ANALISIS SENTIMEN TWITTER TERHADAP PERLINDUNGAN DATA PRIBADI DENGAN PENDEKATAN MACHINE LEARNING

TWITTER SENTIMENT ANALYSIS OF PERSONAL DATA PROTECTION WITH MACHINE LEARNING APPROACH

Joko Ade Nursiyono[1], Qorinul Huda[2]

[1]BPS PROVINSI JAWA TIMUR
[2]POLITEKNIK STATISTIKA STIS
(joko.ade@bps.go.id, 211911012@stis.ac.id)

Abstrak – Seiring kemajuan teknologi dan informasi, aspek pertahanan dan kemanan utamanya dalam perlindungan data pribadi menjadi sangat penting. Perlindungan data pribadi merupakan hak asasi manusia yang wajib dilindungi oleh negara. Digitalisasi data merupakan tuntutan dan tantangan dalam kemajuan informasi. Upaya dalam perlindungan data pribadi pada dasarnya dilaksanakan melalui instrumen kepastian hukum berupa regulasi yang mengatur sebuah sistem agar terwujud sistem yang kuat dalam melindungi *cyber crime*. Berbagai regulasi telah ada dalam tatanan hukum di Indonesia. Meskipun demikian, masih terdapat kasus kebocoran data pribadi masyarakat Indonesia. Tujuan penelitian ini adalah menggambarkan kondisi perlindungan data pribadi di Indonesia dan menganalisis kasus kebocoran data yang dideteksi dalam *tweet* Twitter pada periode 1 Juli 2021 hingga 29 September 2022. Penelitian dilakukan dengan teknik *scrapping tweet* Twitter dan dilakukan pengklasifikasian respons *netizen* berdasarkan sentimen positif, negatif, dan netral. Di setiap sentimen dianalisis dengan *word cloud* dengan menemukan topik apa yang sering dibahas *netizen* terhadap perlindungan data pribadi. Selanjutnya evaluasi klasifikasi dilanjutkan dengan melihat akurasi pada algoritma klasifikasi *machine learning* yaitu *naive bayes* dan *random forest*. Hasil penelitian menyebutkan pada periode 1 Juli 2021 hingga 29 September 2022, respons masyarakat terhadap perlindungan data pribadi masih negatif. Yang mengartikan bahwa sistem perlindungan data di Indonesia masih belum efektif dengan terjadinya berbagai kasus kebocoran data. Berdasarkan nilai akurasinya, algoritma *naive bayes* sangat baik dalam mengkasifikasikan *tweet* berdasarkan sentimennya yaitu sebesar 99,84% dibandingkan algoritma *random forest*.

Kata Kunci: *tweet, twitter, cyber crime, naive bayes, random forest, machine learning*

Abstract *– As technology and information advance, the main defense and security aspects in the protection of personal data become very important. The protection of personal data is a human right that must be protected by the state. Data digitization is a demand and challenge in the advancement of information. Efforts in protecting personal data are basically carried out through legal certainty instruments in the form of regulations that regulate a system in order to realize a strong system in protecting cybercrime. Various regulations already exist in the legal system in Indonesia. Nevertheless, there are still cases of personal data leakage among Indonesians. The purpose of this study is to describe the condition of personal data protection in Indonesia and analyze cases of data leaks detected in Twitter tweets in the period from July 1, 2021, to September 29, 2022. The study was conducted by using Twitter tweet-scrapping techniques and classifying netizen responses based on positive, negative, and negative sentiments. neutral. Each sentiment is analyzed with word cloud by finding what topics are often discussed by netizens on the protection of personal data. Furthermore, the classification evaluation is continued by looking at the accuracy of the machine learning classification algorithm, namely naive bayes and random forest. The results of the study stated that in*

the period from July 1, 2021, to September 29, 2022, the public's response to the protection of personal data was still negative. This means that the data protection system in Indonesia is still not effective with the occurrence of various cases of data leakage. Based on the accuracy value, the Naive Bayes algorithm is very good at classifying tweets based on their sentiments, which is 99.84% compared to the random forest algorithm.

Keywords: *tweet, twitter, cyber crime, naive bayes, random forest, machine learning*

Pendahuluan

Seiring kemajuan teknologi dan informasi, lingkup pertahanan dan keamanan suatu negara semakin meluas, salah satunya perlindungan dan keamanan data pribadi. Dalam Undang-Undang Nomor 23 Tahun 2006 Tentang Administrasi Kependudukan pasal 1 ayat 22, data pribadi merupakan data perorangan yang disimpan, dirawat, dan dijaga kebenarannya dan dilindungi oleh negara.

Digitalisasi data untuk menghadapi tantangan kemajuan teknologi dan informasi terus dilakukan. Untuk itu, sejak Juni 2020, pemerintah mulai menerapkan kebijakan digitalisasi data pribadi (Pane, Siregar, and Sirait, 2021), misalnya melalui KTP elektronik hingga sistem PeduliLindungi selama Pandemi COVID-19. Walau demikian, aspek perlindungan dan keamanan data pribadi masih menjadi persoalan.

Selain dikarenakan rendahnya kesadaran masyarakat (Delpiero et al., 2021), keamanan dan perlindungan data pribadi belum ditopang oleh sistem *cyber* yang baik. Menurut Sanjaya et al. (2022), perlindungan dan keamanan siber merupakan upaya pertahanan siber di Indonesia. Laporan *National Cyber Security Index* (NCSI) menyebutkan skor indeks keamanan siber Indonesia menduduki peringkat 3 terendah di antara negara-negara G20 (Annur ,2022).

