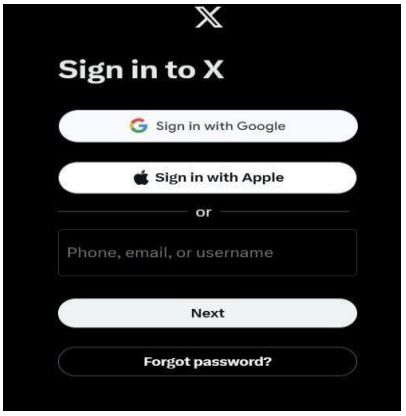
BAB IV
HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Setelah perancangan selesai, langkah selanjutnya adalah melakukan implementasi dari tiap langkah tersebut, yang dimulai dari pengumpulan data, pembersihan data, dan pelabelan data dan menggambail apa saja yang perlu diambil dan di hilangkan.

Berdasarkan penjelasan sebelumnya, proses pengumpulan data atau *crawling* data dilakukan melalui serangkaian langkah berikut:

1. *Registrasi* Akun *Developer Twitter*



Gambar 4.1 *Registrasi Akun*

Mulai menggunakan mode pengembangan gratis, setelah akun disetujui, pihak terkait akan memberikan akses ke mode pengembangan gratis, gunakan *API Twitter* dalam mode ini untuk mengembangkan dan menguji aplikasi.

2. *Acsess* Token

```
#@title Twitter Auth Token

twitter_auth_token = 'e3d19be6a62bdf5d825
```

Gambar 4.2 *Auth Token*

Kode tersebut bersifat sangat rahasia, dengan menggunakan akses web *service* yang diberikan, pengguna dapat mengambil data *tweet*, termasuk tanggal, id, nama pengguna, dan teks *tweet*, menggunakan bahasa pemrograman

atau aplikasi apa pun, dengan mematuhi ketentuan yang berlaku di situs tersebut.

3. *Crawling* Data

```
# Crawl Data

filename = 'debat 18-31_Jan.csv'
search_keyword = 'Debat Capres until:2024-02-07 since:2024-01-18 lang:id'
limit = 500

!npx --yes tweet-harvest@2.2.8 -o "{filename}" -s "{search_keyword}" -l {limit} --token {twitter_auth_token}
```

Gambar 4.3 *Crawling Data*

Gambar 4.3 merupakan proses mengumpulkan data dari *twitter* menggunakan alat *tweet-harvest*. Prosesnya dimulai dengan menentukan kata kunci pencarian dan rentang tanggal yang relevan. Kata kunci tersebut digunakan sebagai filter untuk mengambil *tweet* yang berhubungan dengan topik tertentu, sedangkan rentang tanggal membatasi waktu *tweet-tweet* tersebut diposting.

Setelah parameter pencarian ditetapkan, kode menjalankan perintah untuk memulai pengumpulan data dari *twitter*. *tweet-harvest* bertanggung jawabuntuk mengambil data yang sesuai dengan kriteria pencarian tersebut.Hasilnya disimpan dalam *file CSV*.

4. Hasil Crawling Data

	created_at	id_str	full_text	quote_count	reply_count	retweet_count	favorite_count	lang	user_id_str	conversation_id_str	username	
0	Fri Jan 26 02:39:53 +0000 2024	1750710116578685138	Sebuah ajakan salam 4 jari sebagai ekspresi po...	12	5	99	167	in	57222373	1750710116578685138	gitaputrid	h
1	Thu Jan 25 18:05:24 +0000 2024	1750580645888430516	ternyata jadi anak presiden se-privileged itu ...	19	25	113	823	in	1277749122296983552	1750580645888430516	UGM_FESS	https
2	Fri Jan 26 03:19:37 +0000 2024	1750720117896585466	19 Hari Jelang Pencoblosan TKN Optimistis Pra...	0	0	1	2	in	886878090768424960	1750720117896585466	yosephrosario_	ht
3	Fri Jan 26 02:21:18 +0000 2024	1750705442895085616	Calon Wakil Presiden nomor urut 3 Mahfud MD di...	12	0	6	27	in	1162545299614662656	1750705442895085616	catchmeupid	http
4	Fri Jan 26 02:28:19 +0000 2024	1750707207447150765	Waktu saya kuliah seenggaknya 2-3 semester	13	9	144	380	in	351930241	1750707207447150765	gibranhuzaifah	h

