ANALISIS SENTIMEN MASYARAKAT TERHADAP UNDANG-UNDANG PERLINDUNGAN DATA PRIBADI PADA APLIKASI X DENGAN METODE SUPPORT VECTOR MACHINE

Rayhan Abdul Jabbar Fahmi[1], Wahib Muhibi Nur[2], Dee Canawine[3]
Muhammad Naufal Kusumajaya[4], Ahmad Faris Fadhlillah[5], Nur Aini Rakhmawati[6]
[1,2,3,4,5,6] Departemen Sistem Informasi,
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[1]5026211003@student.its.ac.id, [2]05211940000059@student.its.ac.id, [3]5026211018@student.its.ac.id,
[4]5026201121@student.its.ac.id, [5]5026211115@student.its.ac.id, [6]5026211018@student.its.ac.id

ABSTRACT

Machine Learning plays a significant role in addressing classification and processing issues to predict developments related to the Personal Data Protection Law. Personal data is individual information that must be safeguarded, with protection guaranteed by the state. Unfortunately, Indonesia ranks very low in terms of cybersecurity compared to other countries. This research aims to explore challenges and seek potential solutions to ensure the security and protection of personal data. In this study, the Support Vector Machine (SVM) method is used to analyze and categorize public sentiment regarding the Personal Data Protection Law on the X application platform as positive, neutral, or negative. A sample data set of 372 tweets was obtained from Crawling results. The data was processed using the Python programming language. Before analysis, preprocessing was performed to remove unnecessary words or information so that the accuracy level obtained could approach the real situation. After analysis, the results showed 131 tweets with positive sentiment, 59 tweets with negative sentiment, and 182 tweets with neutral sentiment. Performance evaluation using the SVM algorithm with a 70% training data and 30% testing data comparison yielded an accuracy rate of 75.89%. This research is expected to identify public perceptions and responses to the Personal Data Protection Law and develop more effective strategies to maintain the security of personal data in Indonesia. This effort is important considering the increasing challenges of cyber threats in the current digital era.

***Keywords:** Cyber Security, Personal Data Protection, Sentiment Analysis, Support Vector Machine.*

I. PENDAHULUAN

menjadi isu yang semakin kompleks di era kemajuan teknologi informasi yang berkembang pesat seperti saat ini, terutama yang berkaitan dengan administrasi kependudukan. Menurut hasil survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), jumlah pengguna internet di Indonesia pada periode 2022-2023 mencapai 215,63 juta individu, angka ini menunjukkan peningkatan sebesar 2,67% dibandingkan dengan periode sebelumnya dengan jumlah pengguna internet mencapai 210,03 juta orang. Perlu diketahui bahwa jumlah pengguna internet tersebut setara dengan 78,19% dari total populasi Indonesia, yang pada periode tersebut berjumlah 275,77 juta jiwa [1].

Dengan alat, tujuan, dan topik yang sesuai, media sosial memiliki peran penting dalam konteks pemerintahan, khususnya untuk menggali inovasi dalam mengembangkan layanan pemerintahan berbasis elektronik [2]. Pemerintah telah mengambil langkah-langkah penting dalam menerapkan teknologi informasi dalam administrasi publik, seperti melalui pengenalan Kartu Tanda Penduduk elektronik (e-KTP) dan sistem PeduliLindungi selama pandemi COVID-19 yang dimulai sejak Juni 2020.

Sangat disayangkan masih kurangnya kesadaran masyarakat terhadap pentingnya melindungi data pribadi dan ketidakoptimalan infrastruktur keamanan siber merupakan tantangan utama [3]. Laporan National Cyber Security Index (NCSI) menempatkan Indonesia pada peringkat yang sangat rendah dalam hal keamanan siber di antara negara-negara G20

Perlindungan dan keamanan data pribadi telah

[4]. Dampak negatif dari tantangan ini ini telah sangat terasa dalam beberapa tahun terakhir, terutama melalui serangkaian kasus kebocoran data pribadi yang kerap terjadi sejak tahun 2014. Baru-baru ini, publik sekali lagi dikejutkan oleh insiden kebocoran data yang melibatkan 337 juta rekaman yang merupakan bagian dari basis data Direktorat Jenderal Kependudukan dan Pencatatan Sipil (Dukcapil) Kementerian Dalam Negeri (Kemendagri) [5]. Kebocoran data sebesar ini memiliki implikasi serius, terutama terkait dengan privasi dan keamanan individu yang terkena dampaknya. Tidak hanya merugikan individu secara pribadi, kebocoran data semacam ini juga dapat disalahgunakan oleh pihak-pihak yang tidak bertanggung jawab, yang dapat memanfaatkan data pribadi untuk tujuan kriminal, seperti pencurian identitas, penipuan, atau pemerasan. Selain itu, kasus-kasus seperti ini juga merusak kepercayaan publik terhadap lembaga dan pemerintah yang seharusnya bertanggung jawab atas keamanan data pribadi warganya [6]. Beberapa penelitian sebelumnya telah mengadakan analisis sentimen dari sumber data Twitter. Contohnya, penelitian Wibowo, dkk. yang menganalisis sentimen terkait kasus kebocoran data Tokopedia pada bulan Mei 2020 menggunakan algoritma seperti Random Forest, Logistic Regression, dan Support-Vector Machine [7]. Selain itu, Amal, dkk. juga melakukan analisis klasifikasi sentimen terhadap isu kebocoran data kartu identitas ponsel di Twitter pada bulan Desember 2022 dengan menggunakan algoritma yang sama [2]. Sementara itu, Nursiyono, dkk. mengadakan