Dampak kasus lemahnya kesadaran masyarakat terhadap data pribadi serta lemahnya keamanan *cyber* di Indonesia terlihat pada beberapa kasus kebocoran data pribadi sejak 2020. Data riset dari CNNIndonesia (2021) menyebutkan sebanyak 91 juta data pengguna dan 7 juta mercant Tokopedia bocor dan dijual di Empire Market; sebanyak 1,2 juta data pengguna Bhineka.com bocor, dan dijual di darkweb; data 2,3 juta pemilih 2014 Komisi Pemilihan Umum (KPU) juga bocor; terakhir data peserta BPJS Kesehatan sebanyak 100.002 juga bocor.

Belum ditambah dengan bocornya jutaan data E-KTP dan aplikasi PeduliLindungi. Bahkan, pihak yang

menjadi domain pertahanan *cyber*: Badan Siber dan Sandi Negara (BSSN) (Perpres, 2021) tidak luput dari upaya pembocoran data pribadi yang dilakukan oleh akun twitter bernama Bjorka.

Kasus kebocoran data pribadi ini menuai respon beragam di masyarakat, terutama masyarakat di dunia maya (*netizen*). Salah satu media yang merekam respon tersebut adalah Twitter. Dari twitter inilah, respons masyarakat terhadap sebuah topik menimbulkan sentimen. Sentimen inilah yang menarik untuk selanjutnya dianalisis. Penelitian menggunakan analisis sentimen telah banyak dilakukan.

Sebagaimana penelitian yang telah dilakukan oleh Sholehurrohman and Ilman (2022) mengenai kasus kebocoran data pengguna Facebook. Melalui pendekatan *Naive Bayes Classifier* diperoleh akurasi analisis sentimen sebesar 83,06 persen. Riset sejenis juga dilakukan oleh Aldean, Paradise, and Nugraha (2022) mengenai sentimen vaksinasi COVID-19 di Twitter menggunakan metode Random Forest dengan akurasi 79,00 persen.

Namun, penelitian tersebut hanya menggunakan satu algoritma saja dan belum diterapkan pada kasus sentimen keamanan data pribadi. Untuk itu, penelitian ini dilakukan dengan tujuan menganalisis sentiment *netizen* terhadap data pribadi di Twitter menggunakan *Nave Bayes Classifier* dan *Random Forest*.

Metode Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan algoritma *Machine Learning* (ML). Data yang digunakan dalam penelitian ini berupa cuitan (*tweet*) hasil scraping pengguna twitter aktif sebanyak 85.683 *tweet* selama periode 01 Juli 2021 hingga 29 September 2022. Teknik scraping banyak dilakukan dengan cara mengambil sebagian obyek twitter (Nursiyono, 2021). Untuk mendapatkan *tweet* twitter dilakukan dalam beberapa langkah berikut:

1. Mengaktifkan *library rtweet* dan dplyr;
2. Membuat kresidensial orisinalitas developer twitter untuk memperoleh API *developer* twitter;
3. Melakukan scraping twitter menggunakan API dengan Bahasa *tweet* yang diperlukan adalah Bahasa Indonesia, lang = "id".

Dari sejumlah data *tweet* tersebut, kemudian dibersihkan dan dimodelkan menggunakan perangkat R Studio versi 4.0.2. Proses pembersihan (*pre-*

processing) dilakukan dengan tahapan berikut:

1. Mengaktifkan *library textclean, katadasaR, tokenizers, tm, wordcloud2, ggplot2, plyr, dan stringr*;
2. Melakukan pembersihan teks, membuang karakter yang tidak diperlukan, seperti "#", ":", "/", "\n\r\t", "?","";
3. Menghapus *mention* dan hastags (#);
4. Menghapus kata hubung, misalnya di, ke, dari, oleh, pada, tentang, bahwa, dan lainnya;
5. Melakukan *stemming* untuk mengembalikan kata-kata berimbuhan ke kata dasarnya;
6. Vektorisasi *tweet* dan menghapus kata yang tidak bermakna, termasuk kata kunci "data pribadi";
7. Menghapus duplikasi *tweet*. Pada tahapan ini, dari 85.683 *tweets* diperoleh *tweet* unik (terbebas dari duplikasi) sebanyak 65.241;
8. Vektorisasi *tweet* menjadi untuk melakukan labelisasi setiap *tweet*.

Setelah *labelling* selesai, selanjutnya dilakukan analisis ketepatan sentimen dengan pendekatan dua algoritma ML, yaitu *Naïve Bayes Classifier* dan *Random Forest*.

***Naive Bayes Classifier* (NBC)**

Menurut Rosalina, Purbasari, and Mandyarta (2022), *Naive Bayes Classifier* merupakan algoritma pengklasifikasian berbasis probabilistik untuk melakukan prediksi peluang kejadian masa depan berdasarkan informasi peluang di masa lampau.

