Analisis Sentiment Terhadap Penggunaan Aplikasi Peduli Lindungi pada Media Sosial Twitter dengan Menggunakan Algoritma Naive Bayes dan SVM

Salsabila Oktafani[1], Alvin Putra Perdana[2], Jihan Kamilah[3], Nurhikmah Mawaddah Solin[4]

Program Studi Informatika, Universitas Pembangunan Nasional Veteran Jakarta

[1]2010511001@mahasiswa.upnvj.ac.id
[2]2010511011@mahasiswa.upnvj.ac.id
[3]2010511013@mahasiswa.upnvj.ac.id
[4]2010511026@mahasiswa.upnvj.ac.id

***Abstrak*—PeduliLindungi menjadi aplikasi yang sering digunakan era saat ini. Tujuan dari penggunaan aplikasi PeduliLindungi sebagai bentuk penanganan penyebaran Virus Corona (COVID-19) serta penyakit menular lainnya. Algoritma yang digunakan dalam penelitian analisis sentimen ini yaitu Naive Bayes dan SVM. Dalam jangka waktu 24 Mei 2022 - 01 Juni 2022, didapatkan data sejumlah 985 sebelum dilakukannya *preprocessing*. Setelah *preprocessing* dilakukan, didapatkan data bersih sejumlah 704 data. Dengan menggunakan tiga kategori, yaitu negatif, netral, dan positif. Setelah diterapkannya *Oversampling* menggunakan SMOTE untuk mengatasi imbalance data, didapatkan hasil akurasi sebesar 0.881 untuk model *Naive Bayes Classifier*, dan untuk model SVM didapatkan akurasi sebesar 0.954. Peneliti juga menyimpulkan pengguna twitter lebih sering memberikan sentimen atau tanggapan yang bersifat netral terkait aplikasi PeduliLindungi.**

Kata Kunci—Sentimen analisis, PeduliLindungi, COVID-19, Naive Bayes, SVM.

Abstract—PeduliLindung is an application that is often used in today's era. The purpose of using the PeduliLindung application is as a form of handling the spread of the Corona Virus (COVID-19) and other infectious diseases. The algorithms used in this sentiment analysis research are Naive Bayes and SVM. In the period of 24 May 2022 - 01 June 2022, 985 data were obtained before preprocessing was carried out. After preprocessing is done, the net data obtained is 704 data. By using three categories, namely negative, neutral, and positive. After the implementation of oversampling using SMOTE to overcome the imbalance data, the results obtained accuracy of 0.875 for the Naive Bayes Classifier model, and for the SVM model an accuracy of 0.957 was obtained. The researcher also concludes that Twitter users more often provide neutral sentiments or responses regarding the PeduliLindung application.

Keywords—Sentiment analysis, PeduliLindungi, COVID-19, Naive Bayes, SVM.

I. PENDAHULUAN

PeduliLindungi, ialah aplikasi dan situs web yang digunakan oleh masyarakat dalam guna menangani penyebaran Virus Corona (COVID-19), serta penyakit menular lainnya. Era saat ini, aplikasi seperti PeduliLindungi dibutuhkan, sehingga dapat melakukan tindakan pemutusan rantai penyebaran. Setiap tempat kunjungan, mempersiapkan code QR untuk di-*scan* oleh pengunjung sebelum memasuki area tersebut. Dengan demikian, masyarakat yang berada di area tersebut merasakan aman, disebabkan tersedianya *tracking user* yang sedang/terdampak Covid-1, mauapun yang tidak. Jadi, dibutuhkannya data/informasi pribadi *user* guna membuat proses penggunaan PeduliLindungi lancar. Tak perlu ditakutkan, sudah adanya perizinan, serta ketentuan peraturan perundang-undangan yang berlaku untuk menjamin data *user*. Aplikasi PeduliLindungi merupakan perangkat lunak yang legal untuk digunakan. PeduliLindungi menggunakan infrastruktur milik Pemerintah RI dalam melakukan penyimpanan data pribadi masyarakat.

Twitter merupakan sosial media dimana user bisa mengirim postingan teks, foto, video, ataupun link yang biasa disebut dengan "*Tweet*". Semakin tahun, pengguna twitter ikut bertambah. Dengan begitu, semakin banyak pula postingan *tweet* yang dihasilkan. Twitter juga digunakan oleh user untuk menuangkan sesuatu, bisa merupakan komentar ataupun ulasan mengenai suatu hal. Bisa mengomentari dari sisi positif ataupun sebaliknya. Setiap harinya, berbagai sentimen akan terus bertambah. Dimana pada sentimen tersebut, opini dikenali dan dilakukan ekstraksi dalam bentuk teks. Yang tadinya informasi tidak terstruktur bisa berubah menjadi terstruktur sehingga didapatkannya maksud dari data/informasi tersebut, dan hal inilah yang dikenal dengan analisis sentimen. Dalam penelitian ini, digunakannya algoritma *Naïve Bayes Classifer* dengan metode MultinomialB dan SVM (*Support Vector Machines*) diharapkan untuk dapat melakukan analisis sentimen mengenai penggunaan aplikasi PeduliLindungi, dengan bantuan mendapatkan data dari Twitter API. Untuk tahap pengukuran nilai evaluasi, akan digunakannya akurasi, *recall, precision,* dan *juga f1-score*. Dari nilai evaluasi tersebutlah yang nantinya akan digunakan sebagai penentu, algoritma manakah yang lebih baik diantara Naïve Bayes, dengan SVM.

