


Analisis Sentimen Aplikasi Identitas Kependudukan Digital Menggunakan Metode Support Vector Machine dan Lexicon Based

Bella¹⁾, Wahyuni²⁾, dan Ita Arfyanti³⁾ 

Program Studi Sistem Informasi, STMIK Widya Cipta Dharma
Jl. Prof. M. Yamin No 25, Samarinda 75123
E-mail: bella181704@gmail.com, Wahyuni@wicida.ac.id, ita@wicida.ac.id

ABSTRAK

Aplikasi identitas kependudukan digital menjadi fokus utama dalam penelitian ini, data yang digunakan adalah hasil dari *scraping* ulasan yang ada di *Google Play Store*. Metode SVM digunakan untuk klasifikasi sentimen berdasarkan fitur-fitur teks dari tanggapan pengguna, sementara pendekatan *lexicon based* digunakan untuk memperkuat analisis sentimen dengan memanfaatkan kumpulan kata-kata berlabel sentimen. Hasil penelitian yang dilakukan menunjukkan bahwa metode SVM mampu mengklasifikasikan dengan akurasi sebesar 81%, presisi dan recall untuk sentimen negatif masing-masing 87%, untuk sentimen netral 61% dan 65%, serta untuk sentimen positif 84% dan 80%. Temuan dari penelitian ini diharapkan dapat memberikan wawasan yang berharga bagi pengembang aplikasi identitas kependudukan digital untuk meningkatkan pengalaman pengguna dan menjadi bahan evaluasi untuk kedepannya.

Kata Kunci: Analisis Sentimen, Identitas Kependudukan Digital, Support Vector Machine

SENTIMENT ANALYSIS OF DIGITAL POPULATION IDENTITY APPLICATIONS USING SUPPORT VECTOR MACHINE AND LEXICON BASED METHODS

ABSTRACT

The digital population identity application is the main focus of this research, the data used is the result of scraping reviews on the Google Play Store. The SVM method is used for sentiment classification based on text features from user responses, while a lexicon-based approach is used to strengthen sentiment analysis by utilizing a collection of sentiment-labeled words. The results of the research conducted show that the SVM method is able to classify with an accuracy of 81%, precision and recall for negative sentiments respectively 87%, for neutral sentiments 61% and 65%, and for positive sentiments 84% and 80%. Findings from the research It is hoped that this will provide valuable insight for developers of digital population identity applications to improve user experience and serve as evaluation material for the future.

Keywords: Sentiment Analysis, Digital Population Identity, Support Vector Machine (SVM)

1. PENDAHULUAN

Identitas kependudukan digital menjadi salah satu inovasi teknologi yang krusial dalam transformasi layanan publik dan administrasi di era digital saat ini. Konsep ini tidak hanya menggantikan format tradisional kartu identitas fisik dengan bentuk digital yang lebih fleksibel, tetapi juga membawa implikasi mendalam terkait dengan keamanan, privasi, efisiensi, dan penerimaan masyarakat. Ini tidak hanya memudahkan akses layanan publik tetapi juga meningkatkan keamanan identitas dan efisiensi administrasi. Implementasi identitas kependudukan digital menawarkan berbagai manfaat, termasuk pengurangan birokrasi, peningkatan akurasi data, dan pelayanan yang lebih cepat. Namun, tantangan seperti keamanan data, privasi individu, serta penerimaan masyarakat menjadi isu krusial yang harus diatasi.

Penerimaan masyarakat terhadap identitas kependudukan digital mempengaruhi keberhasilan dan keberlanjutan sistem ini. Respon masyarakat bisa bervariasi dari penerimaan positif hingga kekhawatiran tentang privasi dan penggunaan data pribadi. Memahami berbagai pendapat, dukungan, dan kritik yang terungkap dalam ulasan dalam jumlah yang banyak tidaklah mudah, oleh karena itu dilakukan analisis sentiment pengguna yang berupa ulasan aplikasi di situs Google Play Store. Analisis sentimen adalah teknik untuk mengukur dan menganalisis opini, sikap, dan perasaan yang dinyatakan dalam teks. Dalam konteks identitas kependudukan digital, analisis sentimen dapat digunakan untuk mengeksplorasi respons masyarakat dari berbagai sumber seperti media sosial, survei online, atau platform diskusi publik lainnya.

Hal ini membantu dalam memahami persepsi masyarakat terhadap sistem ini secara lebih mendalam. Metode analisis sentimen yang relevan dalam konteks aplikasi identitas kependudukan digital mencakup teknik-teknik seperti analisis teks berbasis aturan, *machine learning* (termasuk *Support Vector Machine*), dan pendekatan berbasis *lexicon based*. Kombinasi dari pendekatan ini memungkinkan untuk pengumpulan data yang lebih luas dan pemahaman yang lebih baik tentang sentimen masyarakat terhadap aspek-aspek kunci dari sistem identitas digital.

2. RUANG LINGKUP

Dalam penelitian ini permasalahan mencakup:

2.1 Cakupan permasalahan

Cakupan permasalahan berdasarkan latar belakang pada penelitian ini adalah “Bagaimana analisis sentimen aplikasi Identitas Kependudukan Digital menggunakan metode *support vector machine* dan *lexicon based*?”

2.2 Batasan-batasan penelitian

Agar tidak meluasnya permasalahan maka dibuatlah batasan masalah sebagai berikut :

1. Objek pada penelitian ini adalah aplikasi Identitas Kependudukan Digital.
2. Klasifikasi sentimen pada penelitian ini menggunakan kamus *lexicon* dan hasil klasifikasi dibagi menjadi tiga yakni, sentimen positif, negatif dan netral.
3. Sentimen publik yang diambil dari data ulasan *Google Play Store* terhadap aplikasi Identitas Kependudukan Digital dengan cara *scrapping* menggunakan *python*.
4. Pada penelitian ini menggunakan aplikasi *Google Colab* untuk *scrapping* dan untuk melakukan analisis dan pemrosesan data dengan bahasa *python*.

