

**ANALISIS SENTIMEN PENGGUNA MEDIA SOCIAL X  
TERHADAP VAKSINASI COVID 19 MENGGUNAKAN  
*LEXICON BASED DAN NAIVE BAYES***

**SKRIPSI**

**MIFTAH AULIA**

**171402009**



**PROGRAM STUDI S1 TEKNOLOGI INFORMASI  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA  
MEDAN  
2024**

**ANALISIS SENTIMEN PENGGUNA MEDIA SOCIAL X  
TERHADAP VAKSINASI COVID 19 MENGGUNAKAN  
*LEXICON BASED DAN NAIVE BAYES***

**SKRIPSI**

Diajukan untuk melengkapi tugas dan memenuhi syarat memperoleh ijazah  
Sarjana Teknologi Informasi

**MIFTAH AULIA**

**171402009**



**PROGRAM STUDI S1 TEKNOLOGI INFORMASI  
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA  
MEDAN  
2024**

## PERSETUJUAN

Judul : : ANALISIS SENTIMEN PENGGUNA MEDIA SOCIAL X  
TERHADAP VAKSINASI COVID 19 MENGGUNAKAN  
*LEXICON BASED DAN NAIVE BAYES*

Kategori : SKRIPSI

Nama : MIFTAH AULIA

Nomor Induk Mahasiswa : 171402009

Program Studi : SARJANA (S-1) TEKNOLOGI INFORMASI

Fakultas : ILMU KOMPUTER DAN TEKNOLOGI INFORMASI  
UNIVERSITAS SUMATERA UTARA

Medan, 04 Juli 2024

Komisi Pembimbing:

Pembimbing 2



Dr. Erna Budhiarti Nababan M.IT.

NIP. 196210262017042001

Pembimbing 1



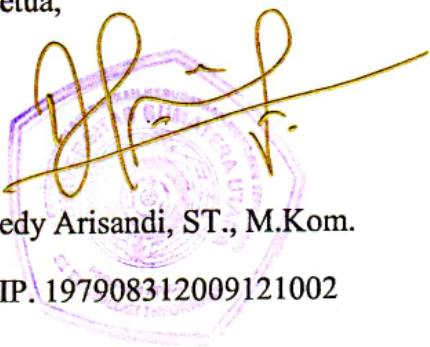
Prof. Dr. Drs. Opim Salim Sitompul., M.Sc.

NIP. 196108171987011001

Diketahui/disetujui oleh

Program Studi S-1 Teknologi Informasi

Ketua,



Dedy Arisandi, ST., M.Kom.

NIP. 197908312009121002

**PERNYATAAN**

ANALISIS SENTIMEN PENGGUNA MEDIA SOCIAL X  
TERHADAP VAKSINASI COVID 19 MENGGUNAKAN  
*LEXICON BASED DAN NAIVE BAYES*

**SKRIPSI**

Saya mengakui bahwa skripsi ini adalah hasil karya saya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing telah disebutkan sumbernya.

Medan, 04 Juli 2024

Miftah Aulia

171402009

## **UCAPAN TERIMA KASIH**

Alhamdulillah, puji dan syukur penulis ucapkan kepada Allah Subhaanahu wa Ta'aala, karena berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi ini. Selesainya skripsi ini juga karena adanya dukungan dan doa dari berbagai pihak. Maka dari itu, penulis mengucapkan terima kasih kepada:

1. Kedua orang tua penulis, Bapak Zul Azmi dan Ibu Siti Aisyah yang selalu memberi semangat, kasih sayang, nasehat, dukungan, doa, dan bersedia untuk bertukar pikiran bahkan sampai saat ini.
2. Syahilda Rahmadani dan Mutia Hafizah selaku adik penulis yang selalu memberi semangat dan dukungan.
3. Bapak Prof. Dr. Muryanto Amin, S. Sos., M. Si. Selaku Rektor Universitas Sumatera Utara.
4. Ibu Dr. Maya Silvi Lydia, M.Sc. selaku Dekan Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara.
5. Bapak Dedy Arisandi, ST., M.Kom. selaku Ketua Program Studi S1 Teknologi Informasi Universitas Sumatera Utara.
6. Bapak Prof. Dr. Drs. Opim Salim Sitompul, M.Sc. selaku Dosen Pembimbing 1 dan Ibu Dr. Erna Budhiarti Nababan M.IT. selaku Dosen Pembimbing 2 yang telah meluangkan waktu dalam membimbing penulis dalam menyelesaikan penelitian dan penulisan skripsi ini.
7. Seluruh Dosen di Fakultas Ilmu Komputer dan Teknologi Informasi USU yang telah memberikan ilmu yang bermanfaat bagi penulis selama masa perkuliahan.
8. Seluruh pegawai Fakultas Ilmu Komputer dan Teknologi Informasi Sumatera Utara yang telah membantu segala urusan administrasi pada masa perkuliahan.
9. Aderay Miraj, Tondi Alfarizi dan Muhammad Syahran selaku sahabat dekat penulis yang selalu memberi semangat dan bersedia mendengar keluh kesah dalam segala hal.
10. Teman-teman yang tergabung dalam Yuan Senari, Muhammad Bagus Syahputra Tambunan dan Ahmad Adil yang paling sering diajak diskusi tentang perkuliahan.
11. Teman-teman Kom C 2017 yang tidak dapat penulis sebutkan satu persatu, yang telah memberi banyak bantuan semasa perkuliahan dan dalam menyelesaikan skripsi ini.

12. Sahabat semasa SMA yang masih dekat sampai saat ini, yang terus memberi dukungan dan semangat hingga skripsi ini selesai.

Semoga Allah Subhaanahu wa Ta'aala selalu memberikan kasih sayang dan berkah kepada semua pihak yang telah ikut dalam mendukung penulis dalam proses pembuatan skripsi ini.

Medan, 04 Juli 2024

Miftah Aulia

**ANALISIS SENTIMEN PENGGUNA MEDIA SOCIAL X  
TERHADAP VAKSINASI COVID 19 MENGGUNAKAN  
*LEXICON BASED DAN NAIVE BAYES***

**ABSTRAK**

Dalam menghadapi pandemi COVID-19, penting untuk memahami pandangan masyarakat terhadap vaksinasi melalui media sosial, terutama Media Social X. Penelitian ini bertujuan untuk menganalisis sentimen pengguna Media Social X terhadap vaksinasi COVID-19 di Indonesia menggunakan metode Lexicon-Based dan Naive Bayes. Data tweet terkait vaksin COVID-19 dikumpulkan dari seluruh provinsi di Indonesia, dengan total dataset sebanyak 2,879 tweet. Dataset ini dibagi menjadi data training sebanyak 2,164 tweet dan data testing sebanyak 542 tweet, dengan rasio pembagian 80% untuk data training dan 20% untuk data testing. Kata kunci yang dijadikan fokus adalah analisis sentimen, vaksinasi COVID-19, Media Social X , metode Lexicon-Based, Naive Bayes, dan Indonesia. Metode Lexicon-Based digunakan untuk mengklasifikasikan sentimen setiap tweet berdasarkan kamus positif-negatif yang disediakan, sementara Naive Bayes digunakan untuk analisis klasifikasi secara supervisi. Tujuan penelitian ini tidak hanya untuk memahami pandangan masyarakat, tetapi juga untuk meningkatkan akurasi hasil analisis sentimen sebelumnya. Diharapkan penelitian ini dapat memberikan wawasan yang lebih dalam terhadap pandangan masyarakat terkait vaksinasi COVID-19 di Indonesia, yang nantinya dapat digunakan untuk meningkatkan strategi komunikasi dan pencegahan penyakit.

**Kata Kunci :** Vaksinasi COVID-19, media sosial, Media Social X, analisis sentimen, metode *Lexicon-Based, Naive Bayes*.

## **ANALYSIS OF MEDIA SOCIAL X USER SENTIMENT TOWARDS COVID 19 VACCINATION USING LEXICON BASED AND NAIVE BAYES**

### **ABSTRACT**

*Facing the COVID-19 pandemic, understanding public sentiment towards COVID-19 vaccination through social media, particularly Social Media X, is crucial. This research aims to analyze Social Media users' sentiments regarding COVID-19 vaccination in Indonesia using Lexicon-Based and Naive Bayes methods. COVID-19 vaccine-related tweets were collected from all provinces in Indonesia, totaling 2,879 tweets. The dataset was divided into 2,164 training tweets and 542 testing tweets, with an 80% to 20% ratio. Keywords focused on are sentiment analysis, COVID-19 vaccination, Social Media X , Lexicon-Based method, Naive Bayes, and Indonesia. The Lexicon-Based method was utilized to classify sentiment in each tweet based on provided positive-negative dictionaries, while Naive Bayes was employed for supervised classification analysis. The objective of this research is not only to understand public perspectives but also to enhance the accuracy of previous sentiment analysis results. It is hoped that this study will provide deeper insights into public views on COVID-19 vaccination in Indonesia, which can be utilized to improve communication strategies and disease prevention efforts.*

**Keywords :** *COVID-19 vaccination, social media, Social Media X , sentiment analysis, Lexicon-Based method, Naive Bayes.*

**DAFTAR ISI**

Persetujuan .....	ii
Pernyataan .....	iii
Ucapan Terima Kasih.....	iv
Abstrak.....	vi
Abstract.....	vii
Daftar isi.....	viii
Daftar tabel.....	xi
Daftar gambar .....	xii
BAB 1 PENDAHULUAN .....	1
1.1 Latar Belakang .....	1
1.2. Rumusan Masalah.....	5
1.3. Batasan Masalah .....	5
1.4. Tujuan Penelitian .....	6
1.5. Manfaat Penelitian .....	6
1.6. Metodologi Penelitian.....	6
1.6.1 Studi Literatur .....	6
1.6.2 Pengumpulan Data .....	6
1.6.3 Analisis Permasalahan.....	7
1.6.4 Perancangan Sistem.....	7
1.6.5 Evaluasi .....	7
1.7 Sistematika Penulisan .....	7
BAB 2 LANDASAN TEORI.....	9
2.1 Covid-19 .....	9
2.2 Vaksin .....	9
2.3 Analisis Sentimen .....	10
2.4 Metode Lexicon Based .....	11
2.5. Metode Naive Bayes .....	13
2.6 <i>Countvectorizer</i> .....	14
2.7 Flask.....	15

2.8 Penelitian Terdahulu .....	15
BAB 3 ANALISIS DAN PERANCANGAN SISTEM .....	19
3.1 Dataset.....	19
3.2 Arsitektur Umum .....	20
3.2.1 <i>Crawling/Scraping Data</i> .....	21
3.2.2 <i>Preprocessing Data</i> .....	22
3.2.2.1 Cleaning.....	22
3.2.2.2. <i>Remove Stopwords</i> .....	23
3.2.2.3 <i>Stemming</i> .....	24
3.2.2.4 <i>Normalisasi</i> .....	25
3.2.3 TF-IDF .....	26
3.2.4 <i>Lexicon Based</i> .....	31
3.2.5 <i>Naive Bayes</i> .....	37
3.2.5.1 Dataset .....	42
3.2.5.2 Train Test Split .....	42
3.2.5.3 Modeling.....	44
3.2.5.4 Implementasi <i>Naive Bayes</i> .....	44
3.2.6 Output.....	45
3.2.6.1 <i>Classification Report</i> .....	45
3.2.6.2 <i>Confusion Matrix</i> .....	45
3.2.7 Test Model.....	47
3.2.7.1 Test Model (Text) .....	47
3.2.7.2 Test Model (csv) .....	48
3.3 Perancangan Sistem .....	49
3.3.1 Rancangan Tampilan Login .....	49
3.3.2 Rancangan Tampilan <i>Dashboard</i> .....	49
3.3.3 Rancangan Tampilan Dataset.....	50
3.3.4 Rancangan Tampilan <i>Pre-Processing</i> Data .....	50
3.3.4.1 Rancangan Tampilan <i>Pre-Processing</i> Data <i>Cleaning</i> .....	50
3.3.4.2 Rancangan Tampilan <i>Pre-Processing</i> Data <i>Stopword</i> .....	51
3.3.4.3 Rancangan Tampilan <i>Pre-Processing</i> Data <i>Stemming</i> .....	51
3.3.4.4 Rancangan Tampilan <i>Pre-Processing</i> Data <i>Normalisasi</i> .....	52

3.3.5 Rancangan Tampilan TF – IDF.....	52
3.3.5.1 Rancangan Tampilan Data TF - IDF .....	52
3.3.5.2 Rancangan Tampilan Data Vocabulary .....	53
3.3.5.3 Rancangan Tampilan Word Cloud .....	53
3.3.6 Rancangan Tampilan Labeling Data.....	54
3.3.7 Rancangan Tampilan <i>Split Data</i> .....	54
3.3.8 Rancangan Tampilan Data <i>Training</i> .....	55
3.3.9 Rancangan Tampilan Data <i>Testing</i> .....	55
3.3.10 Rancangan Tampilan <i>Naïve Bayes</i> .....	56
3.3.11 Rancangan Tampilan Test Model (Text) .....	56
3.3.12 Rancangan Tampilan Test Model (csv) .....	57
BAB 4 IMPLEMENTASI PENGUJIAN SISTEM .....	58
4.1 Implementasi Sistem.....	58
4.1.1. Spesifikasi Perangkati Keras dan Perangkati Lunak.....	58
4.1.2. Implementasi Perancangan Tampilan Antarmuka .....	58
4.2. Implementasi Model Naive Bayes .....	66
4.3 Implementasi Test model ( text ) .....	67
4.4 Implementasi Test model ( csv ).....	68
4.5 Pengujian.....	69
4.4 Pembahasan.....	71
BAB 5 KESIMPULAN DAN SARAN .....	72
5.1 Kesimpulan .....	72
5.2 Saran .....	73
DAFTAR PUSTAKA .....	74

**DAFTAR TABEL**

Tabel 2. 1 Penelitian Terdahulu .....	17
Tabel 2. 1 Penelitian Terdahulu (Lanjutan) .....	18
Tabel 3. 1 Dataset Media Social X .....	20
Tabel 3. 2 Kemunculan Term Padai Dokumen.....	28
Tabel 3. 3 Skor Df.....	28
Tabel 3. 3 Skor Df (Lanjutan).....	28
Tabel 3. 4 Skor IDF .....	29
Tabel 3. 5 Nilai TF-IDF .....	30
Tabel 3. 6 Sample Dataset .....	34
Tabel 3. 7 Kamus Sentimen .....	35
Tabel 3. 8 Tokenisasi Teks .....	36
Tabel 3. 9 Sample Dataset .....	40
Tabel 3. 10 Tabel Matrix Confusion.....	46
Tabel 4. 1 Tabel Hasil Pengujian NB ( Naïve Bayes ) .....	69
Tabel 4. 2 Tabel Confusion Matrix.....	70

## DAFTAR GAMBAR

Gambar 2. 1 Alur Metode Lexicon Based .....	12
Gambar 3. 1 Dataset Media Social X.....	20
Gambar 3. 2 Arsitektur Umum Penelitian .....	21
Gambar 3. 3 Proses Crawling/Scraping Data X .....	21
Gambar 3. 4 Hasil Data Crawling.....	22
Gambar 3. 5 Tahap Cleaning .....	23
Gambar 3. 6 Tahap Stopwords.....	24
Gambar 3. 7 Tahap Stemming .....	25
Gambar 3. 8 Tahap Normalisasi .....	26
Gambar 3. 9 Label TF-IDF.....	27
Gambar 3. 10 Pseudocode Implementasi Lexicon Based .....	31
Gambar 3. 11 Alur saat proses Lexicon Based .....	33
Gambar 3. 12 Proses Dalam Naive Bayes .....	38
Gambar 3. 13 Train Test Split.....	43
Gambar 3. 14 Pseudocode Implementasi Naive Bayes .....	44
Gambar 3. 15 Classification Report.....	45
Gambar 3. 16 Confusion Matrix .....	46
Gambar 3. 17 Test Model (text).....	48
Gambar 3. 18 Test Model (csv) .....	48
Gambar 3. 19 Rancangan Tampilan iLogin.....	49
Gambar 3. 20 Rancangan Tampilan Dashboard .....	49
Gambar 3. 21 Rancangan Tampilan Dataset.....	50
Gambar 3. 22 Rancangan Tampilan Pre-Processing Data Cleaning.....	50
Gambar 3. 23 Rancangan Tampilan Pre-Processing Data Stopword .....	51
Gambar 3. 24 Rancangan Tampilan Pre-Processing Data Stemming.....	51
Gambar 3. 25 Rancangan Tampilan Pre-Processing Data Normalisasi.....	52
Gambar 3. 26 Rancangan Tampilan Data TF – IDF.....	52
Gambar 3. 27 Rancangan Tampilan Data Vocabulary .....	53
Gambar 3. 28 Rancangani Tampilan Word Cloud.....	53
Gambar 3. 29 Rancangan Tampilan Labeling Data .....	54
Gambar 3. 30 Rancangan Tampilan Split Data .....	54
Gambar 3. 31 Rancangan Tampilan Data Training .....	55
Gambar 3. 32 Rancangan Tampilan Data Testing .....	55
Gambar 3. 33 Rancangan Tampilan Naive Bayes .....	56
Gambar 3. 34 Rancangan Tampilan Test Model (Text) .....	56
Gambar 3. 35 Rancangan Tampilan Test Model (csv) .....	57
Gambar 4. 1 Halaman Login.....	59
Gambar 4. 2 Halaman Dashboard.....	59
Gambar 4. 3 Halaman Upload File .....	60
Gambar 4. 4 Halaman Cleaning.....	60

Gambar 4. 5 Halaman Stopword.....	61
Gambar 4. 6 Halaman Stemming.....	61
Gambar 4. 7 Halaman Normalisasi 1 .....	62
Gambar 4. 8 Halaman Normalisasi 2 .....	62
Gambar 4. 9 Halaman TF-IDF.....	63
Gambar 4. 10 Halaman Vocabulary.....	63
Gambar 4. 11 Halaman Wordcloud .....	64
Gambar 4. 12 Halaman Labeling .....	65
Gambar 4. 13 Halaman Split data.....	65
Gambar 4. 14 Halaman training data .....	66
Gambar 4. 15 Halaman testing data.....	66
Gambar 4. 16 Klasifikasi Naïve Bayes 1 .....	67
Gambar 4. 17 Klasifikasi Naive Bayes 2 .....	67
Gambar 4. 18 Implementasi Test model ( text ) 1 .....	68
Gambar 4. 19 Implementasi Test model ( text ) 2 .....	68
Gambar 4. 20 Implementasi Test model ( csv ) .....	69

## **BAB 1**

### **PENDAHULUAN**

#### **1.1 Latar Belakang**

Wabah penyakit virus corona telah ditetapkan sebagai pandemi global oleh *World Health Organization* (WHO) pada tanggal 11 Maret 2020 lalu dalam situs resminya. Banyak aspek kehidupan ekonomi dan sosial masyarakat yang telah berubah akibat wabah virus corona yang disebabkan oleh virus corona baru SARS-CoV-2. Di Negara Indonesia, Presiden Joko Widodo mengumumkan kasus pertama Covid-I9 masuk ke indonesia pada 2 Maret 2020 lalu, dan menjangkit 2 orang WNI asal Depok, Jawa Barat. Sejak kasus tersebut, jumlah orang yang terjangkit virus corona di Indonesia terus meningkat setiap harinya, per tanggal 24 Desember 2021 sudah terjadi total 4.261.12 kasus dan angka kematian sudah mencapai 144.047 Jiwa (Ward & del Rio, 2020). Vaksinasi COVID-19 di Indonesia secara resmi dimulai setelah Presiden Joko Widodo meresmikan Peraturan Presiden Republik Indonesia Nomor 99 Tahun 2020 pada tanggal 5 Oktober 2020, yang mengatur pengadaan vaksin dan pelaksanaan vaksinasi dalam rangka penanggulangan pandemi COVID-19. Pelaksanaan vaksinasi ini merupakan langkah penting untuk menanggulangi penyebaran virus SARS-CoV-2 yang telah menyebabkan perubahan signifikan dalam aspek ekonomi dan sosial masyarakat sejak pertama kali terdeteksi di Indonesia pada tanggal 2 Maret 2020. Upaya vaksinasi ini diharapkan dapat melindungi masyarakat dari COVID-19 serta menghentikan penyebaran virus di masa depan.

Dalam menghadapi penyebaran Covid-19 yang sangat cepat dan bahaya yang muncul jika tidak segera ditangani, salah satu cara efektif untuk mencegah penyebaran virus ini adalah dengan mengembangkan vaksin. Program vaksinasi COVID-19 di Indonesia mulai dilakukan pemerintah pada 13 Januari 2021. Vaksin tidak hanya melindungi yang telah divaksin, tetapi juga masyarakat umum dengan mencegah penyebaran penyakit ke seluruh populasi. Meskipun belum ada vaksin untuk SARS dan MERS yang ditemukan, vaksin Covid-19 dapat ditemukan terlebih dahulu. Produksi vaksin yang aman dan efektif sangat penting karena akan

digunakan untuk menghentikan penyebaran penyakit dan mencegahnya di masa mendatang (Sallam, 2021). Pemerintah Indonesia juga terlibat aktif dalam persiapan vaksinasi untuk masyarakat umum. Presiden Joko Widodo pada 5 Oktober 2020 meresmikan Perpres Republik Indonesia Nomor 99 Tahun 2020 tentang Pengadaan Vaksin dan Pelaksanaan Vaksinasi Dalam Rangka Penanggulangan Pandemi Coronavirus Disease 2019 (COVID-19) untuk mengatur kewenangan pemerintah, kementerian/lembaga dan para pejabatnya dalam rencana kegiatan vaksinasi. Rencana kegiatan vaksinasi tersebut harus mempertimbangkan berbagai faktor, termasuk opini publik dan respons masyarakat (Rachman & Pramana, 2020).

Dalam beberapa waktu terakhir, situasi yang diakibatkan oleh lockdown di beberapa bagian dunia dan penerapan *social distancing* telah meningkatkan penggunaan media sosial secara global. Hal ini disebabkan oleh adanya internet yang menghubungkan orang-orang dari tempat yang berbeda secara geografis, sehingga memungkinkan mereka untuk bertukar ide dan informasi. Terlebih lagi, banyak orang yang lebih mengandalkan media sosial untuk mendapatkan informasi terkini. Oleh karena itu, platform media sosial saat ini menjadi saluran mediator antara setiap individu dan seluruh dunia, dan bahkan menjadi aplikasi sosial yang cepat berkembang. Dalam media sosial, orang dapat menunjukkan pandangan, pendapat, dan emosi yang berbeda terhadap berbagai peristiwa yang terjadi akibat pandemi virus corona, termasuk tentang vaksinasi Covid-19(Yulita, W. (2021). Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma *Naïve Bayes Classifier*. Jurnal *Data Mining* dan Sistem Informasi, 2(2), 1-9). Khususnya, para pengguna media sosial Twitter mendapatkan perhatian khusus karena pengguna Twitter dapat dengan mudah menyajikan informasi tentang pendapat mereka terkait vaksinasi melalui pesan publik yang disebut Tweet (D'Andrea et al., 2019). Seiring waktu, data Twitter telah digunakan dalam berbagai penelitian, termasuk menganalisis opini publik terkait topik tertentu seperti vaksinasi (Cotfas et al., 2021). Vaksinasi menjadi salah satu topik yang menimbulkan sejumlah pertanyaan di media sosial, terutama terkait dengan keamanan keseluruhan proses vaksin. Oleh karena itu, sejumlah penelitian telah menganalisis dampak berbagai inisiatif media sosial yang berbeda

pada keraguan vaksinasi atau persepsi publik tentang proses vaksinasi secara umum. (D'Andrea et al., 2019).