Algoritma ini bekerja dengan menentukan peluang sebuah karakter yang masuk kelas tertentu (sentimen) atau *posterior* kemudian dikalikan dengan kemunculan teks dalam setiap *tweet* pada kelas tertentu (sentimen) atau *likelihood*. Bentuk matematis algoritma ini dituliskan berikut:

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)}$$

Keterangan:

X : data yang belum terklasifikasi sentimennya

H : hipotesis bahwa X masuk dalam klasifikasi sentimen tertentu (positif, netral, negatif)

$P(H|X)$: peluang hipotesis H berdasarkan kondisi X

$P(X|H)$: peluang X berdasarkan hipotesis H

$P(H)$: peluang hipotesis H

Random Forest

Random Forest merupakan salah satu algoritma *Machine Learning* (ML) baik untuk klasifisikasi maupun regresi dengan piranti agregasi pohon keputusan yang terdiri atas *internal node, root node,* dan *leaf node* (Siburian dan Mulyana, 2018) yang dikenalkan oleh Leo Breiman dan Adele Cutler. *Random forest* merupakan algoritma yang menggunakan pemisahan biner rekursif untuk mecapai *node* akhir dalam struktur pohon berdasarkan pohon klasifikasi dan regresi (Religia, Nugroho, and Hadikristanto ,2021). *Random forest* merupakan algoritma gabungan *decision tree* dengan cara pengelompokkan "*weak learners*" bersama-sama untuk membentuk "*strong learners*" melalui teknik gabungan (*ensemble techniques*).

Penggunaan algoritma ini selain memberikan akurasi yang bagus dan dimanfaatkan jika data pelatihan berdimensi besar (Primajaya dan Sari 2018). Breiman memperkenalkan algoritma *random forest* dengan kelebihanya yang mampu menghasilkan *error* relatif rendah, performa baik dalam klasifikasi, mampu mengatasi data *training* berdimensi besar secara efisien, serta metode yang efektif untuk mengestimasi *dataset* yang mengandung *missing* data. Algoritma *Random Forest* dalam penelitian ini diterapkan dengan langkah berikut:

1. Menentukan jumlah *tree* sebanyak *k* dari sebanyak fitur *m*, dengan $k < m$;
2. Menentukan jumlah *k-fold* sebanyak 10 untuk pengacakan dan randomisasi;
3. Pada setiap *tree* yang terbentuk, dilakukan pengambilan subset prediktor (p) dengan $m < p$;
4. Langkah selanjutnya dilakukan perulangan kembali sebanyak *k tree*;
5. Prediksi klasifisikasi *fit* diperoleh dari agregasi *vote* terbanyak dari seluruh *tree* yang terbentuk.

Akurasi Model

Akurasi model merupakan ukuran performa algoritma yang digunakan dalam penelitian. Akurasi model ini merupakan dasar pengambilan keputusan algoritma terbaik dalam analisis. Dalam algoritma klasifikasi *machine learning*, digunakan *confusion matrix* sebagai salah satu metode untuk mengukur kinerja (*performance of a classification model or classifier*). Pada intinya, *confusion matrix* mengandung informasi yang membandingkan hasil

klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya.

Tabel 1. Konfigurasi *Confusion Matrix*

Prediksi	Sebenarnya	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Keterangan:

TP : Positif sebenarnya

TN : Negatif sebenarnya

FP : Positif tidak sebenarnya

FN : Negatif tidak sebenarnya

Secara matematis, ukuran akurasi model dituliskan berikut:

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN}$$

Model klasifikasi mengharapkan nilai akurasi yang tinggi. Kategori akurasi model dapat diklasifikasikan sebagai berikut :

Tabel 2. Kategori keakuratan model

Nilai Akurasi	Kategori
90% - 100%	Sangat baik
80% - 90%	Baik
70% - 80%	Cukup
60% - 70%	Kurang
50% - 60%	Gagal

Sumber: Hasil olah peneliti,2022

Selain ukuran akurasi, terdapat pula ukuran presisi, *recall,* dan *specivity*. Namun, akurasi merupakan ukuran yang paling banyak digunakan untuk pengujian model pada berbagai algorima klasifikasi pada *machine learning* (Purnama , 2019).

Metode Latent Dirichlect Allocation (LDA)

Latent Dirichlet Allocation atau LDA adalah algoritma untuk mendeteksi topik melalui pemodelan probabilistik dalam sekumpulan data. Metode ini digunakan dalam menemukan topik yang terkandung pada *tweet* twest twitter.

Tujuan dari penerapan metode LDA untuk mengetahui topik apa yang ramai diperbincangkan *internet citizen* (*netizen)* terkait keamanan data pribadi. Hal ini dilatar belakangi oleh seringnya masyarakat mengekspresikan segala kondisi pada media sosial.

Bentuk analisis melalui *word cloud* yang merupakan sekumpulan kata-kata dalam sebuah teks yang menggambarkan frekuensi penggunaanya.

Apabila frekuensi penggunaan kata tertentu besar atau dominan, maka ukuran kata tersebut dalam *word cloud* juga besar. Semakin besar ukuran sebuah kata artinya semakin sering digunakan dalam ekspresi di media sosial.

Dalam penelitian ini, metode *Dirichlet Allocation* yang digunakan sebatas dalam menggambarkan kata apa yang sering muncul dalam *tweet* twitter. Frekuensi dari kata yang sering dibahas direpresentasikan dalam *word cloud* berdasarkan sentimen *netizen*. Dari

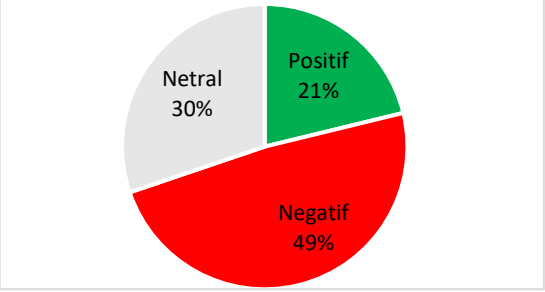
ukuran kata dalam *word cloud* di ambil 20 kata teratas yang sering dipakai *netizen* dalam menanggapi topik terkait perlindungan data pribadi. Kata yang sering muncul tersebut dianalisis untuk menggambarkan kondisi keamanan data pribadi baik dengan respons positif maupun negatif.

