 ISSN xxxx-yyyy

Analisis Sentimen Vaksin Booster Dengan Menggunakan Metode Naïve Bayes

Aprisal zikri ananda [a*], Tedy Setiady[b],

[a] Program Studi Teknik Informatika Universitas Ahmad Dahlan ,Prof Dr Soepomo S.H., Janturan, Umbulharjo, Yogyakarta 55164, Indonesia, Aprisal1700018199@webmail.uad.ac.id

[b] I Program Studi Teknik Informatika Universitas Ahmad Dahlan, Prof Dr Soepomo S.H., Janturan, Umbulharjo, Yogyakarta 55164, Indonesia, tedy.setiady@tif.uad.ac.id

*Corresponding Authors

A B S T R A K

Interaksi sosial media sangat marak dilakukan oleh hampir seluruh orang di Indonesia dengan media sosial seperti Twitter. Pada media ini banyak terdapat pendapat (*sentiment*) terkait vaksin bosster. Penelitian ini menjadikan Vaksin Bosster sebagai objek analisis sentimen pada penelitian ini. Sentimen-sentimen tersebut kemudian diproses untuk menjadi analisis sentimen menggunakan *Naïve Bayes Classifier.*

Analisis sentimen menggunakan klasifikasi *Naïve Bayes* sebagai algoritma penyusunnya. Klasifikasi ini memiliki nilai akurasi tinggi dengan cara kerja yang simpel. Tujuan penelitian ini adalah mengolah data Vaksin Bosster pada Twitter dalam membuat informasi sentimen analisis

Data yang digunakan adalah data Bosster tanggal 22- 27 oktober 2022 pada platform Twitter menggunakan bantuan website Netlytic. Pengecekan akurasi menggunakan *Confusion Matrix* dengan akurasi sebesar 69%. Klasifikasi menghasilkan 3 data negatif, 15 data netral, dan 43 data positif .

Kata kunci: Analisis Sentimen vaksinbosster, *Twitter, Naïve Bayes, Confusion Matrix*

1. Pendahuluan

Infeksi corona adalah penyakit menular yang menyerang sistem pernapasan manusia. Virus ini berasal dari China, khususnya dari kota Wuhan di provinsi Hubei. Pada awal Maret 2020, virus ini masuk ke Indonesia dan sejak itu penyebarannya semakin cepat dan memakan korban. Pemerintah telah mengeluarkan kebijakan dan peraturan untuk mengatasi virus corona, seperti pembatasan kegiatan masyarakat, protokol kesehatan, dan vaksinasi bertahap untuk mencegah penyebaran virus.

Twitter menjadi salah satu media sosial yang populer digunakan saat ini. Pengguna twitter dapat mengunggah dan mengekspresikan pendapat mereka secara bebas. Berbagai informasi penting dapat ditemukan di twitter dan digunakan sebagai sumber data penelitian, terutama untuk data mining.

Salah satu informasi yang dapat ditemukan di twitter adalah tanggapan masyarakat terhadap vaksinasi Covid-19. Saat ini, Covid-19 menjadi trending topik dan banyak opini, rumor, dan informasi yang belum jelas kebenarannya, sehingga menimbulkan pro dan kontra dalam masyarakat terkait dengan vaksinasi dari pemerintah. Analisis sentimen perlu dilakukan untuk mengklasifikasikan tweet terkait dengan vaksinasi Covid-19 dari pemerintah.

Dalam data mining, teknik seperti transformation, tokenizing, stemming, classification, dan lain-lain sangat berpengaruh terhadap tingkat akurasi sentimen analisis.

Penelitian dilakukan untuk menganalisis respon masyarakat tentang vaksinasi booster dengan cara melakukan klasifikasi respon tersebut ke dalam respon positif dan negatif dengan mengambil data dari twitter. Selanjutnya, pengelompokkan opini masyarakat dilakukan dengan menggunakan metode nayve bayes classifier masyarakat terkait dengan wacana vaksinasi tersebut pada media sosial twitter penelitian tentang pengembangan analisis sentimen pada dokumen twitter mengenai dampak Virus Corona menggunakan Metode Naive Bayes

1.1 Batasan Masalah

Dengan mengacu pada latar belakang yang telah dijelaskan sebelumnya, penulis menetapkan beberapa batasan masalah yang perlu disebutkan agar pembahasannya lebih terfokus. Beberapa batasan yang diberlakukan terhadap permasalahan adalah sebagai berikut:

a. Pengumpulan data dilakukan pada tanggal 22 - 23 Oktober 2022 melalui platform Twitter dengan bantuan Tools Netlytic.

b. Data yang digunakan terbatas pada tweet yang hanya mengandung teks tanpa gambar, video, tautan website, dan emotikon.

c. Data tweet yang diambil menggunakan kata kunci "Vaksin OR Bosster" dan berbahasa Indonesia.