2.3 Rencana hasil yang didapatkan

Penelitian ini bertujuan untuk menganalisis sentimen dari ulasan di *Google Play Store* dengan memanfaatkan algoritma *Support Vector Machine* dan selanjutnya untuk mengevaluasi kinerja dari *Support Vector Machine* tersebut.

3. BAHAN DAN METODE

Sebagai landasan dilakukannya penelitian ini, maka perlu dikembangkan teori-teori yang berhubungan dengan analisis sentimen, Identitas Kependudukan Digital, *Support Vector Machine*.

3.1 Analisis Sentimen

Analisis sentimen yaitu metode analisis yang digunakan dalam memilih sentimen audiens mengenai reaksi atau pandangannya terhadap suatu kenyataan yang terjadi (Syarifuddin, 2020). Analisis sentimen memiliki proses yaitu mengelompokkan *teks* yang ada di dalam sebuah kalimat lalu menentukan apakah opini atau pendapat tersebut bersifat positif atau negative.

3.2 Identitas Kependudukan Digital

Aplikasi identitas kependudukan digital adalah sebuah aplikasi yang bertujuan untuk memberikan solusi digital dalam mengelola dan mengakses informasi identitas kependudukan.

Aplikasi ini dirancang untuk membantu warga negara dalam berbagai aspek yang terkait dengan identitas mereka, seperti dokumen identifikasi pribadi, data demografis, dan informasi kependudukan lainnya (Wiranti & Frinaldi, 2023)

3.3 Support Vector Machine

SVM (*Support Vector Machine*) adalah algoritma klasifikasi yang memiliki sifat *supervised learning* (pembelajaran) yang bekerja dengan mencari *hyperplane* optimal yang memisahkan jarak antar kelas (Wulandari, 2020)

3.4 Metode

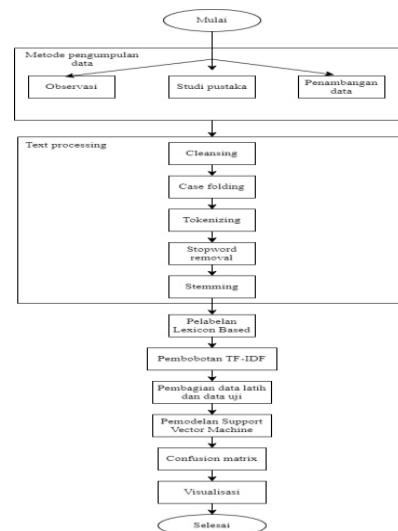
Metode penelitian merupakan sebuah cara yang digunakan untuk mendapatkan dan menganalisa data untuk kegunaan dan tujuan tertentu. Di dalam metode penelitian ini akan ada beberapa hal yang perlu diperhatikan.

3.1 Tempat dan Waktu Penelitian

Tempat penelitian dilakukan di Dinas Kependudukan dan Pencatatan Sipil Kota Samarinda. Penelitian ini dilakukan selama kurang lebih dua bulan yang dimulai pada bulan mei sampai juli 2024.

3.2 Tahapan Penelitian

Diagram Tahapan Penelitian yang dibuat adalah rencana yang akan dilakukan oleh peneliti dalam melakukan analisis sentimen. Pada diagram tahapan penelitian terdapat rincian dalam proses pencarian awal data sampai dengan akhir, pada masing-masing proses yang ditulis pada diagram tahapan akan dijelaskan satu-persatu guna memperjelas diagram yang diolah tersebut. Bisa dilihat pada gambar. Penjelasan rinci ini mencakup semua langkah penting, mulai dari pengumpulan data mentah hingga tahap analisis dan interpretasi hasil, memastikan bahwa setiap tahapan penelitian diuraikan dengan jelas untuk memudahkan pemahaman keseluruhan proses penelitian. Langkah-langkah ini meliputi pengumpulan data, pra-pemrosesan data, metode analisis, dan lain – lain secara terstruktur dalam diagram tersebut.



Gambar 3. 2 Tahapan Penelitian

3.2.1 Pengumpulan Data

Penelitian ini menggunakan data kuantitatif untuk menganalisis data dari teks berbahasa Indonesia berupa ulasan dari Google Play Store. Metode yang digunakan dalam pengumpulan data selama penelitian adalah :

1. Observasi

Observasi dilakukan sebagai studi pendahuluan, peneliti mengamati secara langsung aplikasi Identitas Kependudukan Digital dengan melihat ulasan yang ada pada Google Play Store. Observasi dilakukan peneliti pada bulan Mei 2024. Beberapa hal yang diperoleh peneliti dari observasi yang telah dilakukan adalah memperoleh informasi bahwa terdapat tanggapan-tanggapan yang diunggah pengguna Aplikasi Identitas Kependudukan Digital pada Google Play Store.

2. Studi Pustaka

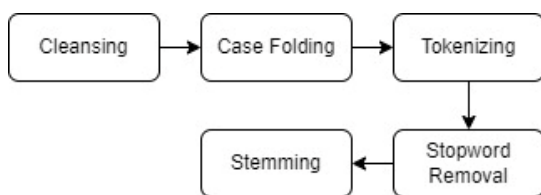
Metode pengumpulan studi pustaka ini dilakukan untuk mencari, mempelajari, dan menggunakan berbagai literatur seperti, buku, jurnal, paper, e-book, atau literatur lain yang berhubungan dengan tema penelitian ini. Informasi yang terkumpul akan dijadikan data pendukung dan pembanding untuk penelitian ini.

3. Penambangan Data

Penambangan data dilakukan menggunakan library Python Scraper. Data review yang akan diambil berupa ulasan yang terdapat pada situs Google Play. Variabel data review yang diambil adalah username pengguna, isi dari ulasan, jumlah bintang ulasan serta tanggal dan waktu ulasan dikirimkan. Data ulasan yang didapat kemudian disimpan dalam Comma Separated Value (CSV).