Dalam penelitian Analisis Sentimen Opini Terhadap Vaksin Covid-19 pada Media Sosial Twitter Menggunakan *Support Vector Machine* dan *Naive Bayes* opini vaksin Covid – 19 di media Twitter memberikan opini positif yang artinya bahwa penerimaan vaksin Covid – 19 lebih besar dibanding penolakannya. Hasil dari analisa tersebut mendapatkan (40%) respon negatif dan (48%) respon positif. Berdasarkan kesimpulan kegiatan hasil analisis pada penelitian tersebut diharapkan dapat melakukan optimasi pada metode yang digunakan (Fitriana, F., Utami, E., & Al Fatta, H. (2021). Analisis Sentimen Opini Terhadap Vaksin Covid-19 pada Media Sosial Twitter Menggunakan *Support Vector Machine* dan *Naive Bayes*. Jurnal Komtika (Komputasi dan Informatika), 5(1), 19-25.)

Pada Penelitian ini, kegiatan analisis dilakukan dengan cara seperti menganalisis pro dan kontra terhadap program vaksinasi yang dilakukan Pemerintah Indonesia dengan menggunakan metode analisis yang lebih tervalidasi, penelitian ini menggunakan metode *Lexicon Based & Naive Bayes*. Berdasarkan hal tersebut, maka penulis mengajukan penelitian dengan judul “ANALISIS SENTIMEN PENGGUNA MEDIA SOCIAL X TERHADAP VAKSINASI COVID 19 MENGGUNAKAN *LEXICON BASED* DAN *NAIVE BAYES*”. Mengapa penulis menggunakan metode *Lexicon Based* dalam penelitian kali ini, karena data yang diambil dan digunakan dari media twitter belum adanya label sentimen, untuk membuat label tersebut dan menghasilkan sentimen adalah dengan menggunakan metode *lexicon based* yang memanfaatkan kamus positif-negatif yang telah dibuat pada penelitian terdahulu dan disediakan langsung oleh *python*. Dengan memanfaatkan kamus *lexicon* tersebut data yang sebelumnya tidak ada label sentimen akan dideteksi menggunakan *lexicon based* setiap kata yang ada pada dataset disesuaikan dengan kamus lexicon yang ada, apakah termasuk positif ataupun negatif. Pada penelitian ini juga penulis menggunakan *Naïve bayes* untuk proses supervised learning. *Naïve bayes classifier* adalah metode klasifikasi yang berdasarkan probabilitas dan Teorema Bayesian dengan asumsi bahwa setiap variabel X bersifat bebas atau berdiri sendiri dan tidak ada kaitannya dengan variabel lainnya. Metode NBC menempuh dua tahap dalam proses klasifikasi teks,

yaitu tahap pelatihan dan tahap klasifikasi (Amir Hamzah, 2012). Probabilitas adalah kemungkinan terjadinya suatu peristiwa antara 0 s/d 1 (wahyudi & fadlil, 2013). Klasifikasi Gaussian Naive Bayes dapat digunakan untuk memproses atribut numerik pada layanan jaringan komputer (Fadlil, Riadi, & Aji, 2017). Pada penelitian ini penulis akan menggunakan data minimal 2000 data tweet dengan kata kunci Vaksin Covid19. Dengan data tersebut maka akan menghasilkan suatu hasil sentimen pro dan kontra akan Vaksin Covid19 dengan memanfaatkan opini publik dari pengguna Twitter(X). Adapun penulis mengambil rujukan dari beberapa paper terkait *lexicon based*.

Pada penelitian yang berjudul “*Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons*” dengan menghasilkan akurasi menggunakan metode *lexicon based* sebesar 84%. Di penelitian ini dijelaskan bahwa penggunaan *lexicon based* dari enam kamus lexicon sebagai bahan uji coba. Penulis berharap disaat pekerjaan penelitian ini akan menghasilkan akurasi yang lebih tinggi lagi (Khoo, C. S., & Johnkhan, S. B. 2018).

Pada penelitian kedua yang berjudul “*An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia*”. Pada penelitian ini dibahas ada dua metode yang tersedia, analisis sentimen berbasis *supervised learning* dan berbasis *lexicon*. Metode *supervised learning* memang memiliki kinerja yang lebih baik daripada metode berbasis *lexicon*. Namun, performa metode *supervised learning* sangat bergantung pada kualitas dan jumlah data yang telah berlabel. Di hasil akhir pada penelitian di bahas tentang uji coba menerapkan analisis sentimen berbasis *lexicon* pada opini data Indonesia. Secara keseluruhan, metode yang digunakan memiliki akurasi 68%. Hasil ini cukup baik sebagai titik awal untuk penelitian selanjutnya yang akan penulis lakukan (Pamungkas, E. W., & Putri, D. G. P. 2016, August).

Penelitian terakhir yang penulis ambil sebagai rujukan yaitu pada judul “*Unsupervised Twitter Sentiment Analysis on The Revision of Indonesian Code Law and the Anti-Corruption Law using Combination Method of Lexicon Based and Agglomerative Hierarchical Clustering*”. Pada penelitian ini dijelaskan bahwa tujuan dari penelitian ini adalah untuk menganalisis sentimen Tweet pengguna Twitter menolak revisi UU tersebut apakah mereka memiliki sentimen positif atau

negatif menggunakan metode *lexicon based*. Fitur ekstraksi Berbasis *lexicon* dan *Frekuensi Term* (TF) yang melakukan proses secara otomatis. Pada tahap perhitungan akurasi, pada penelitian ini *the error ratio, confusion matrix, and silhouette coefficient*. Karena itu, hasilnya cukup baik. Dari 2.408 tweet, akurasi tertinggi hasilnya adalah 61,6% (Prayoga, Etc. 2020).

Namun, penelitian-penelitian yang dijelaskan hanya berfokus pada mencari hasil akhir persentase sentimen yang ada dan belum ada membuat suatu kesimpulan atau atau analisis yang lebih dalam. Metode berbasis lexicon bekerja dengan membuat kamus terlebih dahulu. Kata-kata dalam kamus digunakan untuk mengidentifikasi apakah suatu kalimat mengandung pendapat atau tidak. Diharapkan dengan menggunakan metode *lexicon based* dan *naïve bayes* penelitian yang penulis kerjakan menjadikan akurasi yang lebih tinggi.

## **1.2. Rumusan Masalah**

Program Vaksinasi Covid-19 sudah dilakukan oleh Pemerintah Indonesia. Akan tetapi, Sebagian masyarakat masih ada yang tidak percaya terkait Vaksin Covid-19. Dengan melakukan analisis sentimen menggunakan metode *Lexicon Based & Naive Bayes* maka dapat jumlah Pro dan Kontra terhadap Program Vaksinasi Covid-19 yang dilakukan Pemerintah Indonesia.

## **1.3. Batasan Masalah**

Berdasarkan rumusan masalah yang ada, maka batasan dalam proses penelitian ini adalah :

- 1.** Metode yang dipakai untuk analisis sentimen adalah *Lexicon Based & Naive Bayes*.
- 2.** Data yang digunakan hanya yang di *scrap* melalui media social *Twitter* pada rentang waktu 1 Juni 2021 sampai 31 Januari 2022 dengan menggunakan kata kunci #vaksincovid19.
- 3.** Dataset akan terdiri dari teks-tweets dalam Bahasa Indonesia.

#### **1.4. Tujuan Penelitian**

Tujuan penelitian untuk mengetahui sentimen masyarakat terhadap vaksinasi Covid-19 di Indonesia melalui media social X menggunakan *Lexicon Based & Naive Bayes*. Hasilnya diharapkan dapat memberikan pemahaman yang lebih baik tentang persepsi publik terhadap vaksinasi Covid-19 di Indonesia.

#### **1.5. Manfaat Penelitian**

Manfaat dari penelitian ini diharapkan dapat :

1. Mengetahui jumlah sentimen pro dan negatif terhadap vaksinasi Covid-19 di Indonesia.
2. Membantu pemerintah terhadap program vaksinasi Covid-19 yang diharapkan bisa menjadi rujukan untuk pihak pengambil kebijakan dalam mengetahui persepsi publik apabila ingin mengambil kebijakan serupa.
3. Menjadi referensi dalam penelitian di bidang analisis sentimen berikutnya.

#### **1.6. Metodologi Penelitian**

Adapun tahapan-tahapan yang dilakukan dalam penelitian ini adalah sebagai berikut :

##### **1.6.1 Studi Literatur**

Pada tahap ini dilakukan studi literatur sebagai bahan referensi untuk penggerjaan penelitian. Penulis mencari referensi dari jurnal, buku, artikel, panduan serta sumber referensi lainnya tentang metode *Lexicon Based & Naive Bayes* untuk menganalisis sentimen.

##### **1.6.2 Pengumpulan Data**

Selanjutnya dilakukan proses pengumpulan data, kemudian akan dilakukan sebuah analisa data serta melakukan evaluasi kualitas data yang digunakan dalam penelitian ini. Data yang diambil dalam penelitian ini merupakan data-data yang terdapat dari media social X. Setelah data

tersebut menjadi sebuah dataset maka data tersebut dibersihkan atau proses *cleaning* menjadi data yang berkualitas. Dengan seperti ini proses penelitian ini akan lebih mudah.

#### **1.6.3 Analisis Permasalahan**

Pada tahap ini, dilakukan analisis terhadap permasalahan penelitian yang akan dilakukan sebelumnya. Diharapkan pada tahap ini akan diperoleh pemahaman tentang metode yang sesuai untuk melakukan analisis sentimen dengan menggunakan *Lexicon Based & Naive Bayes*.

#### **1.6.4 Perancangan Sistem**

Tahap ini yang sangat penting karena berbagai macam metode pemodelan dipilih dan diterapkan ke dataset yang sudah disiapkan untuk mengatasi tujuan analisis sentimen. Adapun metode yang digunakan yaitu *Lexicon Based & Naive Bayes*.

#### **1.6.5 Evaluasi**

Tahapan yang terakhir ialah evaluasi, jika model atau program sudah berhasil dibuat maka dilakukan evaluasi apakah sudah akurat atau belum berdasarkan dilakukannya perbandingan dari hasil akurasi penelitian terdahulu, jika hasil akurasi lebih rendah maka dilakukan peninjauan kembali dengan memperhatikan apakah perlu lebih banyak model untuk dibuat dan diukur namun jika hasil akurasi lebih tinggi atau sudah dikatakan seimbang maka evaluasi dikatakan selesai dan program berhasil dibuat.

### **1.7 Sistematika Penulisan**

Struktur penulisan dalam penelitian ini dijelaskan dalam beberapa bagian utama sebagai berikut:

#### **Bab 1: Pendahuluan**

Bab pendahuluan menguraikan penjelasan latar belakang judul penelitian “Analisis Sentimen Pengguna Media Social X Terhadap Vaksinasi Covid 19

Menggunakan *Lexicon Based & Naive Bayes*”, rumusan masalah, tujuan penelitian, batasan masalah, manfaat penelitian, metodologi penelitian dan sistematika penulisan.

### **Bab 2: Landasan Teori**

Bab landasan teori menjelaskan konsep-konsep yang terkait dengan pemahaman masalah yang akan diteliti dalam penelitian ini, termasuk juga konsep teori yang berkaitan dengan klasifikasi sentimen, analisis vaksin, penggunaan Media Social X, dan metode Lexicon Based & Naive Bayes.

### **Bab 3: Analisis dan Perancangan Sistem**

Pada bab ini disajikan informasi mengenai analisis arsitektur umum, penggunaan metode Lexicon Based & Naive Bayes, klasifikasi sentimen, serta perancangan sistem yang telah dilakukan.

### **Bab 4: Implementasi dan Pengujian Sistem**

Pada bagian implementasi dan pengujian sistem, dijelaskan tentang penerapan analisis dan perancangan sistem yang telah diuraikan dalam Bab 3, serta proses pengujian sistem yang telah disiapkan.

### **Bab 5: Kesimpulan dan Saran**

Pada bab ini didapat informasi berupa kesimpulan dari bahasan penelitian yang dilakukan dan saran untuk penelitian yang akan dilakukan berikutnya.

## **BAB 2**

### **LANDASAN TEORI**

#### **2.1 Covid-19**

Covid-19 adalah suatu virus yang pertama kali ditemukan di china. COVID-19 pertama kali dilaporkan di Indonesia pada 2 Maret 2020 dalam jumlah dua kasus. Pada 31 Maret 2020 menunjukkan sejumlah kasus yang dikonfirmasi 1.528 kasus dan 136 kasus kematian. 10 Angka kematian COVID-19 di Indonesia adalah 8,9%, angka tersebut adalah tertinggi di Asia Tenggara (Sagala, 2020). Virus corona ini merupakan virus RNA dengan partikel ukuran 120- 160 nm (Sitepu & Syafril, 2020). Virus ini terutama menginfeksi hewan, termasuk kelelawar dan unta. Sebelum Terjadinya wabah COVID-19 ada 6 jenis virus corona Dapat menginfeksi manusia yaitu *alphacoronavirus* 229E, *alphacoronavirus* NL63, *betacoronavirus* OC43, *betacoronavirus* HKU1, Penyakit Pernapasan Akut Parah Coronavirus (SARS-CoV), dan Pernapasan Timur Tengah Sindrom virus corona (MERS-CoV) (Al-Sharif et al., 2021). Corona virus yang merupakan etiologi COVID-19 termasuk dalam genus betacoronavirus. Hasil analisis filogenetik menunjukkan bahwa virus ini masuk dalam subgenus yang sama dengan virus corona menyebabkan wabah Penyakit Pernapasan Akut Parah (SARS) 2002-2004 yaitu *Sarbecovirus*. Atas dasar ini, Komite Internasional Taksonomi Virus mengajukan nama *SARSCoV-2*. Struktur genom virus ini memiliki pola seperti coronavirus secara umum (Susilo et al., 2020).

#### **2.2 Vaksin**

Vaksin *Sinovac* telah diproduksi menggunakan metode yang baik untuk mematikan virus Covid-19 sehingga vaksin yang dihasilkan tidak mengandung senyawa virus hidup maupun virus yang dilemahkan. Dilihat dari laman WHO atau *World Health Organization*, vaksin corona yang telah dikembangkan sampai saat ini mengandung antigen yang sama dengan antigen yang menyebabkan penyakit. Namun ternyata antigen yang ada di dalam vaksin corona tersebut dikendalikan atau bisa dibilang dilemahkan yang menyebabkan orang yang

divaksin menjadi hilang akan virus corona tersebut. Penelitian sebelumnya telah menunjukkan bahwa keraguan vaksin adalah fenomena umum secara global, dengan variabilitas dalam alasan yang dikutip di balik penolakan penerimaan vaksin. Itu alasan paling umum termasuk risiko yang dirasakan dan manfaat, keyakinan agama tertentu dan kurangnya pengetahuan dan kesadaran. Alasan yang disebutkan di atas dapat diterapkan Keragu-raguan vaksin COVID-19, seperti yang ditunjukkan publikasi terbaru menunjukkan kuatnya korelasi antara niat untuk mendapatkan vaksin virus corona dan persepsi keamanannya sikap negatif terhadap vaksin COVID-19 dan keengganan untuk mendapatkan vaksin. Analisis faktor-faktor tersebut diperlukan untuk mengatasi Keragu-raguan vaksin COVID-19, menyusul penilaian cakupan dan besarnya ini ancaman kesehatan masyarakat. Ini dapat membantu dalam memandu tindakan intervensi yang ditujukan membangun dan memelihara tanggapan untuk mengatasi ancaman ini (Rahayu, 2021). Tentang Program Vaksinasi Covid-19, masih ada pro dan kontra di masyarakat. Beberapa orang bersedia divaksin, tetapi yang lain belum bersedia dengan berbagai alasan seperti riwayat kesehatan, ibu hamil dan menyusui, serta alasan pribadi. Komite Penanganan Covid-19 dan Pemulihan Ekonomi Nasional (KPCPEN) pada laman covid19.go.id mengatakan bahwa hal ini terjadi karena terdapat beberapa informasi keliru yang beredar di masyarakat terkait vaksin, seperti halal-haram vaksin, kandungan berbahaya dalam vaksin, efektivitas serta keamanan vaksin, dan sebagainya. Padahal, pemerintah telah memastikan hanya akan menyediakan vaksin yang terbukti aman dan lolos uji klinis sesuai rekomendasi WHO. Vaksin Covid-19 produksi Sinovac dijamin suci dan halal(Dewi, S. A. E. (2021). Komunikasi Publik Terkait Vaksinasi Covid 19. *Health Care: Jurnal Kesehatan*, 10(1), 162-167)

### 2.3 Analisis Sentimen

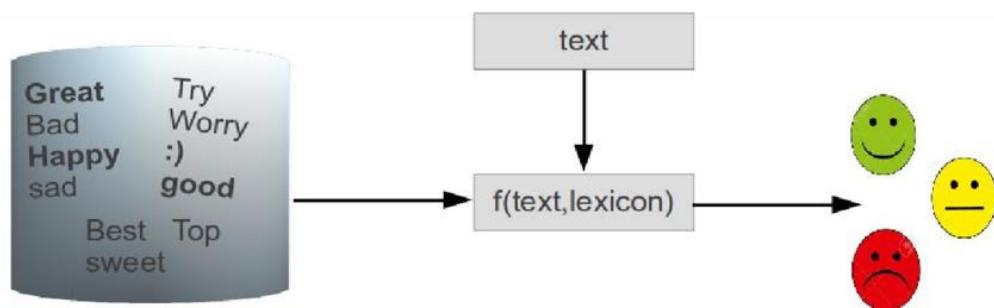
Analisis sentimen menggunakan pemrosesan bahasa alami (NLP), analisis teks, dan teknik komputasi untuk mengotomatiskan ekstraksi atau klasifikasi sentimen dari tinjauan sentimen. Analisis sentimen dan opini ini telah tersebar di banyak bidang seperti informasi konsumen, pemasaran, buku, aplikasi, situs web, dan sosial. Diperlukan pendekatan pemrograman untuk analisis sentimen dengan

menangani banyak submasalah terlibat dalam mengekstraksi makna dan polaritas dari teks. Analisis sentimen atau opini mining adalah bidang studi komputasi yang berkaitan dengan analisis opini, sentimen, emosi, penilaian, dan sikap terhadap berbagai entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik, dan atribut-atribut yang terkait. Awal dan pertumbuhan yang cepat dari bidang ini bertepatan dengan media sosial di Web, misalnya, ulasan, diskusi forum, blog, mikro blog, Twitter, dan jejaring sosial, karena untuk pertama kalinya dalam sejarah manusia. Sejak awal tahun 2000, analisis sentimen telah tumbuh menjadi salah satu wilayah penelitian paling aktif dalam pemrosesan bahasa alami (NLP). Itu juga banyak dipelajari dalam penambangan data, penambangan web, penambangan teks, dan pengambilan informasi. Nyatanya, memang demikian menyebar dari ilmu komputer ke ilmu manajemen dan ilmu sosial seperti pemasaran, keuangan, ilmu politik, komunikasi, ilmu kesehatan, bahkan sejarah, karena kepentingannya untuk bisnis dan masyarakat secara keseluruhan. Perkembangan ini disebabkan oleh fakta bahwa pendapat adalah pusatnya hampir semua aktivitas manusia dan merupakan pengaruh utama dari perilaku kita. Keyakinan dan persepsi tentang realitas, dan pilihan yang kita buat, pada tingkat tertentu, dikondisikan pada bagaimana orang lain melihat dan mengevaluasi dunia. Untuk alasan ini, kapan pun kita perlu membuat keputusan, kita sering mencari pendapat orang lain. Ini tidak hanya berlaku untuk individu tetapi juga untuk organisasi. Oleh karena itu, analisis sentimen secara otomatis seperti ini sangat diperlukan (Qi et al., 2020)

## 2.4 Metode Lexicon Based

Metode *Lexicon Based* adalah cara mengelompokkan kata-kata ke dalam kelompok sentimen positif dan sentimen negatif. Misalnya kata-kata seperti kategori “cantik”, “baik”, “pintar” dalam sekelompok kata yang memiliki sentimen positif, kemudian kata-kata seperti kategori “buruk”, “buruk”, “bodoh” dalam sekelompok kata yang memiliki sentimen negatif. Keberadaan sebuah kata dalam kelompok polaritas sentimental merepresentasikan implikasi emosi yang terkandung dalam kata tersebut. Polaritas sentimen suatu kata dalam kelompok sentimen dinyatakan dengan nilai yang menyatakan bobot hubungan kata tersebut

dengan kelompok sentimennya. Jadi, *lexicon* sentimen adalah sumber leksikal yang berisi informasi tentang emosi yang terkandung dalam beberapa kata (Christina & Ronaldo, 2020). Lexicon berisi daftar kata atau frase dan ini adalah sumber daya penting dalam analisis sentimen. Ada beberapa pendekatan yang benar membangun lexicon secara manual atau otomatis. Lexicon manual adalah waktu yang mahal dan tidak bekerja dengan semua domain sehingga *lexicon* otomatis menjadi topik penelitian yang lagi hangat karena mudah digunakan dan dikerjakan domain apa saja. Metode berbasis *lexicon* menghitung polaritas sentimen sebagai fungsi dari kata-kata yang memiliki sentimen melalui media sosial twitter (Bagheri & Islam, 2017). *Lexicon based* biasanya menggunakan kamus untuk mendukung klasifikasi sentimen yaitu *SentiWordNet* (Kusumawati, 2017). Alur metode *lexicon based* dapat dilihat dari gambar 2.1 dibawah ini :



**Gambar 2. 1 Alur Metode Lexicon Based (Kusumawati, 2017)**

Dalam proses klasifikasi sentimen menggunakan metode *lexicon based* dilakukan pada tiap kata dalam kalimat di dataset dengan *SentiWordNet*. Pemilihan kata yang memiliki lebih dari satu arti maka synset akan dilakukan berdasarkan metode *First Sense* dari *SentiWordNet* tersebut dengan memperhatikan mana yang muncul paling atas atau yang lebih popular. Setelah itu, kata yang berhasil di klasifikasi sesuai yang ada di *SentiWordNet* akan dilakukan pencarian nilai sentimen dalam satu kalimat dengan rumus sebagai berikut :

$$S_{positive} \sum_{i \in t}^{n} positive\ score_i \quad 1)$$

$$S_{negative} \sum_{i \in t}^{n} negative\ score_i \quad 2)$$

Dalam persamaan ini, ( $S_{positive}$ ) dan ( $S_{negative}$ ) mengacu pada bobot yang diberikan pada suatu kalimat berdasarkan jumlah skor polaritas positif dan negatif kata opini di dalamnya. Bobot ini digunakan sebagai acuan untuk membandingkan kalimat-kalimat yang berbeda. Dengan demikian, jumlah nilai positif ( $S_{positive}$ ) dan nilai negatif ( $S_{negative}$ ) dari setiap kata di dalam kalimat dapat dihitung, dan kemudian digunakan untuk menentukan orientasi sentimen suatu kalimat melalui Persamaan 3, yang membandingkan total jumlah nilai positif dan negatif.

$$Sentence_{sentiment} \begin{cases} positive & if S_{positive} > S_{negative} \\ neutral & if S_{positive} = S_{negative} \\ negative & if S_{positive} < S_{negative} \end{cases} 3)$$

Dari rumus diatas dapat ditentukan bahwa jika total jumlah nilai positif lebih besar dari jumlah nilai negatif maka kalimat memiliki sentimen positif. Jika negatif maka sebaliknya.

## 2.5. Metode Naive Bayes

*Bayesian classification* merupakan teknik klasifikasi statistik yang berguna untuk memperkirakan probabilitas keanggotaan suatu kelas. Teknik ini didasarkan pada teorema Bayes dan memiliki kemampuan klasifikasi yang mirip dengan decision tree dan neural network. *Bayesian classification* terbukti memiliki keakuratan dan kecepatan yang tinggi ketika diterapkan pada basis data dengan volume besar. (Purnama, P., & Supriyanto, C. (2013). Deteksi Penyakit Diabetes Tipe II Dengan Naive Bayes Berbasis Particle Swarm Optimization. Jurnal teknologi informasi, 9, 49-53.)

Metode Bayes adalah suatu pendekatan statistik untuk melakukan inferensi induksi pada masalah klasifikasi. Konsep dasar dan definisi pada Teorema Bayes dijelaskan terlebih dahulu, kemudian teorema ini digunakan untuk melakukan klasifikasi dalam Data Mining.