1.2 Rumusan Masalah

a. Berdasarkan latar belakang dan identifikasi masalah di atas, maka dapat dirumuskan masalah:
b. Bagaimana mengetahui sentimen opini masyrakat mengenai Vaksin Bosster di indonesia melalui sosial media twitter.
c. Bagaimana mengukur nilai akurasi dari metode nayve bayes (NBC)

1.3 Tujuan Penelitian

Tujuan penelitian ini adalah untuk mengetahui sentimen opini masyarakat mengenai Vaksin Booster di Indonesia .

1.4 Maanfaat Penelitian

Penelitian ini diharapkan dapat menjadi bahan masukan bagi masyarakat Indonesia pada khusnya masyarakat yang belum melakukan vaksin bosster

2 Kajian dan Pustaka

Pada Penelitian pertama (Zulfikar Firmansyah Firmansyah,

2020) dengan judul “Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter”.”Proses ini dilakukan setelah data didapatkan pada tahapan crawling data, dilanjutkan dengan pembersihan kata pada proses preprocessing, dan pembobotan kata dengan algoritma *TF-IDF*. Kata kunci yang digunakan untuk menjaring respon & opini masyarakat terhadap vaksin COVID-19 dalam proses web scraping tersebut adalah menggunakan dua kata kunci yaitu “Vaksin Covid” dan “Vaksin Corona”. Kata kunci yang digunakan dinilai dapat opini masyarakat Indonesia terhadap vaksin COVID-19 di media sosial twitter. Data tweets yang diambil yaitu tweets yang diposting di media sosial twitter pada rentang tanggal 25 Oktober-3 November 2020 karena adanya keterbatasan pengumpulan data. Hasil analisis menunjukkan bahwa masyarakat lebih banyak memberikan respon positif terhadap wacana tersebut (30%) dibandingkan dengan respon negatifnya (26%). Kata-kata bersentimen yang paling sering muncul juga mengindikasikan lebih banyak kata yang bersentimen positif dibandingkan dengan kata yang bersentimen negatif. (Firmansyah, 2021)

Penelitian kedua (Ni Putu Gita Naraswati, 2021) dengan judul “Analisis Sentimen Pengguna Media Sosial Twitter Terhadap Wabah Covid-19 Dengan Metode Naive Bayes Classifier Dan Support Vector Machine” dengan tujuan menganalisis sentimen publik dari cuitan *Twitter* mengenaiwabah covid-19. Adapun metode yang digunakan *Naive Bayes Classification dan suport vector machine* karena memiliki algoritma yang sederhana dengan akurasi yang tinggi. Berdasarkan hasil evaluasi, menghasilkannilai akurasi pengklasifikasian sebesar 78,3%. Sedangkan nilai akurasi yang diperoleh dari hasil klasifikasi dengan metode Support Vector Machine adalah sebesar 81,6%. Sementara hasil pengujian akurasi menggunakan metode Cross Validationdengan 10 K-Fold CV menghasilkan nilai

rata-rata akurasi pada metode Naïve Bayes Classifier sebesar 69,8% dan nilai rata-rata akurasi pada metode Support Vector Machine sebesar 74,4% (sujadi & sandi fajar)

Penelitian kelima (Yulianita & Tiani wahyu), Eri Zuliarso (2018) judul penelitian “Analisis Sentimen dalam penanganan covid-19 menggunakan metode (NBC)” tujuan dari penelitian ini adalah untuk melihat opnini masyarakat tentang penanganan covid19 di Indonesia, dataset dalam penelitian ini jumlah tweet pada bulan April-Juni 2020 terdapat 10.000 tweet. Dalam tweet tersebut terdapat 6.515 tweet yang mengandung positif dan 3.485 mengandung negatife.Hasil akurasi yang didaptkan (NBC) 89.13%

2.1 Twitter

Twitter merupakan satu dari banyak media sosial yang umum digunakan masyarakat Indonesia termasuk para pengamat. *Twitter* menyediakan berbagai fitur postingan *(tweet)* yang dapat digunakan seperti memposting teks, gambar, video, link website, dan sebagainya. Pada penelitian ini penulis hanya akan menganalisis tweet berbentuk teks

2.1.1 Analisis sentiment

Analisis Sentimen adalah bagian dari text *mining. Analisis sentimen* juga menjadi projek yang paliing utama bagi orang-orang yang baru belajar machine *learning* menggunakan data teks (*Andreas Chandra, 2019:27*).

2.2 Data mining

Data Mining merupakan proses pengumpulan sebuah informasi penting pada suatu data yang berukuran besar. Beberapa fungsi dari data mining sebagai berikut:

• Fungsi untuk memahami secara jauh data yang diteliti disebut fungsi deskripsi dengan tujuan mengetahui fungsi dari data tersebut. Data tersebut bertujuan untuk mengetahui fungsi dari deskripsi.

• Fungsi prediksi menemukan suatu pola tertentu, pola tersebut dapat diketahui dari variabel yang terkandung dalam data.

