel 4 *confusion matrix* masing-masing kolom akan merepresentasikan hasil prediksi model untuk setiap kategori sentimen.

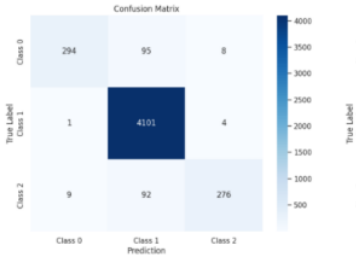
Tabel 4. *Confusion Matrix*

Jumlah data yang benar diprediksi sebagai positif (true positive).	Jumlah data yang sebenarnya positif tetapi salah diprediksi sebagai netral (false negative).	Jumlah data yang sebenarnya positif tetapi salah diprediksi sebagai negatif (false negative).
Jumlah data yang sebenarnya netral tetapi salah diprediksi sebagai positif (false positive).	Jumlah data yang benar diprediksi sebagai netral (true neutral).	Jumlah data yang sebenarnya netral tetapi salah diprediksi sebagai negatif (false negative).
Jumlah data yang sebenarnya negatif tetapi salah diprediksi sebagai positif (false positive).	Jumlah data yang sebenarnya negatif tetapi salah diprediksi sebagai netral (false positive).	Jumlah data yang benar diprediksi sebagai negatif (true negative).

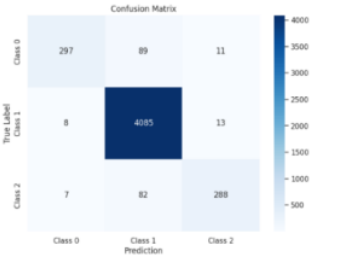
Dengan memperhatikan representasi ini, kita dapat mengukur akurasi, presisi, recall, dan lainnya untuk setiap kategori sentimen dalam evaluasi model analisis sentimen. Hasil *confusion matrix* untuk klasifikasi *Naive Bayes* dapat dilihat pada Gambar 11, sementara *confusion matrix* untuk klasifikasi Support Vector Machine (SVM) dapat ditemukan pada Gambar 12, dan confusion matrix untuk klasifikasi Random Forest terlihat pada Gambar 13.



Gambar 11. *Confusion Matrix* Klasifikasi *Naive Bayes*



Gambar 12. *Confusion Matrix* Klasifikasi *Support Vector Machine*



Gambar 13. *Confusion Matrix* Klasifikasi *Random Forest*

Visualisasi *confusion matriks* menjadi langkah awal untuk menginterpretasikan kinerja model. Gambar *confusion matriks* memberikan pandangan yang lebih intuitif tentang sejauh mana model mampu mengklasifikasikan kelas dengan benar. Setelah melalui proses tersebut, evaluasi lebih lanjut dilakukan dengan merekapitulasi hasil dalam bentuk tabel. Tabel tersebut, yang disajikan dalam Tabel 5, menampilkan parameter evaluasi seperti akurasi, presisi, *recall*, dan nilai *F1-score* untuk memberikan pemahaman yang lebih holistik terhadap performa model klasifikasi yang telah diimplementasikan.

Tabel 5. Matrix Evaluasi Model Algoritma

Model Algoritma	Naive Bayes	Support Vector Machine	Random Forest
Accuracy	89.73%	95.72%	95.7%
Recall	89.73%	95.72%	95.7%
Precision	90.14%	95.74%	95.63%
F1-Score	87.67%	95.45%	95.48%

Dari hasil evaluasi kinerja tiga model algoritma, yakni Naive Bayes, Support Vector Machine (SVM), dan Random Forest, dapat ditarik beberapa kesimpulan yang relevan. Pertama, model SVM dan Random Forest menunjukkan kinerja yang lebih baik daripada Naive Bayes, dengan akurasi masing-masing mencapai 95.72% dan 95.7% dibandingkan dengan 89.73% dari Naive Bayes. Kedua, konsistensi antara recall dan precision pada model SVM dan Random Forest mengindikasikan bahwa keduanya mampu secara efektif mengidentifikasi dan mengklasifikasikan sentimen secara tepat, sedangkan pada Naive Bayes terdapat perbedaan yang cukup signifikan antara recall dan precision. Ketiga, F1-Score, yang mengukur keseimbangan antara recall dan precision, menegaskan bahwa SVM dan Random Forest memiliki performa yang sebanding baik dengan nilai sekitar 95.45% dan 95.48% masing-masing, sementara Naive Bayes menunjukkan kinerja yang sedikit lebih rendah dengan nilai 87.67%. Dalam konteks penelitian ini, kinerja yang lebih tinggi dari SVM dan Random Forest menunjukkan bahwa kedua model tersebut mungkin lebih cocok untuk digunakan dalam analisis sentimen terkait keamanan data perbankan, khususnya dalam mengidentifikasi dan menanggapi kasus kebocoran data. Meskipun demikian, perlu diingat bahwa pemilihan model tergantung pada karakteristik dan tujuan spesifik dari aplikasi analisis sentimen yang diinginkan.