II. STUDI PUSTAKA

A. Penelitian Terkait

Penelitian [1] pada tahun 2018 melakukan pembangunan aplikasi analisis sentimen Twitter menggunakan web *scraping* dari november 2012 sampai januari 2017 guna pengklasifikasian teks terkait Sistem Administrasi Manunggal Satu Atap (SAMSAT) di Malang Kota dengan menggunakan algoritma Naive Bayes. Serta, beberapa tahapan yang dilakukan dalam penelitian ini adalah tahap praproses (*case folding, cleaning, tokenizing, dan stopword*), dan klasifikasi. Kelas label dibagi menjadi tiga yakni netral, negatif, dan positif. Hasil yang didapatkan ialah tingkat

akurasi pada percobaan pertama sebesar positif 81%, negatif 89%, dan netral 80%,. Sedangkan, pada percobaan kedua akurasi yang didapatkan yaitu positif 82%, negatif 92%, dan netral 80%.

Penelitian [2] pada tahun 2018 melakukan analisis sentimen pada media sosial Instagram berkaitan dengan klasifikasi *cyberbullying* menggunakan algoritma SVM (*Support Vector Machine*). Pada penelitian ini data pengkategorian dibagi menjadi 200 komentar positif *cyberbullying*, dan 200 komentar negatif *cyberbullying*. Perbandingan yang terapkan pun 70% untuk data *training* : 30% untuk data *testing*. Tahapan yang dilakukan dalam penelitian ini berupa praproses, pembobotan TF-IDF, dan klasifikasi. Hasil yang didapatkan ialah tingkat akurasi sebesar 90%, *precision* 94,44%, 85% *recall*, dan *fmeasure* 89,47%.

Penelitian [3] pada tahun 2020 melakukan analisis sentimen pada Twitter guna pengkalsifikasian kecenderungan user gojek menggunakan algoritma SVM (*Support Vector Machine*) dari tanggal 12-18 Januari 2019. Untuk sentimen yang dipilih pun hanya 1.500 *tweets* saja, dan akan diklasifikasi menjadi dua jenis yaitu positif dan negatif. Serta, beberapa tahapan yang dilakukan dalam penelitian ini adalah tahap ekstraksi *tweet*, praproses data, pelabelan data secara manual dan dengan *sentiment scoring*, *feature selection*, pembobotan TF-IDF, pembangunan model SVM, perhitungan akurasi, dan penentuan kernel terbaik. Hasil yang didapatkan ialah tingkat akurasi terbaik dari pelabelan manual sebesar 79,19% dan akurasi kappa terbaiknya 16,52% , serta untuk pelabelan *sentiment scoring* memiliki akurasi terbaik sebesar 79,19% dan kappa nya 21%.

Penelitian [4] pada tahun 2020 melakukan analisis sentimen terkait ulasan aplikasi mobile pada Google Play Store menggunakan algoritma Naive Bayes dan C4.5 dengan berbasiskan normalisasi kata menggunakan *Levenshtein Distance*, dan data didapatkan menggunakan extension *web scraper* pada Chrome. Pada penelitian ini terdapat tahapan penginputan *K edit distance*, *split* data latih menjadi data teks dan *rating*, teks praproses, *Levenshtein Distance*, pengklasifikasian Naive bayes dan C4.5, dan pengujian. Hasil akurasi rata-rata yang diperoleh yaitu 85,3% , dan untuk akurasi tertingginya 87,1%.

Penelitian [5] pada tahun 2020 melakukan analisis sentimen terkait komentar yang berhubungan dengan kesehatan mental pada masa pandemi Covid-19 dengan menggunakan algoritma *Naive Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM). Data yang didapatkan sebanyak 498 data tweet pada bulan Agustus 2020. Pada penelitian ini terdapat beberapa tahapan yaitu crawling data, teks praproses, melakuakn analisis sentimen, dibagianya data training dan data testing, klasifikasi, dan interprestasi. Hasil yang didapatkan ialah tingkat akurasi sebesar 80,81% dengan menggunakan SVM kernel Polinomial, 78,79% dan 71,73% dengan kernel RBF dan Linier, serta 70,71% jika menggunakan NBC.

B. Tinjauan Pustaka

1. *Information Retrieval* (Temu Kembali Informasi)

   Temu kembali informasi yakni ilmu pencarian data/informasi dokumen, metadata yang menjelaskan dokumen, ataupun mencarinya dalam *database*. Contohnya, ada tempat yang menyimpan banyak dokumen, dan *user* melakukan perumusan atas sebuah pertanyaan (*request* ataupun *query*), untuk jawabannya sendiri ialah kumpulan dokumen yang mempunyai informasi yang diperlukan dan didapatkannya jawaban itu didasari pertanyaan *user*.