3.2.2 Pre-Processing Data

Pre-processing merupakan tahapan awal yang akan dilalui dalam memproses teks. Pada penelitian ini akan dilakukan tahapan Pre-processing dengan tahapan case folding, data cleaning, tokenisasi, stopword removal/filtering dan stemming. Alur proses pada tahapan ini ditunjukkan pada Gambar 3.2



Gambar 3. 2 Tahapan Text Processing

1. Case Folding

Case folding adalah proses mengubah semua karakter dalam sebuah dokumen menjadi huruf kecil atau huruf besar yang sama, untuk mempercepat perbandingan selama proses pemrosesan data.

2. Cleansing

Cleansing merupakan tahapan yang bertujuan menghilangkan karakter atau symbol link url (<http://situs.com>), username atau mention(@username),

hashtag(#), retweet, dan emoticon. Selain itu, pada tahap ini juga membersihkan duplikasi data.

3. Tokenizing

Pada proses *tokenizing* dilakukan untuk memisahkan teks menjadi bagian-bagian yang lebih terstruktur sehingga mempermudah proses analisis selanjutnya, seperti penghapusan karakter khusus, menghitung frekuensi kata, atau membangun representasi vektor dari teks.

4. Stopword Removal/Filtering

Tahap *stopword removal* membuang kata-kata tidak bermakna yang tidak mempengaruhi klasifikasi sentimen, seperti kata penghubung dan kata depan, berdasarkan kamus *stopword*.

5. Stemming

Tahap ini bertujuan untuk mengembalikan suatu kata ke dalam bentuk dasarnya. Dengan menghilangkan kata imbuhan awalan, akhiran, sisipan atau pun kombinasi imbuhan. Pada penelitian ini akan digunakan library sastrawi untuk melakukan proses *stemming*. Sastrawi merupakan library pada Bahasa pemrograman python yang dibangun dengan algoritma NA. Algoritma tersebut berdasarkan aturan Bahasa Indonesia.

3.2.3 Pelabelan Dengan Lexicon Based

Setelah Pelaksanaan Tahapan Preprocessing dilaksanakan dan menghasilkan preprocessing dengan variable *Stemming_data* yang merupakan data tidak terstruktur dan tidak memiliki kelas/label, agar dapat digunakan untuk proses mining dengan algoritma SVM, maka perlu dilakukan proses pelabelan data untuk menentukan kata-kata yang merepresentasikan kata positif, negatif dan netral dan menghitung jumlah kata positif dan negatif pada setiap satu kalimat dengan bantuan dari kamus positif dan negatif yang bersumber dari penelitian dan repository. Pelabelan diklasifikasikan kedalam 3 kelas yaitu positif, negatif dan netral.

3.2.4 Pembobotan TF_IDF

Setelah tahap preprocessing berikutnya adalah tahap pembobotan kata. Pembobotan bertujuan untuk memberikan bobot pada fitur kata berdasarkan frekuensi kemunculan kata. TF-IDF (Term Frequency-Inverse Document Frequency) memberi nilai penting pada kata dalam satu dokumen dengan mempertimbangkan seberapa sering kata tersebut muncul di dokumen itu dan seberapa jarang kata itu muncul di seluruh kumpulan dokumen. Rumus pada TF-IDF seperti pada dibawah berikut ini :

$$TF_{(t,d)} = \frac{\text{Number of times } t \text{ appear in document } d}{\text{Total number of terms in document } d}$$

$$IDF_{(t,d)} = \frac{\text{Total number of documents } D}{\text{Number of document the term in it}}$$

$$TFI(t, d D) = TF(t,d) * IDF(t,d) \text{ dimana :}$$

ti = Dokumen ke - i,

df = Document frequency,

n = Banyaknya data

idf = Inverse document frequency

3.2.5 Pemodelan Dengan Metode *Support Vector Machine*

Metode *Support Vector Machine* (SVM) digunakan dalam penelitian ini untuk analisis sentimen. Hasilnya adalah klasifikasi teks menjadi kelas positif, negatif dan netral berdasarkan bobot fitur masing-masing dokumen. Proses ini melibatkan pembobotan fitur dengan SVM, menghasilkan pola yang digunakan dalam pengujian. Evaluasi dilakukan dengan memberi skor pada dokumen, dan analisis keputusan didasarkan pada tingkat akurasi serta jumlah dokumen di kelas positif dan negatif. Untuk memastikan hasil yang akurat, data dibagi menjadi dua bagian, yaitu data training dan data testing, menggunakan rasio 0,1. Ini berarti bahwa 90% dari data digunakan untuk proses training, di mana model SVM dilatih untuk mengenali pola dalam teks, sedangkan 10% sisanya digunakan untuk data testing, yang berfungsi untuk menguji seberapa baik model tersebut bekerja pada data yang belum pernah dilihat sebelumnya. Pembagian ini bertujuan untuk menghindari overfitting dan memastikan bahwa model yang dihasilkan memiliki kemampuan generalisasi yang baik pada data baru.

3.2.6 Confusion Matrix

Setelah hasil pengujian menggunakan Support Vector Machine selesai, akan dilakukan confusion matrix untuk menentukan 3 macam hasil performa machine learning yakni akurasi, presisi, dan recall. Hasil klasifikasi yaitu berupa jumlah sentimen positif dan negatif ulasan pengguna terhadap aplikasi Identitas Kependudukan Digital, sehingga akan diperoleh kesimpulan mengenai tanggapan pengguna terhadap aplikasi Identitas Kependudukan Digital.

3.2.7 Visualisasi

Tahap ini merupakan tahap memvisualisasikan kata-kata yang sering muncul pada data ulasan ke dalam bentuk wordcloud. Hal ini untuk mengetahui kata-kata apa saja yang paling banyak terdapat pada data ulasan dan berapa frekuensi kata tersebut muncul.

4. HASIL DAN PEMBAHASAN

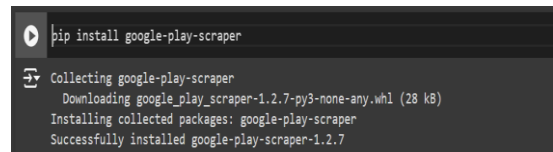
4.1 Hasil Penelitian

Pada penelitian ini dataset yang digunakan ialah data yang berasal dari ulasan Google Play Store, objek penelitiannya sendiri ialah aplikasi Identitas Kependudukan Digital. Adapun tahapan proses pengolahannya hingga menjadi dataset yang nantinya akan digunakan untuk pengujian penelitian ini, berikut tahapan proses pengambilan dataset :

4.2 Pembahasan

4.2.1 Pengumpulan Data

Data pada penelitian ini diambil dari ulasan *Google Play Store* dengan menggunakan bahasa *Python* dan *library Python Scrapper* yang di jalankan di *Google Collaboratory*.