2.3 Preprocessing

Tahap *pre-processing* atau praproses data merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain. Pada umumnya, *pre- processing* data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. Adapun ekstraksi data saat *pre-processing* anatara lain :

2.3.1. Filtering

Filtering adalah proses pembersihan teks dari kata yang tidak diperlukan, ini dilakukan untuk mengurangi *noise*. Kata yang dihilangkan adalah karakter *punctuation*, *hashtag, number, one character, URL, re- tweet, stopword,* dan *negation word*. Setelah filtering berhasil maka karakter-karakter yang memiliki salah satu contoh diatas, seperti @fulan akan dihilangkan.

2.3.2. Tokenizing

Tokenizing adalah proses pemecahan teks menjadi kata tunggal. Menggunakan metode n gram teks penilitian ini akan dipecah menjadi satu kata pertoken. Setelah dilakukan tokenizing maka teks akan terpisah menjadi token satu kata yang terpisah oleh tanda koma dan teks menjadi data *array*.

2.3.3. Case folding

Case folding yaitu mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf “a” sampai dengan “z” yang diterima. Setelah proses ini dilakukan maka seluruh huruf besar pada teks akan digantikan dengan huruf kecil.

2.3.4. Stemming

Stemming merupakan metode yang digunakan untuk mengembalikan kata menjadi kata dasar. Setelah proses

stemming selesai maka setiap kata dalam teks akan kehilangan imbuhannya dan hanya tersisa kata dasar pada teks tersebut.

2.4 Naïve Bayes Classifier

Naïve bayes merupakan metode klasifikasi yang berdasar pada teorema Bayes. Klasifikasi ini sangat memperhatikan tingginya akurasi serta kecepatan dalam memproses suatu data dalam jumlah yang besar. Algoritma *Naïve Bayes* memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *Teorema Bayes*. Ciri utama dari *Naïve Bayes Classifier* ini adalah asumsi yang sangat kuat (naif) akan independensi dari masing-masing kondisi atau kejadian. Adapun rumus perhitungan pada gambar *:*

Diketahui bahwa:

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

• X adalah data sampel dengan kelas (label) yang tidak diketahui

• H adalah hipotesa bahwa X adalah data dengan kelas (label).

• P(H) adalah peluang dari hipotesa H.

• P(X) adalah peluang data sampel yang akan diamati.

• P(X|H) merupakan peluang data sampel X, bila diasumsikan bahwa hipotesa benar (valid).

Untuk masalah klasifikasi, yang dihitung adalah P(H|X), yaitu peluang bahwa hipotesa benar (valid) untuk data sampel X yang diamati

2.5. Confusion Matrix

Confusion Matrix adalah tabel pengujian dengan nilai *True Positif* (TP), *True Negatif* (TN), *False Positif* (FP), dan *False Nrgatif* (FN). 4 item inilah yang akan menampilkan nilai prediksi dan nilai aktual. Dari representasi 4 nilai tersebut dapat dicari juga nilai lain seperti *accuracy, precision*, *recall*, dan nilai *F-1 Score*. Nilai akurasi didapat dari persamaan, sedangkan nilai presisi dari persamaan *recall* adalah dan terakhir persamaan *F-1 Score*.

3.Metode

3.1 Objek Penelitian

Objek penelitian ini adalah opimi masyarakat Indonesia yang menggunakan bahasa Indonesia terkait vaksin bosster yang disampaikan melalu media social twitter

3.2 Crawling Data

Pengumpulan data atau biasa disebut dengan crawling data. Pada penelitian ini dilakukan pada dataset Twitter untuk crawling data pada Twitter yang membutuhkan API Twitter dan juga memberikan pembatasan berjumlah crawling data berjumlah 306 dan batas pengambilan tweets 2 minggu yang lalu. Proses ini menggunakan alat bantu website bernama Netlytic yang dapat diakses pada URL *netlytic.org* data di crawling pada tanggal 22-27 mei 2022 dengan kata kunci “Vaksin OR Bosster” dan filter Bahasa menggunakan Bahasa Indonesia. Crawling data yang dilakukan berhasil mendapatkan 306 data. Dari hasil crawling ini didapat data mentah dengan berbagai atribut seperti nama user, link, tweet, description, user bio dan lain sebagainya. Data yang berhasil diambil tersimpan dalam format xlsx.