Hasil dan Pembahasan

Eksplorasi Data

Indonesia telah memiliki beberapa instrumen dan regulasi dalam upaya perlindungan data pribadi. Meskipun demikian, masih ada beberapa kasus terkait kemanan data yang dapat dinalisis melalui sentimen *tweets* twitter. Analisis ini berbentuk teks yang dapat menggambarkan opini, emosi, dan sikap masyarakat (Medhat, Hassan, and Korashy ,2014). Analisis sentimen dapat menggambarkan permasalahan terkait kemanan data yang sering terjadi di masyarakat.

Setelah dilakukan penambangan (*mining*) data dengan teknik *web scrapping* pada media sosial Twitter, diperoleh *tweets* sebanyak 65241 untuk kata kunci "data pribadi" antara tanggal 1 Juli 2021 sampai 29 September 2022. Tpitter menyediakan fasilitas *Application Programming Interface (API)* sebagai bahan akses data informasi *web* namun dengan sejumlah batasan tertentu. Data mentah hasil *mining tweet* dari Twitter dilakukan penyaringan untuk membatasi hanya opini yang berasal dari akun masyarakat. Setelah itu, dilakukan pelabelan klasifikasi jenis sentimen dari *tweet* yang masuk. Klasifikasi terdiri dari sentimen positif, sentimen negatif, dan sentimen netral. Dimana hasil klasifikasi ini akan diuji dengan metode klasifikasi *Naive Bayes* dan *Random Forest* untuk mengetahui presentasi kebenaran dan evaluasi performanya. Hasil klasifikasi dari sistem sentimen merupakan representasi dari kecenderungan emosi responden di media sosial Twitter.



Gambar 1. Persentase sentimen twitter terhadap keamanan data
Sumber: peneliti, 2022

Gambar 1 menunjukkan respons sentimen masyarakat di Twitter terhadap masalah keamanan data pribadi yang beragam. Pada gambar tersebut terlihat bahwa masyarakat lebih banyak memberikan respons bersentimen negatif yaitu sebesar 49,0% (20.880

tweet), sementara itu masyarakat yang merespon dengan sentimen positif sebesar 21,2% (13.813 *tweet*). Hal tersebut menandakan bahwa masyarakat lebih banyak memberikan respon negatif terkait keamanan data pribadi. Sentimen negatif ditenggarai oleh maraknya kasus kebocoran data yang pada dasarnya karena lemahnya kesadaran masyarakat dan keamanan *cyber* di Indonesia.

Jika mengacu pada riset *Data Breach Investigations Report* (DBIR) dalam Annur (2022) menyebutkan bahwa sepanjang tahun 2021 terdapat 5.212 kasus kebocoran data di seluruh dunia. Kebocoran data tersebut mengacu pada peristiwa gangguan keamanan yang menyebabkan terungkapnya data-data tertentu dan bersifat rahasia.

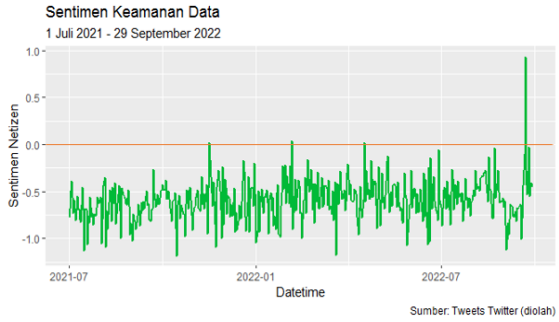
Sektor industri yang mengalami insiden kebocoran tertinggi sepanjang tahun 2021 yaitu sektor keuangan (690 kasus), sektor professional (681 kasus), sektor kesehatan (571 kasus), sektor administrasi publik (537 kasus), dan sektor informasi (378 kasus).

Sistem keamanan data pribadi dalam pengaksesan dunia maya sering menjadi isu dalam peretasan. Peretasan akun di media sosial baik individu maupun sektor publik diakibatkan oleh serangan luar maupun dalam dan lemahnya sistem (*accidental loss*). Kejahatan yang timbul dapat menggunakannya sebagai jual beli data informasi pribadi. Data informasi pribadi dapat digunakan oleh pihak yang tidak bertanggungjawab untuk registrasi ke database lain, misalnya pinjaman online. Oleh karenanya regulasi dan sistem yang terintegrasi perlu dibangun untuk mengatasi ketidaksesuaian data dalam rangka mendeteksi penyalahgunaan data pribadi.

Setiap waktu, masyarakat melakukan transaksi dan registrasi di berbagai platform digital yang terus memperbesar *database*. Kondisi demikian yang berpotensi dalam masalah keamanan data. Atribut pribadi seperti nama, tanggal lahir, nomor telepon, alamat email dan lainnya menjadi atribut yang sering menjadi sasaran dalam serangan *cyber*. Serangan *cyber* tidak hanya menyerang dalam lingkup mikro, namun multiregion bahkan multi instansi. Tentu hal ini memunculkan permasalahan dalam bidang hukum.