Gambar 4.1 Menginstall google Play Scrapper

Selanjutnya pada gambar 4.2 adalah hasil *scraping* data bentuk *soft file* csv dari jumlah keseluruhan data yang diperoleh berjumlah 1000 data yang berjumlah 4 atribut diantaranya *username*, *score*, *at* dan *content*. Dari 4 atribut yang ditampilkan, data yang akan dilakukan preprocessing adalah *content*. setelah itu proses pembersihan data dan menyeleksi atribut teks.

username	score	at	content
Jagat Mahmood	1	7/8/2024 7:50	ga bisa masuk "perika koneksi mula" padahal data alif...olong di perbaiki.
Amawati	1	7/8/2024 6:34	Saya ingin cetak ulang KTP suami yang rusak, tapi sangat menyulitkan harus melalui verifikasi petugas ketika mendaftar di IKD. Berarti saya harus membusa s
Arohman	1	7/8/2024 6:28	Terbelakang
Dibot Pro 2	1	7/8/2024 5:53	tolol ngapain sih lu msh buat aplikasi, aplikasi lu urus ga becus bobrok, rakyat mula yang sangara ditanya di hack lagi blog
Ibony Kemud	5	7/8/2024 5:12	Bagus
Daryati Daryati	5	7/8/2024 4:41	terima kasih sudah mempermudah perindahan saya dari IKD ke jgga serta data saya sudah benar.
Helmi	1	7/8/2024 3:58	Bintang 1 dulu soalnya daftar online masa masih tetap minta kode QR nya ke dakuipal kan IKDCAK
gerti Laji	1	7/8/2024 2:37	tidak bisa masuk ke apk
Indah Gusardi	5	7/8/2024 2:23	sangat bermanfaat
Yudhi A.W. (A.W.)	3	7/8/2024 1:07	Sewa download, tapi tak bisa di pakai
Wiona Arista	1	7/8/2024 0:56	Gk bisa di buka...Daftar juga gk bisa masuk...Setelah di alihkan koneksi internet nya padahal jaringan lancar
J'hong Misen	5	7/7/2024 22:56	Sudah di buka bro.
Dik andi Huda	5	7/7/2024 18:53	Mantap
Tibon Iskandar	4	7/7/2024 15:38	Berbeli baik
dikha .a	2	7/7/2024 15:13	Tidak efisien dan konyol, masa ia mau daftar di aplikasi harus ttp ke dakuipal untuk dapatkan barcodemnya, banyak aplikasi yg ngasi pinjamn dari cukup dafa
elia ningih	3	7/7/2024 14:40	Setelah instal dan dibuka ada tulisan "terjadi kesalahan koneksi" udah klik "toba lag" berkali kali tetap tdk bisa, setting dipergantiin "on aplikasi" udah dibu
Bejo	3	7/7/2024 12:50	udah daftar tidak bisa di buka
Lupiah GSS	5	7/7/2024 10:43	Mantap
ghifur office	1	7/7/2024 10:36	ga guna ni aplikasi, mau registrasi aja harus di dakuipal nya, ya ngapain di buat digital, ngemotin
Gurawan	2	7/7/2024 6:48	Scan kode Q nya gimana?...daftar lewat aplikasi...tp scan kode Q harus dari petugas dakuipal...gimana di...peruma daftar lewat aplikasi
Almad Arifa	5	7/7/2024 5:33	Good job

Gambar 4.2 Data CSV hasil Scapping

4.2.2 Pre-Processing Data

Sebelum melakukan tahapan *preprocessing* terlebih dahulu import library yang dibutuhkan untuk memanggil dataset yang akan di proses di *google collab*. Setelah library berhasil di import, selanjutnya memanggil dataset yang telah didapat dari proses *scraping*.

1. *Cleansing*

	content	cleansing
0	ga bisa masuk "perika koneksi mula" padahal da...	ga bisa masuk perika koneksi mula padahal da...
1	Saya ingin cetak ulang KTP suami yang rusak, t...	Saya ingin cetak ulang KTP suami yang rusak ta...
2	Terbelakang	Terbelakang
3	tolol ngapain sih lu msh buat aplikasi, apika...	tolol ngapain sih lu msh buat aplikasi apika...
4	Bagus	Bagus
...
995	Gk bisa bisa daftar	Gk bisa bisa daftar
996	Kenapa gak bisa dibuka...muter muter terus pad...	Kenapa gak bisa dibuka muter muter terus padah...
997	Apa mau daftar pindai code Qr apa membunggan	Apa mau daftar pindai code Qr apa membunggan
998	Disuruh daftar aplikasi ikd waktu pembuatan e...	Disuruh daftar aplikasi ikd waktu pembuatan e...
999	Semakin memberikan kemudahan masyarakat dalam ...	Semakin memberikan kemudahan masyarakat dalam ...

Gambar 4.3 Hasil *Cleansing*

Pada gambar 4.3 menunjukkan hasil *cleansing* data pada atribut *teks* untuk menghilangkan atau menghapus bagian yang tidak diperlukan seperti *url*, *hashtag*, *mention*, *emoticon* dan karakter. Pada data yang sudah bersih maka akan melanjutkan ketahap *case folding*.