2. Peduli Lindungi

   PeduliLindungi merupakan salah satu aplikasi yang dikembangkan untuk menghentikan penularan *Corona virus Disease* (COVID-19). Kegunaan PeduliLindungi yaitu terdapatnya pengidentifikasian status riwayat/histori kontak fisik terhadap orang yang positif atau dicurigai positif Covid-19, dan informasi Covid-19 user itu sendiri. Hingga, menampilkan status vaksinasi *user* meng-*scan QR* guna *Check-In/Out* ke suatu tempat. Semua fungsi tersebut, terdapat di satu aplikasi, yaitu PeduliLindungi. Tujuannya guna mengarapkan kepedulian, serta partisipasi masyarakat dalam saling menjaga, dan meminimalisir resiko penyebaran Covid-19.

3. Twitter

   Twitter didirikan pada Maret 2006 oleh Jack Dorsey. Twitter merupakan salah satu media sosial yang digunakan untuk berkomunikasi dimana pengguna dapat mengirim dan membaca pesan dengan maksimal 140 karakter, yang saat ini kita kenal sebagai *tweet*. Twitter juga biasanya digunakan oleh masyarakat untuk mengirimkan opini singkat melalui tweet mengenai suatu produk ataupun layanan.

4. *Crawling Data*

   Crawling Data yaitu teknik yang dilakukan untuk mengumpulkan suatu data informasi yang ada didalam web secara otomatis berdasarkan kata kunci yang diberikan oleh pengguna.

5. *Text Preprocessing*

   Text Preprocessing atau Praproses Teks adalah tahapan awal dimana data yang berupa teks akan diterapkan berbagai persiapan, seperti ekstraksi data penting pada teks agar lebih rapi, dan terstruktur. Contoh proses dalam praproses teks ialah *case holding*, *Tokenizing*, *Filtering*, *dan Stemming*. [6] Karena pada nyatanya, teks dapat berisi huruf, angka, ataupun simbol. Namun, umumnya yang dibutuhkan hanyalah beberapa kata penting yang terdiri dari kumpulan huruf saja, sehingga diperlukannya tahap praproses.

6. *Stopword Removal*

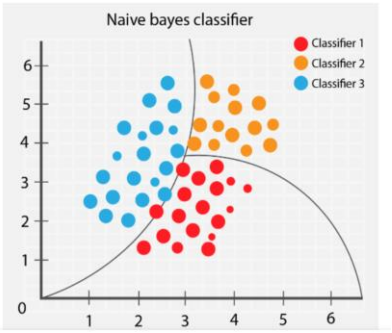
   Stopword Removal termasuk dalam tahapan *filtering*. Proses ini, dimaksudkan untuk pemilihan kosa kata penting dari hasil tokenisasi, yang mana merupakan kata yang digunakan untuk mewakili dokumen. Jika, tidak merepresentasikan dokumen, atau merupakan kata hubung dan kata yang umum maka akan dilakukannya penghapusan kata tersebut.

7. Klasifikasi

   Klasifikasi merupakan suatu pengelompokan data yang memiliki label. Klasifikasi, merupakan

suatu proses guna menemukan suatu model yang bisa membedakan antara kelas data atau konsep. Adapun metode klasifikasi pada machine learning seperti *Support Vector Machines* (SVM), Naive Bayes, dan lainnya.

8. Naive Bayes

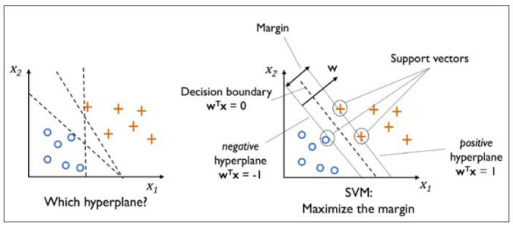


Gambar. 1 Ilustrasi Klasifikasi Naive Bayes

Naive Bayes menjadi salah satu contoh metode dari klasifikasi. Naive Bayes menggunakan teori probabilitas dalam melakukan klasifikasi data. Contoh kasus dalam Naive Bayes ini yaitu ketika melakukan *filtering email spam*. Algoritma Naive Bayes pada penelitian ini menggunakan kategori negatif, netral, dan positif. Berikut rumus terkait Naive Bayes.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

9. SVM (*Support Vector Machine*)



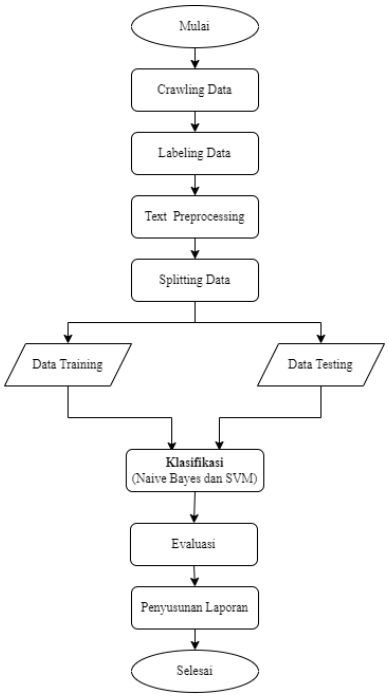
Gambar. 2 Ilustrasi Klasifikasi SVM

SVM juga menjadi contoh metode dari klasifikasi *machine learning*. SVM adalah algoritma pembelajaran mesin yang menerapkan pendekatan *supervised learning*. SVM akan mengelompokkan data dengan melakukan pemisahan dengan didasarkan *hyperplane*-nya. Contoh dari algoritma SVM seperti diaplikasikan ketika melakukan analisis sentimen.