Teorema Bayes memiliki bentuk umum yang diterapkan dalam metode klasifikasi:

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

4)

Keterangan :

X = Data dengan class yang belum diketahui

H = Hipotesis data X merupakan suatu class spesifik

$P(H|X)$  = Probabilitas hipotesis H berdasarkan kondisi x (posteriori prob.)

$P(H)$  = Probabilitas hipotesis H (prior prob.)

$P(X|H)$  = Probabilitas X berdasarkan kondisi tersebut

$P(X)$  = Probabilitas dari X

Model *Naïve Bayes* adalah model pembelajaran mesin statistik tradisional yang banyak digunakan dalam banyak tugas penambangan data teks termasuk: klasifikasi sentimen ulasan produk. Meskipun cukup sederhana, ini menunjukkan efek kompetitif dibandingkan dengan beberapa model yang kompleks(Xu, F., Pan, Z., & Xia, R. (2020). E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. *Information Processing & Management*, 57(5), 102221.

## 2.6 Countvectorizer

Prosedur mengubah data teks mentah menjadi angka yang dapat dipahami disebut rekayasa fitur teks data. Pembelajaran mesin dan kinerja algoritma pembelajaran mendalam dan akurasi pada dasarnya tergantung pada jenis rekayasa fitur teknik yang digunakan(Kulkarni, A., & Shivananda, A. (2019). *Converting text to features. In Natural language processing recipes* (pp. 67-96). Apress, Berkeley, CA.).

*Count-Vectorizer* (CV) digunakan untuk mengubah kumpulan kalimat menjadi vektor. Setiap kata dalam grup sentimen akan dihitung. Kata dominan yang condong ke satu jenis sentimen akan menjadi anggota grup sentimen. Itu Metode *count vectorizer* hanya menganggap jumlah kata sebagai nilai fitur. Metode ini tidak mencerminkan dominasi kata dalam sebuah kalimat. Dominasi sebuah kata

dalam sebuah kalimat tidak hanya dihitung pada jumlah kejadian di kalimat, tetapi juga jumlah penampilan pada semua pendapat dalam kumpulan data (Aribowo, A. S., Basiron, H., Herman, N. S., & Khomsah, S. (2020). *An evaluation of preprocessing steps and tree-based ensemble machine learning for analysing sentiment on Indonesian youtube comments. International Journal of Advanced Trends in Computer Science and Engineering*, 7078-7086).

## 2.7 Flask

Framework Flask adalah kerangka kerja web dari Python bahasa. Flask menyediakan perpustakaan dan kumpulan kode yang dapat digunakan untuk membangun situs web, tanpa perlu melakukannya semuanya dari awal. Karena fitur-fiturnya yang sederhana, flask akan lebih ringan dan tidak tergantung pada banyak eksternal perpustakaan yang perlu diperhatikan. Secara umum, flask menyediakan ‘Werkzeug’ yang berguna untuk menerima permintaan (url) dan menanggapi (Mufid, M. R., Basofi, A., Al Rasyid, M. U. H., & Rochimansyah, I. F. (2019, September). Design an MVC model using python for flask framework development. In 2019 International Electronics Symposium (IES) (pp. 214-219). IEEE.) Flask memberikan kesederhanaan dan fleksibilitas dengan menerapkan server web minimal sebagai *micro framework*. Dasbor Flask mengumpulkan informasi tambahan tentang outlier seperti: jejak tumpukan Python, beban CPU, parameter permintaan, dll. Untuk memungkinkan pengelola menyelidiki penyebab waktu respons yang lambat (Vogel, P., Klooster, T., Andrikopoulos, V., & Lungu, M. (2017, September).

## 2.8 Penelitian Terdahulu

Penelitian ini merujuk pada studi yang dilakukan oleh Fajar Fathur Rachman dan Setia Pramana tentang analisis respon masyarakat terhadap wacana vaksinasi dengan menggunakan metode Latent Dirichlet Allocation (LDA). Dalam penelitian tersebut, respon masyarakat terhadap wacana vaksinasi diklasifikasikan menjadi respon positif atau negatif, dengan hasil menunjukkan bahwa sekitar 30% memberikan respon positif dan 26% memberikan respon negatif (Rachman & Pramana, 2020). Berdasarkan penelitian tersebut, penulis akan mengembangkan

penelitian lebih lanjut dengan melakukan analisis sentimen pengguna Twitter terhadap vaksin Covid-19. Dalam penelitian ini, respon positif dan negatif terhadap vaksinasi Covid-19 akan diklasifikasikan menggunakan Metode Lexicon Based dan naïve bayes. Metode Lexicon Based adalah suatu cara klasifikasi sentimen dengan membuat kamus data opini terlebih dahulu. Kata-kata yang ada pada kamus tersebut akan digunakan untuk proses identifikasi apakah suatu kalimat mengandung opini atau tidak. Seluruh proses pengerjaan dalam penelitian ini menggunakan Bahasa Pemrograman Python. Output Hasil yang diharapkan dari penelitian ini adalah untuk mengetahui jumlah opini positif atau negatif dari pengguna Twitter dalam menanggapi vaksin Covid-19. Adapun penulis mengambil beberapa penelitian terdahulu yang terkait dengan analisis sentimen diantaranya :

- Rachman, F. F., & Pramana, S. (2020) pada penelitian ini dijelaskan bahwa Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin *COVID-19* pada Media Sosial *Twitter*. Metode yang digunakan pada penelitian ini adalah metode LDA atau Latent Dirichlet Allocation.
- Raghupathi, V., Ren, J., & Raghupathi, W. (2020) pada penelitian ini dijelaskan terkait Mempelajari Persepsi Publik tentang Vaksinasi: Analisis Sentimen *Tweet*. Metode yang digunakan yaitu TF-IDF.
- Sharma, C., Whittle, S., Haghghi, P. D., Burstein, F., & Keen, H. (2020) pada penelitian ini dijelaskan tentang Analisis sentimen posting media sosial tentang farmakoterapi: Tinjauan pelingkupan. Adapun metode yang dipakai ialah gabungan dari *lexicon based* dan *machine learning*.
- Matošević, G., & Bevanda, V. (2020) pada penelitian ini membahas tentang Analisis sentimen tweet tentang penyakit COVID-19 selama pandemi dengan mengklasifikasikan sentimen di berbagai macam negara yaitu USA, UK, Italy, Spain, Germany and Sweden.
- Rahayu, R. N. & S. (2021) pada penelitian ini dijelaskan bahwa vaksin Covid 19 Di Indonesia terhadap Analisis Berita Hoax. Penelitian ini menganalisis dari berita-berita hoax yang ada di Indonesia terkait vaksin Covid-19.

Penulis mereferensikan beberapa penelitian sebelumnya secara singkat pada tabel 2.1 Penelitian Terdahulu berikut ini :

**Tabel 2. 1 Penelitian Terdahulu**

No	Peneliti	Tahun	Metode	Keterangan
1	Rachman, F., & Pramana, S	2020	<i>Latent Dirichlet Allocation</i> ( LDA )	Berdasarkan yang penulis baca hasil dari analisis sentimen yang dapat ditarik dari penelitian tersebut bahwa masyarakat lebih banyak memberikan respon yang positif terhadap vaksin <i>COVID-19</i> dibandingkan dengan respon yang negatif. Pengembangan yang akan penulis lakukan untuk penelitian ini adalah penulis akan mengklasifikasikan sentiment. Dengan seperti itu kita dapat mengetahui jumlah yang responnya positif atau negatif.
2	Raghupathi, V., Ren, J., & Raghupathi, W.	2020	<i>Lexicon Based</i>	Hasil pada penelitian ini adalah mengklasifikasikan analisis sentimen terkait vaksin dari media sosial X. Pengembangan yang akan penulis lakukan adalah melakukan metode lain untuk proses analisis sentimen yaitu dengan menggunakan metode Lexicon Based.
3	Sharma, C., Whittle, S., Haghghi, P. D., Burstein, F., & Keen, H.	2020	<i>A Scoping Review</i>	Hasilnya ialah analisis sentimen lebih cenderung mengarah ke respon masyarakat terkait adanya obat-obatan tertentu yang ada pada vaksin. Pengembangan yang penulis lakukan ialah dengan melakukan ulang scraping dataset pada Media Social X.

**Tabel 2. 1 Penelitian Terdahulu ( Lanjutan )**

No	Penulis	Tahun	Metode	Keterangan
4	Matošević, G., & Bevanda, V	2020	<i>Naive Bayes</i>	Hasil yang didapat pada penelitian ini ialah dengan mengklasifikasikan sentimen di berbagai macam negara yaitu USA, UK, Italy, Spain, Germany and Swede. Pengembangan yang penulis lakukan setelah membaca penelitian ini adalah dengan klasifikasi sentimen terkait Vaksin.
5	Rahayu, R. N. & S.	2021	Literature Rivew	Hasilnya adalah sebuah Analisa dari data media sosial X tentang berita hoax terkait vaksin. Pengembangan yang penulis lakukan yaitu dengan menggunakan hasil Analisa berita hoax tersebut dilakukan penulis mendapatkan ide untuk membuat analisis sentimen positif atau negatif respon masyarakat terhadap vaksin covid-19.
6	Xu, F., Pan, Z., & Xia, R	2020	<i>Naïve Bayes</i>	Hasilnya adalah menyajikan kerangka pembelajaran Naive Bayes berkelanjutan untuk produk e-commerce skala besar dan multi-domain review klasifikasi sentimen.

## **BAB 3**

### **ANALISIS DAN PERANCANGAN SISTEM**

#### **3.1 Dataset**

Dataset yang digunakan dalam penelitian ini diperoleh dari Media Social X dan terdiri dari 2879 komentar mengenai vaksin COVID-19. Data ini dikumpulkan dengan melakukan crawling Media Social X menggunakan kata kunci #vaksincovid19. Selanjutnya, data tersebut telah dibagi dengan rasio 80:20 untuk pengujian dan pelatihan, dengan 2164 data digunakan untuk pelatihan dan 542 data untuk pengujian. Dalam proses pengujian, 542 data akan digunakan untuk analisis lexicon-based, sedangkan 2164 data akan digunakan untuk proses pembelajaran menggunakan naive bayes. Pada tahap train-test split, data dibagi dengan perbandingan 80% untuk data pelatihan dan 20% untuk data pengujian, sesuai dengan rasio yang telah ditentukan.

Dataset yang digunakan dalam penelitian ini diperoleh dari Media Social X, yang terdiri dari 2879 komentar mengenai vaksin COVID-19. Data ini dikumpulkan dengan melakukan crawling pada Media Social X menggunakan kata kunci #vaksincovid19. Karena dataset ini berkenaan dengan Media Social X, basis datanya berupa teks yang mencerminkan berbagai tanggapan masyarakat terhadap vaksin COVID-19. Komentar-komentar ini dikategorikan secara manual berdasarkan sentimen yang diekspresikan dalam teks tersebut.

Input dalam penelitian ini adalah tanggapan masyarakat terhadap vaksin yang dikumpulkan dari tweet. Data telah dikategorikan secara manual ke dalam dua kelompok, yaitu positif dan negatif, berdasarkan sentimen yang diungkapkan dalam review. Review yang mengungkapkan kebahagiaan dan persetujuan dianggap positif (diberi label 1), sementara yang mengekspresikan kekecewaan, kemarahan, dan penolakan dianggap negatif (diberi label 0). Data ini akan digunakan untuk melatih sistem.

Contoh dataset dari Media Social X dapat dilihat pada Gambar 3.1 berikut ini.

No	Text
1	pasukan lemvar dara babie udah blom wkwkwk yg syahid lawan negara kt najiiiz 😊😊
2	program utk warga emas batu pahat khemah bangsa cina je melayu 🇲🇾
3	glat pawas ipda jsiburian beserta piked fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp
4	glat pawas ipda jsiburian beserta piked fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp
5	az lsinya virus nonaktif kalo demam habis emg antibodi be
6	setakat pekerja genting je staff gomen berkaitan dgn frontl
7	semoga cepet dapet semoga cepet bebas main main
8	gk btuh cman btuh kehadiran
9	juta data website jkjav sync dgn data mysejahtera orang dah tap
10	yampun nder buruan obatin ya bersihin lukanya megangin ya huhu biar aja
11	bangunan parliment tu dah subrental at least jana pendapatan dibiarka
12	menyediakan yg keje casino dpt tp cikgu baki frontliner x dicucuk
13	ayahanda kepala tiger abanglong abangah ayah cik tokngh meon gangster melayu d
14	bahas sinovac mantan menkes siti fadilah divaksin gak divaksin kena
15	jujur kudu dikaji ulang pemberiannya yg meninggal msh dah ratusan juta
16	islam tu kiamatyg pas 🌟 cabutan khas judi drpd dinai
17	organisasi medis n nakes tlk dilitik penentuan jenis
18	mandiri dok yg udah karyawan bumn anak perusahaan
19	petugas kesehatan berpesan menjaga kesehatan menerapkan protokol kesehatan dianjurkan pe
20	tensi gue trus tenggepan orangnya rendah ya tensinya gatau aja normal gue wkkwkwk wkt pas
21	sbb jenis ni kerajaan beli then diorsing sambung program vaksinasi so lebihan vak

**Gambar 3. 1 Dataset Media Social X**

Contoh dari sebuah review diilustrasikan dalam Tabel 3.1 berikut ini.

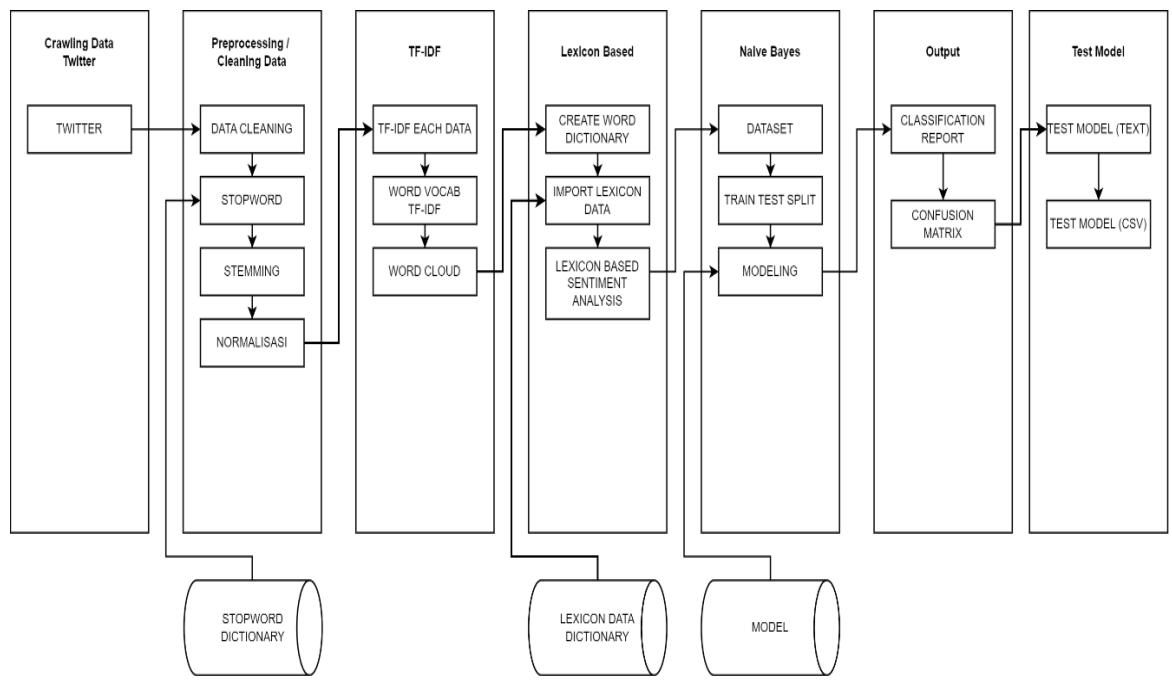
**Tabel 3. 1 Dataset Media Social X**

Kalimat	Label
pasukan lemvar dara babie udah blom wkwkwk yg syahid lawan negara kt najiiiz	0
semoga cepet dapet semoga cepet bebas main main	1

### 3.2 Arsitektur Umum

Langkah kerja penelitian ini dilakukan dengan mengikuti alur kerangka kerja yang telah dibuat agar penelitian ini lebih terstruktur. Tahapan penelitian dimulai dengan mengumpulkan dataset melalui proses crawling data di Media Social X. Setelah dataset didapatkan, langkah selanjutnya adalah preprocessing/cleaning data yang terdiri dari beberapa tahap, yaitu data cleaning, stopword removal, stemming, dan normalisasi. Setelah itu, data masuk ke proses TF-IDF untuk pembobotan kata. Selanjutnya, dilakukan word vocabulary TF-IDF dan word cloud TF-IDF. Setelah itu, data masuk ke proses lexicon based, yang terdiri dari pembuatan kamus kata, impor data lexicon, dan analisis sentimen berbasis lexicon. Setelah proses ini selesai, langkah berikutnya adalah proses naive bayes, dimana dataset dibagi menjadi data training dan data testing, dilanjutkan dengan proses modeling.

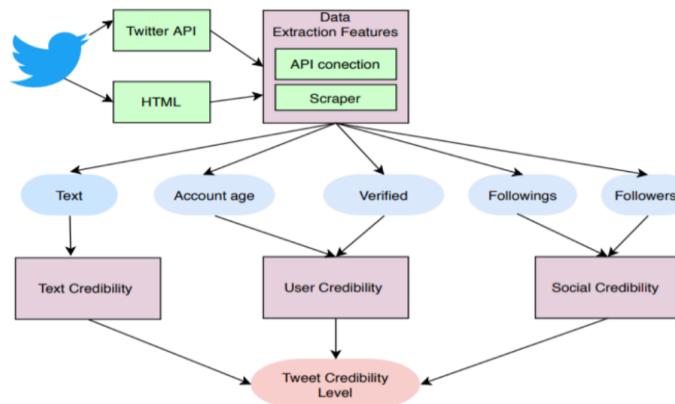
Setelah itu, langsung dilakukan proses output, yaitu classification report dan confusion matrix. Setelah mendapatkan hasil dari proses tersebut, masuk ke proses test model, baik melalui uji model teks maupun uji model CSV. Arsitektur umum pada penelitian ini dapat dilihat pada gambar 3.2 berikut ini.



**Gambar 3. 2 Arsitektur Umum Penelitian**

### 3.2.1 *Crawling/Scraping Data*

Pada tahap ini, dilakukan pengumpulan data dari media sosial X menggunakan API dan pustaka Python yang disediakan. Data yang diambil meliputi username, teks tweet, lokasi, dan lain-lain. Berikut proses crawling/scraping data X dapat dilihat pada gambar 3.3.



**Gambar 3. 3 Proses Crawling/Scraping Data X (Dongo et al., 2020).**

Pada situs twitter, para developer diberikan akses untuk dapat mengambil data twitter yang ada dengan menggunakan API. Dengan adanya API maka akan terhubung ke server database twitter untuk bisa melakukan *scraping* data. Selain menggunakan API juga terdapat cara lain yaitu menggunakan Selenium, cara ini disebut teknik *crawling*. Pada penelitian ini penulis hanya menggunakan data yang penting yaitu teks tweet dalam media social X. Contoh hasil data crawling bisa dilihat pada Gambar 3.4 berikut ini.

1	text								
2	pasukan lemvar dara babie udah blom wkwkwk yg syahid lawan negara kt najiiiz öÿ^n öÿ^n								
3	program utk warga emas batu pahat khemah bangsa cina je melayu öÿ^n c								
4	giat pawas ipda jsiburian beserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp								
5	giat pawas ipda jsiburian beserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp								
6	semoga kenangan								
7	az isinya virus nonaktif kalo demam habis emg antibodi be								
8	setakat pekerja genting je staff gomen berkaitan dgn frontl								
9	semoga cepet dapet semoga cepet bebas main main								
10	gk btuh cman btuh kehadiran								
11	juta data website jkjav sync dgn data mysejahtera orang dah tap								

**Gambar 3.4 Hasil Data Crawling**

Gambar tersebut mengilustrasikan proses data crawling untuk menganalisis sentimen pengguna X terhadap vaksinasi COVID-19. *Metode Lexicon Based* dan *Naive Bayes* digunakan untuk menganalisis teks dari X. Dengan demikian, dapat diperoleh pemahaman tentang pandangan masyarakat terhadap vaksinasi COVID-19 berdasarkan data X.

### 3.2.2 Preprocessing Data

#### 3.2.2.1 Cleaning

Pada tahap awal pembersihan data, langkah pertama adalah melakukan proses pembersihan. Pembersihan data merupakan tahap penting dalam analisis data yang bertujuan untuk meningkatkan kualitas dan kehandalan data yang digunakan. Proses ini melibatkan berbagai teknik untuk menangani masalah umum dalam dataset, seperti data yang hilang, tidak valid, atau outlier.

Hasil tahap Cleaning dapat dilihat pada gambar 3.5 berikut ini.

Cleaning
pasukan lemvar dara babie udah blom wkwkwk yg syahid lawan negara kt najiiiz
program utk warga emas batu pahat khemah bangsa cina je melayu
giat pawas ipda jsiburian beserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp
giat pawas ipda jsiburian beserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp
az isinya virus nonaktif kalo demam habis emg antibodi be
setakat pekerja genting je staff gomen berkaitan dgn frontl
semoga cepet dapet semoga cepet bebas main main
gk btuh cman btuh kehadiran
juta data website jkjav sync dgn data mysejahtera orang dah tap
yampun nder buruan obatin ya bersihin lukanya megangin ya huhu biar aja

**Gambar 3. 5 Tahap Cleaning**

### 3.2.2.2. *Remove Stopwords*

Ini adalah teknik yang menghilangkan kata-kata yang sering digunakan yang tidak berarti dan tidak berguna untuk klasifikasi teks. Ini mengurangi ukuran korpus tanpa kehilangan informasi penting.

Hasil tahap Stopwords dapat dilihat pada gambar 3.6 berikut ini.

Stopwords	↑↓
pasukan lemvar dara babie udah blom wkwkwk yg syahid lawan negara kt najiiiz	
program utk warga emas batu pahat khemah bangsa cina je melayu	
giat pawas ipda jsiburan beserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp	
giat pawas ipda jsiburan beserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp	
az isinya virus nonaktif kalo demam habis emg antibodi be	
setakat pekerja genting je staff gomen berkaitan dgn frontl	
semoga cepet dapet semoga cepet bebas main main	
gk btuh cman btuh kehadiran	
juta data website jkjav sync dgn data mysejahtera orang dah tap	
yampun nder buruan obatin bersihin lukanya megangin huhu biar aja	

**Gambar 3. 6 Tahap Stopwords**

### 3.2.2.3 *Stemming*

Stemming adalah proses menghapus awalan atau akhiran dari sebuah kata yang terdapat imbuhan untuk mengubah ke bentuk kata dasarnya. Stemming bekerja dengan membuang akhiran kata, contoh kata ('beserta' jadi 'serta') menurut beberapa tata bahasa aturan dan mendapatkan kata dasar dari kata tersebut.

Tahap *stemming* dapat dilihat pada gambar 3.7.

Stemming	↑↓
pasu lemvar dara babie udah blom wkwkwk yg syahid lawan negara kt najiiiz	
program utk warga emas batu pahat khemah bangsa cina je melayu	
giat pawas ipda jsiburian serta piket fungsi laksana patroli monitoring kantor dpcipk dpcpp	
giat pawas ipda jsiburian serta piket fungsi laksana patroli monitoring kantor dpcipk dpcpp	
az isi virus nonaktif kalo demam habis emg antibodi be	
takat kerja genting je staff gomen kait dgn frontl	
moga cepet dapet mogea cepet bebas main main	
gk btuh cman btuh hadir	
juta data website jkjav sync dgn data mysejahtera orang dah tap	
yampun nder buru obatin bersihin luka megangin huhu biar aja	

**Gambar 3.7 Tahap Stemming**

#### **3.2.2.4 Normalisasi**

Normalisasi berfungsi dengan mengubah kata ke bentuk standar atau dasar sesuai dengan aturan tata bahasa tertentu, kata yang tidak normal/tidak baku dihilangkan agar memudahkan saat dilakukan proses analisis data. Tujuan normalisasi data untuk meningkatkan kualitas data dengan menyusun data sedemikian rupa sehingga dapat meningkatkan akurasi, konsistensi, dan keandalan.