2. *Case Folding*

	content	cleansing	case_folding
0	ga bisa masuk "perika koneksi mula" padahal da...	ga bisa masuk perika koneksi mula padahal da...	ga bisa masuk perika koneksi mula padahal da...
1	Saya ingin cetak ulang KTP suami yang rusak, t...	Saya ingin cetak ulang KTP suami yang rusak ta...	saya ingin cetak ulang ktp suami yang rusak ta...
2	Terbelakang	Terbelakang	terbelakang
3	tolol ngapain sih lu msh buat aplikasi, apika...	tolol ngapain sih lu msh buat aplikasi apika...	tolol ngapain sih lu msh buat aplikasi apika...
4	Bagus	Bagus	bagus
...
995	Gk bisa bisa daftar	Gk bisa bisa daftar	gk bisa bisa daftar
996	Kenapa gak bisa dibuka...muter muter terus pad...	Kenapa gak bisa dibuka muter muter terus padah...	kenapa gak bisa dibuka muter muter terus padah...
997	Apa mau daftar pindai code Qr apa membunggan	Apa mau daftar pindai code Qr apa membunggan	apa mau daftar pindai code qr apa membunggan
998	Disuruh daftar aplikasi ikd waktu pembuatan e...	Disuruh daftar aplikasi ikd waktu pembuatan e...	disuruh daftar aplikasi ikd waktu pembuatan e...
999	Semakin memberikan kemudahan masyarakat dalam ...	Semakin memberikan kemudahan masyarakat dalam ...	semakin memberikan kemudahan masyarakat dalam ...

Gambar 4.4 Hasil *Case Folding*

Pada gambar 4.4 menunjukkan hasil yang setelah dilakukan *Cleansing* dilanjut dengan *Case Folding* pada atribut teks untuk merubah semua huruf besar menjadi huruf kecil. Contoh kalimat yang ada pada gambar "Gk bisa bisa daftar", menjadi "gk bisa bisa daftar".

3. Tokenizing

	content	cleansing	case_folding	tokenize	filtering/stopword_removal
0	ga bisa masuk "perika koneksi mulu" padahal da...	ga bisa masuk perika koneksi mulu padahal da...	ga bisa masuk perika koneksi mulu padahal data...	[ga, bisa, masuk, perika, koneksi, mulu, data, adef, ...]	[ga, masuk, perika, koneksi, mulu, data, adef, ...]
1	Saya ingin ostak ulang KTP suami yang rusak l...	Saya ingin ostak ulang KTP suami yang rusak l...	saya ingin ostak ulang ktp suami yang rusak la...	[saya, ingin, ostak, ulang, ktp, suami, yang, rusak, menyulikan, ...]	[ostak, ulang, ktp, suami, rusak, menyulikan, ...]
2	Terbelakang	Terbelakang	terbelakang	[terbelakang]	[terbelakang]
3	tolol ngapan sih lu mah buat aplikasi apika...	tolol ngapan sih lu mah buat aplikasi apika...	tolol ngapan sih lu mah buat aplikasi apika...	[tolol, ngapan, sih, lu, mah, buat, aplikasi, apika, ...]	[tolol, ngapan, sih, lu, mah, buat, aplikasi, apika, ...]
4	Bagus	Bagus	bagus	[bagus]	[bagus]
995	Gk bisa bisa daftar	Gk bisa bisa daftar	gk bisa bisa daftar	[gk, bisa, bisa, daftar]	[gk, bisa, bisa, daftar]
996	Kemapa gak bisa dibuka. muler muler muler terus padah...	Kemapa gak bisa dibuka. muler muler muler terus padah...	kemapa gak bisa dibuka muler muler terus padah...	[kemapa, gak, bisa, dibuka, muler, muler, terus, padah, ...]	[gak, dibuka, muler, muler, terus, padah, ...]
997	Apa mau daftar pinda code qr apa membingungkan	Apa mau daftar pinda code qr apa membingungkan	apa mau daftar pinda code qr apa membingungkan	[apa, mau, daftar, pinda, code, qr, apa, membingungkan]	[daftar, pinda, code, qr, membingungkan]
998	Dsuruh daftar aplikasi id waktu perbituan e ...	Dsuruh daftar aplikasi id waktu perbituan e ...	dsuruh daftar aplikasi id waktu perbituan e ...	[dsuruh, daftar, aplikasi, id, waktu, perbituan, e, ...]	[dsuruh, daftar, aplikasi, id, waktu, perbituan, e, ...]

Gambar 4.5 Hasil *Tokenizing*

Pada gambar 4.5 merupakan tahapan *tokenizing* setelah dilakukan *case folding* kalimat "gk bisa bisa daftar" yang jika ditokenisasikan menjadi ['gk', 'bisa', 'bisa', 'daftar'] dari kalimat tersebut telah terbagi menjadi kata token yang terpisah.

4. Stopword Removal

	content	cleansing	case_folding	tokenize	filtering/stopword_removal
0	ga bisa masuk "perika koneksi mulu" padahal da...	ga bisa masuk perika koneksi mulu padahal data...	ga bisa masuk perika koneksi mulu padahal data...	[ga, bisa, masuk, perika, koneksi, mulu, data, adef, ...]	[ga, masuk, perika, koneksi, mulu, data, adef, ...]
1	Saya ingin ostak ulang KTP suami yang rusak l...	Saya ingin ostak ulang KTP suami yang rusak l...	saya ingin ostak ulang ktp suami yang rusak la...	[saya, ingin, ostak, ulang, ktp, suami, yang, rusak, menyulikan, ...]	[ostak, ulang, ktp, suami, rusak, menyulikan, ...]
2	Terbelakang	Terbelakang	terbelakang	[terbelakang]	[terbelakang]
3	tolol ngapan sih lu mah buat aplikasi apika...	tolol ngapan sih lu mah buat aplikasi apika...	tolol ngapan sih lu mah buat aplikasi apika...	[tolol, ngapan, sih, lu, mah, buat, aplikasi, apika, ...]	[tolol, ngapan, sih, lu, mah, buat, aplikasi, apika, ...]
4	Bagus	Bagus	bagus	[bagus]	[bagus]
995	Gk bisa bisa daftar	Gk bisa bisa daftar	gk bisa bisa daftar	[gk, bisa, bisa, daftar]	[gk, daftar]
996	Kemapa gak bisa dibuka. muler muler muler terus padah...	Kemapa gak bisa dibuka. muler muler muler terus padah...	kemapa gak bisa dibuka muler muler terus padah...	[kemapa, gak, bisa, dibuka, muler, muler, terus, padah, ...]	[gak, dibuka, muler, muler, terus, padah, ...]
997	Apa mau daftar pinda code qr apa membingungkan	Apa mau daftar pinda code qr apa membingungkan	apa mau daftar pinda code qr apa membingungkan	[apa, mau, daftar, pinda, code, qr, apa, membingungkan]	[daftar, pinda, code, qr, membingungkan]
998	Dsuruh daftar aplikasi id waktu perbituan e ...	Dsuruh daftar aplikasi id waktu perbituan e ...	dsuruh daftar aplikasi id waktu perbituan e ...	[dsuruh, daftar, aplikasi, id, waktu, perbituan, e, ...]	[dsuruh, daftar, aplikasi, id, waktu, perbituan, e, ...]