Kumpulan data atau biasa disebut data indeks. Penelitian ini dilakukan dengan menggunakan database Twitter dari data perayapan Twitter , yang membutuhkan API Twitter dan juga memiliki batas perayapan data sebanyak 306 dan batas pengambilan tweet dari dua minggu lau. Proses ini menggunakan alat web bernama Netlytic, yang dapat ditemukan di netlytic.org tanggal indeks dari 22 mei hingga 27 oktobeter 2022 dengan kata kunci “VAKSIN OR Bosster dan filter bahasa ke Indonesia pengindeksan data yang dilakukan menghasilkan 306 data. Hasil pengindeksan ini memberikan data mentah dengan berbagai atributseperti nama, tautan, tweet, deskripsi, bios pengguna, dll. Data yang berhasil diambil disimpan dalam format XLSX

3.3 Analisis Kebutuhan

Spesifikasi kebutuhan pada penilaian ini diantaranya :

3.3.1 Perangkat Lunak (Software)

Perangkat lunak kebutuhan pada penelitian ini diantaranya :

a. Microsoft exel 2010
b. Microsoft word 2010
c. Python
d. Netlytic
e. Jupyter Notebook

3.3.2 Perangkat Keras (Hardware)

Perangkat keras yang dibutuhkan pada penelitian ini yaitu laptop dengan spesifikasi sebagai berikut:

- Processor :AMD Ryzen ™ 7 5800H Mobile Processor up to 4.4 GHz maxboost)
- RAM 8GB DDR4 on board
- Operating System : compatibility with the Windows 8.1

3.4 Tahapan Penelitian

Adapun tahapan-tahapan yang akan dilakukan pada penelitian ini adalah sebagai berikut :

- Crawling data tweets

Data yang diambil merupakan data yang ter-*mention*-kan berdasarkan akun-akun elit politik dan pengamatan politik, yang kemudian *data tweet* yang telah diambil disimpan ke dalam *database*.

- Seleksi Data

  Proses memilih kolom dari file *csv* yang telah didapatkan. Pada penelitian ini hanya akan menggunakan data opini masyarakat terkait vaksin bosster berupa komentar, maka kolom yang dipilih adalah kolom “text” .

- Pengolahan data

  Proses pengolahan data dilakukan untuk mengetahui nilai sentimen dari opini masyarakat Indonesia yang disampaikan melalui *tweet* di media sosial *twitter*, tahapan yang dilakukan adalah sebagai berikut:

  - Text Processing

    Pada tahap ini terdapat beberapa langkah proses untuk melakukan penelitian guna memproses data yang ada, yaitu mulai dengan membersihan kalimat dari kata yang tidak diperlukan guna mengurangi *noise*. Kemudian dilakukan proses *case folding* untuk menyamakan kata menjadi huruf kecil semua. Selanjutnya, *tokenizing* untuk memecah kalimat menjadi kata serta menghilangkan tanda baca dan *delimeter.* Setelah melakukan proses *tokenizing* langkah selanjutnya adalah *stopword removal* dengan cara menghilangkan kata yang tidak penting dan menghilangkan kata depan, kata belakang, kata sambung, dan sebagainya. Terakhir, dilakukan proses *stemming* untuk mengubah bentuk kata menjadi kata dasar serta mengubah kata yang tidak baku menjadi kata baku.

  - Pembobotan

    Tahapan ini memberikan nilai pada setiap *term* yang ada setelah sebelumnya melewati tahap *preprocessing* terlebih dahulu. Pada penelitian ini sistem pembobotan hanya sampai pada proses *term frequency,*

dikarenakan algoritma *Naïve Bayes* hanya memperhatikan apakah suatu kata atau *term* ada atau tidak di dalam dokumen.

- Pelabelan

  Proses selanjutnya yaitu memberikan label pada data yang sudah dilakukan proses *cleaning* sebelumnya. Pada tahap ini, untuk pemberian label pada data yang digunakan untuk membedakan komentar dengan sentimen positif dan negatif dengan metode pengambilan keputusan secara intuisi dengan mempertimbangkan konotasi kata.

- Perhitungan Sentimen Analisis Nayve bayes

  Setelah mengetahui nilai bobot dari *term frequency,* proses selanjutnya adalah melakukan analisis sentimen dengan menggunakan metode *Naïve Bayes* untuk mengetahui apakah opini termasuk ke dalam kelas sentimen yang mana.

- Pengajuan Akurasi

  Pada proses ini dilakukan pengujian untuk mengetahui nilai akurasi pada penelitian ini dengan menggunakan metode *confusion matrix* yang memberikan informasi perbandingan hasil klasifikasi yang dilakukan pada penelitian dengan hasil klasifikasi sebenarnya. Representasi hasil proses klasifikasi pada *confusion matrix* terdapat 4 istilah yaitu *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN).* Pengukuran *performance matrix* dari *confusion matrix* untuk mengukur kinerja model yang telah dibuat dengan metode *Naïve Bayes* sehingga diperoleh nilai akurat .