analisis sentimen pada perlindungan data pribadi dengan pendekatan Machine Learning menggunakan metode Naive Bayes [6]. Dalam tugas analisis sentimen, model IndoBERT menunjukkan kinerja terbaik dalam hal F1 Score dibandingkan dengan berbagai model lainnya, termasuk Naive Bayes, Logistic Regression, BiLSTM, dan lainnya [8].

Dengan demikian, perlu dilakukan penelitian dengan metode *Support Vector Machine* (SVM) untuk menganalisis sentimen netizen terhadap disahkannya UU PDP di platform X. Upaya ini diharapkan akan memberikan wawasan yang berharga untuk mengatasi isu yang sangat krusial ini dan memberikan panduan untuk meningkatkan kesadaran masyarakat serta memperkuat infrastruktur keamanan siber. Dalam penelitian ini, metode tersebut akan menampilkan sebuah algoritma *Machine Learning* yang dikembangkan untuk klasifikasi dan regresi untuk memilah data sehingga akan didapatkan data berupa konteks analisis sentimen masyarakat tentang UU Perlindungan Data Pribadi pada platform aplikasi X. Konsep dasar dari penggunaan SVM adalah pendekatan berbasis pengambilan sampel mengubah distribusi pelatihan data, sehingga kedua kelas (negatif dan positif) dapat disajikan dengan baik melalui data pelatihan [9].

II. TINJAUAN PUSTAKA

Data Crawling

Data *Crawling*, atau perayapan data adalah sebuah proses otomatis untuk mengumpulkan dan melakukan indeks data dari berbagai sumber tertentu atau sistem komputer yang menjadi target tertentu untuk dilakukan pengambilan data. Aktivitas ini biasanya dilakukan oleh program komputer, *bot* atau *software* yang biasa dikenal sebagai "web crawler" atau "spider" [10].

Preprocessing Data

Data *Preprocessing* merupakan tahapan awal yang diimplementasikan pada database untuk menghapus *noise*, *missing value*, dan data yang tidak konsisten. Tahapan ini bertujuan melengkapkan database sehingga menjadi konsisten [11]. Agar data tersebut menjadi tepat sasaran dan menghasilkan analisis yang akurat, berikut langkah-langkah dalam Data Preprocessing:

1. *Data Cleaning*

Proses pembersihan *tweet* dari kata yang tidak diperlukan seperti HTML, *emoticons, hashtag, username,* dan *url*.

2. *Case Folding*

Tahapan dalam *text preprocessing* yang dilakukan untuk menyeragamkan bentuk atau karakter huruf besar pada data diubah seluruhnya menjadi huruf kecil [12]. Pada proses ini karakter-karakter 'A'-'Z' yang terdapat pada data diubah ke dalam karakter 'a'-'z'.

3. *Stopword Removal*

Proses menghapus kata yang tidak relevan, bersifat general dan sering muncul dalam suatu kalimat berdasarkan daftar *stopword*. Daftar *stopword* yang biasa digunakan bisa berbentuk *digital library* yang sudah pernah tersedia sebelumnya, namun tidak semua kata-kata yang termasuk di dalamnya tergolong kata yang tidak relevan dalam suatu data tertentu .

Support Vector Machine (SVM)

Support Vector Machine merupakan teknik prediksi pada mesin pembelajaran universal dan bisa diterapkan pada klasifikasi, regresi maupun pengenalan pola (pattern recognition) [7]. Metode SVM mampu menyelesaikan permasalahan menggunakan fungsi linear dalam suatu ruang fitur (feature space) yang dikembangkan untuk mengatasi permasalahan dengan didasarkan pada teori optimasi dan mengimplementasikan learning bias dari teori statistik klasik.

Natural Language Processing *(*NLP*)*

NLP merupakan bidang yang berkaitan dengan komputasi dan kecerdasan buatan yang berfokus pada interaksi antara komputer dan bahasa manusia yang alami atau alamiah. Tujuan utama NLP adalah untuk memungkinkan komputer untuk memahami, memproses, dan menghasilkan bahasa manusia dengan cara yang bermakna. Beberapa bagian pada teks yang dianggap kurang relevan seringkali diabaikan karena sistem hanya mentitikberatkan pada bagian spesifik domain teks, masalah pada NLP seringkali diperingkas [13].