III. METODE PENELITIAN

A. *Kerangka Berpikir*

Berdasarkan penelitian terkait, tinjauan penelitian, dan permasalahan yang telah dijelaskan. Kerangka berpikir dari penelitian digambarkan melalui bagan dibawah ini:



Gambar. 3 Kerangka Berpikir

B. *Instrumen Penelitian*

Berikut alat bahan perangkat yang digunakan untuk melakukan penelitian dalam melakukan pengklasifikasian *sentiment*:

1. Bahan

   Dalam penelitian ini, bahan yang digunakan adalah tanggapan masyarakat pengguna twitter dari postingan *tweet* guna untuk melakukan analisis sentimen.

2. Perangkat yang digunakan dan dibutuhkan dalam penelitian ini diantaranya:
   a. Perangkat lunak
      - Google Colab

        Perangkat lunak Google Colab dipakai untuk menginput dan menjalankan kode program tanpa instalasi teks editor.

      - Spreadsheet

        Untuk input label manual bersama.

      - Google Document

        Digunakan sebagai tempat mengetik laporan secara bersama dalam satu waktu.

      - Microsoft Office Word

        Ms. Word sebagai tempat untuk melakukan finishing pada laporan.

   b. Perangkat keras
      - Laptop dengan spesifikasi RAM 4GB-8GB
      - *Processor i3-i7 core*
      - *Mouse*

C. *Pengumpulan Data*

Pada penelitian ini, pengumpulan data dilaksanakan menggunakan beberapa cara yaitu :

1. Studi Pustaka

Pada tahapan ini dilakukan guna mendapatkan penjelasan terkait teori-teori yang berhubungan dengan penelitian.

2. *Crawling* Data pada Twitter API

Pada penelitian ini dilakukannya dua kali *crawling* data menggunakan Twitter API dengan *keyword* "Peduli Lindungi". Pertama kali diambil pada tanggal 24 Mei 2022 – 31 Mei 2022, dan didapatkannya data sebanyak 984. Namun, data awal tersebut masih terdapat duplikasi sehingga dilakukannya penghapusan tweet yang sama sebelum praproses teks dilakukan sehingga total data didapatkan sebanyak 693. Untuk crawling data yang kedua, dilakukan pada tanggal 24 Mei 2022 - 01 Juni 2022. Dengan total data bersih yang didapat yaitu 797.

D. Teknik Analisis Data

1. Analisis sentimen

Merupakan tahapan proses melakukan identifikasi dan melakukan pengelompokkan suatu opini dari suatu data, dimana kategori yang digunakan bisa negatif, netral, atau positif. Karena banyaknya data yang dikeluarkan pada sosial media mengenai suatu opini dari produk atau layanan, maka analisis menjadi peran mengelola data tersebut. Dengan melakukan analisis sentimen, informasi yang tidak terstruktur akan menjadi data yang terstruktur.

Adapun hal pertama yang dilakukan, yaitu menarik data dengan menghubungkan twitter API menggunakan python dengan bantuan dari library *tweepy*. Dibutuhkan *api_key*, *api_secret_key*, *acceess_token*, dan *access_token_secret* agar bisa terhubung dengan twitter API. Pada penelitian ini menggunakan keyword "peduli lindungi". Didapat data bersih sejumlah 797 dengan melakukan dua kali *crawling*.

2. Preprocessing

Hal-hal yang dilakukan pada tahap preprocessing sebagai berikut:

a. Cleaning

Tahapan ini dilakukan guna untuk melakukan *replace* menjadi spasi, menghapus *mention*, simbol, emoji, *hyperlink*, angka, *retweet*-an, dan *hastag*.

b. Case Folding

Yaitu, melakukan perubahan dimana sebelumnya huruf kapital akan berubah menjadi non kapital/kecil.

c. Tokenizing

Merupakan tahap melakukan pemotongan string input berdasarkan tiap kata yang menyusunnya.

d. Filtering

Pada tahap *filtering* satu ini, dilakukan dengan melakukan pengambilan kata penting pada token yang dihasilkan pada tahap sebelumnya. Dengan membuang atau menghapus *stopword* yang ada, akan membantu pengumpulan kata penting. Serta, menerapkan *stemming* untuk menghilangkan kata himbuan.

e. Filtering duplicate tweet

Tweet yang sama akan dihapus, sehingga mengurangi adanya *duplicate tweet* yang ada pada isi data.

IV. HASIL PENELITIAN DAN PEMBAHASAN

A. Data yang Digunakan

Seperti yang telah dijelaskan sebelumnya, bahwa peneliti melakukan *crawling* data sebanyak dua kali dalam rentang waktu yang berdekatan. Selama *crawling* data pun, peneliti melakukan *filtering duplicate tweet* guna mempermudah *labeling*. Namun, karena masih banyaknya *noise* pada teks maka terdapat kemungkinan masih banyaknya *duplicate* data di dalamnya.