4. Hasil dan Pembahasan

4.1 Pengumpulan data

Data yang digunakan dalam penelitian ini adalah data opini masyarakat Indonesia terkait vaksin bosster diindonesia yang disampaikan melalui *tweet* di media sosial *twitter*. Data didapatkan dari proses *crawling* menggunakan *Netlytic twitter* yang kemudian diekspor ke dalam bentuk file *csv*

4.2 Crawling data

Proses ini menggunakan alat bantu website bernama Netlytic yang dapat diakses pada URL *netlytic.org* data di crawling pada tanggal 22-27 Oktober 2022 dengan kata kunci “Vakisn OR Bosster” dan filter Bahasa menggunakan Bahasa Indonesia

4.3 Seleksi Data

Dari data yang telah diperoleh sebelumnya kemudian akan dilakukan proses seleksi data, yaitu memilih kolom mana saja yang akan digunakan dalam penelitian. Pada penelitian ini kolom yang dipilih adalah kolom ”text” yang mengandung *tweet* terkait Vaksin Bosster

4.4 Preprocessing

Tahapan pertama yang dilakukan adalah melakukan *preprocessing* terlebih dahulu pada teks untuk mempersiapkan dokumen teks yang tidak terstruktur menjadi data terstruktur untuk siap digunakan pada proses selanjutnya. Tahapan *preprocessing* meliputi

a. **Cleaning**

Pada tahap ini dilakukan pembersihan kalimat dari kata yang tidak diperlukan guna mengurangi *noise* seperti menghilangkan karakter ikon ekspresi, hashtag (#), *username* (@), *url* (*http://situs.com*), *e-mail* (nama@situs.com), symbol, dan tanda baca. Pada Tabel 4.1. menunjukkan contoh proses *cleaning*

Sebelum Cleaning	Sesudah Cleaning
Indonesia sudah mulai	indonesia suntik vaksin

menyuntikkan vaksin booster kedua https://t.co/vsMEskvYdt	booster

b. Case Folding

Tahapan untuk mengubah semua huruf menjadi *lowercase* atau huruf kecil agar semua sama bentuknya. Contoh dari proses *case folding* dapat dilihat pada tabel 4.2 dibawah ini.

Sebelum Cleaning	Sesudah Cleaning
Indonesia sudah mulai menyuntikkan vaksin booster kedua https://t.co/vsMEskvYdt	indonesia suntik vaksin booster

a. Tokenizing

Selanjutnya pada tahap ini dilakukan pemotongan *string input* menjadi *token* berdasarkan tiap kata yang menyusunnya untuk contoh prosesnya dapat dilihat pada Tabel 4.3. dibawah ini.

Tabel 4.3 ProsesTokenizing

Sebelum Tokenizing	Sesudah Tokenizing
indonesia suntik vaksin booster	[Indonesia] , [Suntik] , [Vaksin] , [Bosster]

b. Stemming

Proses selanjutnya adalah dengan membuat kata berimbuhan kembali menjadi kata dasar dan apabila masih terdapat kata tidak baku maka akan diubah menjadi kata baku. Pada tabel 4.4. menunjukkan proses *stemming.*

Tabel 4.4 Proses Stemming

Sebelum Stemming	Sesudah Stemming
indonesia suntik vaksin booster	Indonesia suntik vaksin bosster

c. Pembobotan

Setelah dilakukan proses *preprocessing*, setiap kata akan dihitung nilai bobotnya dengan memperhatikan ada tidaknya *term* di dalam data. Adapun pada Tabel 4.5. untuk contoh perhitungan *term frequency* (TF).

Tabel 4.5 Pembobotan Term Frequency

Term	Term Frequency (TF)
Indonesia	1
Suntik	1
Vaksin	1
bosster	1

d. Pelabelan

Pemberian label pada data yang digunakan untuk membedakan komentar dengan sentimen positif dan negatif secara manual dengan metode pengambilan keputusan secara intuisi dengan mempertimbangkan konotasi kata.

e. Perhitungan Sentimen dengan Nayve Bayes

 ISSN xxxx-yyyy

Langkah selanjutnya setelah mendapatkan nilai bobot untuk setiap kata adalah dengan menghitung nilai *similaritas* antara kata atau dokumen dengan menggunakan algoritma *Naïve Bayes.*

f. Pengujian Akurasi

Langkah terakhir adalah melakukan pengujian terhadap data yang telah diproses klasifikasi sebelumnya untuk mendapatkan nilai akurasi. Pada penelitian ini menggunakan pengujian *confusion matrix* untuk mengetahui tingkat keakurasiannya.