Gambar 4.6 Hasil *Stopword Removal*

Pada gambar 4.6 merupakan hasil *stopword* menghilangkan kata yang tidak berguna seperti dan, pada, sebagai dll. Seperti hasil dari kalimat berikut "apa mau daftar pinda code qr apa membingungkan" yang jika dilakukan *stopword removal* menjadi ["daftar", 'pinda', 'code', 'qr', 'membingungkan'] kata yang dihilangkan yaitu 'apa' dan 'mau' karena dianggap tidak terlalu memiliki arti penting dalam proses *text mining* dan kalimat tersebut telah terbagi menjadi kata token yang terpisah.

5. Stemming

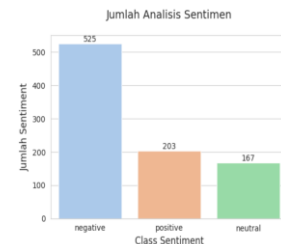
	content	cleansing	case_folding	tokenize	filtering/stopword_removal	stemming_data
0	ga bisa masuk "perika koneksi mulu" padahal da...	ga bisa masuk perika koneksi mulu padahal data...	ga bisa masuk perika koneksi mulu padahal data...	[ga, bisa, masuk, perika, koneksi, mulu, data, adef, ...]	[ga, masuk, perika, koneksi, mulu, data, adef, ...]	ga masuk perika koneksi mulu data adef lolong...
1	Saya ingin ostak ulang KTP suami yang rusak l...	Saya ingin ostak ulang KTP suami yang rusak l...	saya ingin ostak ulang ktp suami yang rusak la...	[saya, ingin, ostak, ulang, ktp, suami, yang, rusak, menyulikan, ...]	[ostak, ulang, ktp, suami, rusak, menyulikan, ...]	ostak ulang ktp suami rusak suit verifikasi l...
2	Terbelakang	Terbelakang	terbelakang	[terbelakang]	[terbelakang]	belakang
3	tolol ngapan sih lu mah buat aplikasi apika...	tolol ngapan sih lu mah buat aplikasi apika...	tolol ngapan sih lu mah buat aplikasi apika...	[tolol, ngapan, sih, lu, mah, buat, aplikasi, apika, ...]	[tolol, ngapan, sih, lu, mah, buat, aplikasi, apika, ...]	tolol ngapan sih lu mah aplikasi aplikasi lu ...
4	Bagus	Bagus	bagus	[bagus]	[bagus]	bagus
5	terima kasih sudah mempermudah perpindahan saya...	terima kasih sudah mempermudah perpindahan saya...	terima kasih sudah mempermudah perpindahan saya...	[terima, kasih, sudah, mempermudah, perpindahan, saya, ...]	[terima, kasih, sudah, mempermudah, perpindahan, ...]	terima kasih sudah pindah jkt jogya data

Gambar 4.7 Hasil *Stemming*

Pada gambar 4.7 menunjukkan proses hasil *stemming* dengan kata imbuhan menjadi kata dasar pada dataset, seperti "terima kasih sudah mempermudah perpindahan saya dari JKT ke jogya serta data saya sudah benar" menjadi "terima kasih mudah pindah jkt jogya data".

4.2.3 Pelabelan Dengan Lexicon

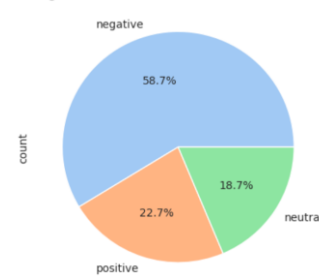
Setelah Pelaksanaan Tahapan Preprocessing dilaksanakan dan menghasilkan preprocessing dengan variable *Stemming_data* yang merupakan data tidak terstruktur dan tidak memiliki kelas/label, agar dapat digunakan untuk proses mining dengan algoritma SVM, maka perlu dilakukan proses pelebelan data untuk menentukan kata-kata yang merepresentasikan kata positif dan negatif dan Menghitung jumlah kata positif dan negatif pada setiap satu kalimat dengan bantuan dari kamus lexicon positif dan negatif yang bersumber dari penelitian dan repository. Pelabelan diklasifikasikan kedalam 3 kelas yaitu positif, negatif dan netral.



Gambar 4.8 Hasil Jumlah Sentimen

Pada gambar 4.8 di atas, ditunjukkan bahwa jumlah sentimen negatif lebih banyak, dengan total 525 dokumen. Sentimen positif berjumlah 203 dokumen, sementara sentimen netral berjumlah 167 dokumen. Hal ini menunjukkan bahwa sentimen negatif lebih dominan dibandingkan dengan sentimen lainnya. Jika dipresentasikan, hasilnya adalah 58.7% negatif, 22.7% positif, dan 18.7% netral, seperti yang ditampilkan pada gambar 4.9 Dominasi sentimen negatif ini mengindikasikan bahwa sebagian besar teks yang dianalisis cenderung memiliki pandangan atau opini yang negative.