Jika mengacu pada Pidato Kenegaraan Presiden tahun 2019 terkait bidang pertahanan dan kemanan. Presiden menginstruksikan tanggap dan siap dalam menghadapi perang siber yang mana didalamnya termasuk penyalahgunaan data. Karena data

adalah jenis kekayaan baru bangsa, oleh karenanya kedaulatan data harus diwujudkan atas data pribadi yang harus dilindungi dan regulasi harus dipersiapkan. Meskipun demikian, masih terdapat kasus kebocoran data pribadi di periode tersebut.



Gambar 2. Sentimen keamanan data
Sumber: peneliti, 2022

Gambar 2 menunjukkan rata-rata sentimen masyarakat terhadap masalah keamanan data pada periode 1 Juli 2021 hingga 29 September 2019 dan didominasi oleh respon sentimen negatif. Sentimen yang bertahan di angka negatif ini dikarenakan kasus kebocoran data pada periode tersebut. Dalam periode tersebut, rata-rata sentimen positif hanya terjadi 4 kali, yaitu tanggal 16 Nopember 2021, 5 Februari 2022, 17 April 2022, dan 22 September 2022. Sentimen positif mengacu pada respons positif terkait masalah keamanan data. Sentimen positif terjadi karena ada upaya pemerintah untuk melindungi dan mengupayakan keamanan siber. Misalnya pada 16 November 2021, Kepala Bidang Pengaduan Advokasi dan Informasi Perizinan menghadiri peluncuran *(launching)* Kaltim-CSIRT oleh Badan Siber dan Sandi Negara (BSSN) untuk menghadapi insiden keamanan ciber (Admin DPMPTSP Kaltim, 2021).

Sentimen positif juga diduga oleh seringnya klarifikasi isu hoaks oleh Kementerian Komunikasi dan Informatika (KOMINFO). Klarifikasi ini terhadap isu hoaks pesan berantai di media sosial yang dengan mudahnya tersebar di masyarakat. Isu hoaks berantai intinya memuat dua hal, yaitu politik dan keuntungan pribadi (penipuan). Hoaks politik ditujukan untuk menyerang instansi tertentu, misalnya pada tanggal 16 November 2021 terdapat berita hoaks "Istri CEO Pfizer meninggal dunia akibat komplikasi vaksin COVID-19" (Daffiandra, 2021). Sedangkan hoaks keuntungan pribadi ditujukan untuk melakukan penipuan terhadap masyarakat untuk memperkaya diri sendiri. Misalnya pada tanggal 16 November 2021 terdapat berita hoaks "HUT ke -64 BCA dan pembagian hadiah Rp5 Juta" (Daffiandra ,2021) .

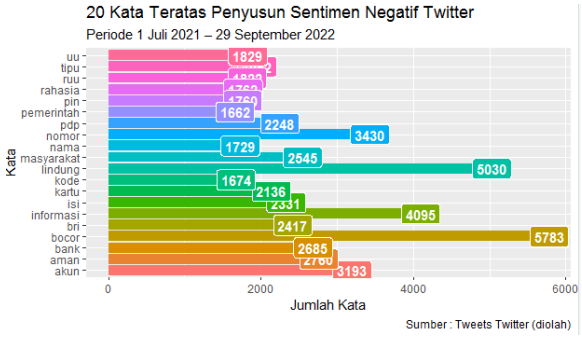
Sentimen netral memiliki porsi 32% (30548 tweets) yang memiliki arti bahwa *tweets* masyarakat tidak hanya mengandung respons positif dan negatif,

namun respons tersebut lebih ke harapan, pengetahuan, definisi, atau pendapat secara umum. Masyarakat dengan respons netral memandang masalah keamanan data sebagai sesuatu yang wajar karena dinamisnya perkembangan teknologi.

Latent Dirichlet Allocation

Dari sebanyak 65.241 *tweets* dikelompokkan menjadi 3 jenis sentimen sebagaimana di gambar 1 dan dianalisis topik atau kata apa yang sering muncul. Sentimen negatif dapat menggambarkan respons masyarakat terkait kemanan data. Jika mengacu pada kata atau topik pada *tweets* dengan kata kunci data pribadi sebagaimana pada gambar 3 pada periode 1 Juli 2021 sampai 29 September 2022 yang menunjukkan respons negatif, masyarakat dominan membahas mengenai perlindungan dan kebocoran data. Kebocoran dan perlindungan data menjadi kata atau topik dominan yang dibahas. Mengingat dalam pasal 28G UUD Negara RI tahun 1945 yang menjelaskan setiap orang berhak atas perlindungan diri pribadi, keluarga, kehormatan, martabat dan harta benda. Pemerintah wajib menjaga kebenaran dan melindungi kerahasiaan terhadap data pribadi.

Gambar 3. *Word cloud* sentimen negatif
Sumber: peneliti, 2022



Gambar 4. 20 kata teratas penyusun sentimen negatif Twitter
Sumber: peneliti, 2022

Berdasarkan gambar 3 dan gambar 4, topik yang sering dibahas masyarakat di media sosial twitter adalah kasus bocornya data. Topik tersebut digambarkan dengan banyaknya kata "bocor","lindung","informasi"," nomor" , "aman", dan "masyarakat". Masyarakat merespons negatif kasus bocornya data sebagai bentuk ekspresi kekecewaan terhadap upaya perlindungan data di Indonesia. Selain itu kasus bocornya data di Indonesia dikarenakan oleh lemahnya kesadaran masyarakat dan lemahnya keamanan *cyber* (CNNIndonesia, 2021).