Hasil proses normalisasi dapat dilihat pada gambar 3.8 berikut ini.

Normalisasi	↑↓
pasu lemvar dara syahid lawan negara	
program warga emas batu pahat bangsa cina melayu	
giat serta piket fungsi laksana patroli kantor	
giat serta piket fungsi laksana patroli kantor	
isi virus nonaktif kalo demam habis antibodi	
takat kerja genting staff kait	
moga mogga bebas main main	
hadir	
juta data data orang dah tap	
buru luka biar aja	

**Gambar 3. 8 Tahap Normalisasi**

### 3.2.3 TF-IDF

TF-IDF merupakan salah satu teknik pembobotan kata yang digunakan untuk menunjukkan tingkat relevansi sebuah kata dalam sebuah dokumen. Pembobotan dilakukan dengan menggunakan kata-kata unik dalam dokumen untuk menghasilkan model yang nantinya digunakan dalam klasifikasi Naive Bayes. Nilai bobot TF semakin besar jika sebuah kata muncul dalam dokumen secara banyak, sedangkan nilai IDF semakin besar jika sebuah kata muncul dalam jumlah dokumen yang sedikit.

Nilai minimal pembobotan kata dapat diatur dengan menggunakan parameter min\_df=1 dan ngram\_range=(1,1). Penentuan nilai bobot yang lebih akurat dapat dilakukan dengan menentukan parameter tertentu (G.A. Dalaorao, et al., 2019). Pada penelitian ini, teknik TF-IDF diterapkan dengan menggunakan unigram ( $n=1$ ). N-gram adalah teknik ekstraksi fitur TF-IDF yang menggabungkan  $n$  kata menjadi satu bagian dengan memisahkan kata-kata dalam kalimat. Sebagai contoh, kata "semoga cepat dapet semoga cepat bebas main-main" akan diubah menjadi ['semoga', 'cepat', 'dapet', 'bebas', 'main'] menggunakan teknik unigram TF-IDF. Gambar 3.9, Tabel 3.2, Tabel 3.3, Tabel 3.4 dan Tabel 3.5 menampilkan contoh penerapan teknik unigram TF-IDF untuk menghitung skor TF, DF, dan IDF pada dua dokumen review yang berbeda.

Dokumen 1 : pasukan lemvar dara babie udah blom wkwkwk yg syahid lawan negara kt najiiiz

Dokumen 2 : semoga cepet dapet semoga cepet bebas main main

Label TF-IDF dapat dilihat pada gambar 3.9.

### Gambar 3.9 Label TF-IDF

Term Frequency adalah ukuran seberapa sering sebuah kata muncul dalam sebuah dokumen. Dihitung sebagai rasio jumlah kali suatu kata muncul dalam suatu dokumen terhadap total jumlah kata dalam dokumen tersebut.

Berikut perhitungan TF :

Rumus:

$$TF(t, d) = \frac{\text{Jumlah kali kata } t \text{ muncul dalam dokumen } d}{\text{total jumlah kata dalam dokumen } d}$$

$$\begin{aligned} TF \text{ Pasukan} &= \frac{D1+D2}{\text{total jumlah kata}} \\ &= \frac{1}{18} \\ &= 0,05 \end{aligned}$$

**Tabel 3. 2 Kemunculan Term Pada Dokumen**

Term	D1	D2	Skor TF
pasukan	1	0	1/18 = 0.05
lemvar	1	0	1/18 = 0.05
dara	1	0	1/18 = 0.05
babie	1	0	1/18 = 0.05
udah	1	0	1/18 = 0.05
blom	1	0	1/18 = 0.05
wkwickw	1	0	1/18 = 0.05
yg	1	0	1/18 = 0.05
syahid	1	0	1/18 = 0.05
lawan	1	0	1/18 = 0.05
negara	1	0	1/18 = 0.05
kt	1	0	1/18 = 0.05
najiiiz	1	0	1/18 = 0.05
semoga	0	2	2/18 = 0.11
cepet	0	2	2/18 = 0.11
dapet	0	1	1/18 = 0.05
bebas	0	1	1/18 = 0.05
main	0	2	2/18 = 0.11

Jumlah dokumen di mana suatu term muncul dijadikan acuan dalam menghitung nilai df, yang ditunjukkan dalam tabel 3.3

**Tabel 3. 3 Skor Df**

Term	D1	D2	Skor DF
pasukan	1	0	1
lemvar	1	0	1
dara	1	0	1
babie	1	0	1
udah	1	0	1
blom	1	0	1
wkwickw	1	0	1
yg	1	0	1
syahid	1	0	1

**Tabel 3.3 Skor Df (Lanjutan)**

<b>Term</b>	<b>D1</b>	<b>D2</b>	<b>Skor DF</b>
lawan	1	0	1
negara	1	0	1
kt	1	0	1
najiiiz	1	0	1
semoga	0	2	2
cepet	0	2	2
dapet	0	1	1
bebas	0	1	1
main	0	2	2

Perhitungan nilai idf, ditunjukkan tabel 3.4

Inverse Document Frequency adalah ukuran seberapa penting sebuah kata di seluruh korpus dokumen. Dihitung sebagai logaritma dari rasio total jumlah dokumen terhadap jumlah dokumen yang mengandung kata tersebut, dengan tambahan nilai penyelarasan untuk menghindari pembagian dengan nol.

Berikut perhitungan skor IDF :

$$IDF(t, D) = \log\left(\frac{\text{Total jumlah dokumen dalam korpus } D}{\text{jumlah dokumen yang mengandung kata } t+1}\right) + 1$$

$$IDF \text{ Pasukan} = \log\left(\frac{18}{1+1}\right) + 1$$

**Tabel 3.4 Skor IDF**

<b>Term</b>	<b>Skor DF</b>	<b>Skor IDF</b>
pasukan	1	$\log(18/(1+1)) + 1 = 1$
lemvar	1	$\log(18/(1+1)) + 1 = 1$
dara	1	$\log(18/(1+1)) + 1 = 1$
babie	1	$\log(18/(1+1)) + 1 = 1$
udah	1	$\log(18/(1+1)) + 1 = 1$
blom	1	$\log(18/(1+1)) + 1 = 1$
wkwkwk	1	$\log(18/(1+1)) + 1 = 1$
yg	1	$\log(18/(1+1)) + 1 = 1$
syahid	1	$\log(18/(1+1)) + 1 = 1$
lawan	1	$\log(18/(1+1)) + 1 = 1$
negara	1	$\log(18/(1+1)) + 1 = 1$
kt	1	$\log(18/(1+1)) + 1 = 1$
najiiiz	1	$\log(18/(1+1)) + 1 = 1$
semoga	2	$\log(18/(2+1)) + 1 = 0.84$
cepet	2	$\log(18/(2+1)) + 1 = 0.84$
dapet	1	$\log(18/(1+1)) + 1 = 1$
bebas	1	$\log(18/(1+1)) + 1 = 1$
main	2	$\log(18/(2+1)) + 1 = 0.84$

Berikut adalah skor tf-idf yang dihasilkan bisa dilihat pada tabel 3.5 :

**Tabel 3. 5 Nilai TF-IDF**

Term	Skor TF	Skor IDF	Skor TF – IDF
pasukan	0.05	1	1.05
lemvar	0.05	1	1.05
dara	0.05	1	1.05
babie	0.05	1	1.05
udah	0.05	1	1.05
blom	0.05	1	1.05
wkwkwk	0.05	1	1.05
yg	0.05	1	1.05
syahid	0.05	1	1.05
lawan	0.05	1	1.05
negara	0.05	1	1.05
kt	0.05	1	1.05
najiiiz	0.05	1	1.05
semoga	0.11	0.84	0.95
cepet	0.11	0.84	0.95
dapet	0.05	1	1.05
bebas	0.05	1	1.05
main	0.11	0.84	0.95

Proses melabelkan pada TF-IDF merupakan langkah kritis dalam analisis teks yang melibatkan penghitungan berbagai metrik untuk setiap kata atau term dalam sebuah dokumen. Pertama-tama, TF atau Term Frequency mengukur seberapa sering sebuah kata muncul dalam dokumen tertentu, dihitung sebagai rasio jumlah kemunculan kata tersebut terhadap total kata dalam dokumen. Misalnya, kata "pasukan" muncul 1 kali dalam Dokumen 1 dari total 18 kata, sehingga memiliki skor TF 0.05. Selanjutnya, Document Frequency (DF) menghitung berapa banyak dokumen dalam korpus yang mengandung kata tertentu. Kata "pasukan" memiliki DF 1 karena muncul dalam D1. IDF atau Inverse Document Frequency mengukur pentingnya sebuah kata di seluruh korpus dokumen dengan mempertimbangkan berapa banyak dokumen yang mengandung kata tersebut. Nilai IDF untuk kata "semoga" adalah 0.84, yang menunjukkan bahwa kata tersebut penting dalam dokumen yang mengandungnya. Terakhir, TF-IDF adalah hasil perkalian antara TF dan IDF untuk setiap kata, digunakan untuk menimbang pentingnya kata dalam konteks dokumen dan korpus secara keseluruhan. Proses ini menghasilkan skor TF-IDF untuk setiap term dalam

dokumen-dokumen yang dianalisis, memberikan landasan bagi model klasifikasi untuk membedakan dan memahami pola sentimen atau topik dari teks yang dianalisis.

### **3.2.4 Lexicon Based**

Fitur berbasis *lexicon*. Fitur-fitur ini dirancang berdasarkan pada intuisi bahwa kata-kata yang mengandung sentimen/emosi yang diidentifikasi oleh *lexicon* dapat membentuk pengetahuan yang berguna untuk mewakili dokumen untuk klasifikasi emosi (Bandhakavi, A., Wiratunga, N., Padmanabhan, D., & Massie, S. (2017). *Lexicon based feature extraction for emotion text classification. Pattern recognition letters*, 93, 133-142.). Klasifikasi sentimen dengan metode *Lexicon Based* dilakukan dengan mempertimbangkan kata-kata positif, negatif, dan netral yang terdapat pada tweet setelah melalui proses pembersihan. Dalam metode ini, digunakan kamus *lexicon* Bahasa Indonesia untuk mencocokkan kata-kata dalam tweet dengan entri yang ada dalam kamus tersebut. Jika tweet mengandung kata-kata yang termasuk dalam kategori positif, maka tweet tersebut akan dianggap memiliki sentimen positif. Sebaliknya, jika tweet mengandung kata-kata negatif, maka tweet tersebut akan dianggap memiliki sentimen negatif. Namun, jika jumlah kata-kata positif dan negatif dalam tweet sama, maka tweet tersebut akan diklasifikasikan sebagai sentimen netral.

Pseudocode saat proses *Lexicon Based* dapat dilihat di gambar 3.10 berikut.

```
$kata = str_word_count($ulasan, 1);
foreach ($kata as $kata_evaluasi) {
    if (isset($lexicon[$kata_evaluasi])) {
        $skor_sentiment +=
$lexicon[$kata_evaluasi];
        $jumlah_kata++;

        if ($lexicon[$kata_evaluasi] > 0) {
            $kata_positif++;
        } elseif ($lexicon[$kata_evaluasi] < 0) {
            $kata_negatif++;
        }
    } else {
        $jumlah_kata++;
    }
}
```

**Gambar 3. 10 Pseudocode Implementasi Lexicon Based**

Pada penelitian ini, fokus utama terletak pada analisis sentimen pengguna X terkait vaksinasi COVID-19. Memahami respon dan pendapat masyarakat terhadap isu kesehatan global ini menjadi krusial dalam konteks pandemi. Langkah pertama dalam menyusun pemahaman tersebut adalah melalui proses pengumpulan data dari platform X, di mana beragam tweet yang membahas vaksinasi COVID-19 dikumpulkan sebagai sumber informasi utama.

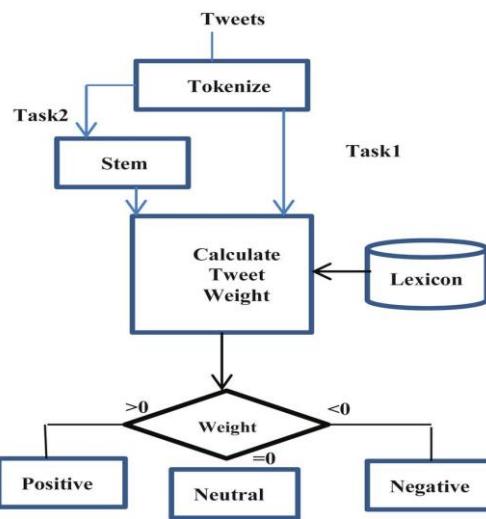
Setelah berhasil mengumpulkan data, langkah berikutnya adalah proses tokenisasi, yang bertujuan untuk merinci setiap tweet ke dalam kata-kata individual. Dengan kata-kata sebagai unit dasar, analisis sentimen dapat lebih terfokus pada makna dan konteks kata dalam menyampaikan pesan. Proses ini menjadi fondasi penting dalam memahami nuansa sentimen yang terkandung dalam setiap tweet.

Tahap selanjutnya melibatkan analisis lexicon based, di mana setiap kata dalam tweet diberi bobot berdasarkan nilai sentimen positif dan negatif yang terdapat dalam lexicon. Task 1 dan Task 2 memperlihatkan dua pendekatan: satu tanpa stemming dan satu dengan proses stemming. Task 1 menghitung bobot kata langsung dari lexicon, sementara Task 2 melibatkan proses stemming sebelum menghitung bobot kata. Kedua pendekatan ini memberikan pemahaman lebih mendalam tentang pengaruh proses stemming terhadap hasil analisis sentimen.

Setelah mendapatkan bobot total tweet, langkah selanjutnya adalah pengklasifikasian sentimen menggunakan bobot tersebut. Pada tahap ini, kriteria yang digunakan sangat sederhana: jika bobot total sama dengan 0, tweet dianggap netral; jika lebih dari 0, tweet dianggap positif; dan jika kurang dari 0, tweet dianggap negatif. Tahap akhir melibatkan evaluasi performa dengan menghasilkan hasil klasifikasi dan akurasi model Naive Bayes.

Dengan rinciannya, alur analisis sentimen ini diharapkan dapat memberikan gambaran yang komprehensif tentang bagaimana metode lexicon based dan Naive Bayes digunakan untuk menganalisis sentimen pengguna X terkait vaksinasi COVID-19. Keseluruhan proses ini dirancang untuk menggali dan memahami respons masyarakat melalui medium media sosial, memberikan kontribusi berharga dalam memahami dinamika opini dan sikap terkini dalam menghadapi tantangan kesehatan global.

Alur saat proses Lexicon Based dapat dilihat pada gambar 3.11 berikut ini.



**Gambar 3. 11 Alur saat proses Lexicon Based**

Penjelasan alur saat proses Lexicon Based

1. Pendahuluan:

Memperkenalkan topik penelitian dan latar belakang pentingnya menganalisis sentimen pengguna X terkait vaksinasi COVID-19.

2. Pengumpulan Data:

Mengumpulkan data tweets terkait vaksinasi COVID-19 dari platform X.

3. Tokenisasi:

Melakukan tokenisasi pada setiap tweet untuk memecah kalimat menjadi kata-kata individual.

4. Task 1: Calculate Tweet Weight (Tanpa Stemming):

- Menghitung bobot setiap kata dalam tweet berdasarkan lexicon. Lexicon berisi kata-kata bersama dengan nilai sentimen positif dan negatifnya.
- Setiap kata dalam tweet diberi bobot berdasarkan lexicon, dan bobot tersebut dijumlahkan untuk mendapatkan bobot total tweet.

5. Task 2: Stemming ke Calculate Tweet Weight:

- Melakukan proses stemming pada kata-kata dalam tweet untuk mengubah kata-kata menjadi bentuk dasarnya.
- Menghitung bobot setiap kata yang telah distem, menggunakan lexicon yang sama seperti pada Task 1.
- Menjumlahkan bobot kata-kata untuk mendapatkan bobot total tweet yang telah distem.

## 6. Lexicon ke Calculate Tweet Weight:

- Menggunakan lexicon untuk menghitung bobot kata-kata dalam tweet.
- Masing-masing kata dalam tweet diberi bobot berdasarkan nilai sentimen yang terdapat dalam lexicon.
- Menjumlahkan bobot kata-kata untuk mendapatkan bobot total tweet.

## 7. Weight:

- Mengklasifikasikan sentimen tweet berdasarkan bobot total yang telah dihitung sebelumnya.
- Jika bobot total sama dengan 0, tweet dianggap netral.
- Jika bobot total lebih dari 0, tweet dianggap positif.
- Jika bobot total kurang dari 0, tweet dianggap negatif.

## Perhitungan Lexicon Based

### 1. Persiapkan Data

Persiapan Data bisa dilihat pada tabel 3.6 berikut :

**Tabel 3. 6 Sample Dataset**

Kalimat	Label
pasukan lemvar dara babie udah blom wkwkwk yg syahid lawan negara kt naiiiiz	Negatif
semoga cepet dapet semoga cepet bebas main main	Positif
semoga kenangan	Netral

### 2. Kamus Sentiment

Kamus sentimen tersebut merupakan hasil dari proses pembuatan kamus sentimen berdasarkan analisis atau pendekatan tertentu dalam penelitian terkait. Kamus sentimen ini tidak ditunjukkan secara spesifik dalam percakapan kita sebelumnya, tetapi secara umum, kamus sentimen seperti ini dibuat dengan mengumpulkan kata-kata serta penilaian sentimen dari sumber-sumber yang relevan.

Misalnya, kata-kata seperti "pasukan", "lemvar", "dara", "babie", dan lainnya diberi nilai -1 karena dalam konteks yang dianalisis, kata-kata ini cenderung memiliki konotasi negatif atau dapat mengindikasikan

sentimen negatif terkait dengan topik tertentu, misalnya vaksinasi COVID-19. Sebaliknya, kata-kata seperti "semoga", "cepet", "dapat", "bebas", dan "main" diberi nilai +1 karena cenderung memiliki konotasi positif atau mendukung terhadap topik yang sama. Penentuan nilai sentimen untuk setiap kata dalam kamus tersebut dapat berdasarkan analisis manual oleh peneliti atau menggunakan algoritma pemrosesan bahasa alami (natural language processing) yang mengklasifikasikan kata-kata berdasarkan konteksnya. Namun, perlu dicatat bahwa kamus sentimen seperti ini dapat bervariasi tergantung pada konteks dan tujuan penelitian, serta dapat disesuaikan atau diperbarui berdasarkan hasil analisis dan umpan balik dari data yang dianalisis (Bandhakavi, A., Wiratunga, N., Padmanabhan, D., & Massie, S. (2017). *Lexicon based feature extraction for emotion text classification. Pattern recognition letters*, 93, 133-142.).

Contoh kamus sentimen dari sample dataset dapat dilihat pada tabel 3.7 berikut :

**Tabel 3. 7 Kamus Sentimen**

Kamus Sentiment	
pasukan	-1
lemvar	-1
dara	-1
babie	-1
udah	-1
blom	-1
wkwkwk	-1
yg	-1
syahid	-1
lawan	-1
negara	-1
kt	-1
najiiiz	-1

Kamus Sentiment	
semoga	+1
cepet	+1
dapet	+1
bebas	+1
main	+1

### 3. Tokenisasi Teks

Tokenisasi dapat dilihat pada tabel 3.8 berikut :

**Tabel 3. 8 Tokenisasi Teks**

Teks	Token
semoga cepet dapet semoga cepet bebas main main	"semoga", "cepet", "dapet", "bebas", "main".

### 4. Menghitung Skor Sentimen

$$\text{Skor Sentimen} = \text{Sentimen}(\text{"semoga"}) + \text{Sentimen}(\text{"cepet"}) + \text{Sentimen}(\text{"dapet"}) + \text{Sentimen}(\text{"bebas"}) + \text{Sentimen}(\text{"main"})$$

### 5. Akumulasi Skor Sentimen

Misalkan X adalah skor sentimen

$$X = \text{Sentimen}(\text{"semoga"}) + \text{Sentimen}(\text{"cepet"}) + \dots + \text{Sentimen}(\text{"main"})$$

$$X = 0+0+0+1+0+0+0 = 1$$

### 6. Interpretasi Hasil

- Jika  $X > 0$ , prediksi dapat dianggap sebagai sentimen positif.
- Jika  $X < 0$ , prediksi dapat dianggap sebagai sentimen negatif.
- Jika  $X = 0$ , prediksi dapat dianggap sebagai sentimen netral.

Dengan hasil  $X=1$  (positif), dapat disimpulkan bahwa sentimen dari teks baru " semoga cepet dapet semoga cepet bebas main main " adalah positif berdasarkan perhitungan Lexicon-Based.

### 3.2.5 Naive Bayes

Klasifikasi Naive Bayes adalah jenis klasifikasi yang termasuk dalam supervised learning, di mana ada pengajar atau supervisor yang melakukan klasifikasi manual pada data pelatihan (Gunawan, B., Sastypratiwi, H., & Pratama, E. E. (2018). Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes. JEPIN (Jurnal Edukasi dan Penelitian Informatika), 4(2), 113-118.) Konsep dasar dan definisi Teorema Bayes dibahas terlebih dahulu sebelum digunakan untuk melakukan klasifikasi dalam Data Mining. Teorema Bayes memiliki bentuk umum tertentu, yang digunakan dalam Naive Bayes untuk memprediksi kategori atau label pada data yang belum dikenal. Teorema Bayes memiliki bentuk umum sebagai berikut :

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

Keterangan :

X = Data dengan class yang belum diketahui

H = Hipotesis data X merupakan suatu class spesifik

$P(H|X)$  = Probabilitas hipotesis H berdasarkan kondisi x (posteriori prob.)

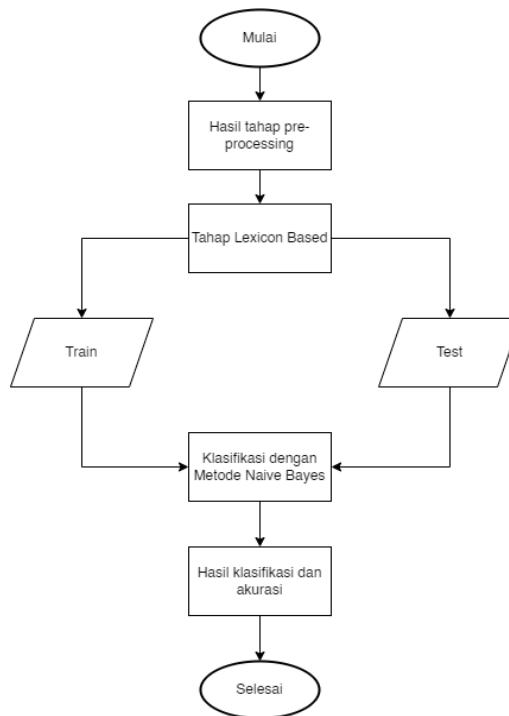
$P(H)$  = Probabilitas hipotesis H (prior prob.)

$P(X|H)$  = Probabilitas X berdasarkan kondisi tersebut

$P(X)$  = Probabilitas dari X

Naive Bayes adalah tahap klasifikasi untuk menentukan apakah data yang akan diuji termasuk dalam sentimen positif atau negatif. Pada tahap ini, menggunakan metode Naïve Bayes Classifier dengan melakukan pelabelan terlebih dahulu. Metode ini terdiri dari dua proses, yaitu proses pelatihan dan pengujian. Naive Bayes adalah metode klasifikasi yang digunakan dalam berbagai aplikasi seperti analisis sentimen, deteksi spam, klasifikasi dokumen, dan diagnosa medis. Dalam analisis sentimen, Naive Bayes mengklasifikasikan teks seperti ulasan produk atau tweet menjadi sentimen positif atau negatif. Untuk deteksi spam, metode ini menilai email berdasarkan kata-kata di dalamnya untuk menentukan apakah email tersebut adalah spam. Dalam klasifikasi dokumen, Naive Bayes mengelompokkan artikel berita atau dokumen lain ke dalam kategori

tertentu seperti politik, olahraga, atau teknologi. Di bidang medis, metode ini membantu mendiagnosis penyakit berdasarkan gejala yang ada. Naive Bayes bekerja dengan menghitung probabilitas suatu kelas berdasarkan data pelatihan yang telah diberi label, dan kemudian menggunakan model tersebut untuk memprediksi kelas data baru. Meskipun asumsi independensinya jarang sepenuhnya benar, metode ini tetap efektif dan mudah diimplementasikan. Alur proses yang terjadi pada Naive Bayes dapat dilihat pada gambar 3.12 berikut.