Akurasi Model dan Klasifikasi

Dasar pengambilan keputusan algoritma terbaik dalam analisis menggunakan akurasi model yang digunakan dalam penelitian. Dalam algoritma klasifikasi *Machinie Learning*, digunakan *confusion matrix* sebagai metode untuk mengukur dan mengevaluasi kinerja.

Tabel 1. Konfigurasi *Confusion Matrix*

Prediksi	Sebenarnya		
	Positif	Netral	Negatif
Positif	TP	FP	FP
Netral	-	TP	-
Negatif	FP	FP	TP

Dalam mengukur ketepatan klasifikasi, perlu diketahui jumlah pada setiap kelas prediksi dan aktual, terdiri atas:

- *True Positif* (TP) yaitu jumlah ulasan bersentimen positif yang tepat terprediksi dalam kelas positif
- *True Negatif* (TN) yaitu jumlah ulasan yang bersentimen negatif tepat terprediksi dalam kelas negatif
- *False Positif* (FP) yaitu jumlah ulasan bersentimen negatif yang terprediksi dalam kelas positif
- *False Negatif* (FN) yaitu jumlah ulasan bersentimen positif yang terprediksi dalam kelas negatif.

Secara matematis, pengukuran akurasi model dapat dituliskan sebagai 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

Akurasi merupakan ukuran yang paling banyak digunakan dalam melakukan pengujian model algoritma klasifikasi pada *Machine Learning* [14].

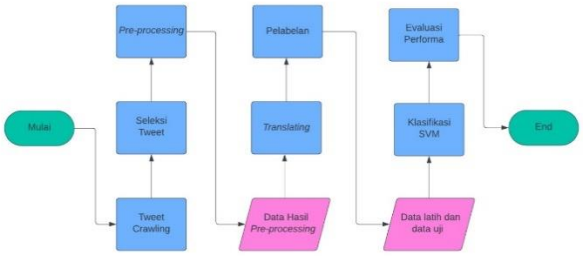
Word Cloud

Sebuah sistem yang memunculkan kata sebagai citra visual

terkait frekuensi kemunculan kata dalam suatu teks [15]. Visualisasi *word cloud* akan memudahkan pengamat dalam melihat gagasan sehingga dapat menjadi alat bantu dalam melakukan analisis terhadap sebuah wacana tertulis.

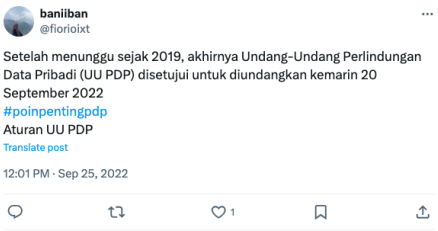
III. METODE PENELITIAN

Dalam penelitian ini, terdapat tahapan-tahapan yang perlu dilewati sebelum dapat mengetahui seberapa akurat algoritma yang digunakan yang dirancang dalam suatu alur proses tahapan penelitian agar sesuai dan terstruktur, *flowchart* alur proses penelitian dapat dilihat seperti pada Gambar 1 sebagai berikut.



Gambar 1. Alur Penelitian

Tools yang kami gunakan, yaitu Google Colaboratory yang merupakan layanan *Machine Learning* yang disediakan oleh *Google*. Platform ini memungkinkan pengguna untuk menelusuri dan menjalankan kode *Python*, mengunggah dan menyimpan kode dengan data, dan berkolaborasi dengan pengguna lain. Selain itu, Google Colaboratory menyediakan akses ke *Virtual Machine* dengan CPU, GPU, dan Tensor Processing Unit (TPU) yang mampu menjalankan kode *Python* [16]. Langkah awal penelitian kami dilakukan dengan mengumpulkan data menggunakan teknik *crawling* dari platform media sosial X dengan bantuan tool Harvest [17]. Teknik *Crawling* melibatkan pengumpulan data sekunder secara otomatis dengan menggunakan kata kunci tertentu [18]. *Keyword* yang kami gunakan dalam penelitian ini adalah "UU PDP" untuk mencari tweet berbahasa Indonesia dari rentang waktu 15 September 2022 hingga 1 Oktober 2023. Data yang terkumpul kemudian disaring secara manual untuk memastikan hanya tweet tunggal yang diposting oleh satu akun pada satu waktu yang disimpan [19].



Gambar 2. Salah satu tweet

Tahap berikutnya adalah pra-pemrosesan data. Langkah pertama melibatkan pembersihan dan penyesuaian huruf, termasuk penghapusan *mention*, tanda kutip ganda, retweet, hyperlink, emotikon, dan pengkonversian seluruh teks menjadi huruf kecil. Selanjutnya, dilakukan Formalisasi Slang Words untuk menerjemahkan kata-kata slang Indonesia ke dalam bentuk standarnya. Proses ketiga melibatkan Formalisasi Singkatan, yang mengubah singkatan-singkatan Indonesia menjadi bentuk lengkapnya. Langkah keempat adalah Formalisasi Akronim, di mana singkatan-singkatan Indonesia spesifik digunakan sebagai kunci dan bentuk lengkapnya sebagai nilai. Proses kelima adalah Stemming, yang mengurangi kata-kata menjadi bentuk akar atau dasarnya. Terakhir, langkah keenam adalah Remove Stopword, yang menghapus kata-kata yang tidak memberikan kontribusi penting dalam analisis.