4.5 Implementasi

Sumber data yang akan diproses berasal dari hasil *crawling* data dari *twitter* yang kemudian di simpan dalam format *csv*. Dari sekian banyak data hasil *crawling* dari *twitter* hanya data bagian *text* saja yang akan diproses. Untuk lebih jelasnya dapat dilihat pada gambar

Gambar 4.2 Dataset Awal

Id	Tweetid	Guild	Link	Author	title
1	1572420457034579970	https://twitter.com/MarkEsso8/statuses/1572420457034579970	https://twitter.com/MarkEsso8/statuses/1572420457034579970	MarkEsso8	"RT @Anggita_lung: Vaksinasi adalah kunci utama untuk melindungi diri dan orang terdekat dari bahaya covid-19 varian baru BA.4 dan BA.5. Se..."
2	1572420373278502914	https://twitter.com/Anggita_lung/statuses/1572420373278502914	https://twitter.com/Anggita_lung/statuses/1572420373278502914	Anggita_lung	RT @roseuune: Keberadaan vaksin Booster serta penerapan disiplin prokes merupakan 2 kunci yg tdk terpisahkan dan saling berkaitan dgn erat....
3	1572419414104109056	https://twitter.com/garuda080/statuses/1572419414104109056	https://twitter.com/garuda080/statuses/1572419414104109056	garuda080	RT @roseuune: Keberadaan vaksin Booster serta penerapan disiplin prokes merupakan 2 kunci yg tdk terpisahkan dan saling berkaitan dgn erat....
4	1572418966819340294	https://twitter.com/detik_jatim/statuses/1572418966819340294	https://twitter.com/detik_jatim/statuses/1572418966819340294	detik_jatim	"Di Jember, pelayanan vaksinasi COVID-19 dilakukan di beberapa tempat. Saat ini tersedia vaksin jenis pfizer, baik dosis 1, 2, dan 3 atau booster nakes. Berikut jadwal dan lokasi vaksinasinya rek.
5	1572418381814591493	https://twitter.com/pooryyi/statuses/1572418381814591493	https://twitter.com/pooryyi/statuses/1572418381814591493	pooryyi	RT @roseuune: Keberadaan vaksin Booster serta penerapan disiplin prokes merupakan 2 kunci yg tdk terpisahkan dan saling berkaitan dgn erat....
6	1572418082194460672	https://twitter.com/Revenge00810455/statuses/1572418082194460672	https://twitter.com/Revenge00810455/statuses/1572418082194460672	Revenge00810455	RT @roseuune: Keberadaan vaksin Booster serta penerapan disiplin prokes merupakan 2 kunci yg tdk terpisahkan dan saling berkaitan dgn erat....
7	1572417865005043712	https://twitter.com/chinta_chintata/statuses/1572417865005043712	https://twitter.com/chinta_chintata/statuses/1572417865005043712	chinta_chintata	"RT @Anggita_lung: Vaksinasi adalah kunci utama untuk melindungi diri dan orang terdekat dari bahaya covid-19 varian baru BA.4 dan BA.5. Se..."

 ISSN xxxx-yyyy

4.6 Seleksi data

Seleksi data dibutuhkan karena data hasil crawling masih mentah sehingga banyak atribut data tidak diperlukan yang masih menempel. Pada penelitian ini data hasil crawling memiliki atribut sebanyak 29 atribut data dan setelah di seleksi tersisa 1 atribut data.

# Memasukkan dataframe df yang dipilih (seleksi) ke dalam variable d
df = df[1:][[5]] # Lihat nomor kolom
df.columns = ['tweet']
df

Gambar 4.2 Hasil Seleksi data

Seleksi data

```
data diseleksi adalah data yang digunakan dalam proses klasifikasi yang merupakan sentimen twitter
Atribut yang digunakan hanya satu yaitu deskripsi tweet
```

```
# Memasukkan dataframe df yang dipilih (seleksi) ke dalam variable d

df = df[1:][[5]] # Lihat nomor kolom
df.columns = ['tweet']
df
```

	tweet
1	RT @Anggita_lung: Vaksinasi adalah kunci utama...
2	RT @roseuune: Keberadaan vaksin Booster serta ...
3	RT @roseuune: Keberadaan vaksin Booster serta ...
4	Di Jember, pelayanan vaksinasi COVID-19 dilaku...
5	RT @roseuune: Keberadaan vaksin Booster serta ...
...	...
301	Sobat BUMN,\n\nPandemi Covid-19 masih berlangs...
302	Update Vaksinasi Covid-19 di Kabupaten Humbang...
303	@msaid_didu Semoga yg belum menerima vaksin bo...
304	Dalam menekan penyebaran COVID-19 di Rutan sek...
305	@INALawanCovid19 Dengan sudah dilakukannya vak...

305 rows × 1 columns

 ISSN xxxx-yyyy

4.7 Text Processing

Setelah dilakukan seleksi data maka selanjutnya adalah melakukan pembersihan pada data yang masih bercampur dengan URL, karakter tidak penting, emotikon dan lain-lain. Preprocessing data sangat dibutuhkan dalam melakukan analisis sentimen untuk mendaptakan hasil yang baik. Pada proses ini data akan mengalami pembersihan sehingga hanya bagian yang dibutuhkan yang ada. Proses ini melibatkan beberapa langkah sebagai berikut, seleksi data, penghilang karakter tidak penting, tokenizing, case folding, dan stemming