**Gambar 3. 12 Proses Dalam Naïve Bayes**

Penjelasan Alur Proses dalam Naïve Bayes :

1. Mulai:

Pengenalan topik penelitian dan tujuan analisis sentimen terhadap pengguna X terkait vaksinasi COVID-19.

2. Tahap Preprocessing:

a. Pengumpulan Data:

Mengumpulkan data tweets terkait vaksinasi COVID-19 dari platform X.

b. Tokenisasi:

- Memecah kalimat-kalimat dalam setiap tweet menjadi kata-kata individual.
- Menghapus karakter khusus, tanda baca, dan tautan yang tidak relevan.

c. Stemming:

Mereduksi kata-kata dalam tweet ke bentuk dasarnya untuk menyederhanakan analisis.

3. Tahap Lexicon Based:

a. Calculate Tweet Weight (Lexicon Based):

- Menggunakan lexicon (kamus kata-kata dengan nilai sentimen) untuk menghitung bobot setiap kata dalam tweet.
- Menjumlahkan bobot kata-kata untuk mendapatkan bobot total tweet.

4. Tahap Test dan Train:

a. Pembagian Dataset:

Membagi dataset menjadi dua bagian: data pelatihan (train set) dan data pengujian (test set).

b. Pelatihan Model Naive Bayes:

- Menggunakan data pelatihan untuk melatih model Naive Bayes.
- Model akan belajar distribusi probabilitas dari kata-kata terhadap sentimen positif atau negatif.

5. Tahap Klasifikasi menggunakan Metode Naive Bayes:

a. Klasifikasi Tweet:

- Menggunakan model Naive Bayes yang telah dilatih untuk mengklasifikasikan sentimen tweet pada data pengujian.
- Menghitung probabilitas sentimen positif dan negatif untuk setiap tweet.

b. Penentuan Kelas Sentimen:

Menentukan kelas sentimen tweet berdasarkan probabilitas tertinggi (sentimen positif atau negatif).

6. Menghasilkan Hasil Klasifikasi dan Akurasi:

a. Hasil Klasifikasi:

Menyajikan hasil klasifikasi sentimen untuk setiap tweet pada data pengujian.

b. Akurasi Model:

Menghitung akurasi model Naive Bayes dengan membandingkan hasil klasifikasi dengan label sentimen sebenarnya pada data pengujian.

7. Selesai:

- Kesimpulan dan interpretasi hasil analisis sentimen.

Perhitungan Naïve Bayes :

1. Persiapkan Data

Persiapan data dapat dilihat pada tabel 3.9 berikut :

**Tabel 3.9 Sample Dataset**

Kalimat	Label
pasukan lemvar dara babie udah blom wkwkwk yg syahid lawan negara kt najiiiz	Negatif
semoga cepet dapet semoga cepet bebas main main	Positif

2. Hitung Probabilitas Prior:

Jumlah instance kelas Positif ( $P(\text{Positif})$ ): 1

Jumlah instance kelas Negatif ( $P(\text{Negatif})$ ): 1

Probabilitas prior  $P(\text{Positif}) = 1/10 = 0.1$

Probabilitas prior  $P(\text{Negatif}) = 1/10 = 0.1$

3. Hitung Probabilitas Likelihood:

Kata kunci: "semoga" dan "cepet".

Hitung probabilitas bahwa kata "semoga" muncul dalam teks positif ( $P(\text{"semoga"} | \text{Positif})$ ):

- Jumlah kata "semoga" dalam teks positif = 2

- Jumlah kata total dalam teks positif = 8

- $P(\text{"semoga"} | \text{Positif}) = 2/8 = 0.25$

Hitung probabilitas bahwa kata "semoga" muncul dalam teks negatif ( $P(\text{"semoga"} | \text{Negatif})$ ):

- Jumlah kata "semoga" dalam teks negatif = 0

- Jumlah kata total dalam teks negatif = 13

- $P(\text{"semoga"} | \text{Negatif}) = 0/13 = 0$

Hitung probabilitas bahwa kata "cepet" muncul dalam teks positif ( $P("cepet" | \text{Positif})$ ):

- Jumlah kata "cepet" dalam teks positif = 2
- Jumlah kata total dalam teks positif = 8
- $P("cepet" | \text{Positif}) = 2/8 = 0.25$

Hitung probabilitas bahwa kata "cepet" muncul dalam teks negatif ( $P("cepet" | \text{Negatif})$ ):

- Jumlah kata "cepet" dalam teks negatif = 0
- Jumlah kata total dalam teks negatif = 13
- $P("cepet" | \text{Negatif}) = 0/13 = 0$

#### 4. Hitung Probabilitas Posterior:

Misalkan kita memiliki teks baru: " semoga cepet dapet semoga cepet bebas main main "

$$\begin{aligned} P(\text{Positif} | \text{" semoga cepet dapet semoga cepet bebas main main"}) &\propto \\ P("semoga cepet dapet semoga cepet bebas main main" | \text{Positif}) * \\ P(\text{Positif}) \end{aligned}$$

$$\begin{aligned} P(\text{Negatif} | \text{" semoga cepet dapet semoga cepet bebas main main"}) &\propto \\ P("semoga cepet dapet semoga cepet bebas main main" | \text{Negatif}) * \\ P(\text{Negatif}) \end{aligned}$$

Hitung probabilitas posterior:

$$\begin{aligned} P(\text{Positif} | \text{" semoga cepet dapet semoga cepet bebas main main"}) &\propto 0.1 * \\ 0.25 &= 0.025 \\ P(\text{Negatif} | \text{" semoga cepet dapet semoga cepet bebas main main"}) &\propto 0.1 * \\ 0 &= 0 \end{aligned}$$

#### 5. Normalisasi Probabilitas:

$$\begin{aligned} \text{Normalisasi: } P(\text{Positif} | \text{" semoga cepet dapet semoga cepet bebas main main"}) + P(\text{Negatif} | \text{" semoga cepet dapet semoga cepet bebas main main"}) &= 0.1 + 0.025 = 0.125 \end{aligned}$$

Normalisasi:  $P(\text{Negatif} | \text{" semoga cepet dapet semoga cepet bebas main main"}) = 1 - P(\text{Positif} | \text{" semoga cepet dapet semoga cepet bebas main main"}) = 1 - 0.1 = 0.9$

#### **6. Prediksi:**

Karena  $P(\text{Negatif} | \text{" Belasan juta dosis vaksin Sinovac akan kembali"}) > P(\text{Positif} | \text{" Belasan juta dosis vaksin Sinovac akan kembali"}),$  prediksi = Negatif.

##### **3.2.5.1 Dataset**

Dataset yang digunakan berasal dari X dan terdiri dari total 2.879 komentar tentang vaksin COVID-19. Data diperoleh dengan cara crawling X menggunakan kata kunci #vaksincovid19. Selanjutnya, data tersebut telah dibagi dengan rasio 80:20 untuk pengujian dan pelatihan, dengan 2.164 data digunakan untuk pelatihan dan 542 data untuk pengujian. Dalam proses pengujian, 590 data akan digunakan untuk analisis berbasis leksikon, sedangkan 1.882 data akan digunakan untuk proses pembelajaran menggunakan metode naive Bayes. Naive Bayes adalah metode klasifikasi yang didasarkan pada teorema Bayes dengan asumsi bahwa setiap fitur dalam dataset adalah independen. Metode ini memprediksi kelas dari suatu data berdasarkan distribusi probabilistik dari fitur-fitur yang terdapat dalam data tersebut. Pada tahap pembagian data pelatihan dan pengujian, data dibagi dengan perbandingan 80% untuk data pelatihan dan 20% untuk data pengujian, sesuai dengan rasio yang telah ditentukan.

##### **3.2.5.2 Train Test Split**

Pada tahap train-test split dalam analisis sentimen pengguna X terhadap vaksinasi COVID-19 menggunakan metode Naive Bayes, dataset yang telah dikumpulkan dibagi menjadi dua subset utama: data pelatihan (training data) dan data pengujian (testing data). Pendekatan ini dilakukan dengan rasio pembagian yang telah ditentukan sebelumnya, yaitu 80% data untuk pelatihan dan 20% untuk pengujian. Data pelatihan digunakan untuk melatih model, sementara data pengujian digunakan untuk menguji

performa model yang telah dilatih. Proses ini bertujuan untuk mengukur seberapa baik model dapat menggeneralisasi pola dari data yang tidak terlihat sebelumnya, sehingga model dapat diharapkan mampu memberikan prediksi yang akurat terhadap data baru. Train Test Split dapat dilihat pada gambar 3.13 berikut.

No	Text	Sentiment	TF-IDF
1	pasu lemvar dara syahid lawan negara	Negatif	[3.131297796597623,3.4323277922616042,3.131297796597623,3.4323277922616042]
2	program warga emas batu pahat bangsa cina melayu	Netral	[0,0,0,0,0,2.017354444290786,1.9552065375419418,2.3183844399547673,2.83021]
3	giat serta piket fungsi laksana patroli kantor	Positif	[0,0,0,0,0,0,0,0,0,0,0,2.4780852828222795,3.131297796597623,3.4323277922616042]
4	isi virus nonaktif kalo demam habis antibodi	Negatif	[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2.5872297522473473,1.5292378052696607,2.1]

**Gambar 3. 13 Train Test Split**

Gambar tersebut menggambarkan proses train test split yang melibatkan tahap pengambilan data, penentuan rasio pembagian data, proses pembagian data, dan pembuatan model. Selain itu, terdapat tabel yang menampilkan kolom-kolom untuk teks, sentimen, dan nilai TF-IDF (Term Frequency-Inverse Document Frequency). Proses ini merupakan bagian dari tahap pra-pemrosesan data yang penting dalam analisis sentimen, di mana data teks yang dikumpulkan dari sumber seperti X dibagi menjadi subset untuk pelatihan dan pengujian model analisis sentimen.

### 3.2.5.3 Modeling

Pada tahap modeling dengan menggunakan metode Naive Bayes, dataset pelatihan digunakan untuk mengembangkan model klasifikasi sentimen. Naive Bayes adalah metode klasifikasi probabilistik yang didasarkan pada teorema Bayes dengan asumsi bahwa setiap fitur dalam dataset adalah independen. Dalam konteks analisis sentimen, model Naive Bayes memanfaatkan distribusi probabilitas dari kata-kata atau fitur-fitur yang muncul dalam setiap kelas sentimen (positif, negatif, atau netral) untuk melakukan prediksi terhadap sentimen dari tweet yang belum dilihat sebelumnya.

### 3.2.5.4 Implementasi *Naïve Bayes*

Setelah selesai melakukan tahap transformasi, analisis dilanjutkan ke tahap Implementasi Algoritma Pengklasifikasi Naïve Bayes. Pada tahap processing diimplementasikan algoritma pengklasifikasi Naïve Bayes. Tahap ini diawali dengan membagi data train dan data tes, kemudian melakukan train pada data train dan terakhir ke tahap klasifikasi.

Adapun pseudocode implementasi naïve bayes dapat dilihat pada Gambar 3.14

```
# Membangun model Naive Bayes
clf = MultinomialNB()
clf.fit(X_train, y_train)

# Load data uji yang telah diproses sebelumnya
df_test = pd.read_csv('test_data.csv')
X_test = tfidf_vectorizer.transform(df_test['Stemming'])
y_test = df_test['Label']

# Melakukan prediksi menggunakan model Naive Bayes
y_pred = clf.predict(X_test)

# Menghitung metrik evaluasi
accuracy = accuracy_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred, average='weighted')
recall = recall_score(y_test, y_pred, average='weighted')
precision = precision_score(y_test, y_pred,
average='weighted')
confusion_mat = confusion_matrix(y_test, y_pred)
```

**Gambar 3. 14 Pseudocode Implementasi Naive Bayes**

### 3.2.6 Output

Pada tahap output dalam analisis sentimen menggunakan metode Naive Bayes, terdapat dua metode umum yang digunakan untuk mengevaluasi performa model, yaitu classification report dan confusion matrix.

#### 3.2.6.1 Classification Report

Classification report menyediakan informasi terperinci mengenai kinerja model klasifikasi. Laporan ini umumnya mencakup beberapa metrik evaluasi seperti akurasi, presisi, recall, dan F1-score untuk setiap kelas sentimen (positif, negatif, atau netral). Akurasi mengukur seberapa baik model dapat mengklasifikasikan semua kelas dengan benar, presisi mengukur proporsi prediksi positif yang benar dari semua prediksi positif yang dibuat, recall mengukur proporsi dari semua kelas positif yang berhasil diidentifikasi oleh model, dan F1-score merupakan rata-rata harmonik dari presisi dan recall. Classification report memberikan gambaran yang jelas tentang kinerja model dalam mengklasifikasikan sentimen dari tweet-tweet yang belum pernah dilihat sebelumnya.

*Classification Report* dapat dilihat pada gambar 3.15 berikut.



Gambar 3. 15 Classification Report

#### 3.2.6.2 Confusion Matrix

Confusion matrix adalah tabel yang menggambarkan kinerja model dengan membandingkan prediksi kelas yang dihasilkan oleh model dengan kelas

sebenarnya dari data uji. *Confusion matrix* terdiri dari empat sel: *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). *True positive* (TP) mewakili jumlah data yang diklasifikasikan dengan benar sebagai kelas positif, *true negative* (TN) mewakili jumlah data yang diklasifikasikan dengan benar sebagai kelas negatif, *false positive* (FP) mewakili jumlah data yang salah diklasifikasikan sebagai kelas positif, dan *false negative* (FN) mewakili jumlah data yang salah diklasifikasikan sebagai kelas negatif. *Confusion matrix* memberikan wawasan tentang jenis kesalahan yang dilakukan oleh model dan membantu dalam mengevaluasi kinerja model secara lebih rinci.

Hasil dan tabel *Confusion matrix* dapat dilihat pada gambar 3.16 dan tabel 3.10 berikut ini.

Confusion Matrix		
	Predict Negatif	Predict Positif
Actual Negatif	283	2
Actual Positif	55	120

**Gambar 3. 16 Confusion Matrix**

**Tabel 3. 10 Tabel Matrix Confusion**

		Kelas prediksi	
		Yes	No
Kelas	Yes	TP	FN
	No	FP	TN

Keterangan:

*True Positive (TP)* : Jumlah data dari kelas *Positive* yang benar diklasifikasikan sebagai kelas *Positive*

*True Negative (TN)* : Jumlah data dari kelas *Negative* yang benar diklasifikasikan sebagai kelas *Negative*

*False Positive (FP)* : Jumlah data dari kelas *Negative* yang salah diklasifikasikan sebagai kelas *Positive*

*False Negative (FN)* : Jumlah data dari kelas *Positive* yang salah diklasifikasikan sebagai kelas *Negative*

Rumus perhitungan nilai akurasi, *recall*, *Specificity*, *Precision* dan *F1-Score* pada evaluasi model menggunakan *confusion matrix*, ditunjukan pada persamaan 3.1, 3.2, 3.3, dan 3.4 sebagai berikut:

$$Akurasi_i = \frac{TP_i + TN_i}{TP_i + TN_i + FN_i + FP_i} \quad (3.1)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (3.2)$$

$$Specificity_i = \frac{TN_i}{TN_i + FP_i} \quad (3.3)$$

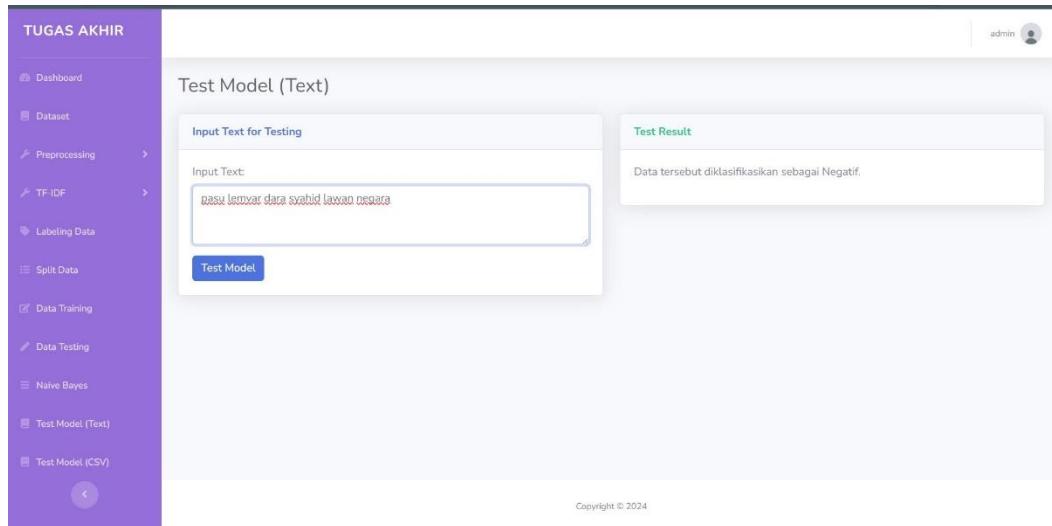
$$F1_i = \frac{2 * Presisi_i * Sensitivitas_i}{Presisi_i + Sensitivitas_i} \quad (3.4)$$

### 3.2.7 Test Model

#### 3.2.7.1 Test Model (Text)

Pada tahap pengujian model dengan menggunakan data teks, model Naive Bayes akan digunakan untuk memprediksi sentimen dari tweet-tweet yang belum pernah dilihat sebelumnya. Data teks yang digunakan untuk pengujian berisi tweet-tweet baru yang tidak pernah dipertimbangkan oleh model selama tahap pelatihan. Setelah model melakukan prediksi sentimen untuk setiap tweet dalam data pengujian, hasil prediksi akan dievaluasi menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score. Langkah ini membantu dalam menilai seberapa baik model dapat mengklasifikasikan sentimen dari data teks yang baru dan belum pernah dilihat sebelumnya. Evaluasi ini juga memberikan gambaran tentang keandalan dan keakuratan model dalam memprediksi sentimen terkait vaksinasi COVID-19 di Indonesia menggunakan data teks.

Test model (text) dapat dilihat pada gambar 3.17 berikut.



**Gambar 3. 17 Test Model (text)**

### 3.2.7.2 Test Model (csv)

Pada tahap pengujian model menggunakan data dalam format CSV (comma-separated values), dataset pengujian yang berisi tweet-tweet baru yang belum terlihat sebelumnya akan dimasukkan ke dalam model Naive Bayes. Setelah proses prediksi selesai, hasil prediksi sentimen untuk setiap tweet akan disimpan dalam format CSV yang baru. File CSV ini akan berisi tweet-tweet beserta prediksi sentimen yang sesuai dengan masing-masingnya. Penggunaan format CSV memudahkan dalam menyimpan dan mengelola hasil prediksi model, sehingga dapat dengan mudah diakses dan dianalisis lebih lanjut oleh peneliti atau pengguna lainnya. Selain itu, penggunaan data dalam format CSV juga memungkinkan untuk integrasi dengan aplikasi atau sistem lain yang membutuhkan output prediksi sentimen dalam bentuk yang terstruktur. Test model (csv) dapat dilihat pada gambar 3.18.

No	Text	Sentimen
1	pasukun lemaru dara bable udah blom wkwkw yg syahid Lawan negara kt najiliz ???????	Negatif
2	program utk warga emas batu pahat khemah bangsa cina je melayu????	Negatif
3	giat pawas ipda jaluruan beserta pihet fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp	Negatif
4	giat pawas ipda jaluruan beserta pihet fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp	Negatif
5	az isinya virus nonaktif kalo demam habis emg antibodi be	Negatif
6	setakat pekerja genting je staff gomen berkaitan dgn frontl	Negatif
7	semoga cepet dapet semoga cepet betas main main	Negatif
8	gk tahu cuman btuh kehadiran	Netral
9	juta data website jkjav sync dgn data mysejohtra orang dah tap	Positif
10	yampun under bursa obatin ya bersihin lukanya megangin ya huuu biar aja	Positif

**Gambar 3. 18 Test Model (csv)**

### 3.3 Perancangan Sistem

Dalam penelitian ini sistem dibuat berbasis Flask Python sebagai media antarmuka sistem dengan pengguna yang bertujuan untuk mempermudah pengguna dalam menjalankan sistem. Adapun penjelasan rancangan untuk setiap bagian yang diterapkan pada sistem adalah sebagai berikut.

#### 3.3.1 Rancangan Tampilan Login

Berikut rancangan tampilan login dapat dilihat pada gambar 3.19.

Rancangan tampilan login yang ditampilkan dalam browser. Judul halaman adalah "LOGIN". Terdapat dua input text berturut-turut untuk "Username" dan "Password", serta satu tombol "Login" di bawahnya.

**Gambar 3. 19 Rancangan Tampilan Login**

#### 3.3.2 Rancangan Tampilan Dashboard

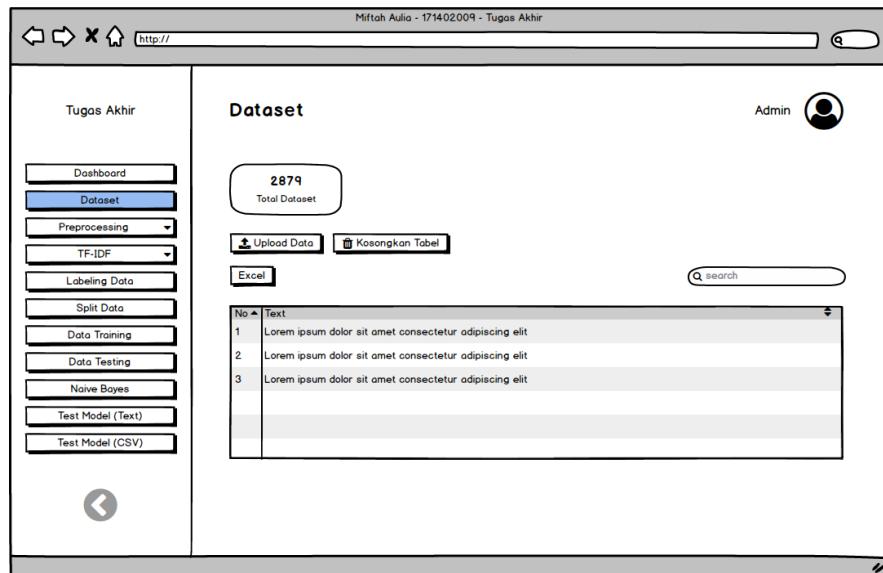
Berikut rancangan tampilan dashboard dapat dilihat pada gambar 3.20.

Rancangan tampilan dashboard yang ditampilkan dalam browser. Di sebelah kiri terdapat sidebar menu dengan opsi: Tugas Akhir, Dashboard, Dataset, Preprocessing, TF-IDF, Labeling Data, Split Data, Data Training, Data Testing, Naive Bayes, Test Model (Text), dan Test Model (CSV). Di sebelah kanan terdapat area "Dashboard" yang menunjukkan statistik "2879 Total Dataset". Di bawahnya terdapat judul "ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP VAKSINASI COVID 19 MENGGUNAKAN LEXICON BASED DAN NAIVE BAYES" dan informasi penulis: Nama : Miftah Aulia dan NIM : 171402009.

**Gambar 3. 20 Rancangan Tampilan Dashboard**

### 3.3.3 Rancangan Tampilan Dataset

Berikut rancangan tampilan dataset dapat dilihat pada gambar 3.21.