Selanjutnya, akan diperoleh data hasil yang telah dipersiapkan dan diproses dengan menerjemahkan data hasil preprocessing dari bahasa Indonesia ke bahasa Inggris karena library TextBlob hanya dapat mengenali Bahasa Inggris dan masih belum terdapat Adaptor TextBlob untuk Bahasa Indonesia [20].

Tahap selanjutnya adalah melakukan pelabelan data secara otomatis menggunakan TextBlob. Pada tahap ini nilai *subjectivity* dan *polarity* dari setiap tweet diketahui, yang selanjutnya dapat digunakan untuk melakukan pelabelan sentimen positif, netral, dan negatif.

Pada algoritma SVM, perlu dilakukan pendefinisian persamaan suatu hyperplane pemisah yang dituliskan dengan:

$$W \cdot X + b = 0 \qquad (1)$$

W adalah suatu bobot vektor, yaitu W = {W1, W2,...,Wn}; n adalah jumlah atribut dan b merupakan suatu skalar yang disebut dengan bias. Jika berdasarkan pada atribut A1, A2 dengan permisalan tupel pelatihan X = (x1, x2), x1 dan x2 merupakan nilai dari atribut A1 dan A2, dan jika b dianggap sebagai suatu bobot tambahan w0, maka persamaan suatu hyperplane pemisah dapat ditulis ulang sebagai berikut:

$$w_0 + w_1 w_1 + w_2 w_2 = 0 \qquad (2)$$

Setelah persamaan dapat didefinisikan, nilai x1 dan x2 dapat dimasukkan ke dalam persamaan untuk mencari bobot w1, w2, dan w0 atau b. SVM menemukan *hyperplane* pemisah maksimum dengan mempunyai jarak maksimum antara tupel pelatihan terdekat. *Support vector* ditunjukkan dengan batasan tebal pada titik tupel. Dengan demikian, setiap titik yang terletak di atas *hyperplane* pemisah memenuhi rumus:

$$w_0 + w_1 w_1 + w_2 w_2 > 0 \qquad (3)$$

Sedangkan, titik yang terletak di bawah hyperplane pemisah memenuhi rumus:

$$w_0 + w_1 w_1 + w_2 w_2 < 0 \qquad (4)$$

Melihat dua kondisi di atas, maka didapatkan dua persamaan *hyperplane* yaitu:

$$H_1 : w_0 + w_1 w_1 + w_2 w_2 \geq 1 \quad \text{untuk } y_i = +1 \qquad (5)$$

$$H_2 : w_0 + w_1 w_1 + w_2 w_2 \leq -1 \quad \text{untuk } y_i = -1 \qquad (6)$$

Perumusan model SVM menggunakan trik matematika yaitu formula Lagrangian. Berdasarkan Lagrangian *formulation*, *Maksimum Margin Hyperplane* (MMH) dapat ditulis ulang sebagai suatu batas keputusan (decision boundary) yaitu:

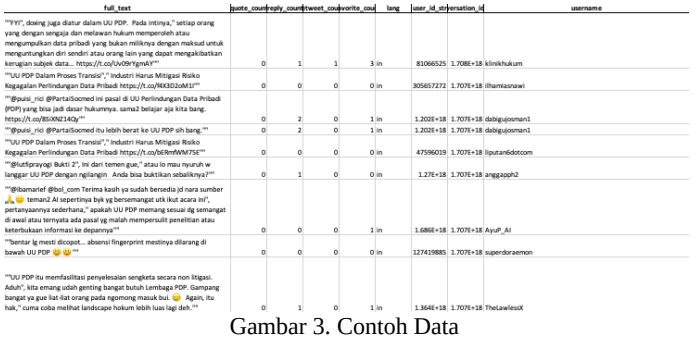
$$d(X^T) = \sum_{i=1}^{l} y_i a_i X_i X^T + b_0 \qquad (7)$$

yi adalah label kelas support vector Xi. XT adalah suatu tupel test. αi dan b0 adalah parameter numerik yang ditentukan secara otomatis oleh optimalisasi algoritma SVM dan l adalah jumlah *vector support* [21].