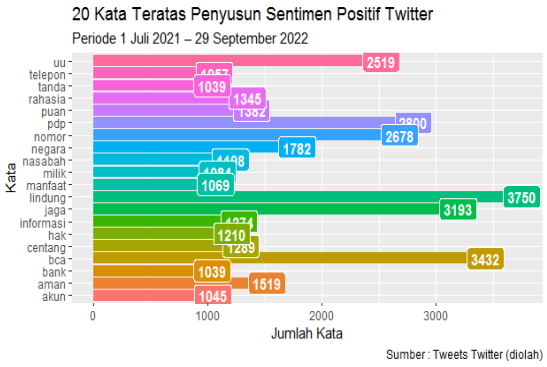
Masyarakat memberikan respons negatif pada kasus penipuan yang mengatasnamakan pihak bank perihal pencurian data nasabah. Masyarakat menkhawatirkan data nasabah yang bocor untuk upaya penggelapan rekening nasabah, jual beli data pribadi, dan penipuan yang menggunakan data pribadi. Tentu kasus kebocoran data pribadi memberikan dampak merugikan yang signifikan bagi masyarakat.

Munculnya ancaman berupa pencurian data mestinya mendorong kesadaran untuk meningkatkan perlindungan data. Bentuk perlindungan tersebut sebagai upaya dalam pertahanan negara. Dalam Undang-Undang Nomor 3 Tahun 2022 tentang Pertahanan Negara semesta yang bersifat tidak hanya fisik semata, namun juga menjacup kategori non fisik utamanya dalam internet, media sosial, *digital*, dan dunia maya. Terkait perlindungan data pribadi menjadi strategi nasional untuk upaya perlindungan.

Menurut Ardiyanti (2014), terdapat empat pondasi untuk membangun pertahanan *cyber* di Indonesia, meliputi pengembangan *software* (sistem dan aplikasi) dan *hardware*, pengembangan saranan dan prasaranan teknologi informasi, manajemen isi (*content management*), *telecommunication and networking*, dan pengembangan keamanan transaksi online.

Pembicaraan bersentimen negatif masyarakat di media sosial twitter perihal keamanan data pribadi mengaitkan dengan isu regulasi yaitu Rancangan Undang–Undang Perlindungan Data Pribadi (RUU PDP). Pembicaraan ini digambarkan melalui kata "pdp","uu", "pemerintah" dan "ruu". Masyarakat memberikan respons negatif sebagai bentuk lambatnya pengesahan RUU ini. RUU PDP ditujukan untuk menjamin hak warga negara atas perlindungan diri pribadi dan menumbuhkan kesadaran masyarakat serta menjamin pengakuan dan penghormatan atas pentingnya perlindungan data pribadi (MENKOMINFO ,2022). Regulasi RUU PDP ini sangat penting mengingat pengguna internet dan munculnya perusahaan bidang teknologi (*tech startups*) yang pesat di Indonesia. Dengan disahkan RUU ini, masyarakat berharap agar data pribadi nasabah dapat dijaga dan dilindungi dari kebocoran. Hal ini terlhat dari kata "rahasia", "aman", "akun", "bank", "bri".

Gambar 5. *Word cloud* sentimen positif
Sumber: peneliti, 2022



Gambar 6. 20 kata teratas penyususn sentimen positif Twitter
Sumber: peneliti, 2022

Pembicaraan masyarakat di media sosial *Twitter* perihal keamanan data pribadi juga memberikan respon positif. Berdasarkan gambar 5 dan gambar 6 topik yang mendominasi dalam sentimen positif terkait perlindungan data BCA dan regulasi perlindungan data pribadi. Pada periode tersebut, terdapat penipuan berupa pesan berantai yang pada intinya meminta data nasabah. Hal itu digambarkan dalam kata "bca", "nomor", "1500888", dan "08111500998". Penipu menggunakan nomor panggilan fake caller yang menyerupai nomor resmi Halo BCA 1500888 dan *WhatsApp* BCA 08111500998. Namun, melalui kanal resminya BCA dan KOMINFO langsung mengonfirmasi bahwa ciri-ciri nomor penipu menngunakan embel-embel +62, 021, dan lainnya. Konfirmasi juga dilanjutkan agar para nasabah menjaga informasi data pribadi seperti nomor ATM, PIN, OTP, Respon KeyBCA, dan lainnya. Cepatnya konfirmasi di kanal resmi BCA ini membuat respons positif masyarakat di kanal twitter. Masyarakat dapat meminimalisir upaya penipuan pencurian data dengan konfirmasi cepat layanan nasabah dan upaya pemblokiran nomor penipu.

Keyakinan masyarakat dalam upaya perlindungan data di Indonesia juga memberikan respons positif di kanal Twitter. Upaya perlindungan data Indonesia dicerminkan dengan adanya RUU PDP. Hal itu digambarkan melalui kata "pdp", "uu", "informasi", dan "indonesia". Masyarakat memberikan pendapat dan harapan agar RUU ini dapat memberikan landasan hukum bagi Indonesia untuk menjaga kedaulatan, perlindungan, dan kemanan data baik dari segi penggunaan dalam negeri dan serangan luar negeri.