TABEL I
DATA CRAWLING

No	Username	Date	Tweet
1.	Qshareina	5/27/2022 11:21:00 PM	@t3luuur @mpraldo @gratisterbaik @human_b bikin ribet urusan aja aplikasi peduli lindungi, apa coba fungsinya.
2.	ruuubiii7	5/29/2022 11:20:00 AM	@andrekelv @whenscarves @chiw mending ke peduli lindungi sih
3.	itsmejojo98	5/28/2022 11:12:00 AM	@vyantraa @txtdriormas qr codenya pasti buat scan peduli lindungi
4.	fluffy_hehe	5/30/2022 6:37:00 AM	Ini ad tante" bilang peduli lindungi jdi peduli lingkungan, aku tak sanggup menahan ktwa😭😭 https://t.co/Cm5JWJAR42
5.	Fyeahfifi	5/29/2022 4:48:00 PM	@fahmihasannn kang temen saya belum vaksin booster, terus iseng cek peduli lindungi, ternyata sertifikatnya udah adâ€¦ https://t.co/cwlbwnljzh

B. Labeling Data

Tahap melakukan pelabelan menggunakan 3 kategori, yaitu negatif, netral, dan positif. Data yang dilakukan pelabelan yakni data yang diambil dari *tweet* sebelumnya. Pelabelan dilakukan secara manual oleh setiap anggota.

Setelah melakukan pelabelan, dilakukan perhitungan fleiss kappa, dengan rumus dari fleiss kappa sebagai berikut:

$$K = \frac{\bar{p}-\bar{p}e}{1-\bar{p}e} \quad (2)$$

$$p_j = \frac{1}{mn}\sum_{i=1}^{n} x_{ij} \quad (3)$$

$$\bar{p} = \frac{1}{mn(m-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{k} {x_{ij}}^2 - mn\right) \quad (4)$$

$$\bar{p}e = \sum_{j}^{k} = 1{p_j}^2 \quad (5)$$

Keterangan:

K = Kappa Fleiss.

$\bar{p}$ = Tingkat kesepakatan keseluruhan.

$\bar{p}\varepsilon$ = *Error*.

p_j = proporsi yang termasuk dalam kategori j.

m = jumlah pelabel.

n = jumlah *data tweet*.

i = indeks perepresentasi total *data tweet* $(1,2,.,n)$.

j = indeks perepresentasi total *data tweet* $(1,2,.,k)$.

k = jumlah kelas.

x_{ij} = jumlah nilai data tweet ke-i dan label ke j.

Berikut merupakan interpretasi dari nilai kappa yang telah dilakukan perhitungan.

TABEL II
INTERPRETASI KAPPA

Kappa	Interpretasi
Null = Hypothesis Kappa = 0	Agreement is due to chance
0.01-0.02	Slight agreement
0.21-0.40	Fair Agreement
0.41-0.60	Moderate Agreement
0.61-0.80	Substantial Agreement
0.81-1.00	Almost Perfect Agreement
Negative (Kappa<0)	Agreement less than that expected by chance

Adapun yang diperoleh dari perhitungan kappa fleiss termasuk dalam interpretasi *Moderate Agreement*. Nilai kappa yang dihasilkan berada pada interpretasi pertengahan, yaitu 0,509.

Berikut contoh label akhir dari hasil kappa persetujuan yang dilakukan pada penelitian ini:

TABEL III
LABELING DATA

No	Tweet	Label
1.	@t3luuur @mpraldo @gratisterbaik @human_b bikin ribet urusan aja aplikasi peduli lindungi, apa coba fungsinya.	Negatif
2.	@andrekelv @whenscarves @chiw mending ke peduli lindungi sih	Positif
3.	@vyantraa @txtdriormas qr codenya pasti buat scan peduli lindungi	Netral
4.	Ini ad tante" bilang peduli lindungi jdi peduli lingkungan, aku tak sanggup menahan ktwa😭😭 https://t.co/Cm5JWJAR42	Netral
5.	@fahmihasannn kang temen saya belum vaksin booster, terus iseng cek peduli lindungi, ternyata sertifikatnya udah adâ€¦ https://t.co/cwlbwnljzh	Negatif

C. Teks Praproses

Teks praproses dilakukan dengan beberapa tahap, seperti *cleaning*, *case folding*, *tokenizing*, *dan filtering*. Berikut merupakan contoh dari melakukan tahapan *cleaning* data pada penelitian ini:

TABEL IV
CLEANING DATA

No	Awal	Hasil
1.	Ini ad tante" bilang peduli lindungi jdi peduli lingkungan, aku tak sanggup menahan ktwa😭😭 https://t.co/Cm5JWJAR42	Ini ad tante bilang peduli lindungi jdi peduli lingkungan, aku tak sanggup menahan ktwa

Berikut merupakan contoh dari melakukan tahapan *case folding* pada penelitian ini:

TABEL V
CASE FOLDING

No	Awal	Hasil
1.	Ini ad tante bilang peduli lindungi jdi peduli lingkungan, aku tak sanggup menahan ktwa	ini ad tante bilang peduli lindungi jdi peduli lingkungan, aku tak sanggup menahan ktwa

Berikut merupakan contoh dari melakukan tahapan *tokenizing* pada penelitian ini:

TABEL VI
TOKENIZING

No	Awal	Hasil
1.	Ini ad tante bilang peduli lindungi jdi peduli lingkungan, aku tak sanggup menahan ktwa	[ini, ad, tante, bilang, peduli, lindungi, jdi, peduli, lingkungan, aku, tak, sanggup, menahan, ktwa]

Berikut merupakan contoh dari melakukan tahapan *tokenizing* pada penelitian ini. Dimana menghapus *stopword* akan membantu menemukan dan mengumpulkan kata penting pada tahapan *stemming* nantinya karena sudah hilangnya kata himbuhan.