IV. HASIL DAN PEMBAHASAN

Pengumpulan Data Tweet

Pengumpulan data menggunakan teknik *tweet crawling*. Teknik ini melibatkan pengambilan data dengan menggunakan kata kunci "UU PDP" untuk mencari tweet berbahasa Indonesia dari rentang waktu 15 September 2022 hingga 1 Oktober 2023. Data yang terkumpul kemudian disaring secara manual untuk memastikan hanya tweet tunggal yang diposting. Hasilnya dapat terihat dalam Gambar 3.



full_text	quote_count	reply_count	tweet_count	favorite_count	lang	user_id_str	conversation_id	username
""FYI", doxing juga diatur dalam UU PDP. Pada intinya," setiap orang yang dengan sengaja dan melawan hukum memperoleh atau mengumpulkan data pribadi yang bukan miliknya dengan maksud untuk menguntungkan diri sendiri atau orang lain yang dapat mengakibatkan kerugian subjek data... https://t.co/UvO9rYgmAY""	0	1	1	3	in	81066525	1.708E+18	klinikhukum
""UU PDP Dalam Proses Transisi"," Industri Harus Mitigasi Risiko Kegagalan Perlindungan Data Pribadi https://t.co/RK3D2oM1I""	0	0	0	0	in	3056572 72	1.707E+18	ilhamiasnawi
""@puisi_nici @PartaiSocmed ini pasal di UU Perlindungan Data Pribadi (PDP) yang bisa jadi dasar hukumnya. sama2 belajar aja kita bang. https://t.co/BSiXNZ14Qy""	0	2	0	1	in	1.202E+18	1.707E+18	dabiguiosman1
""@puisi_nici @PartaiSocmed itu lebih berat ke UU PDP sih bang.""	0	2	0	1	in	1.202E+18	1.707E+18	dabiguiosman1
""UU PDP Dalam Proses Transisi"," Industri Harus Mitigasi Risiko Kegagalan Perlindungan Data Pribadi https://t.co/bERnfWM7SE""	0	0	0	0	in	47596039	1.707E+18	liputan6dotcom
""@lutfipryogi Bukti 2", Ini dari temen gue," atau lo mau nyuruh w langgar UU PDP dengan ngilangin Anda bisa buktikan sebaliknya?""	0	1	0	0	in	1.27E+18	1.707E+18	anggapph2
""@ibamarief @bol_com Terima kasih ya sudah bersedia jd nara sumber 🙏🤝 teman2 AI sepertinya byk yg bersemangat utk ikut acara ini", pertanyaannya sederhana," apakah UU PDP memang sesuai dg semangat di awal atau ternyata ada pasal yg malah mempersulit penelitian atau keterbukaan informasi ke depannya""	0	0	0	1	in	1.686E+18	1.707E+18	AyuP_AI
""bentar lg mesti dicopot... absensi fingerprint mestinya dilarang di bawah UU PDP 😂😂""	0	0	0	0	in	127419885	1.707E+18	superdoraemon
""UU PDP itu memfasilitasi penyelesaian sengketa secara non litigasi. Aduh", kita emang udah genting bangat butuh Lembaga PDP. Gampang bangat ya gue liat-liat orang pada ngomong masuk bui. 😂 Again, itu hak," cuma coba melihat landscape hokum lebih luas lagi deh.""	0	1	0	1	in	1.364E+18	1.707E+18	TheLawlessX

Gambar 3. Contoh Data

Preprocessing Data

Preprocessing Data merupakan tahapan Dimana data yang diperoleh dengan cara crawling sebanyak 372 data, selanjutnya dilakukan pra-pemrosesan data. Teknik yang digunakan pada tahap ini yaitu *Cleaning* dan *Case Folding*, Formalisasi *Slang Words*, Formalisasi Singkatan, Formalisasi Akronim, *Stemming*, dan *Remove Stopword.*

Tabel 3. Contoh Data Hasil Preprocessing Data

Data Tweet	Hasil Preprocessing
Lagian selow amat hidupnya kayak ga ada kerjaan pake nyerang ke WA orang. Serem banget. Selain UU ITE ada juga UU PDP. Pada punya smartphone yg konek ke internet bs digunakan utk cari tau soal aturan hukum ttg teror online.	selow hidupnya kayak nggak kerjaan pakai serang wa orang. serem banget. undang-undang ite undang-undang pdp. smartphone konek internet cari aturan hukum teror online.
Tahu ngga si lur perbuatan yang melanggar UU perlindungan data pribadi beserta sanksinya menurut UU PDP No.27 Tahun 2022 Coba lurrr disimak dulu postingan mimin 👇👇👇	perbuatan melanggar undang-undang perlindungan data pribadi beserta sanksinya undang-undang perlindungan data pribadi nomor tahun coba lurrr disimak postingan mimin
""Mana nih badan penegak UU PDP □□□""	badan penegak undang-undang perlindungan data pribadi
""@airsyadNoi Hahaha tp kebayang gak kalo waktu itu udah ada UU PDP? Bisa digugat itu ngespill"	hahaha kebayang gak waktu undang-undang pdp? digugat ngespill
""@gurnadi @slaksmi @yanuarnugroho @ngabdul @sofiesyarief @icipg @nalarinstitute @dinitaputri @safenetvoice @ainunnajib @ConversationIDN Nah", ini. Sudah di-anonimisasi," jadi perhatian Mas Yan tentang kemungkinan pelanggaran UU PDP tersebut bisa diminimalisir.""	nah, ini. di-anonimisasi, perhatian mas yan pelanggaran undang-undang perlindungan data pribadi diminimalisir.