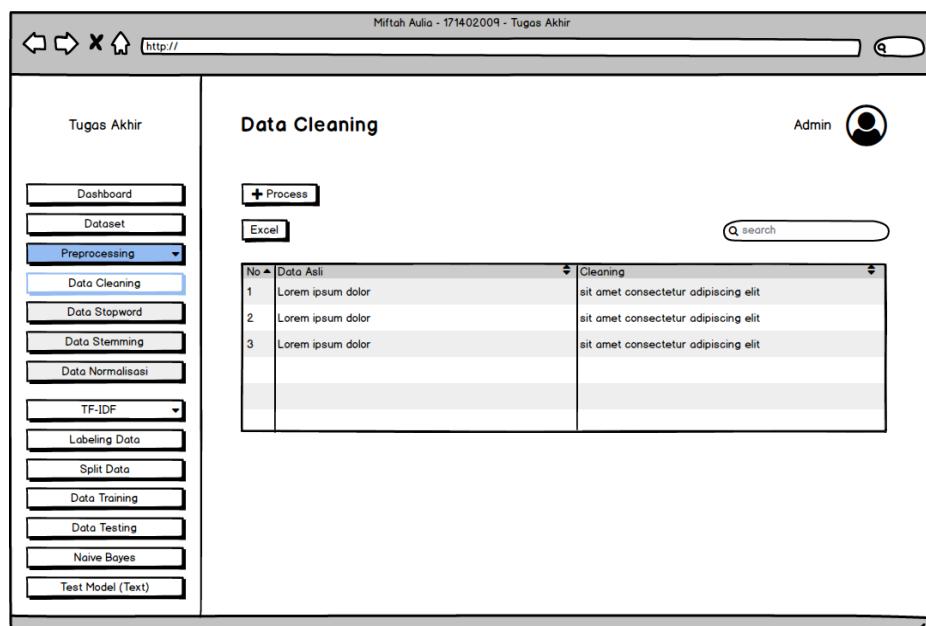


Gambar 3. 21 Rancangan Tampilan Dataset

### 3.3.4 Rancangan Tampilan *Pre-Processing* Data

#### 3.3.4.1 Rancangan Tampilan *Pre-Processing* Data Cleaning

Berikut rancangan tampilan *Pre-processing* data *cleaning* dapat dilihat pada gambar 3.22.



Gambar 3. 22 Rancangan Tampilan Pre-Processing Data Cleaning

### 3.3.4.2 Rancangan Tampilan Pre-Processing Data Stopword

Berikut rancangan tampilan *Pre-processing* data *stopword* dapat dilihat pada gambar 3.23.

No	Stopword	Cleaning
1	Lorem ipsum dolor	sit amet consectetur adipiscing elit
2	Lorem ipsum dolor	sit amet consectetur adipiscing elit
3	Lorem ipsum dolor	sit amet consectetur adipiscing elit

Gambar 3. 23 Rancangan Tampilan Pre-Processing Data Stopword

### 3.3.4.3 Rancangan Tampilan Pre-Processing Data Stemming

Berikut rancangan tampilan *Pre-processing* data *stemming* dapat dilihat pada gambar 3.24.

No	Stopword	Stemming
1	Lorem ipsum dolor	sit amet consectetur adipiscing elit
2	Lorem ipsum dolor	sit amet consectetur adipiscing elit
3	Lorem ipsum dolor	sit amet consectetur adipiscing elit

Gambar 3. 24 Rancangan Tampilan Pre-Processing Data Stemming

### 3.3.4.4 Rancangan Tampilan Pre-Processing Data Normalisasi

Berikut rancangan tampilan *Pre-processing* data *normalisasi* dapat dilihat pada gambar 3.25.

No	Stemming	Normalisasi
1	1. Lorem ipsum dolor	sit amet consectetur adipiscing elit
2	2. Lorem ipsum dolor	sit amet consectetur adipiscing elit
3	3. Lorem ipsum dolor	sit amet consectetur adipiscing elit

Gambar 3. 25 Rancangan Tampilan Pre-Processing Data Normalisasi

### 3.3.5 Rancangan Tampilan TF – IDF

#### 3.3.5.1 Rancangan Tampilan Data TF - IDF

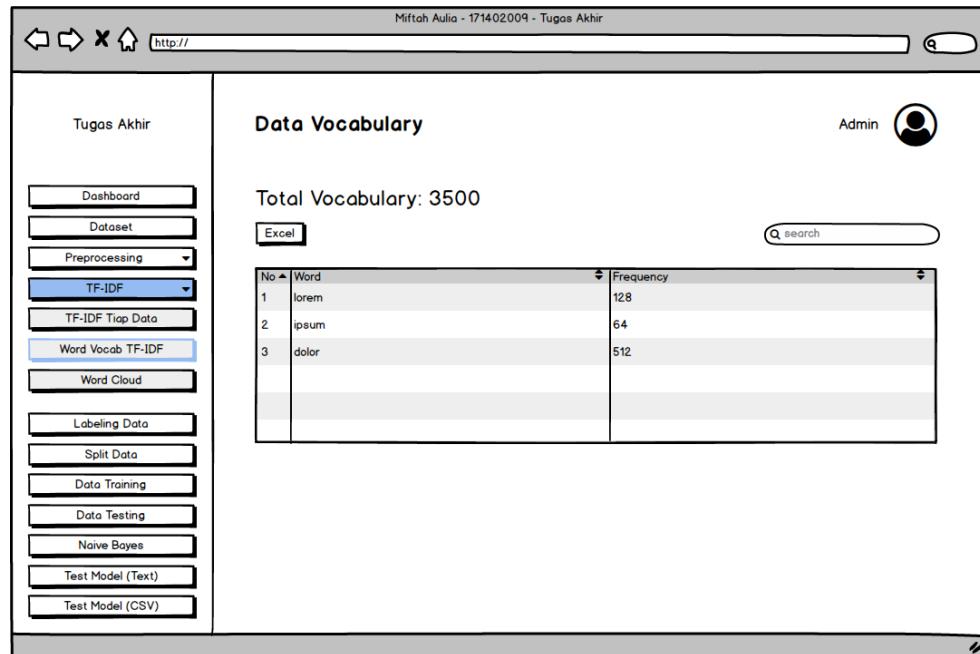
Berikut rancangan tampilan data TF-IDF dapat dilihat pada gambar 3.26.

No	Text	TF-IDF
1	1. Lorem ipsum dolor	3. 13023818239. 3. 13023815486. 3. 13023812381
2	2. Lorem ipsum dolor	3. 13023818239. 3. 13023815486. 3. 13023812381
3	3. Lorem ipsum dolor	3. 13023818239. 3. 13023815486. 3. 13023812381

Gambar 3. 26 Rancangan Tampilan Data TF – IDF

### **3.3.5.2 Rancangan Tampilan Data Vocabulary**

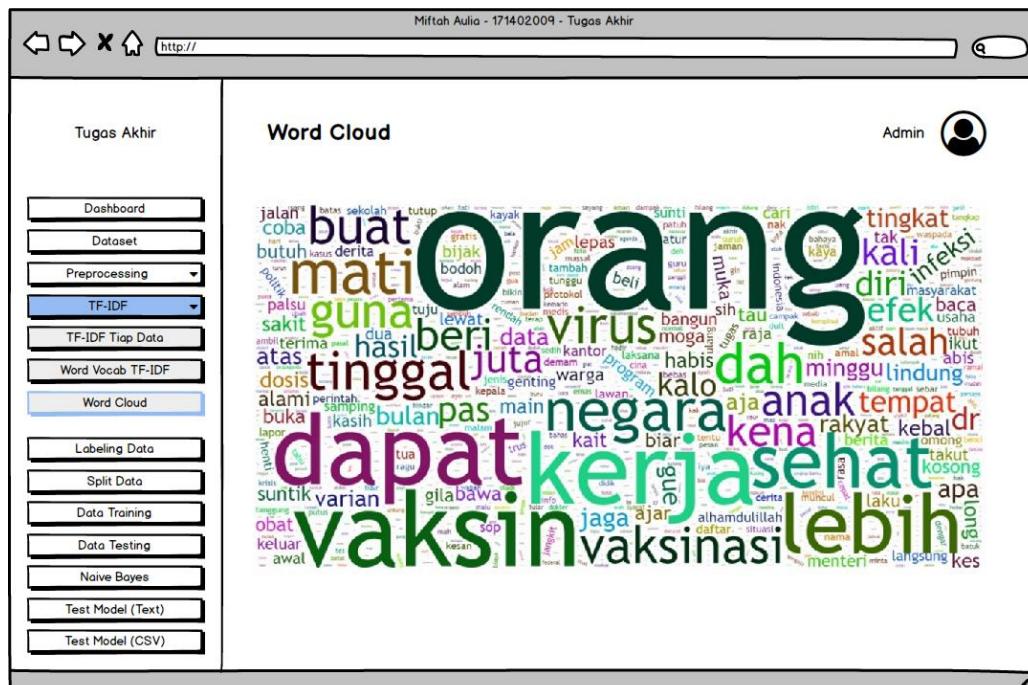
Berikut rancangan tampilan data vocabulary dapat dilihat pada gambar 3.27.



**Gambar 3.27 Rancangan Tampilan Data Vocabulary**

### **3.3.5.3 Rancangan Tampilan Word Cloud**

Berikut rancangan tampilan wordcloud dapat dilihat pada gambar 3.28.



### Gambar 3. 28 Rancangan Tampilan Word Cloud

### 3.3.6 Rancangan Tampilan Labeling Data

Berikut rancangan tampilan labeling data dapat dilihat pada gambar 3.29.

The screenshot shows a web application interface titled 'Labeling Data'. On the left sidebar, under the 'Tugas Akhir' section, the 'Labeling Data' option is highlighted. The main content area displays a table with five rows of data. The columns are labeled 'No', 'Text', 'Sentimen', and 'TF-IDF'. The data is as follows:

No	Text	Sentimen	TF-IDF
1	lorem ipsum	Negatif	3.13023818239.3.13023815486.3.13023812381
2	dolor sit amet	Positif	3.13023818239.3.13023815486.3.13023812381
3	consectetur adipiscing elit	Negatif	3.13023818239.3.13023815486.3.13023812381
4	sed do eiusmod tempor	Negatif	3.13023818239.3.13023815486.3.13023812381
5	incidunt ut labore et dolore magna aliqua	Positif	3.13023818239.3.13023815486.3.13023812381

Gambar 3. 29 Rancangan Tampilan Labeling Data

### 3.3.7 Rancangan Tampilan Split Data

Berikut rancangan tampilan *split data* dapat dilihat pada gambar 3.30.

The screenshot shows a web application interface titled 'Split Data'. On the left sidebar, under the 'Tugas Akhir' section, the 'Split Data' option is highlighted. The main content area displays a table with five rows of data. The columns are labeled 'No', 'Text', 'Sentimen', and 'TF-IDF'. The data is as follows:

No	Text	Sentimen	TF-IDF
1	lorem ipsum	Negatif	3.13023818239.3.13023815486.3.13023812381
2	dolor sit amet	Positif	3.13023818239.3.13023815486.3.13023812381
3	consectetur adipiscing elit	Negatif	3.13023818239.3.13023815486.3.13023812381
4	sed do eiusmod tempor	Negatif	3.13023818239.3.13023815486.3.13023812381
5	incidunt ut labore et dolore magna aliqua	Positif	3.13023818239.3.13023815486.3.13023812381

Gambar 3. 30 Rancangan Tampilan Split Data

### 3.3.8 Rancangan Tampilan Data Training

Berikut rancangan tampilan data *training* dapat dilihat pada gambar 3.31.

No	Text	Sentimen
1	lorem ipsum	Negatif
2	dolor sit amet	Positif
3	consectetur adipiscing elit	Negatif
4	sed do eiusmod tempor	Negatif
5	incididunt ut labore et dolore magna aliqua	Positif

Gambar 3. 31 Rancangan Tampilan Data *Training*

### 3.3.9 Rancangan Tampilan Data *Testing*

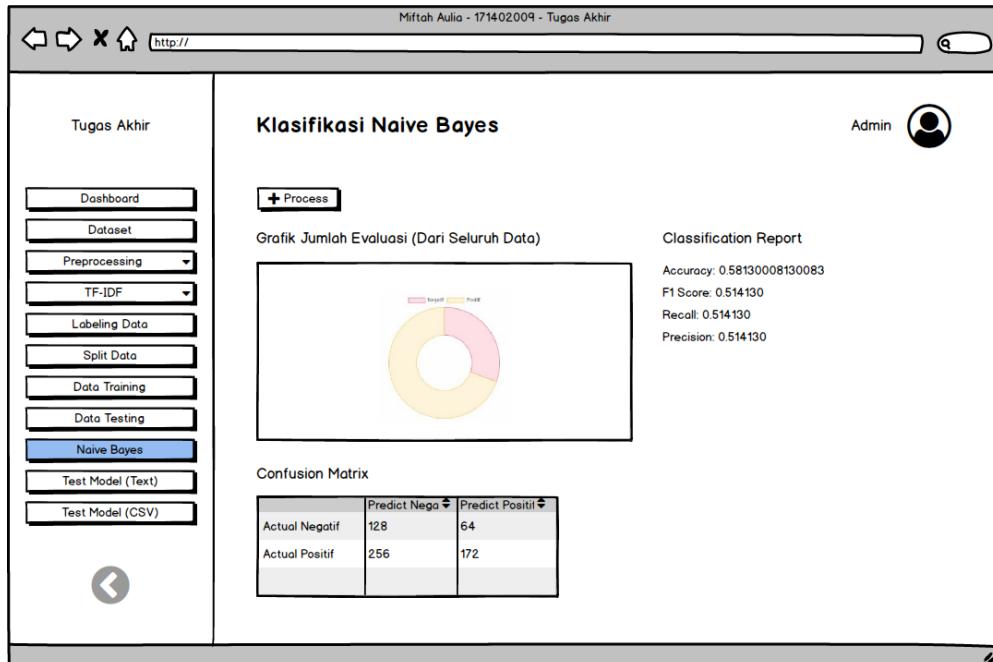
Berikut rancangan tampilan data *testing* dapat dilihat pada gambar 3.32.

No	Text	Sentimen
1	lorem ipsum	Negatif
2	dolor sit amet	Positif
3	consectetur adipiscing elit	Negatif
4	sed do eiusmod tempor	Negatif
5	incididunt ut labore et dolore magna aliqua	Positif

Gambar 3. 32 Rancangan Tampilan Data *Testing*

### 3.3.10 Rancangan Tampilan *Naïve Bayes*

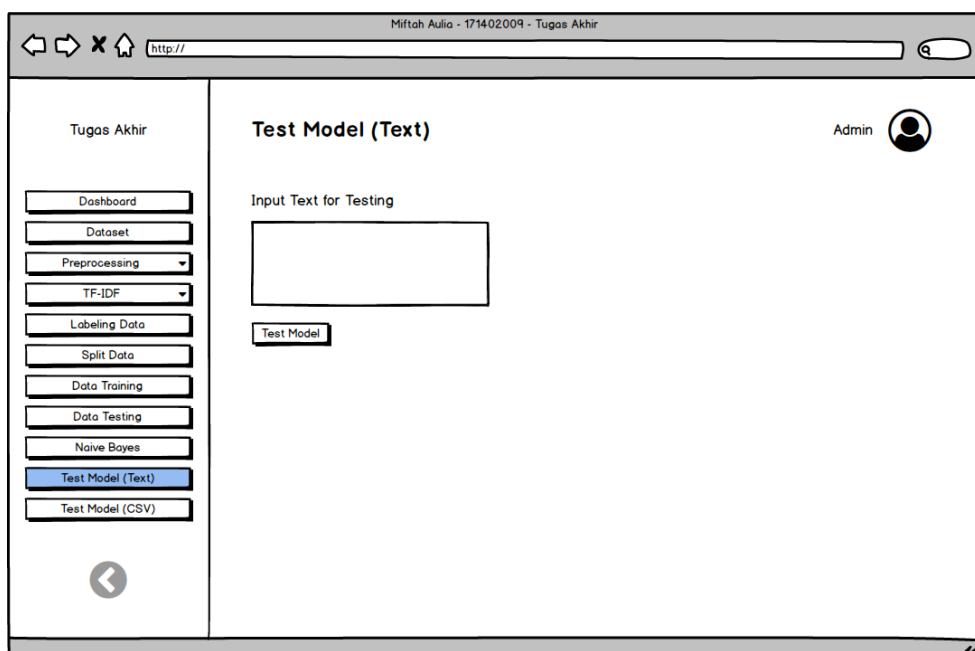
Berikut rancangan tampilan *Naïve Bayes* dapat dilihat pada gambar 3.33.



Gambar 3. 33 Rancangan Tampilan *Naïve Bayes*

### 3.3.11 Rancangan Tampilan Test Model (Text)

Berikut rancangan tampilan test model (*text*) dapat dilihat pada gambar 3.34.



Gambar 3. 34 Rancangan Tampilan Test Model (Text)

### 3.3.12 Rancangan Tampilan Test Model (csv)

Berikut rancangan tampilan test model (csv) dapat dilihat pada gambar 3.35.

The screenshot shows a web application interface. At the top, there's a header bar with browser controls, a URL field containing 'http://', and a title 'Miftah Aulia - 171402009 - Tugas Akhir'. On the right side of the header is a user profile icon labeled 'Admin'. The main content area has a left sidebar with a vertical list of menu items: 'Tugas Akhir' (selected), 'Dashboard', 'Dataset', 'Preprocessing' (with dropdown options 'TF-IDF' and 'Labeling Data'), 'Split Data', 'Data Training', 'Data Testing', 'Naive Bayes', 'Test Model (Text)', and 'Test Model (CSV)' (selected). The main panel is titled 'Test Model (CSV)'. It contains a section for 'Input CSV for Testing' with buttons for 'Upload Data' and '+ Process', and a dropdown menu set to 'Excel'. Below this is a table showing five rows of text samples and their corresponding sentiment scores:

No	Text	Sentimen
1	lorem ipsum	Negatif
2	dolor sit amet	Positif
3	consectetur adipiscing elit	Negatif
4	sed do eiusmod tempor	Negatif
5	incididunt ut labore et dolore magna aliqua	Positif

Gambar 3. 35 Rancangan Tampilan Test Model (csv)

## **BAB 4**

### **IMPLEMENTASI PENGUJIAN SISTEM**

#### **4.1 Implementasi Sistem**

Pada proses pembuatan analisis sentimen pengguna twitter terhadap vaksinasi Covid-19 menggunakan *lexicon based & naive bayes* digunakan perangkat keras dan perangkat lunak yang mendukung yakni:

##### **4.1.1. Spesifikasi Perangkat Keras dan Perangkat Lunak**

Adapun perincian jenis perangkat keras yang digunakan dalam membentuk sistem yaitu:

1. Laptop OMEN by HP
2. Processor Intel Core i7 – 9750H, up to 2.59GHz
3. Kapasitas memory (RAM) yang digunakan 8GB
4. HDD 500 GB

Adapun perincian jenis perangkat lunak yang digunakan dalam membentuk sistem yakni:

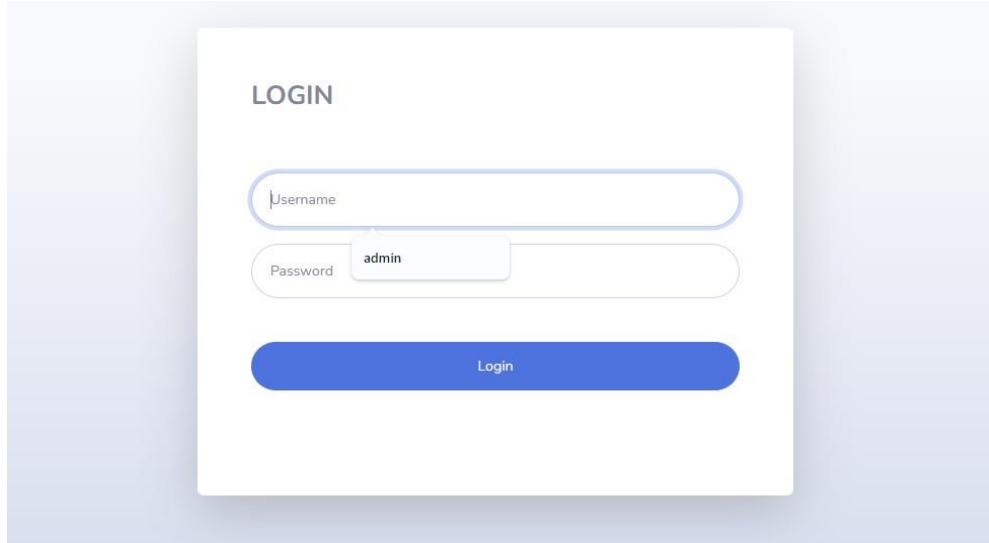
1. Laragon
2. Bahasa pemrograman *Python* versi 3.8 dengan library xgboost, library Sastrawi versi 1.0.1, pandas versi 1.2.4, scikit-learn versi 0.24.2, Flask versi 2.0.0, NLTK versi 3.6.2, numpy versi 1.20.3, dan matplotlib versi 3.4.2

##### **4.1.2. Implementasi Perancangan Tampilan Antarmuka**

Penerapan pengambilan data, teks preprocessing hingga Naivebayes pada web aplikasi yang dibuat. Hasil keluaran berupa aplikasi berbasis website untuk mengalisis sentimen persepsi masyarakat terhadap vaksinasi COVID-19 di Indonesia.

## 1. Implementasi Halaman Login

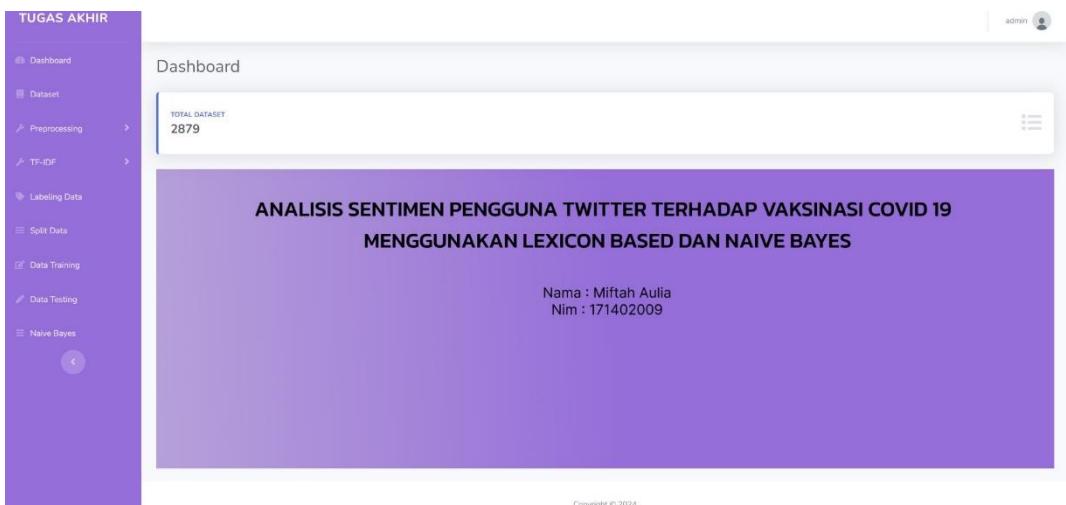
Page Login adalah halaman pertama yang diakses oleh pengguna, di mana mereka diminta untuk memasukkan username dan password. Tampilan halaman login dapat dilihat pada Gambar 4.1 berikut.



**Gambar 4. 1 Halaman Login**

## 2. Implementasi Halaman Dashboard

Halaman Dashboard adalah halaman pertama yang diakses oleh pengguna, di mana terdapat judul penelitian. Tampilan halaman dashboard dapat dilihat pada Gambar 4.2 berikut.



**Gambar 4. 2 Halaman Dashboard**

### 3. Implementasi Halaman Upload File

Halaman unggah file merupakan tempat untuk memasukkan file yang telah diperoleh dari pengambilan data (scrapping data). Data tersebut kemudian diunggah dan ditampilkan. Tampilan halaman unggah data dapat dilihat pada Gambar 4.3 berikut.