Klasifikasi *Naïve Bayes* dan *Random Forest*

Pengelompokan opini masyarakat berupa *tweet* twitter dalam klasifikasi sentimen positif, sentimen negatif, dan sentimen netral dapat dievaluasi melalui akurasi dalam model machine learning. Algoritma *classification machine learning* yang digunakan adalah *random forest* dan *naive bayes*. Akurasi menggambarkan seberapa tepat dari *tweet* masyarakat di kanal twitter yang diklasifikasikan secara tepat berdasarkan sentimennya.

Dalam penelitian ini dari 65.241 *tweet*, sebanyak 75% digunakan sebagai data *training* dan 25% sebagai data *testing*. Data *training* sebagai dasar pembuatan model *random forest* dan *naive bayes*. Dan data *testing* sebagai evaluasi model untuk mengukur tingkat akurasi melalui *confusion matrix*.

Tabel 3. *Confusion matrix naïve bayes*

Prediction	Aktual		
	Negatif	**Netral**	**Positif**
Negatif	7595	0	1
Netral	15	5287	0
Positif	10	0	3428

Sumber: diolah dari Twitter, 2022

Tabel 4. *Confusion matrix random forest*

Prediction	Aktual		
	Negatif	**Netral**	**Positif**
Negatif	7614	0	0
Netral	15	5267	4
Positif	0	20	3425

Sumber: diolah dari Twitter, 2022

Berdasarkan tabel 3, performa model *naive bayes* cukup seimbang untuk ketiga model. Berdasarkan tabel diatas, model dapat memprediksi dengan tepat sebanyak 3.428 dari 3.438 *tweet* dengan sentimen positif (99,7%), 5287 dari 5307 *tweet* dengan sentimen netral (99,6%), dan 7595 dari 7596 *tweet* dengan sentimen negatif (99,9%).

Model diatas memiliki tingkat akurasi 99,84% yang tergolong sangat baik dengan interval kepercayaan 95% antara 99,7% hingga 99,9%. Sehingga dapat disimpulkan bahwa model *naive bayes* mampu mengklasifikasikan secara sangat baik *tweet* Twitter ke dalam tiga kelompok sentimen.

Sedangkan pada tabel 4 menunjukkan hasil *confussion matrix* dari model *random forest*. Performa model *random forest* juga cukup seimbang untuk ketiga model. Berdasarkan tabel diatas, model dapat memprediksi dengan tepat sebanyak 3.425 dari 3.445 *tweet* dengan sentimen positif (99,41%), 5267 dari 5286 *tweet* dengan sentimen netral (99,6%), dan 7614 dari 7614 tweet dengan sentimen negatif (100%).

Model diatas memiliki tingkat akurasi 99,82% yang tergolong sangat baik dengan interval kepercayaan 95% antara 99,7% hingga 99,8%. Sehingga

dapat disimpulkan bahwa model *random forest* mampu mengklasifikasikan secara baik *tweet* twitter kedalam tiga kelompok sentimen.

Kesimpulan, Rekomendasi, dan Pembatasan

Respon masyarakat terkait kasus perlindungan data pribadi dilihat dari sentimen di twitter. Rata-rata sentimen dalam periode 1 Juli 2021 hingga 29 September menunjukkan rata-rata sentimen negatif. Hal ini menandakan masih seringnya kasus kebocoran data pribadi di Indonesia. Sentimen negatif terjadi oleh kasus bocornya data pribadi dan lemahnya pertahanan *cyber* di Indonesia. Sentimen positif terjadi karena seringnya instansi mengkonfirmasi berita hoax dan melakukan upaya perbaikan regulasi. Hasil klasifikasi sentimen dari *tweet* twitter tersebut sangat baik, dikarenakan dalam metode naïve bayes dan random forest memberikan akurasi sebesar 99,84% dan 99,82%.

Hasil penelitian ini memberikan saran kepada pemerintah, instansi, dan lembaga untuk bersinergi melindungi data pribadi masyarakat, tidak cukup menjadi wilayah tugas BSSN saja. Selain itu masyarakat perlu memiliki kesadaran terhadap keamanan data pribadi.

Pembatasan penelitian ini terletak pada bagaimana respon masyarakat maya terhadap topik keamanan data pribadi yang dikaitkan dengan sejumlah fenomena kebocoran data yang terjadi di Indonesia selama 1 Juli 2021 hingga 29 September 2022. Oleh karena adanya keterbatasan dalam beberapa aspek penelitian ini seperti banyaknya *package* R yang digunakan dalam pemodelan *machine learning*, penelitian selanjutnya disarankan menggunakan algoritma *Deep Learning*. Sebab, pemanfaatan algoritma *Deep Learning* diduga mampu meningkatkan akurasi pemodelan.

Daftar Pustaka

Admin DPMPTSP Kaltim. 2021. "Tanggap Hadapi Insiden Keamanan Siber, Kabid Pengaduan Advokasi Dan Informasi Perizinan Hadiri Launching Kaltim-CSIRT Oleh Badan Sandi Siber Negara." Retrieved (https://dpmptsp.kaltimprov.go.id/index.php/single-berita/104).