Visualisasi Word Cloud

Visualisasi Word Cloud merupakan salah satu cara untuk menemukan kata-kata yang populer digunakan dalam penelitian UU PDP. Jadi kata yang lebih banyak digunakan masyarakat adalah "data pribadi", "UU PDP, dan "Jejak Serta Digital". Gambar 4 menunjukkan kata-kata yang sering digunakan dalam opini netizen tentang Pengesahan UU PDP.

Gambar 4. Visualisasi Word Cloud UU PDP

Word Cloud ini dapat menginterpretasikan bahwa sebagian besar sentimen saat ini terkait undang-undang perlindungan data pribadi cenderung netral dan jarang sekali ditemukan *keyword* sentimen yang negatif.

Translating

Data yang sudah melalui proses cleaning pada preprocessing data, selanjutnya diterjemahkan ke dalam bahasa Inggris terlebih dahulu. Karena menggunakan library textblob, sebelumnya library yang digunakan harus di-*install* dengan perintah "!pip install deep-translator". Setelah library diinstall, library harus diimport dan diletakkan pada section dimana seluruh *library* disatukan.

Tabel 4. Contoh Data Hasil Translating

Hasil Preprocessing	Hasil Translating
selow hidupnya kayak nggak kerjaan pakai serang wa orang. serem banget. undang-undang ite undang-undang pdp. smartphone konek internet cari aturan hukum teror online.	in his low life it's like he doesn't have any work to do to attack people. really scary. ite law pdp law. smartphone connected to the internet, look for terror laws online.
si lur perbuatan melanggar undang-undang perlindungan data pribadi beserta sanksinya undang-undang perlindungan data pribadi nomor tahun coba lurrr disimak postingan mimin selengkapnya : betterbanyumas kabbanyumas bijakmenggunakanmediadigital	sir, the act of violating the personal data protection law and its sanctions, the personal data protection law, year number, try to listen to Mimin's post in full: betterbanyumas kabbanyumas wisely use digital media
badan penegak undang-undang perlindungan data pribadi	personal data protection law enforcement bodies
hahaha kebayang gak waktu undang-undang pdp? digugat ngespill	hahaha can you imagine when the PDP law was passed? sued for spilling
nah, ini. di-anonimisasi, perhatian mas yan pelanggaran undang-undang perlindungan data pribadi diminimalisir.	Well, this one. anonymized, the attention of the public regarding violations of personal data protection laws is minimized.

Labeling

Untuk mengkategorikan label tweet tergolong positif, netral, atau negatif, maka nilai *polarity* setiap kata dihitung dan hasilnya akan seperti pada Gambar 5.

```
0      fyi, doxing diatur undang-undang pdp. intinya,...
1      undang-undang perlindungan data pribadi proses...
2      pasal undang-undang perlindungan data pribadi ...
3      undang-undang perlindungan data pribadi proses...
4      bukti , teman gue, nyuruh langgar undang-undan...
                           ...
367    rancangan undang-undang perlindungan data prib...
368    dpr resmi mengesahkan rancangan undang-undang ...
369    rapat paripurna dpr ri digelar ini, selasa, se...
370    dpr mengesahkan ruu perlindungan data pribadi ...
371    dpr mengesahkan ruu perlindungan data pribadi ...
Name: stop_word_removed, Length: 372, dtype: object
```

Gambar 5. Hasil Subjectivity dan Polarity

Setelah data dibersihkan, dilakukan pelabelan data dengan teknik *semi-supervised learning.* Gambar 6 menunjukkan contoh data yang sudah diberi label

	translated_tweet	subjectivity	polarity	Analysis
0	FYI, doxing is regulated by PDP law. In essenc...	0.316667	-0.083333	Negative
1	personal data protection law transition proces...	0.300000	-0.105556	Negative
2	articles of the personal data protection law (...	0.250000	0.100000	Positive
3	personal data protection law transition proces...	0.300000	-0.105556	Negative
4	evidence, my friend, asking to violate the law...	0.300000	0.000000	Neutral
...	...	...	...	...
367	The draft personal data protection law (ruu pd...	0.300000	0.000000	Neutral
368	The DPR officially ratified the draft law (RUU...	0.300000	0.000000	Neutral
369	The plenary session of the Indonesian House of...	0.300000	0.000000	Neutral
370	The DPR passed the Personal Data Protection Bi...	0.300000	0.000000	Neutral
371	The DPR passed the UU Personal Data Protection...	0.300000	0.000000	Neutral

372 rows × 4 columns

Gambar 6. Hasil Label Polarity

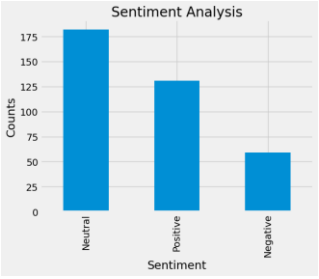
Sehingga ketika dihitung kategori sentimen pada label, diperoleh hasil yang dapat ditunjukkan pada Gambar 7.

```
Neutral     182
Positive    131
Negative     59
Name: Analysis, dtype: int64
```

Gambar 7. Hasil Hitung Label

Analisis Sentimen

Proses Analisis Sentimen dilakukan dengan menggunakan NLP dan algoritma Support Vector Machine (SVM) untuk melakukan klasifikasi. Berikut ini merupakan Unigram yang memvisualisasikan kata yang paling sering muncul per kategori sentimen dalam bentuk diagram bar yang ditunjukkan pada Gambar 8.