IV. SIMPULAN

Dalam penelitian ini, terkumpul 24.401 data *tweet* dari media sosial Twitter. Melalui proses *preprocessing*, data diolah menjadi 24.400 *tweet*. Sentimen dari *tweet-tweet* tersebut dilabeli secara otomatis menggunakan *VADER Sentiment Analysis*, menghasilkan 1.882 sentimen positif, 2.057 sentimen negatif, dan 20.461 sentimen netral. Data set dibagi dengan rasio 80:20, di mana 80% menjadi data latih dan 20% menjadi data uji. Ketiga algoritma, yaitu *Naive Bayes*, *Support Vector Machine*, dan *Random Forest*, digunakan untuk klasifikasi. Hasil evaluasi menunjukkan bahwa algoritma *Support Vector Machine* memiliki nilai akurasi tertinggi, yakni 95.72%, nilai *recall* sebesar 95.72%, nilai presisi sebesar 95.74% dan nilai *f1-score* sebesar 95.45%. *Random Forest* mencapai 95.7%, nilai *recall* sebesar 95.7%, nilai presisi sebesar 95.63% dan nilai *f1-score* sebesar 95.48%. Sementara *Naive Bayes* memiliki akurasi sebesar 89.73%, nilai *recall* sebesar 89.73%, nilai presisi sebesar 90.14% dan nilai *f1-score* sebesar 87.67%. Namun, penggunaan data yang lebih baik atau lebih terstruktur dapat diterapkan untuk mencapai tingkat akurasi maksimal. Alternatif lain yang dapat diambil adalah melibatkan penyatuan lebih banyak data atau menerapkan metode lainnya, dengan tujuan untuk memperoleh hasil yang lebih tepat dan akurat.

UCAPAN TERIMA KASIH

Semoga isi artikel ini memberikan manfaat dan pemahaman yang berharga. Terima kasih juga kepada semua pihak yang telah turut serta dalam proses penulisan artikel ini. Dukungan dan kontribusi dari berbagai pihak sangat berarti, dan tanpa mereka, artikel ini tidak akan terwujud. Teruslah mendukung dan membagikan pengetahuan, karena bersama-sama kita dapat menciptakan ruang diskusi yang bermanfaat.

REFERENSI

[1] A. Faulina, E. Chatra, and S. Sarmiati, "Peran buzzer dan konstruksi pesan viral dalam proses pembentukan opini publik di new media," *JRTI (Jurnal Riset Tindakan Indonesia)*, vol. 7, no. 1, p. 1, Jan. 2020, doi: 10.29210/30031390000.

[2] R. Milafebina, I. Putra Lesmana, and M. R. Syailendra, "Perlindungan Data Pribadi terhadap Kebocoran Data Pelanggan E-commerence di Indonesia." [Online]. Available: https://ojs.staialfurqan.ac.id/jtm/

[3] D. Yanti Liliana, R. Arnanda, A. Ilham Adnan, and dan Hilda Yuliastuti, "Policy Brief-Penguatan Implementasi Regulasi Perlindungan Data Pribadi Bagi Pelanggan Lokapasar di Indonesia," 2023.

[4] H. Atsqalani, N. Hayatin, and C. S. K. Aditya, "Sentiment Analysis from Indonesian Twitter Data Using Support Vector Machine And Query Expansion Ranking," *Jurnal Online Informatika*, vol. 7, no. 1, p. 116, Jun. 2022, doi: 10.15575/join.v7i1.669.

[5] A. Rahman Isnain, A. Indra Sakti, D. Alita, and N. Satya Marga, "SENTIMEN ANALISIS PUBLIK TERHADAP KEBIJAKAN LOCKDOWN PEMERINTAH JAKARTA MENGGUNAKAN ALGORITMA SVM," *JDMSI*, vol. 2, no. 1, pp. 31–37, 2021, [Online]. Available: https://t.co/NfhnfMjtXw

[6] D. Sepri, P. Algoritma, N. Bayes, U. Analisis, K. Penggunaan, and A. Bank, "Penerapan Algoritma Naïve Bayes Untuk Analisis Kepuasan Penggunaan Aplikasi Bank," *Journal of Computer System and Informatics (JoSYC*, vol. 2, no. 1, pp. 135–139, 2020.