Aldean, M. Y., P. Paradise, and N. A. S. Nugraha. 2022. "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac)." *INISTA: Journal of Informatics, Information System, Software Engineering and Applications* 4(2):64–72. https://doi.org/10.20895/inista.v4i2.575

Annur, Cindy Mutia. 2022. "Kebocoran Data Sering Terjadi Di 10 Sektor Industri Ini." *Katadata*. Retrieved (https://databoks.katadata.co.id/datapublish/2022/09/06/kebocoran-data-sering-terjadi-di-10-sektor-industri-ini).

Ardiyanti, H. 2014. "Cyber-Security Dan Tantangan Pengembangannya Di Indonesia." *Jurnal Politica* V(1):95–110. doi: 10.22212/jp.v5i1.336

Arifah, DIsta Amalia. 2011. "Kasus Cybercrime Di Indonesia." *Jurnal Bisnis Dan Ekonomi* 18(2). https://www.unisbank.ac.id/ojs/index.php/fe3/article/view/2099/767

Aswandi, Ririn. 2020. "Perlindungan Data Dan Informasi Pribadi Melalui Indonesian Data Protection System (IDPS)." *Lembaga Penalaran Dan Penulisan Karya Ilmiah Fakultas Hukum Universitas Hasanuddin* 3(3). https://doi.org/10.20956/jl.v3i2.14321

CNNIndonesia. 2021. "Rentetan Kebocoran Data Di Indonesia Sejak 2020." *CNN Indonesia*. Retrieved (https://www.cnnindonesia.com/teknologi/20210523132216-188-645888/infografis-rentetan-kebocoran-data-di-indonesia-sejak-2020).

Daffiandra. 2021. "Klarifikasi Isu Hoaks 16 November 2021." *Beritamagelang*. Retrieved (http://www.beritamagelang.id/hoax/klarifikasi-isu-hoaks-16-november-2021).

Delpiero, M., F. A. Reynaldi, I. U. Ningdiah, and N. Muthmainnah. 2021. "Analisis Yuridis Kebijakan Privasi Dan Pertanggungjawaban Online Marketplace Dalam Perlindungan Data Pribadi Pengguna Pada Kasus Kebocoran Data." *Padjadjaran Law Review* 9(1). http://jurnal.fh.unpad.ac.id/index.php/plr/article/view/509

KOMINFO. 2016. "Indonesia Sudah Miliki Aturan Soal Perlindungan Data Pribadi." Retrieved (https://www.kominfo.go.id/content/detail/8621/indonesia-sudah-miliki-aturan-soal-perlindungan-data-pribadi/0/sorotan_media).

Medhat, W., A. Hassan, and H. Korashy. 2014. "Sentiment Analysis Algorithms and Applications : A Survey." *Ain Shams Engineering Journal* 5(4):1093–1113. doi: http://doi.org/10.1016/j.asej.2014.04.011.

MENKOMINFO. 2022. "Perlindungan Data Pribadi." Retrieved (https://jdih.kominfo.go.id/infografis/view/19).

Pane, A. A., F. A. Siregar, and A. S. Sirait. 2021. "Efektivitas Pelaksanaan Silayda E-KTP." *El-Thawalib* 2(4):216–25. https://doi.org/10.24952/el-thawalib.v2i4.4231

Perpres. 2021. *Peraturan Presiden Republik Indonesia Nomor 28 Tahun 2021 Tentang Badan Siber Dan Sandi Negara (BSSN)*. https://peraturan.bpk.go.id/Home/Details/165493/perpres-no-28-tahun-2021

Primajaya, A., and B. N. Sari. 2018. "Random Forest Algorithm for Prediction of Precipitation." *Indonesian Journal of Artificial Intelligence and Data Mining* 1(1):27. doi: https://doi.org/10.24014/ijaidm.v1i1.4903.

Purnama, Bedy. 2019. *Pengantar Machine Learning*. Bandung: Informatika.

Religia, Yoga, Agung Nugroho, and Wahyu Hadikristanto. 2021. "Klasifikasi Analisis Perbandingan Algoritma Optimasi Pada Random Forest Untuk Klasifikasi Data Bank Marketing." *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* 5(1):187–192. doi: https://doi.org/10.29207/resti.v5i1.2813.

Rosalina, S. D., I. Y. Purbasari, and E. P. Mandyarta. 2022. "IMPLEMENTASI NAÏVE BAYES CLASSIFIER UNTUK MENDIAGNOSIS PENYAKIT INTELLECTUAL DISABILITY." *Jurnal Informatika Dan Sistem Informasi* 3(2):115–24. https://doi.org/10.33005/jifosi.v3i2.489

Sanjaya, Bram Ronald, D. Efrianti, M. Ali, T. Prasetyo, M. Mukhtadi, K. Y. Widyasari, and Z. Khumairoh. 2022. "Pengembangan Cyber Security Dalam Menghadapi Cyber Warfare Di Indonesia." *Journal of Advanced Research in Defense and Security Studies* 1(1):19–34. https://ejournal.hakhara-institute.org/index.php/JARDS/article/view/3

Sholehurrohman, R., and I. S. Ilman. 2022. "Analisis Sentimen Tweet Kasus Kebocoran Data Penggunaan Facebook Oleh Cambrigde Analytica." *Jurnal Pepadun* 3(1):140–47. https://doi.org/10.23960/pepadun.v3i1.108

Siburian, V. W., and I. E. Mulyana. 2018. "Prediksi Harga Ponsel Menggunakan Metode Random Forest." *Pros. Annu. Res. Semin* 4(1):144–147. https://seminar.ilkom.unsri.ac.id/index.php/ars/article/view/1992