Gambar 8. Hasil Klasifikasi

Berdasarkan hasil dari perhitungan nilai polarity yang diperoleh, maka hasil analisis sentimen masyarakat terhadap UU Perlindungan Data Pribadi (PDP) pada aplikasi X dapat dilihat pada Tabel 5.

Tabel 5. Persentase Hasil Analisis Sentimen

Sentimen (%)		
Positif	Netral	Negatif
35.2	48.9	15.9

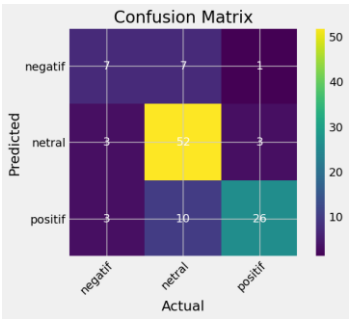
Hasil yang diperoleh untuk analisis sentiment paling tinggi adalah sentiment netral dengan presentase 48.9%, diikuti dengan sentimen positif dengan presentase 35.2%, dan sentiment negatif dengan presentase 12.9%.

Evaluasi

Evaluasi kinerja dilakukan dengan menggunakan perbandingan 70% data latih dan 30% data uji. Di bawah ini adalah gambar laporan klasifikasi yang diperoleh.

	precision	recall	f1-score	support
0	0.54	0.47	0.50	15
1	0.75	0.90	0.82	58
2	0.87	0.67	0.75	39
accuracy			0.76	112
macro avg	0.72	0.68	0.69	112
weighted avg	0.76	0.76	0.75	112

Gambar 9. Hasil Perhitungan Precision, Recall, dan F1-Score



Gambar 10. Kinerja Hasil Klasifikasi

Dari confusion matrix tersebut, dapat diketahui bahwa:

- True Netral, Positif, dan Negatif
  Prediksi tweet netral dengan aktual netral sebanyak 52 tweet; prediksi tweet positif dengan aktual positif sebanyak 26 tweet; dan prediksi tweet negatif dengan aktual negatif sebanyak 7 tweet.
- False Positif
  Tweet aktual netral dengan prediksi positif sebanyak 10 tweet dan prediksi aktual negatif dengan prediksi positif sebanyak 3 tweet.
- False Negatif
  Tweet aktual netral dengan prediksi negatif sebanyak 7 tweet dan aktual positif dengan prediksi negatif sebanyak 1 tweet.

V. KESIMPULAN

Analisis sentimen terkait undang-undang perlindungan data pribadi pasca diresmikan yang diambil dari aplikasi X, diperoleh sebanyak 372 tweet pada sebuah postingan tentang UU PDP menggunakan teknik *crawling*. Proses analisis data tweet menghasilkan sebanyak 131 tweet yang termasuk kategori tweet positif, 182 tweet netral, dan 59 tweet negatif. Evaluasi kinerja menggunakan algoritma SVM dengan perbandingan 70% data pelatihan dan 30% pengujian data menghasilkan akurasi data sebesar 75.89%. Dengan presentase tersebut, maka teknik yang digunakan ini termasuk kategori cukup akurat. Dapat disimpulkan, sentimen netizen Indonesia terhadap UU PDP sebagian besar netral dengan jumlah 182 dari 372 tweet. Meskipun keakuratan datanya sebesar 75.89%, penelitian kami masih terdapat kekurangan. Salah satu kekurangan dalam penelitian ini terjadi pada saat pelabelan. Dalam proses ini, ada kemungkinan perbedaan persepsi terhadap suatu tweet. Selain itu, penelitian ini hanya menggunakan kurang dari seribu tweet. Kedepannya, penelitian ini dapat dilakukan dengan audiens yang lebih besar dan metode yang lebih baik untuk memastikan dapat memperoleh hasil presentase yang lebih akurat lagi.

REFERENSI

[1] R. Yati, “Survei APJII: Pengguna Internet di Indonesia Tembus 215 Juta Orang,” teknologi.bisnis.com. Accessed: Feb. 08, 2024. [Online].

Available: https://teknologi.bisnis.com/read/20230308/101/1635219/survei-apjii-pengguna-internet-di-indonesia-tembus-215-juta-orang

[2] M. Ichlasul Amal, E. Syafira Rahmasita, E. Suryaputra, and N. Aini Rakhmawati, "Analisis Klasifikasi Sentimen Terhadap Isu Kebocoran Data Kartu Identitas Ponsel di Twitter," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, Dec. 2022.

[3] A. Kharis Almasyhari, Y. Priatna Sari, and F. Sukesti, "Edukasi Literasi Digital: Peningkatan Kesadaran Masyarakat Dalam Perlindungan Data Pribadi dan Kaitannya Terhadap Financial Technology," 2022.