TABEL VII
STOPWORD REMOVAL

No	Awal	Hasil
1.	[ini, ad, tante, bilang, peduli, lindungi, jdi, peduli, lingkungan, aku, tak, sanggup, menahan, ktwa]	[ad, tante, bilang, peduli, lindungi, jdi, peduli, lingkungan, sanggup, menahan, ktwa]

Berikut merupakan contoh dari melakukan tahapan *stemming* pada penelitian ini:

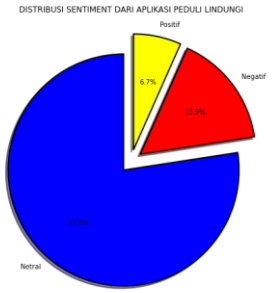
TABEL VIII
STEMMING

No	Awal	Hasil
1.	[ad, tante, bilang, peduli, lindungi, jdi, peduli, lingkungan, sanggup, menahan, ktwa]	[ad, tante, bilang, peduli, lindung, jdi, peduli, lingkung, sanggup, tahan, ktwa]

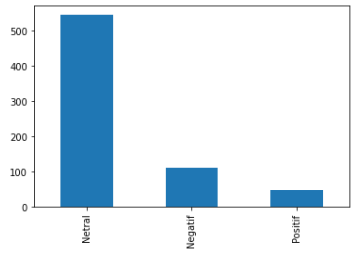
Hasil *stemming* lalu dijadikan teks kembali, dan dilakukannya *filtering duplicate data*. Yang mulanya data ketika diberi label manual jumlahnya 797 data. Setelah *preprocessing data*, terjadi pengurangan data menjadi 704 data.

D. Visualisasi

Setelah proses labeling dan praproses data, peneliti rasa perlu melakukan visualisasi data guna memperjelas data pada tiap kelasnya. Ketika dilakukan pengecekkan menggunakan *function .value_counts()* didapati data Netral sebanyak 545 data, Negatif sebanyak 112 data, dan Positif sebanyak 47 data. Maka, untuk diagram distribusi tiap kelasnya dapat dilihat pada gambar berikut.



Gambar. 4 Diagram distribusi data per kelas



Gambar. 5 Grafik distribusi data per kelas

- *Word Cloud*

Visuali data dengan menggunakan *word cloud*. *Word cloud* itu sendiri merupakan bentuk visualisasi dengan menampilkan kumpulan kata dari teks. Dimana semakin tinggi frekuensi kemunculan kata, maka semakin besar pula ukuran kata yang akan tampil di gambar visualisasinya. Hasil visualisasi yang ditampilkan menggunakan *word cloud* akan lebih bagus dan menarik untuk dipresentasikan.

Gambar. 6 Visualisasi *Word Cloud* keseluruhan

Gambar. 7 Visualisasi *Word Cloud* negatif

Gambar. 8 Visualisasi *Word Cloud* netral

Gambar. 9 Visualisasi *Word Cloud* positif

E. Splitting Data

Dalam penelitian ini kami menetapkan bahwa variabel X yang diperoleh dari keseluruhan *Tweet*, dan variabel y yang merupakan label. Sebelum dilakukan *splitting data*, peneliti menerapkan seleksi fitur pembobotan kata, dengan digunakannya TF-IDF (*Term Frequency Inverse Document Frequency*) pada tweet. TF-IDF menjadi salah satu metode statistik guna untuk melakukan pengukuran pentingnya suatu kata terhadap suatu dokumen atau corpus. Adapun rumusnya dipaparkan dibawah ini:

$$\text{TF} = \begin{cases} 1 + Log10\,(ft,d), & ft,d > 0 \\ 0, & ft,d = 0 \end{cases} \qquad (6)$$

Lalu, variabel X yang telah diterapkannya TF-IDF dan variabel y akan dipecah menjadi *X_train*, *X_test*, *y_train*, dan *y_test*, dengan perbandingan 80% *data training*, dan 20% *data testing*. Serta, ditetapkannya *stratify* dari variabel y, dan *random_state* sebesar 42. Dengan demikian, *data training* sebanyak 563 data, dan *data testing* sebanyak 141 data.

F. Perhitungan Model Algoritma Klasifikasi

1. Naive Bayes

Model algoritma *Naive Bayes Classifier* yang peneliti gunakan adalah MultinomialNB. MultinomialNB yakni termasuk dalam metode *supervising learning*. Dimana akan memerlukan label dalam data sebelum dilakukan *training*. MultinomialNB dinilai memiliki tingkat akurasi yang tinggi dengan perhitungan yang dinilai sederhana. Karena alasan umumnya model MultinomialNB digunakan untuk pengklasifikasian dokumen atau teks, peneliti pun mencobanya pada penelitian kali ini.