[7] D. Alita and R. A. Shodiqin, "Sentimen Analisis Vaksin Covid-19 Menggunakan Naive Bayes Dan Support Vector Machine," *Journal of Artificial Intelligence and Technology Information (JAITI)*, vol. 1, no. 1, pp. 1–12, Feb. 2023, doi: 10.58602/jaiti.v1i1.20.

[8] N. Faridhotun, E. Haerani, and R. M. Candra, "Analisis Sentimen Ulasan Aplikasi WeTV Untuk Peningkatan Layanan Menggunakan Metode K-Nearst Neighbor," *Journal of Information System Research (JOSH)*, vol. 4, no. 3, pp. 855–864, Apr. 2023, doi: 10.47065/josh.v4i3.3349.

[9] I. S. Dianita, H. Irawan, and A. Deah, "PERAN BANK SYARIAH INDONESIA DALAM PEMBANGUNAN EKONOMI NASIONAL," vol. 3, no. 2, p. 2021, [Online]. Available: http://journal.iaimsinjai.ac.id/index.php/asy-syarikah

[10] V. Marcelliana *et al*., "PENERAPAN PERLINDUNGAN KONSUMEN TERHADAP NASABAH PT. BANK SYARIAH INDONESIA DALAM KASUS KEBOCORAN DATA NASABAH," vol. 1, no. 2, pp. 180–194, 2023, doi: 10.59581/deposisi.v1i2.562.

[11] A. T. Zy and W. Hadikristanto, "Implementasi Algoritma Metode Naive Bayes dan Support Vector Machine Tentang Pembobolan dan Kebocoran Data di Twitter," *Bulletin of Information Technology (BIT)*, vol. 4, no. 1, pp. 49–56, 2023, doi: 10.47065/bit.v3i1.

[12] I. Najiyah and I. Haryanti, "SENTIMEN ANALISIS COVID-19 DENGAN METODE PROBABILISTICNEURAL NETWORKDAN TF-IDF," *JURNAL RESPONSIF*, vol. 3, no. 1, 2021, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti

[13] M. H. Ferdiansyah, A. Rosid, Y. Findawati, and A. Eviyanti, "Implementation of the Naive Bayes Method for Sentiment Analysis in the 2024 Presidential Election [Implementasi Metode Naïve Bayes untuk Analisis Sentimen pada Pemilihan Presiden 2024]."

[14] H. Tuhuteru, "Analisis Sentimen Masyarakat Terhadap Pembatasan Sosial Berksala Besar Menggunakan Algoritma Support Vector Machine," 2020.

[15] D. Alita and A. Rahman, "Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier," 2020.

[16] A. Rahman Hakim, W. Gata, A. Zevana Putri Widodo, O. Kurniawan, and A. Rama Syarif, "Analisis Perbandingan Algoritma Machine Learning Terhadap Sentimen Analis Pemindahan Ibu Kota Negara," *Jurnal Teknologi Informasi dan Komunikasi)*, vol. 7, no. 2, 2023, doi: 10.35870/jti.

[17] J. W. Iskandar and Y. Nataliani, "Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 6, pp. 1120–1126, Dec. 2021, doi: 10.29207/resti.v5i6.3588.

[18] T. Terttiaavini and T. S. Saputra, "LITERASI DIGITAL UNTUK MENINGKATKAN ETIKA BERDIGITAL BAGI PELAJAR DI KOTA PALEMBANG," *JMM (Jurnal Masyarakat Mandiri)*, vol. 6, no. 3, p. 2155, Jun. 2022, doi: 10.31764/jmm.v6i3.8203.

[19] D. Apriliani, A. Susanto, M. F. Hidayattullah, and G. W. Sasmito, "Sentimen Analisis Pandangan Masyarakat Terhadap Vaksinasi Covid 19 Menggunakan K-Nearest Neighbors," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 8, no. 1, pp. 34–37, Jan. 2023, doi: 10.30591/jpit.v8i1.4759.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Article - M. Diki Armanda.docx

ORIGINALITY REPORT

14%	14%	8%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	jutif.if.unsoed.ac.id Internet Source	5%
2	journal.irpi.or.id Internet Source	2%
3	publishing-widyagama.ac.id Internet Source	1%
4	jurnal.darmajaya.ac.id Internet Source	1%
5	Submitted to Sriwijaya University Student Paper	1%
6	www.kompasiana.com Internet Source	1%
7	jurnal.undhirabali.ac.id Internet Source	1%
8	www.researchgate.net Internet Source	1%
9	rekayasa.nusaputra.ac.id Internet Source	1%

10	ejurnal.seminar-id.com Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		