No	Text	Cleaning
1	pasukan lembar dara babie udah blom wkwkwk yg syahid lawan negara kt najiz???????	pasukan lembar dari babie udah blom wkwkwk yg syahid lawan negara kt najiz
2	program utk warga emas batu pahat khemah bangsa cina je melayu?????	program utk warga emas batu pahat khemah bangsa cina je melayu
3	giat pawas ipda jsliburian beserta pikef fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp	giat pawas ipda jsliburian beserta pikef fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp
4	az isinya virus nonaktif kalo demam habis emg antibodi be	az isinya virus nonaktif kalo demam habis emg antibodi be
5	setakat pekerja genting je staff gomen berkaitan dgn frontl	setakat pekerja genting je staff gomen berkaitan dgn frontl
6	semoga cepet dapet semoga cepet bebas main main	semoga cepet dapet semoga cepet bebas main main
7	ek btruh eman btruh kahadir	

Gambar 4. 3 Halaman Upload File

### 4. Implementasi Halaman Preprocessing

Halaman prapemrosesan adalah tempat untuk melakukan prapemrosesan data. Data yang diunggah dari file akan melalui tahap-tahap prapemrosesan seperti pembersihan data, penghapusan kata penghubung, pemangkasan hingga normalisasi. Tampilan halaman pembersihan data dapat dilihat pada Gambar 4.4 hingga Gambar 4.8 berikut.

No	Data Asli	Cleaning	Search
1	pasukan lembar dara babie udah blom wkwkwk yg syahid lawan negara kt najiz???????	pasukan lembar dari babie udah blom wkwkwk yg syahid lawan negara kt najiz	
2	program utk warga emas batu pahat khemah bangsa cina je melayu?????	program utk warga emas batu pahat khemah bangsa cina je melayu	
3	giat pawas ipda jsliburian beserta pikef fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp	giat pawas ipda jsliburian beserta pikef fungsi melaksanakan patroli monitoring kantor dpcipk dpcpp	
4	az isinya virus nonaktif kalo demam habis emg antibodi be	az isinya virus nonaktif kalo demam habis emg antibodi be	
5	setakat pekerja genting je staff gomen berkaitan dgn frontl	setakat pekerja genting je staff gomen berkaitan dgn frontl	
6	semoga cepet dapet semoga cepet bebas main main	semoga cepet dapet semoga cepet bebas main main	
7			

Gambar 4. 4 Halaman Cleaning

## 5. Implementasi Halaman (stopword)

Implementasi halaman stopword bisa dilihat pada gambar 4.5 berikut :

No	Cleaning	Stopwords
1	pasukan lemvra darababie udahblom wkwkwk yg syahid lawan negara kt najiz	pasukan lemvra darababie udahblom wkwkwk yg syahid lawan negara kt najiz
2	program utk warga emas batu pahat khemah bangsa cina je melayu	program utk warga emas batu pahat khemah bangsa cina je melayu
3	giat pawas ipda jsburian berserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpccp	giat pawas ipda jsburian berserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpccp
4	giat pawas ipda jsburian berserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpccp	giat pawas ipda jsburian berserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpccp
5	az isinya virus nonaktif kalo demam habis emg antibodi be	az isinya virus nonaktif kalo demam habis emg antibodi be
6	setakat pekerja genting je staff gomen berkaitan dgn frontl	setakat pekerja genting je staff gomen berkaitan dgn frontl
7	semoga cepet dapat semoga cepet bebas main main	semoga cepet dapat semoga cepet bebas main main
8	gk btuh cman btuh kehadiran	gk btuh cman btuh kehadiran
9	iuta data website ikaiav svnc don data mvseiahtera orano dah tao	iuta data website ikaiav svnc don data mvseiahtera orano dah tao

Gambar 4. 5 Halaman Stopword

## 6. Implementasi Halaman (stemming)

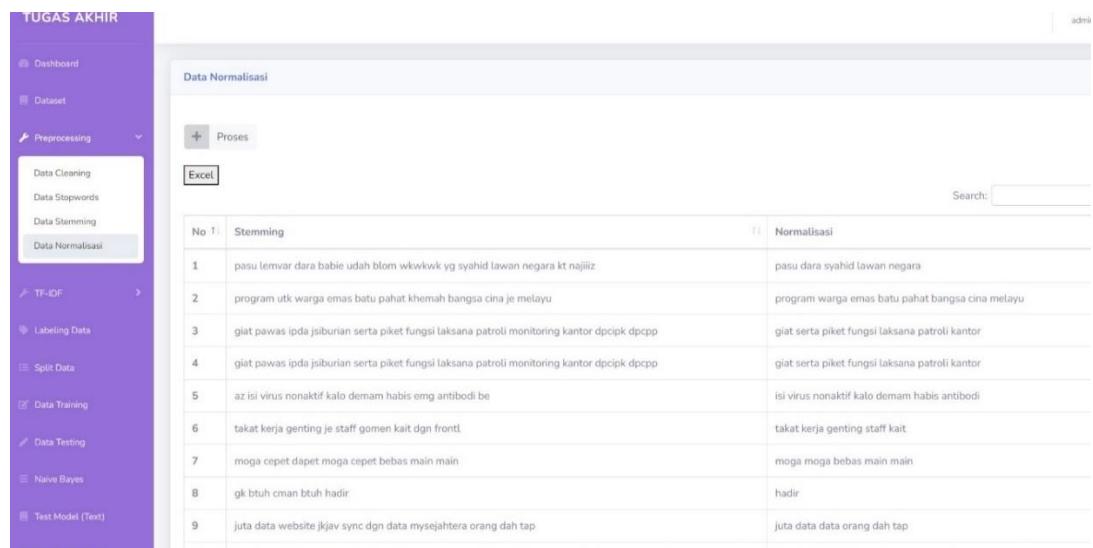
Implementasi halaman stemming dapat dilihat pada gambar 4.6 berikut :

No	Stopwords	Stemming
1	pasukan lemvra darababie udahblom wkwkwk yg syahid lawan negara kt najiz	pasu lemvra darababie udahblom wkwkwk yg syahid lawan negara kt najiz
2	program utk warga emas batu pahat khemah bangsa cina je melayu	program utk warga emas batu pahat khemah bangsa cina je melayu
3	giat pawas ipda jsburian berserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpccp	giat pawas ipda jsburian serta piket fungsi laksana patroli monitoring kantor dpcipk dpccp
4	giat pawas ipda jsburian berserta piket fungsi melaksanakan patroli monitoring kantor dpcipk dpccp	giat pawas ipda jsburian serta piket fungsi laksana patroli monitoring kantor dpcipk dpccp
5	az isinya virus nonaktif kalo demam habis emg antibodi be	az isi virus nonaktif kalo demam habis emg antibodi be
6	setakat pekerja genting je staff gomen berkaitan dgn frontl	takat kerja genting je staff gomen kait dgn frontl
7	semoga cepet dapat semoga cepet bebas main main	moga cepet dapat mogo cepet bebas main main
8	gk btuh cman btuh kehadiran	gk btuh cman btuh hadir
9	iuta data website ikaiav svnc don data mvseiahtera orano dah tao	iuta data website ikaiav svnc don data mvseiahtera orano dah tao

Gambar 4. 6 Halaman Stemming

## 7. Implementasi Halaman ( normalisasi )

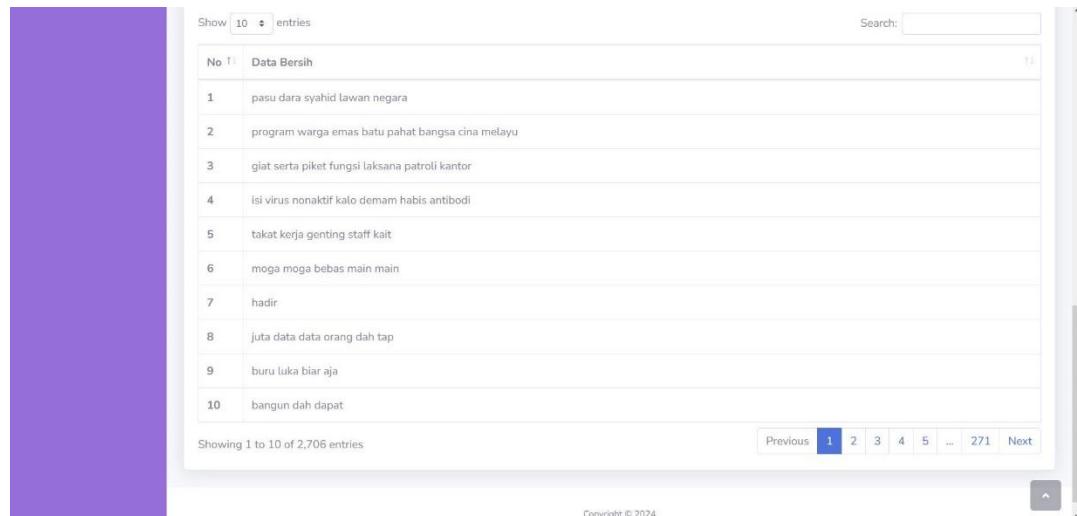
Implementasi halaman normalisasi dapat dilihat pada gambar 4.7 dan 4.8 berikut :



The screenshot shows a web-based text processing tool titled "TUGAS AKHIR". On the left sidebar, under the "Preprocessing" section, "Data Normalisasi" is selected. The main area is titled "Data Normalisasi" and contains a table with two columns: "Stemming" and "Normalisasi". The table lists nine rows of input text and its corresponding normalized output.

No	Stemming	Normalisasi
1	pasu lemvar dara babie udah blom wkwkwk yg syahid lawan negara kt najiiiz	pasu dara syahid lawan negara
2	program utk warga emas batu pahat khemah bangsa cina je melayu	program warga emas batu pahat bangsa cina melayu
3	giat pawas ipda jasberian serta piket fungsi laksana patroli monitoring kantor dpcpkp dpcpp	giat serta piket fungsi laksana patroli kantor
4	giat pawas ipda jasberian serta piket fungsi laksana patroli monitoring kantor dpcpkp dpcpp	giat serta piket fungsi laksana patroli kantor
5	az isi virus nonaktif kalo demam habis emg antibodi be	isi virus nonaktif kalo demam habis antibodi
6	takat kerja genting je staff gomen kait dgn frontl	takat kerja genting staff kait
7	moga cepet dapet mogga cepet bebas main main	moga mogga bebas main main
8	gk btuh cman btuh hadir	hadir
9	juta data website jkjav sync dgn data mysejahtera orang dah tap	juta data data orang dah tap

Gambar 4. 7 Halaman Normalisasi 1



The screenshot shows a web-based text processing tool with a purple sidebar. In the center, there is a table titled "Data Bersih" with two columns: "Data Bersih" and "Normalisasi". Below the table, it says "Showing 1 to 10 of 2,706 entries". At the bottom right, there is a navigation bar with page numbers from 1 to 271.

No	Data Bersih
1	pasu dara syahid lawan negara
2	program warga emas batu pahat bangsa cina melayu
3	giat serta piket fungsi laksana patroli kantor
4	isi virus nonaktif kalo demam habis antibodi
5	takat kerja genting staff kait
6	moga mogga bebas main main
7	hadir
8	juta data data orang dah tap
9	buru luka biar aja
10	bangun dah dapat

Gambar 4. 8 Halaman Normalisasi 2

## **8. Implementasi Halaman TF-IDF**

Halaman TF-IDF adalah halaman untuk melakukan pembobotan TF-IDF, data yang digunakan merupakan hasil *preprocessing*. Berikut tampilan halaman TF-IDF dapat dilihat pada gambar 4.9.

**Gambar 4.9 Halaman TF-IDF**

## 9. Implementasi Halaman Vocabulary and Word Frequencies

Implementasi Halaman Vocabulary and Word Frequencies bertujuan untuk menampilkan daftar kosakata (vocabulary) dan frekuensi kemunculan kata (word frequencies) dari suatu teks atau korpus tertentu. Tampilan halaman data Vocabulary dapat dilihat pada Gambar 4.10.

TUGAS AKHIR

admin

Dashboard

Dataset

Preprocessing

TF-IDF

Labeling Data

Split Data

Data Training

Data Testing

Naive Bayes

Test Model (Text)

Test Model (CSV)

Total Vocabulary: 3499

Vocabulary and Word Frequencies:

Excel

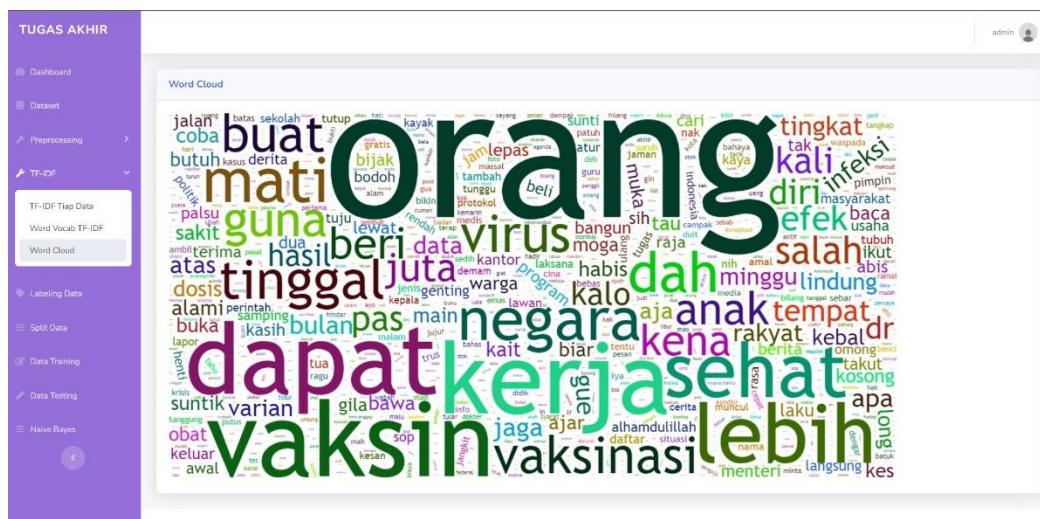
Search:

No	Word	Frequency
1	orang	538
2	jadi	275
3	tahun	252
4	lebih	240
5	vaksin	230
6	banyak	225
7	milik	224

#### **Gambar 4. 10 Halaman Vocabulary**

## 10. Implementasi Halaman Wordcoud

Wordcloud adalah representasi visual dari data teks di mana kata-kata yang paling sering muncul dalam teks tersebut akan ditampilkan lebih besar dan lebih mencolok daripada kata-kata yang jarang muncul. Hal ini dapat dilihat pada Gambar 4.11.



#### **Gambar 4. 11 Halaman Wordcloud**

## **11. Implementasi Halaman Labeling**

Klasifikasi sentimen dengan metode Lexicon Based dilakukan dengan mempertimbangkan kata-kata positif dan negatif yang terdapat pada tweet setelah melalui proses pembersihan. Dalam metode ini, digunakan kamus Lexicon Bahasa Indonesia untuk mencocokkan kata-kata dalam tweet dengan entri yang ada dalam kamus tersebut. Jika tweet mengandung kata-kata yang termasuk dalam kategori positif, maka tweet tersebut akan dianggap memiliki sentimen positif. Sebaliknya, jika tweet mengandung kata-kata negatif, maka tweet tersebut akan dianggap memiliki sentimen negatif. Halaman implementasi dari labeling data dapat dilihat pada Gambar 4.12 berikut.

TUGAS AKHIR			
		admin	
		Data Sentimen	
<a href="#">Dashboard</a>		<span>+ Proses</span> <span>Excel</span>	
<a href="#">Dataset</a>		<input type="text" value="Search:"/>	
<a href="#">Preprocessing</a>			
<a href="#">TF-IDF</a>			
<a href="#">Labeling Data</a>			
<a href="#">Split Data</a>			
<a href="#">Data Training</a>			
<a href="#">Data Testing</a>			
<a href="#">Naive Bayes</a>			
<a href="#">Test Model (Text)</a>			

#### **Gambar 4. 12 Halaman Labeling**

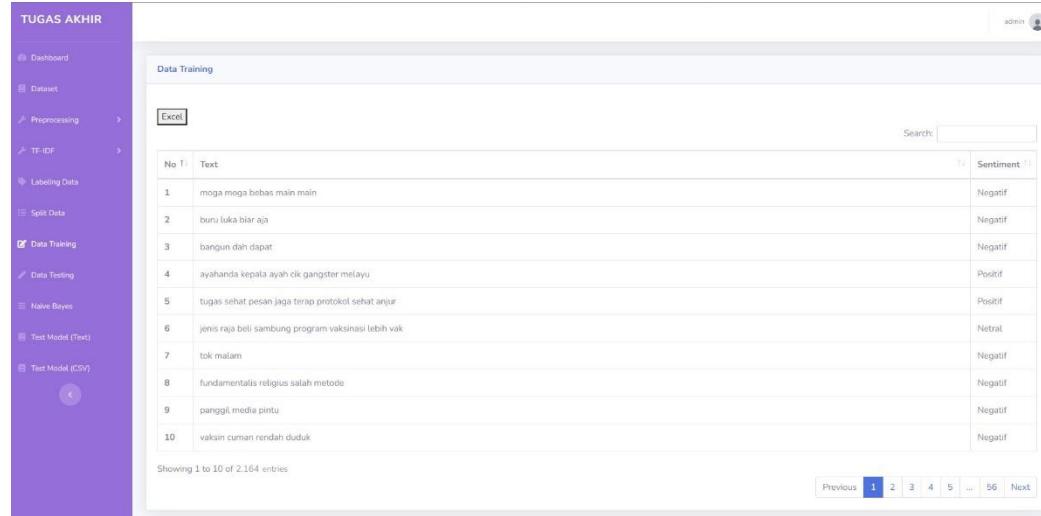
## **12. Implementasi Halaman split data**

Halaman Split Data 80:20 adalah proses pembagian dataset menjadi dua bagian, di mana 80% dari data digunakan sebagai data latih (training data) dan 20% sisanya digunakan sebagai data uji (tes data). Tampilan halaman split data dapat dilihat pada gambar 4.13 berikut.

#### **Gambar 4. 13 Halaman Split data**

### 13. Halaman training data

Halaman training data menampilkan data yang digunakan untuk melatih model, termasuk data tweet yang akan digunakan untuk pengembangan model klasifikasi. Tampilan halaman training data dapat dilihat pada Gambar 4.14 berikut.

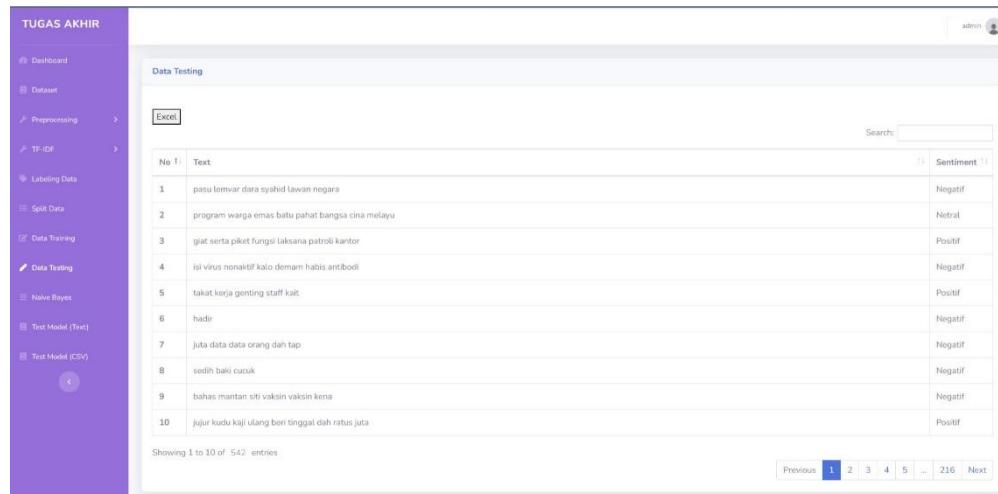


No	Teks	Sentimen
1	moga mogga bebas main main	Negatif
2	buru luka biar aja	Negatif
3	bangun dan dapat	Negatif
4	ayahanda kepala ayah cik gangster melayu	Positif
5	tugas sehat pesan jaga terap protokol sehat anjur	Positif
6	jenis raja belli sambung program vaksinasi lebih vak	Neutra
7	tok malam	Negatif
8	fundamentalis religius salah metode	Negatif
9	panggil media pintu	Negatif
10	vakan cuman rendah duduk	Negatif

Gambar 4. 14 Halaman training data

#### 14. Halaman testing data

Halaman testing data menampilkan data yang digunakan untuk menguji model yang telah dilatih. Data ini berbeda dari data latih dan digunakan untuk mengevaluasi kinerja model. Tampilannya dapat dilihat pada Gambar 4.15.



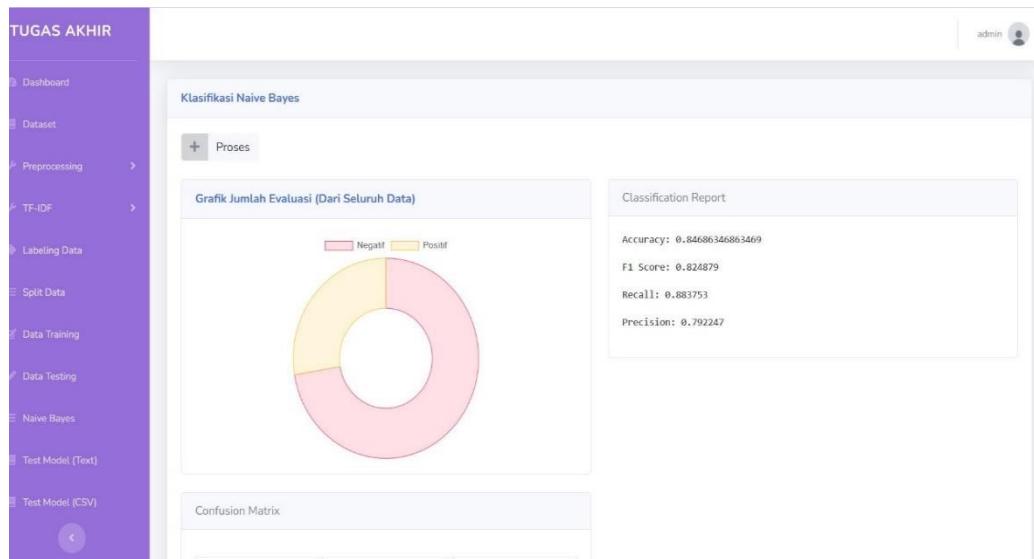
No	Teks	Sentimen
1	pasu fermvar dara syahid lawan negara	Negatif
2	program warga emas batu pahat bangsa cina melayu	Neutra
3	grat serta paket fungsi laksana patroli kantor	Positif
4	isi virus nonaktif kalau demam habis antibodi	Negatif
5	takut kerja genting staff kait	Positif
6	hadir	Negatif
7	juta data data orang dah tap	Negatif
8	sedih baki curuk	Negatif
9	bahan mentan sti vaksin vaksin kena	Negatif
10	jujur kudu kaji ulang beri tinggal dih ratus juta	Positif

Gambar 4. 15 Halaman testing data

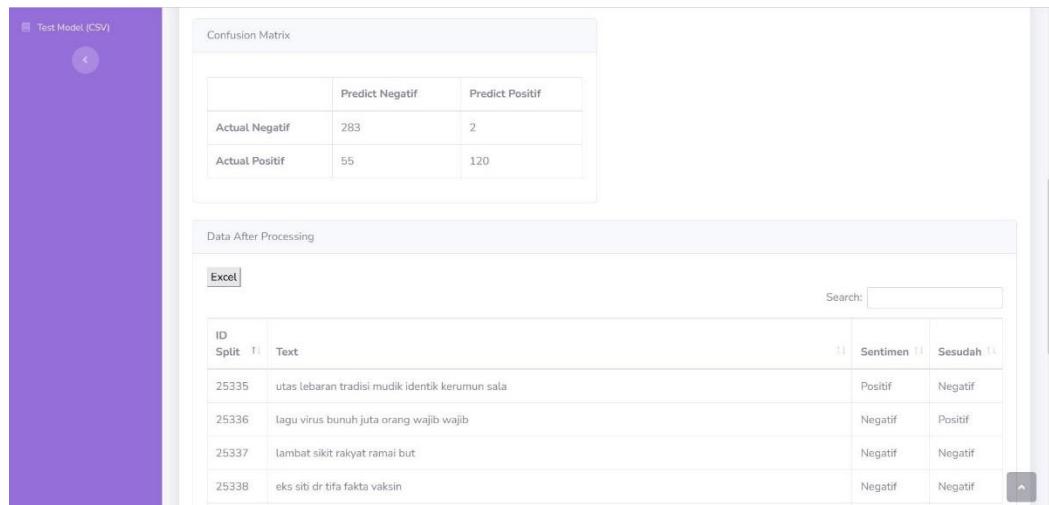
#### 4.2. Implementasi Model Naive Bayes

Implementasi Model Naive Bayes adalah bagian yang menjelaskan implementasi model klasifikasi Naïve Bayes. Ini termasuk langkah-langkah dan proses yang

dilakukan untuk mengembangkan model tersebut. Tampilan implementasi model naïve bayes dapat dilihat pada gambar 4.16 dan 4.17 berikut.