Setelah model MultinomialNB dibangun maka data *train* yang terdiri dari 563 data training akan dilakukan proses *fit* atau pelatihan. Untuk variabel yang akan menyimpan label prediksi berasal dari nilai *predict()* variabel *X_test* diberi nama *y_pred*.

2. SVM (*Support Vector Machine*)

Dalam penelitian ini peneliti membangun model SVM menggunakan beberapa parameter, yakni jenis kernel adalah rbf yang merupakan default kernel dalam SVM model. Nilai C yang didefinisikan yaitu 0.5, *gamma* sebesar 1.0, *decision_function_shape* yang digunakan adalah ovr, dan ditetapkannya *class_weighted* yaitu *balance*. Penetapan beberapa parameter sebelumnya telah diketahui menghasilkan hasil evaluasi terbaik.

G. Evaluasi

Setelah diterapkannya algoritma klasifikasi model Naive Bayes dan SVM, peneliti melakukan evaluasi terhadap model tersebut. Evaluasi tersebut dilakukan dengan melakukan perhitungan nilai Akurasi, *recall*, *precision*, dan *f1-score*. Serta, akan ditampilkannya *confusion matrix* dari nilai aktual berupa *data test* dan nilai prediksi hasil pemodelan.

$$Accuracy = \frac{(TP + TN)}{(TP+FP+FN+TN)} \quad (7)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (8)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (9)$$

$$F1 - Score = \frac{2\ x\ Recall\ x\ Precision}{Recall + Precision} \quad (10)$$

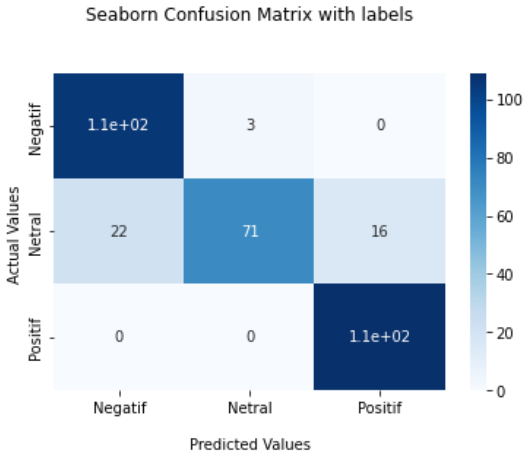
Keterangan :

- *TP (True Positive Correct result)*
- *TN (True Negative Correct absence of result)*
- *FP (False Positive Unexpected result)*
- *FN (False Negative Missing result)*

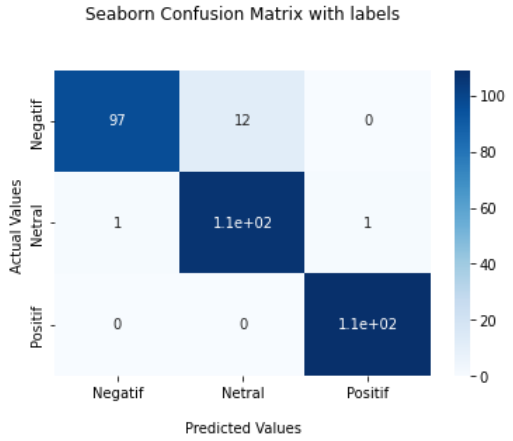
Untuk model NBC sendiri didapatkan akurasi sebesar 0.773, *recall* sebesar 0.773, *precision* 0.333, dan *f1-score* 0.291. Dan, untuk model SVM didapatkan akurasi sebesar 0.794, *recall* sebesar 0.528, *precision* 0.423, dan *f1-score* 0.439.

Dikarenakan, nilai akurasi yang lumayan namun nilai evaluasi poin lainnya kurang memuaskan, maka peneliti melakukan *oversampling*. Oversampling digunakan untuk menangani *imbalance data*, mengingat sebaran kelas pada data yang digunakan peneliti tidak seimbang. Metode yang digunakan ialah *oversampling* menggunakan SMOTE. Untuk jumlah data terbanyak terdapat pada kelas netral yaitu sebanyak 545 data, dengan demikian kelas negatif dan positif pun akan dilakukan *resampling* hingga masing-masingnya berjumlah 545 data.

Setelah ditangani nya *imbalance data* menggunakan *oversampling* SMOTE, evaluasi kembali digunakan dengan menggunakan model serta parameter yang sama dengan tahap evaluasi pertama. Lalu, didapatkannya akurasi sebesar 0.881, *recall* sebesar 0.899, *precision* 0.881, dan *f1-score* 0.873 untuk model NBC. Dan, untuk model SVM didapatkan akurasi sebesar 0.954, *recall* sebesar 0.958, *precision* 0.954, dan *f1-score* 0.954.