Diagram Persentase Hasil Klasifikasi Sentimen



Gambar 4.9 Diagram Hasil Klasifikasi sentimen Dalam Bentuk Persen

4.2.4 Pembobotan TF-IDF

Data ulasan yang telah melalui *preprocessing* teks masih berupa teks, namun ketika menganalisis klasifikasi data harus berbentuk numerik atau angka. Data harus terlebih dahulu dikonversi ke bentuk numerik dengan membobot kata dengan menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) yang menggunakan perhitungan modul *scikit-learn* yaitu *TfidfVectorizer*. Dimana hasil dari tahap TF-IDF yaitu dari 895 dataset setelah dilakukan pembobotan kata dan kemunculan kata berjumlah 1686 kata. Jumlah kata berasal dari *vocabulary* yang telah dipelajari oleh *CountVectorizer*.

```
[ ] cv = CountVectorizer()
X = cv.fit_transform(stemming_data)

print(len(cv.vocabulary_))

1686

[ ] cv.vocabulary_

{'ga': 504,
'masuk': 964,
'perika': 1205,
'koneksi': 812,
'mulu': 1025,
'data': 354,
'aktif': 57,
'tolong': 1565,
'baik': 147,
'cetak': 318,
'ulang': 1603,
'ktp': 836,
'suami': 1469,
'rusak': 1318,
'sulit': 1473,
'verifikasi': 1635,
```

Gambar 4.10 kamus (*dictionary*) yang menghubungkan kata-kata (*terms*) dengan indeks uniknya

Pada gambar 4.10 menjelaskan kamus ini digunakan dalam *CountVectorizer* atau *TfidfVectorizer* dari library *scikit-learn*, yang mengubah koleksi dokumen *teks* menjadi matriks istilah-dokumen. Setiap kata atau *term* dalam *teks* diberikan indeks unik untuk mengidentifikasi kata tersebut di dalam vektor atau matriks. Misalnya, kata "ga" memiliki indeks 504, "masuk" memiliki indeks 964 dan seterusnya.

```
[ ] #contoh memanggil salah satu kata pada kolom 1
stemming_data_tf = stemming_data[892]
print(stemming_data_tf)

['daftar', 'pindai', 'code', 'qr', 'bingung']

[ ] #hasil tf dalam dokumen 1
term_frequency = cv.transform([stemming_data_tf])
print(term_frequency)

(0, 222) 1
(0, 326) 1
(0, 341) 1
(0, 1216) 1
(0, 1270) 1
```

Gambar 4.11 Hasil TF

Pada gambar 4.11 merupakan hasil TF dari kalimat "daftar pindai code qr bingung", terdapat ada 5 kata dengan masing-masing nomor indeksnya. Contoh jika dalam hasil yang diberikan (0, 222) 1, ini berarti kata dengan indeks 222 dalam vocabulary *CountVectorizer* muncul satu kali di dokumen pertama dan begitupun seterusnya.

word	Frequency
daftar	87
pindai	2
code	47
qr	43
bingung	7

Gambar 4.12 Hasil *Frequency* kata

Gambar 4.12 di atas merupakan tabel "*Term Frequency*" yang menunjukkan frekuensi kemunculan kata-kata tertentu dalam sebuah korpus atau dataset teks.

Dl	Tf	Idf	Tf-Idf
daftar	0,2	1,012303781	0,202460756
pindai	0,2	2,65079303	0,530158606
code	0,2	1,279725177	0,25594504
qr	0,2	1,318354579	0,2636709158
bingung	0,2	2,1067249953	0,42134499906

Gambar 4. 13 Hasil TF-IDF

Pada gambar 4.13 menunjukkan hasil nilai yang diperoleh dari perhitungan *tf-idf* yang menggunakan *tools google collab*. Kata "daftar" memiliki nilai *tfidf* 0.202460756 dan seterusnya menunjukkan pasangan indeks term dan nilai TF-IDFnya yang hasilnya dikalikan. Nilai TF-IDF memberikan bobot pada setiap *term* berdasarkan frekuensi kemunculannya dalam dokumen dan seberapa jarang *term* tersebut muncul di dokumen lain.

4.2.5 Pemodelan Dengan Metode SVM

Setelah melakukan pembobotan TF-IDF, selanjutnya tahapan pembagian data dilakukan untuk mengevaluasi performa model dari SVM, yang dibagi menjadi data *training* dan data *testing* dimana didalam penelitian ini menggunakan pembagian data 90:10. Hasil dari Nilai X train : (805,1668), Nilai X Test : (90,1668) Nilai Y train (805) dan nilai dari Y Test (90).

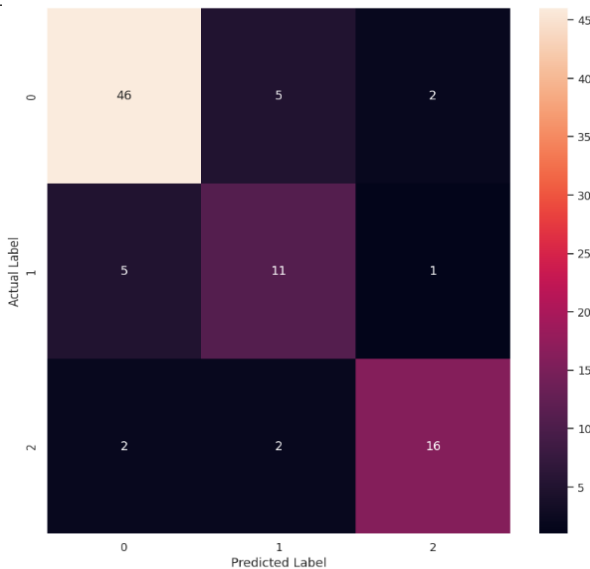
	precision	recall	f1-score	support
negative	0.87	0.87	0.87	53
neutral	0.61	0.65	0.63	17
positive	0.84	0.80	0.82	20
accuracy			0.81	90
macro avg	0.77	0.77	0.77	90
weighted avg	0.81	0.81	0.81	90

Gambar 4.14 Hasil dari pemodelan algoritma SVM

Pada gambar 4.14 diatas mengindikasikan bahwa nilai akurasi dengan pembagian 90% data training dan 10% data testing adalah 81%. Pada kelas "negative" memiliki *precision* 0.87, *recall* 0.87, dan *F1-score* 0.87. Kelas "neutral" memiliki *precision* 0.61, *recall* 0.65, dan *F1-score* 0.63, sedangkan kelas "positive" memiliki *precision* 0.84, *recall* 0.80, dan *F1-score* 0.82.