Gambar 4.3 Dataset Awal

```
def remove(tweet):
    tweet = re.sub(r'https?:\/\/.*[\r\n]*', '', tweet)
    #remove angka
    tweet = re.sub('[0-9]+', '', tweet)

    #remove stockmarket ticker like $GE
    tweet = re.sub(r'\$\w*', '', tweet)

    #remove old style retweet text 'RT'
    tweet = re.sub(r'^RT[\s]+', '', tweet)

    #remove Hashtags (Only remove hashtags from the word)
    tweet = re.sub(r'#', '', tweet)

    tweet = re.sub(r':', '', tweet)

    #tweet = re.sub('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))'+, '', tweet, flags=re.MUL
    #tweet = re.sub(r'https', '', tweet)

    return tweet
df['remove_RT'] = df['remove_user'].apply(lambda x: remove(x))
```

```
#import stopword

from nltk.corpus import stopwords
stopwords_indonesia = stopwords.words('indonesian')

#import sastrawi
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory = StemmerFactory()
stemmer = factory.create_stemmer()

#tokenize
from nltk.tokenize import TweetTokenizer

# Happy Emoticons
emoticons_happy = set([
    ':-)', ':)', ';)', ':o)', ':]', ':3', ':c)', ':>', '=]', '8)', '=)', ':}',
    ':^)', ':-D', ':D', '8-D', '8D', 'x-D', 'xD', 'X-D', 'XD', '=-D', '=D',
    '=-3', '=3', ':-))', ":'-)", ":')", ':*', ':^*', '>:P', ':-P', ':P', 'X-P',
    'x-p', 'xp', 'XP', ':-p', ':p', '=p', ':-b', ':b', '>:)', '>;)', '>:-)',
    '<3'
    ])

# Sad Emoticons
emoticons_sad = set([
    ':L', ':-/', '>:/', ':S', '>:[', ':@', ':-(', ':[', ':-||', '=L', ':<',
    ':-[', ':-<', '=\\', '=/', '>:(', ':(', '>.<', ":'-(", ":'(", ':\\', ':-c',
    ':c', ':{', '>:\\', ';('
    ])

# all emoticons (happy + sad)
emoticons = emoticons_happy.union(emoticons_sad)

def clean_tweets(tweet):
    # remove stock market tickers like $GE
    tweet = re.sub(r'\$\w*', '', tweet)

    # remove old style retweet text "RT"
    tweet = re.sub(r'^RT[\s]+', '', tweet)

    # remove hyperlinks
    #tweet = re.sub(r'https?:\/\/.*[\r\n]*', '', tweet)

    # remove hashtags
    # only removing the hash # sign from the word
    tweet = re.sub(r'#', '', tweet)

    #remove coma
    tweet = re.sub(r',','',tweet)

    #remove angka
    tweet = re.sub('[0-9]+', '', tweet)

    # tokenize tweets
    tokenizer = TweetTokenizer(preserve_case=False, strip_handles=True, reduce_len=True)
    tweet_tokens = tokenizer.tokenize(tweet)

    tweets_clean = []
    for word in tweet_tokens:
        if (word not in stopwords_indonesia and # remove stopwords
              word not in emoticons and # remove emoticons
                 word not in string.punctuation): # remove punctuation
            #tweets_clean.append(word)
            stem_word = stemmer.stem(word) # stemming word
            tweets_clean.append(stem_word)

    return tweets_clean
```

4.8 Penerapan Nayve bayes

Setelah data mengalami pembersihan dan labeling maka selanjutnya data akan diproses dengan algoritma *Naïve Bayes*. Pada *Algoritma Naïve Bayes* data akan dipecah menjadi data training dan data testing. Pada penelitian ini penulis akan membaginya menjadi 8:2 untuk data *training* dan data *testing* atau dapat dikatakan 80% *training* dan 20% *testing*. Untuk penerapannya dilakukan dengan kodingan pada gambar

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(v_data, df['sentiment'], test_size=0.2)
#Gunakan random_state = 0 untuk hasil yang sama, atau integer apapun
model_g.fit(X_train,y_train)
```

```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
y_preds = model_g.predict(X_test)
print(confusion_matrix(y_test,y_preds))
print(classification_report(y_test,y_preds))
print('nilai akurasinya adalah ',accuracy_score(y_test, y_preds))
```

Hasil dari perhitungan *Naïve Bayes* adalah 61 data negatif, 61 data netral, dan 61 data positif. Hasil ini dapat dilihat pada gambar Hasil Klasifikasi.

Gambar 4. 4 Penerapan *Naïve Bayes*

 ISSN xxxx-yyyy

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(v_data, df['sentiment'], test_size=0.2)
#Gunakan random_state = 0 untuk hasil yang sama, atau integer apapun
model_g.fit(X_train,y_train)
```

```
▾ GaussianNB
GaussianNB()
```

```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

y_preds = model_g.predict(X_test)

print(confusion_matrix(y_test,y_preds))
print(classification_report(y_test,y_preds))
print('nilai akurasinya adalah ',accuracy_score(y_test, y_preds))
```

```
[[ 2  0  1]
 [ 2  4  9]
 [ 2  5 36]]
              precision    recall  f1-score   support

    negative       0.33      0.67      0.44         3
     neutral       0.44      0.27      0.33        15
    positive       0.78      0.84      0.81        43

    accuracy                           0.69        61
   macro avg       0.52      0.59      0.53        61
weighted avg       0.68      0.69      0.67        61

nilai akurasinya adalah  0.6885245901639344
```

Hasil dari perhitungan *Naïve Bayes* adalah 3 data negatif, 15 data netral, dan

43 data positif. Hasil ini dapat dilihat pada gambar 4.4 .