Gambar. 10 *Confusion Matrix* model Naive Bayes



Gambar. 11 *Confusion Matrix* model SVM

IV. KESIMPULAN DAN SARAN

A. *Kesimpulam*

Dari penelitian analisis sentimen terhadap penggunaan aplikasi PeduliLindungi pada media sosial twitter dengan digunakannya algoritma naive bayes dan SVM memperoleh hasil akurasi sebesar 0.881, *recall* sebesar 0.899, *precision* 0.881, dan *f1-score* 0.873 untuk model NBC. Dan, untuk model SVM didapatkan akurasi sebesar 0.954, recall sebesar 0.957, precision 0.954, dan f1-score 0.954.

Pada penelitian ini, jika dilihat para pengguna twitter lebih sering memberikan sentimen atau tanggapan yang bersifat netral terkait aplikasi PeduliLindungi, seperti halnya beberapa *tweet* berisikan berita mengenai aplikasi PeduliLindungi itu sendiri, pertanyaan terkait cara menggunakannya atau adanya kewajiban menggunakan PeduliLindungi di berbagai kegiatan. Serta mengingat belum lama ini, kegiatan UTBK berlangsung. Jadi, kebanyakan *tweet* sekedar memberitahukan ketentuan penggunaan PeduliLindungi dalam pelaksanaan UTBK. Hal ini direpresentasikan dengan kata “utbk”, “ui” ataupun “masuk” yang mempunyai frekuensi tinggi pada distribusi *tweet*. Namun, beberapa *tweet* menunjukan bahwa penggunaan PeduliLindungi cukup membebani mereka karena seringnya permintaan *updating*, dengan representasi kata “ngeselin”, “capek”, “*update*” atauapun “ribet” yang mempunyai

frekuensi tinggi. Dan, sangat sedikit tweet berisi sentimen positif terhadap aplikasi PeduliLindungi ataupun penggunaannya.

B. Saran

Peneliti mengetahui bahwa adanya kekurangan dalam penelitian ini. Maka dari itu, kritik ataupun saran yang membangun dari para pembaca dan pengamat sangat diperlukan guna perbaikan penelitian-penelitian selanjutnya. Adapun saran untuk penelitian selanjutnya, lebih baik menggunakan data dengan jumlah yang lebih banyak ketika melakukan pengklasifikasian sentimen analisis, ataupun menetapkan rentang waktu yang lebih lama dalam proses crawling data agar lebih banyaknya data yang bisa dipelajari oleh sistem, dan dapat mencoba menggunakan algoritma lainnya.

REFERENCES

[1] I. F. Rozi, E. N. Hamdana and M. B. I. Alfahmi, "PENGEMBANGAN APLIKASI ANALISIS SENTIMEN TWITTER MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER (Studi Kasus SAMSAT Kota Malang)," *Jurnal Informatika Polinema,* vol. 4, no. 2, pp. 149-154, 2018.

[2] W. A. Luqyana, "Analisis Sentimen Cyberbullying Pada Komentar Instagram Dengan Metode Klasifikasi Support Vector Machine," Universitas Brawijaya, Malang, 2018.

[3] N. Fitriyah, B. Warsito and D. A. I. Maruddani, "ANALISIS SENTIMEN GOJEK PADA MEDIA SOSIAL TWITTER DENGAN KLASIFIKASI SUPPORT VECTOR MACHINE (SVM)," *Jurnal Gaussian,* vol. 9, no. 3, pp. 376-390, 2020.

[4] A. R. Satria, S. Adinugroho and S. Suprapto, "Analisis Sentimen Ulasan Aplikasi Mobile menggunakan Algoritma Gabungan Naïve Bayes dan C4.5 berbasis Normalisasi Kata Levenshtein Distance," *JPTIIK (Jurnal Pengembangan Teknologi Informasi dan Ilmu Kompoter),* vol. 4, no. 11, pp. 4154-4163, 2020.

[5] D. A. Pangestu, "Analisis Sentimen Terhadap Opini Publik Tentang Kesehatan Mental Selama Pandemi Covid-19 Di Media Sosial Twitter Menggunakan Naive Bayes Classifier Dan Support Vector Machine," UII Yogyakarta, Yogyakarta, 2020.

[6] L. Hermawan and M. B. Ismiati, "Pembelajaran TextPreprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval," *JURNAL TRANSFORMATIKA,* vol. 17, no. 2, pp. 188-199, 2020.

KONTRIBUSI

A. Pencarian Studi Literatur.
Alvin Putra Perdana, Jihan Kamilah, Nurhikmah Mawaddah Solin, Salsabila Oktafani.

B. Crawling Data
Jihan Kamilah, Salsabila Oktafani.

C. Labeling Data
Alvin Putra Perdana, Jihan Kamilah, Nurhikmah Mawaddah Solin, Salsabila Oktafani.

D. Pembuatan Kode Program.
Alvin Putra Perdana, Jihan Kamilah, Nurhikmah Mawaddah Solin, Salsabila Oktafani.

E. Penyusunan Laporan.
Alvin Putra Perdana, Jihan Kamilah, Nurhikmah Mawaddah Solin, Salsabila Oktafani.

F. Pembuatan PPT.
Jihan Kamilah, Nurhikmah Mawaddah Solin, Salsabila Oktafani.

G. Presentasi Materi (Video).
Alvin Putra Perdana, Jihan Kamilah, Nurhikmah Mawaddah Solin, Salsabila Oktafani.

H. Upload Laporan, dan Link Video.
Alvin Putra Perdana.