4.2.6 Confusion Matrix

Setelah didapatkannya hasil klasifikasi sentimen menggunakan SVM perlu dilakukan pengujian menggunakan *confusion matrix* untuk mengetahui informasi yang diprediksi dan hasil yang sebenarnya dan yang digunakan adalah data testing yaitu 90 data testing. Perhitungan dari *confusion matrix* adalah *accuracy*, *recall*, dan *precision*. Berikut merupakan hasil pengujian menggunakan *confusion matrix*.



Gambar 4.15 Hasil Confusion Matrix

Dari Gambar 4.15 dapat dijelaskan bahwa 46 data sebagai negatif, 11 data sebagai data netral dan 16 data sebagai positif. Pada perhitungan manual akurasi, presisi dari matriks diatas sebagai berikut. Pada perhitungan manual akurasi dari matriks diatas sebagai berikut:

$$Accuracy = \frac{\text{Jumlah prediksi benar}}{\text{Jumlah Prediksi}} \times 100\%$$

$$Accuracy = \frac{46+5+2+5+11+1+2+2+16}{73} \times 100\%$$

$$Accuracy = \frac{90}{90} \times 100\%$$

$$Accuracy = 0.811 \times 100\% = 81\%$$

4.2.7 Visualisasi



Gambar 4.16 WordCloud ulasan negative

Pada gambar 4.16 merupakan hasil visualisasi dari ulasan *negative* terlihat bahwa frekuensi kemunculan sebuah kata meningkat seiring dengan ukurannya di dalam *word cloud*. Menunjukkan bahwa pada kelas sentimen *negative* kata yang sering muncul adalah aplikasi, ribet, daftar, susah, masuk, data, aja, *online*, dukcapil dan *download*.



Gambar 4.17 WordCloud ulasan positif

Pada gambar 4.17 menunjukkan bahwa pada kelas sentimen positif kata yang sering muncul adalah mudah, aplikasi, mantap, bagus, baik, ktp dan dukcapil.



Gambar 4.18 WordCloud ulasan netral

Pada gambar 4.18 menunjukkan bahwa pada kelas sentimen netral kata yang sering muncul adalah buka, *online*, aplikasi, apk, login, kantor dan hp.

Hasil dari penelitian ini menggunakan data sebanyak 1000 yang sudah diolah dari mengecek data duplikat sebanyak 105, menghapus data komentar yang kosong, dilakukan pre-processing sehingga menjadi 895. Kemudian dilakukan pelabelan dengan hasil negative 525 data, positive 203 data, dan neutral 167. Kemudian dilakukan TF-IDF yang memiliki 1686 kata unik, dilanjutkan melakukan pemodelan SVM yang memperoleh hasil accuracy 81% dengan presisi dan recall untuk sentimen negatif masing-masing 87%, untuk sentimen netral 61% dan 65%, serta untuk sentimen positif 84% dan 80%. Ini menunjukkan bahwa model SVM yang dikembangkan mampu mengklasifikasikan sentimen dengan akurasi yang baik.

5. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan maka diperoleh beberapa kesimpulan sebagai berikut:

1. Dari 1000 sampel ulasan dan melalui proses *preprocessing*, didapatkan 895 gambaran data ulasan Identitas Kependudukan Digital yang digunakan dan dibagi menjadi tiga kelas yaitu data kelas positif sebanyak 203 ulasan, kelas negatif sebanyak 525 ulasan dan kelas netral sebanyak 167 ulasan.
2. Hasil penerapan *Support Vector Machine* dengan kernel linear pada klasifikasi ulasan Identitas Kependudukan Digital mendapatkan akurasi pada pembagian data latihan & data uji 90:10 sebesar 81%, dengan presisi dan recall untuk sentimen negatif

masing-masing 87%, untuk sentimen netral 61% dan 65%, serta untuk sentimen positif 84% dan 80%.

3. Dari klasifikasi dan visualisasi didapatkan pengetahuan berupa kata yang sering muncul di ulasan positif seperti “mudah”, “aplikasi”, “mantap” dan “baik”, sementara di ulasan negatif terdapat kata “aplikasi”, “daftar”, “susah”, “ribet”. Beberapa penilaian positif dari pengguna di antaranya tentang mudah, aplikasi, mantap dan baik yang dapat diartikan sebagai kemudahan dalam memuat identitas pribadi dan anggota keluarga. Sedangkan penilaian negatif dari pengguna di antaranya adalah pendaftaran yang susah dikarenakan memindai Code QR dan verifikasi wajah harus secara *offline*.

6. SARAN

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran diantaranya sebagai berikut:

1. Pada penelitian ini, *emoticon*, bahasa asing, bahasa daerah diabaikan, pada penelitian selanjutnya dapat diharapkan dapat memasukan *emoticon*, bahasa asing dan bahasa daerah ke dalam perhitungan agar menghasilkan nilai yang lebih akurat.
2. Pada penelitian selanjutnya dapat menggunakan metode klasifikasi lainnya yang belum digunakan pada penelitian ini.
3. Dalam penelitian ini menggunakan data ulasan sebanyak 1000 data. Untuk penelitian selanjutnya, dapat meningkatkan kualitas dan jumlah sampel data yang terbaru.

7. DAFTAR PUSTAKA

- Syarifuddin, M. (2020). Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn. *INTI Nusa Mandiri*, 15(1), 23–28. <https://doi.org/10.33480/inti.v15i1.1347>
- Wiranti, N. E., & Frinaldi, A. (2023). Meningkatkan Efisiensi Pelayanan Publik dengan Teknologi di Era Digital. *JIM: Jurnal Ilmiah Mahasiswa Pendidikan Sejarah*, 8(2), 748–754. <https://doi.org/10.24815/jimps.v8i2.24833>
- Wulandari, A. (2020). Aplikasi Support Vector Machine (SVM) untuk Pencarian Binding Site Protein-Ligan. *MATHunesa: Jurnal Ilmiah Matematika*, 8(2), 157-161. <https://doi.org/10.26740/mathunesa.v8n2.p157-161>