[4] C. Mutia Annur, "Indeks Keamanan Siber Indonesia Peringkat ke-3 Terendah di Antara Negara G20," databoks.katadata.co.id. Accessed: Apr. 03, 2024. [Online]. Available: https://databoks.katadata.co.id/datapublish/2022/09/13/indeks-keamanan-siber-indonesia-peringkat-ke-3-terendah-di-antara-negara-g20

[5] I. Basyari, "Kemendagri Investigasi Dugaan Kebocoran 337 Juta Data Dukcapil," kompas.id. Accessed: Apr. 03, 2024. [Online]. Available: https://www.kompas.id/baca/polhuk/2023/07/17/337-juta-data-dukcapil-diduga-bocor

[6] J. Ade Nursiyono and Q. Huda, "Analisis Sentimen Twitter Terhadap Perlindungan Data Pribadi dengan Pendekatan Machine Learning," 2023.

[7] N. Ikbar Wibowo, T. Andika Maulana, H. Muhammad, and N. Aini Rakhmawati, "Perbandingan Algoritma Klasifikasi Sentimen Twitter Terhadap Insiden Kebocoran Data Tokopedia," MEI, 2021. [Online]. Available: https://github.com/nadhifikbarw/ep-scrapper

[8] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.00677

[9] H. Sain and S. Wulan Purnami, "Combine Sampling Support Vector Machine for Imbalanced Data Classification," in *Procedia Computer Science*, Elsevier, 2015, pp. 59–66. doi: 10.1016/j.procs.2015.12.105.

[10] F. Amalia Mahmud, "Data Crawling: Fungsi & Perbedaannya dengan Data Scraping," cmlabs.co.

[11] G. Nurvinda and A. Widya Davita, "Langkah Awal dalam Pemrosesan Data: Data Preprocessing dalam Data Mining," dqlab.id. Accessed: Apr. 03, 2024. [Online]. Available: https://dqlab.id/langkah-awal-dalam-pemrosesan-data-dalam-data-mining

[12] S. Multi Fani, R. Santoso, and Suparti, "PENERAPAN TEXT MINING UNTUK MELAKUKAN CLUSTERING DATA TWEET AKUN BLIBLI PADA MEDIA SOSIAL TWITTER MENGGUNAKAN K-MEANS CLUSTERING," vol. 10, pp. 583–593, 2021, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/

[13] K. R. Chowdary, "Natural Language Processing," in *Fundamentals of Artificial Intelligence*, 2020, pp. 603–649.

[14] B. Purnama, *Pengantar Machine Learning: Konsep dan Praktikum dengan Contoh Latihan Berbasis R dan Python*. Penerbit Informatika Bandung, 2019.

[15] I. Tri Julianto and Lindawati, "Analisis Sentimen Terhadap Sistem Informasi Akademik Mahasiswa Institut Teknologi Garut," 2022. [Online]. Available: https://jurnal.itg.ac.id/

[16] W. Vallejo, C. Díaz-Uribe, and C. Fajardo, "Google Colab and Virtual Simulations: Practical e-Learning Tools to Support the Teaching of Thermodynamics and to Introduce Coding to Students," *ACS Omega*, vol. 7, no. 8, pp. 7421–7429, Mar. 2022, doi: 10.1021/acsomega.2c00362.

[17] H. Satria, "Crawl Data Twitter Menggunakan Tweet Harvest," github.com. Accessed: Oct. 02, 2023. [Online]. Available: https://github.com/helmisatria/tweet-harvest

[18] A. Upreti, "Hands-on Web Scraping: Building your Twitter Dataset with Python and Scrapy," towardsdatascience.com. Accessed: Apr. 03, 2024. [Online]. Available: https://towardsdatascience.com/hands-on-web-scraping-building-your-own-twitter-dataset-with-python-and-scrapy-8823fb7d0598

[19] W. Muhibi Nur, M. Naufal Kusumajaya, D. Canawine, R. Abdul Jabbar Fahmi, A. Faris Fadhlillah, and N. Aini Rakhmawati, "Preprocessed Indonesian Twitter Dataset on UU Perlindungan Data Pribadi for Sentiment Analysis Research," zenodo.org. Accessed: Oct. 03, 2023. [Online]. Available: https://zenodo.org/records/8411378

[20] K. Arun and A. Srinagesh, "Multi-lingual Twitter sentiment analysis using machine learning," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 6, pp. 5992–6000, Dec. 2020, doi: 10.11591/ijece.v10i6.pp5992-6000.

[21] M. P. Dwi Cahyo, Widodo, and B. Prasetya Adhi, "Kinerja Algoritma Support Vector Machine dalam Menentukan Kebenaran Informasi Banjir di Twitter," *PINTER : Jurnal Pendidikan Teknik Informatika dan Komputer*, vol. 3, no. 2, pp. 116–121, Dec. 2019, doi: 10.21009/pinter.3.2.5.