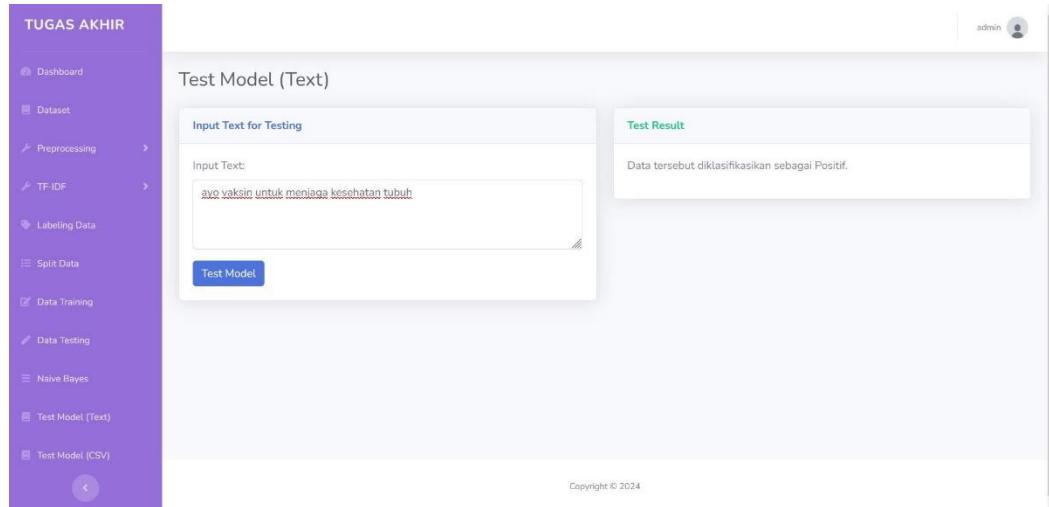
**Gambar 4. 16 Klasifikasi Naïve Bayes 1**



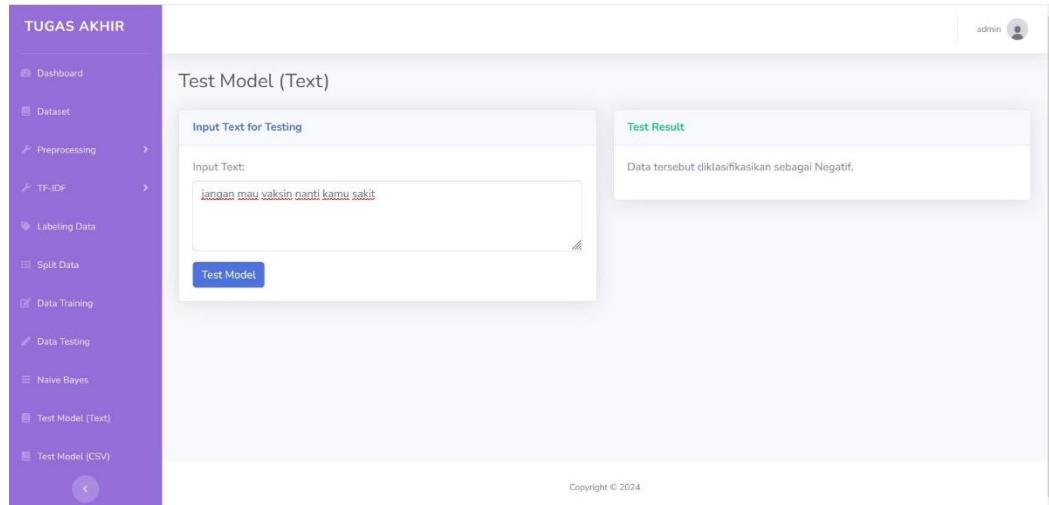
**Gambar 4. 17 Klasifikasi Naïve Bayes 2**

### 4.3 Implementasi Test model ( text )

Implementasi Test model (text) adalah bagian yang menjelaskan proses pengujian model menggunakan data teks, termasuk langkah-langkah yang dilakukan dalam proses pengujian. Tampilan implementasi test model dapat dilihat pada Gambar 4.18 dan 4.19 berikut.



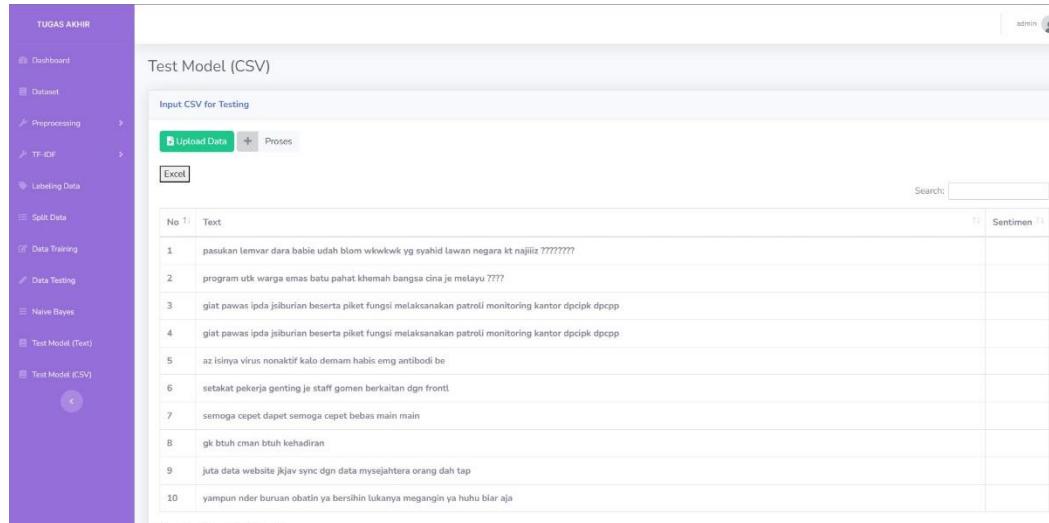
**Gambar 4. 18 Implementasi Test model ( text ) 1**



**Gambar 4. 19 Implementasi Test model ( text ) 2**

#### 4.4 Implementasi Test model ( csv )

Implementasi Test model (CSV) adalah bagian yang menjelaskan proses pengujian model menggunakan data dalam format CSV, termasuk langkah-langkah yang dilakukan dalam proses pengujian. Tampilan implementasi test model (csv) dapat dilihat pada Gambar 4.20 berikut.



**Gambar 4. 20 Implementasi Test model ( csv )**

#### 4.5 Pengujian

Tahap pengujian bertujuan untuk mengevaluasi keberhasilan proses klasifikasi yang telah dibangun. Implementasi pengujian sistem dalam penelitian ini melibatkan pembuatan confusion matrix. Melalui confusion matrix, hasil evaluasi akurasi, presisi, dan recall dapat diperoleh. Pengujian dilakukan untuk membandingkan nilai akurasi antara dua metode, yaitu Metode Naïve Bayes. Total 2.879 data digunakan dalam pengujian ini, dengan 80% digunakan sebagai data training dan 20% sebagai data testing, menggunakan metode evaluasi confusion matrix. Selain itu, untuk memastikan keandalan model, pengujian metode Naïve Bayes juga dilakukan menggunakan cross-validation. Cross-validation adalah teknik validasi model yang membagi dataset menjadi beberapa subset yang saling bergantian untuk digunakan sebagai data training dan testing. Dengan demikian, setiap data dalam dataset digunakan untuk pengujian (testing) sekaligus, menghindari kemungkinan bias dan menghasilkan evaluasi yang lebih konsisten terhadap kinerja model. Hasil pengujian metode Naïve Bayes menggunakan cross-validation dapat dilihat dalam Tabel 4.1 dan Tabel 4.2 berikut.

**Tabel 4. 1 Tabel Hasil Pengujian NB ( Naïve Bayes )**

Pembagian Data	Akurasi	Presisi	Recall
Data Training 80%	84%	79%	88%
Data testing 20%			

**Tabel 4. 2 Tabel Confusion Matrix**

	Predict Negative	Predict Positive
Actual Negative	283	2
Actual Positive	55	120

Akurasi sebesar 84% yang disebutkan dalam konteks pengujian menggunakan metode Naïve Bayes mengacu pada kemampuan model untuk memprediksi sentimen terhadap vaksinasi COVID-19 berdasarkan data testing. Proses pengujian dimulai dengan membagi dataset menjadi data training (80%) untuk melatih model dan data testing (20%) untuk menguji performa model yang telah dilatih. Selama pengujian, model Naïve Bayes mengklasifikasikan setiap data testing dan membandingkan prediksinya dengan kategori sebenarnya dari sentimen vaksinasi. Akurasi dihitung dengan membagi jumlah prediksi yang benar dengan total jumlah data testing, kemudian dikonversi menjadi persentase. Dengan akurasi sebesar 84%, hasil ini menunjukkan bahwa model mampu memprediksi dengan benar 84% dari keseluruhan data testing, memberikan gambaran yang baik tentang kinerja model dalam analisis sentimen terhadap vaksinasi COVID-19.

Akurasi sebesar 84% yang dicapai dalam penelitian ini mengindikasikan bahwa model memiliki kinerja yang baik dalam mengklasifikasikan sentimen terhadap vaksinasi COVID-19 berdasarkan data yang digunakan. Namun, untuk menilai apakah model tersebut mengalami overfitting atau underfitting, perlu dilakukan evaluasi lebih lanjut terhadap kurva belajar (learning curve) dan analisis terhadap data validasi. Jika kurva belajar menunjukkan bahwa kedua kurva (training dan validation) konvergen dan tidak terdapat perbedaan yang signifikan antara keduanya, serta nilai akurasi pada data validasi tetap tinggi, ini mengindikasikan bahwa model tidak mengalami overfitting. Sebaliknya, jika terdapat perbedaan yang besar antara akurasi pada data training dan data validasi, bisa jadi model mengalami overfitting terhadap data training. Selain itu, evaluasi terhadap metrik lain seperti presisi, recall, dan F1-score juga penting untuk memperoleh gambaran yang lebih komprehensif tentang kinerja model. Dengan demikian, untuk memastikan bahwa akurasi 84% tersebut valid dan tidak terdapat

overfitting atau underfitting, diperlukan analisis lebih mendalam terhadap berbagai aspek performa model dan penggunaan teknik evaluasi yang tepat.

#### 4.4 Pembahasan

Dari hasil implementasi dan pengujian yang telah dilakukan, dapat disimpulkan bahwa sistem analisis sentimen yang dikembangkan mampu mengklasifikasikan sentimen positif dan negatif terkait vaksinasi COVID-19 menggunakan data sentimen yang diperoleh dari media sosial X.

Pengujian sistem menunjukkan bahwa metode Naïve Bayes mencapai tingkat akurasi sebesar 84%, presisi sebesar 79%, dan recall sebesar 88% dengan menggunakan pembagian data training sebesar 80% dan data testing sebesar 20%. Tingkat akurasi metode Naïve Bayes sangat dipengaruhi oleh jumlah data dalam kamus sentimen yang digunakan. Semakin besar jumlah data dalam kamus, semakin tinggi nilai akurasi yang dihasilkan, dan sebaliknya, semakin sedikit jumlah data dalam kamus, semakin rendah nilai akurasi yang diperoleh.

Tabel 4.1 menyajikan hasil pengujian metode Naïve Bayes yang mencakup akurasi, presisi, dan recall. Selain itu, Tabel 4.2 menampilkan confusion matrix yang menggambarkan hasil prediksi dari metode Naïve Bayes terhadap data testing. Dari tabel tersebut, dapat dilihat bahwa metode Naïve Bayes berhasil memprediksi sejumlah data dengan benar sebagai sentimen negatif maupun positif, meskipun terdapat beberapa kasus di mana prediksi tidak sesuai dengan kenyataan yang sebenarnya.

## **BAB 5**

### **KESIMPULAN DAN SARAN**

#### **5.1 Kesimpulan**

Berdasarkan penelitian yang telah dilakukan, kesimpulan yang dapat diambil pada penelitian mengenai sentimen pengguna media social X terhadap vaksinasi COVID-19 menggunakan Lexicon-Based dan Naive Bayes adalah sebagai berikut:

1. Hasil penelitian menunjukkan bahwa penggunaan metode Lexicon-Based dan Naive Bayes dalam analisis sentimen pengguna X terhadap vaksinasi COVID-19 memiliki kinerja yang cukup baik dalam mengklasifikasikan dan memprediksi sentimen. Dengan akurasi sebesar 84%, f1-score sebesar 82%, presisi sebesar 79%, dan recall sebesar 88%, metode ini mampu memberikan gambaran yang baik tentang pandangan masyarakat terhadap vaksinasi.
2. Model terbaik yang dihasilkan dari penelitian ini adalah model yang menggunakan metode Naive Bayes dengan pembagian data training sebesar 80% dan data testing sebesar 20%. Model ini memiliki tingkat akurasi sebesar 84%, f1-score sebesar 82%, presisi sebesar 79%, dan recall sebesar 88%. Hal ini menunjukkan bahwa model tersebut cukup efektif dalam memprediksi sentimen terkait vaksinasi COVID-19 berdasarkan data yang dianalisis.
3. Model Lexicon-Based dan Naive Bayes mampu memprediksi sentimen positif dan negatif terkait vaksinasi COVID-19, menunjukkan kemampuannya dalam memahami beragam pandangan masyarakat terhadap vaksinasi. Dengan menggunakan confusion matrix, kita dapat melihat bahwa model mampu memprediksi dengan baik, baik sentimen positif maupun negatif, meskipun terdapat beberapa kesalahan dalam prediksi.
4. Berdasarkan jumlah dataset penelitian yang telah dilakukan mendapatkan hasil sentimen yang lebih cenderung mengarah pada sentimen negatif.

## 5.2 Saran

Berikut adalah beberapa saran untuk penelitian selanjutnya guna meningkatkan kualitas dan efisiensi analisis sentimen terkait vaksinasi COVID-19:

1. Diharapkan pada penelitian berikutnya dapat memperbaiki proses pada pembersihan data agar lebih baik dan akurat dalam pemrosesan data untuk menghasilkan analisis sentimen yang lebih berkualitas.
2. Mempercepat proses training model agar lebih cepat dalam menghasilkan hasil analisis sentimen dan meningkatkan responsifitas sistem.
3. Memperluas dan memperkaya dataset yang digunakan agar mencakup berbagai sudut pandang masyarakat terhadap vaksinasi COVID-19, sehingga hasil analisis lebih representatif dan akurat.
4. Mengeksplorasi metode atau algoritma lain yang mungkin memberikan kinerja lebih baik dalam analisis sentimen untuk mencapai tingkat akurasi dan keandalan yang optimal.
5. Menggunakan berbagai hashtag lain yang relevan selain #vaksincovid19 agar data lebih banyak dan beragam, sehingga hasil analisis lebih representatif.

## DAFTAR PUSTAKA

- Abd, D. H., Abbas, A. R., & Sadiq, A. T. (2021). Analyzing sentiment system to specify polarity by lexicon-based. *Bulletin of Electrical Engineering and Informatics*, 10(1), 283–289. <https://doi.org/10.11591/eei.v10i1.2471>
- Arief, R., & Imanuel, K. (2019). Analisis Sentimen Topik Viral Desa Penari Pada Media Sosial Twitter Dengan Metode Lexicon Based. *Jurnal Ilmiah Matrik*, 21(3), 242–250. <https://doi.org/10.33557/jurnalmatrik.v21i3.727>
- Chi, G., & Cushman, M. (2020). Attention and citation: Common interests of researchers and journals. *Research and Practice in Thrombosis and Haemostasis*, 4(3), 353–356. <https://doi.org/10.1002/rth2.12322>
- Christina, S., & Ronaldo, D. (2020). Studi Literatur Sistematis Terhadap Pengembangan lexicon Sentiment. *Jurnal ELTIKOM*, 4(2), 121–131. <https://doi.org/10.31961/eltikom.v4i2.211>
- Cotfas, L. A., Delcea, C., Roxin, I., Ioanas, C., Gherai, D. S., & Tajariol, F. (2021). The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month following the First Vaccine Announcement. *IEEE Access*, 9, 33203–33223. <https://doi.org/10.1109/ACCESS.2021.3059821>
- D'Andrea, E., Ducange, P., Bechini, A., Renda, A., & Marcelloni, F. (2019). Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, 116, 209–226. <https://doi.org/10.1016/j.eswa.2018.09.009>
- Dongo, I., Cadinale, Y., Aguilera, A., Martínez, F., Quintero, Y., & Barrios, S. (2020). Web Scraping versus Twitter API. 263–273. <https://doi.org/10.1145/3428757.3429104>
- Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4), 491-511.
- Kusumawati, I. 2017. (2017). Analisa Sentimen Menggunakan Lexicon Based Kenaikan sentimen Rokok Pada Media Sosial Twitter. *Analisa Sentimen Menggunakan Lexicon Based Untuk Melihat Persepsi Masyarakat Terhadap Kenaikan harga Rokok Pada Media Sosial Twitter*.
- Matošević, G., & Bevanda, V. (2020). Sentiment analysis of tweets about COVID-19 disease during pandemic. In 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO) (pp. 1290-1295). IEEE.

Pamungkas, E. W., & Putri, D. G. P. (2016, August). An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia. In 2016 6th International Annual Engineering Seminar (InAES) (pp. 28-31). IEEE.

Prayoga, N. R., Fahrudin, T. M., Kamisutara, M., Rahagiyanto, A., Alfath, T. P., Winardi, S., & Susilo, K. E. (2020). Unsupervised Twitter Sentiment Analysis on The Revision of Indonesian Code Law and the Anti-Corruption Law using Combination Method of Lexicon Based and Agglomerative Hierarchical Clustering. *EMITTER International Journal of Engineering Technology*, 8(1), 200-220.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python natural language processing toolkit for many human languages. *ArXiv*. <https://doi.org/10.18653/v1/2020.acl-demos.14>

Rachman, F. F., & Pramana, S. (2020). Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter. *Health Information Management Journal*, 8(2), 100–109. <https://inohim.esaunggul.ac.id/index.php/INO/article/view/223/175>

Raghupathi, V., Ren, J., & Raghupathi, W. (2020). Studying public perception about vaccination: A sentiment analysis of tweets. *International journal of environmental research and public health*, 17(10), 3464.

Rahayu, R. N. & S. (2021). Vaksin covid 19 di indonesia : analisis berita hoax. *Jurnal Ekonomi, Sosial & Humaniora*, 2(07), 39–49. <https://www.jurnaltelektiva.com/index.php/jurnal/article/view/422>

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information (Switzerland)*, 11(4). <https://doi.org/10.3390/info11040193>

Sharma, C., Whittle, S., Haghghi, P. D., Burstein, F., & Keen, H. (2020). Sentiment analysis of social media posts on pharmacotherapy: A scoping review. *Pharmacology research & perspectives*, 8(5), e00640.

Sallam, M. (2021). Covid-19 vaccine hesitancy worldwide: A concise systematic review of vaccine acceptance rates. *Vaccines*, 9(2), 1–15. <https://doi.org/10.3390/vaccines9020160>

Susilo, A., Rumende, C. M., Pitoyo, C. W., Santoso, W. D., Yulianti, M., Herikurniawan, H., Sinto, R., Singh, G., Nainggolan, L., Nelwan, E. J., Chen, L. K., Widhani, A., Wijaya, E., Wicaksana, B., Maksum, M., Annisa, F., Jasirwan, C. O. M., & Yunihastuti, E. (2020). Coronavirus Disease 2019: Tinjauan Literatur Terkini. *Jurnal Penyakit Dalam Indonesia*, 7(1), 45. <https://doi.org/10.7454/jpdi.v7i1.415>

Ward, J. W., & del Rio, C. (2020). The COVID-19 Pandemic: An Epidemiologic, Public Health, and Clinical Brief. *Clinical Liver Disease*, 15(5), 170–174.  
<https://doi.org/10.1002/cl>



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN,  
RISET, DAN TEKNOLOGI**  
**UNIVERSITAS SUMATERA UTARA**  
**FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI**

Jalan Universitas No. 9A Gedung A, Kampus USU Medan 20155, Telepon: (061) 821007  
Laman: <http://Fasilkomti.usu.ac.id>

**KEPUTUSAN  
DEKAN FAKULTAS ILMU KOMPUTER  
DAN TEKNOLOGI INFORMASI  
NOMOR :2359/UN5.2.14.D/SK/SPB/2024**

**DEKAN FAKULTAS ILMU KOMPUTER  
DAN TEKNOLOGI INFORMASI UNIVERSITAS SUMATERA UTARA**

Membaca : Surat Permohonan Mahasiswa Fasilkom-TI USU tanggal 3 Juli 2024 perihal permohonan ujian skripsi:

    Nama : Miftah Aulia  
    NIM : 171402009  
    Program Studi : Sarjana (S-1) Teknologi Informasi  
    Judul Skripsi : Analisis Sentimen Pengguna Media Social X Terhadap Vaksinasi Covid 19 Menggunakan Lexicon Based Dan Naive Bayes

Memperhatikan : Bahwa Mahasiswa tersebut telah memenuhi kewajiban untuk ikut dalam pelaksanaan Meja Hijau Skripsi Mahasiswa pada Program Studi Sarjana (S-1) Teknologi Informasi Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara TA 2023/2024.

Menimbang : Bahwa permohonan tersebut diatas dapat disetujui dan perlu ditetapkan dengan surat keputusan

Mengingat :  
1. Undang-undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional.  
2. Peraturan Pemerintah Nomor 17 tahun 2010 tentang pengelolaan dan penyelenggara pendidikan.  
3. Keputusan Rektor USU Nomor 03/UN5.1.R/SK/SPB/2021 tentang Peraturan Akademik Program Sarjana Universitas Sumatera Utara.  
4. Surat Keputusan Rektor USU Nomor 1876/UN5.1.R/SK/SDM/2021 tentang pengangkatan Dekan Fasilkom-TI USU Periode 2021-2026

**MEMUTUSKAN**

Menetapkan Pertama :

: Membentuk dan mengangkat Tim Penguji Skripsi mahasiswa sebagai berikut:

    Ketua : Dr. Romi Fadillah Rahmat, B.Comp.Sc., M.Sc.  
                  NIP: 198603032010121004

    Sekretaris : Rossy Nurhasanah S.Kom., M.Kom  
                  NIP: 198707012019032016

    Anggota Penguji : Prof. Dr. Drs. Opim Salim Sitompul M.Sc  
                          NIP: 196108171987011001

    Anggota Penguji : Dr. Erna Budhiarti Nababan M.IT  
                          NIP: 196210262017042001

    Moderator : -

    Panitera : -

Kedua : Segala biaya yang diperlukan untuk pelaksanaan kegiatan ini dibebankan pada Dana Penerimaan Bukan Pajak (PNPB) Fasilkom-TI USU Tahun 2024.

Ketiga : Keputusan ini berlaku sejak tanggal ditetapkan dengan ketentuan bahwa segala sesuatunya akan diperbaiki sebagaimana mestinya apabila dikemudian hari terdapat kekeliruan dalam surat keputusan ini.

Tembusan :

1. Ketua Program Studi Sarjana (S-1) Teknologi Informasi
2. Yang bersangkutan
3. Arsip

Medan, 04 Juli 2024

Ditandatangani secara elektronik oleh:  
Dekan



Maya Silvi Lydia  
NIP 197401272002122001