Gambar 4. 4 Gambar hasil klasifikasi

```
              precision    recall  f1-score   support

    negative       0.33      0.67      0.44         3
     neutral       0.44      0.27      0.33        15
    positive       0.78      0.84      0.81        43

    accuracy                           0.69        61
   macro avg       0.52      0.59      0.53        61
weighted avg       0.68      0.69      0.67        61

nilai akurasinya adalah  0.6885245901639344
```

4.9 Pengujian Confusion Matrix

Pengujian dilakukan dengan Confusion Matrix denan kodingan seperti digambar

Gambar 4.5 Hasil Pengajuan Confusion Matrix

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(v_data, df['sentiment'], test_size=0.2)
#Gunakan random_state = 0 untuk hasil yang sama, atau integer apapun
model_g.fit(X_train,y_train)
```

```
▾ GaussianNB
GaussianNB()
```

```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

y_preds = model_g.predict(X_test)

print(confusion_matrix(y_test,y_preds))
print(classification_report(y_test,y_preds))
print('nilai akurasinya adalah ',accuracy_score(y_test, y_preds))
```

```
[[ 2  0  1]
 [ 2  4  9]
 [ 2  5 36]]
              precision    recall  f1-score   support

    negative       0.33      0.67      0.44         3
     neutral       0.44      0.27      0.33        15
    positive       0.78      0.84      0.81        43

    accuracy                           0.69        61
   macro avg       0.52      0.59      0.53        61
weighted avg       0.68      0.69      0.67        61

nilai akurasinya adalah  0.6885245901639344
```

Pengujian ini menghasilkan akurasi sebesar 69.0%.

5.. Kesimpulan

Berdasarkan penelitian ini yang telah dilakukan maka penulis memperoleh kesimpulan sebagai berikut :

1. Penerapan Naïve Bayes Classifier Menghasilkan data negatif, 3 data netra 15, dan 43 data positif
2. Mengunakan confusion matrix menghasilkan tingkat akurasi sebesar 68% dengan split 8:2 data latih berbanding data uji atau dapat dikatan 306 data latih berbanding beberapa data

 ISSN xxxx-yyyy

DAFTAR PUSTAKA

1) Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M.M. & Williams, H.E. 2007. Stemming Indonesian : A Confix-Stripping Approach. Transaction on Asian Langeage Information Processing. Vol. 6, No. 4, Articel 13. Association for Computing Machinery : New York .
2) Alexander, P & Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the Seventh conference on International Language Resources and Evaluation pp: 1320-1326.
3) Baeza-Yates, R.A. & Ribeiro-Neto, B. 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing: Boston.
4) Farber, D. 2012. Twitter hits 400 million tweets per day, mostly mobile. http://www.cnet.com/news/twitter-hits-400-million-tweets-per-day-mostly- mobile/
5) Feldman, R & Sanger, J. 2007. The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press : New York
6) Kanakaraj, M., Mohana, R. & Guddeti, R. 2015. NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers. International Conference on Signal Processing, Communication and Networking.
7) Liu, B. 2010. Sentiment Analysis A Multi-Faceted Problem. IEEE Intelligent Systems.
8) Liu, B. 2012. Opinion Mining. Chicago, United States of America.
9) Luo, F., Li, C. & Cao, Z. 2016. Affective-feature-based sentiment analysis using SVM classifier. Proceedings of 20rd International Conference on Computer Supported Cooperative Work in Design
10) Maarif, A. A. (2015) ‘Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah’, Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah, (5), p. 4.
11) Felix Fridom Mailo , Lutfan Lazuardi, “Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia”, Journal of Information Systems for Public Health Volume IV No. 1 Desember 2019.
12) Najib, A. C., Irsyad, A., Qandi, G. A., & Rakhmawati, N. A. (2019). Perbandingan Metode Lexicon-based dan SVM untuk Analisis Sentimen Berbasis Ontologi pada Kampanye Pilpres Indonesia Tahun 2019 di Twitter. Fountain of Informatics Journal, 4(2), 41.
13) Siti, M. (2016). Pre-Processing Text Mining pada Data Twitter. Universitas Islam Lamongan, Lamongan.

 ISSN xxxx-yyyy