Gambar 4.4 Hasil *Crawling Data*

Hasil crawling pada gambar 4.4 menunjukan data yang berhasil diambildari *twitter*. Setiap baris data mencakup informasi tentang *tweet*, termasuk waktu posting, teks *tweet*, dan jumlah interaksi seperti balasan, *retweet*, dan *like.*

5. Simpan Data

Setelah data dikumpulkan, terkadang perlu disimpan dalam format yang sesuai agar dapat diakses dan digunakan kembali dengan mudah. Penyimpanan data ini dapat dilakukan dalam berbagai format, seperti file CSV, *Excel*, JSON, atau yang lainya.

4.2 *Preprocessing* Data

Berikut adalah hasil dari langkah - langkah *prapocessing data* :

4.2.1 Hasil *Import* Data

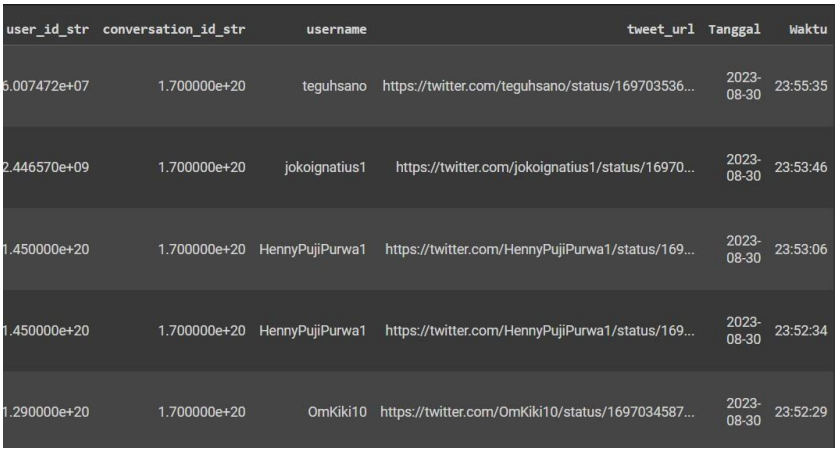
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8876 entries, 0 to 8875
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   created_at           8876 non-null   object
 1   id_str               8876 non-null   float64
 2   full_text            8876 non-null   object
 3   quote_count          8875 non-null   float64
 4   reply_count          8875 non-null   float64
 5   retweet_count        8875 non-null   float64
 6   favorite_count       8875 non-null   float64
 7   lang                 8875 non-null   object
 8   user_id_str          8855 non-null   float64
 9   conversation_id_str  8875 non-null   float64
 10  username             8875 non-null   object
 11  tweet_url            8875 non-null   object
dtypes: float64(7), object(5)
memory usage: 832.2+ KB
```

Gambar 4.5 Hasil Import Data

Dataset pada gambar 4.5 yang diimpor berisi informasi tentang *tweet* dengan total 8876 entri dan terdiri dari 12 kolom. Kolom-kolom tersebut mencakup waktu pembuatan *tweet* (`*created_at*`), ID *tweet* (`*id_str*`), teks lengkap tweet (`*full_text*`), jumlah kutipan (`*quote_count*`), jumlah balasan (`*reply_count*`), jumlah *retweet* (`*retweet_count*`), jumlah favorit (`*favorite_count*`), bahasa *tweet* (`*lang*`), ID pengguna (`*user_id_str*`), ID percakapan (`*conversation_id_str*`), nama pengguna (`*username*`), dan URL *tweet* (`*tweet_url*`). Sebagian kolom memiliki nilai yang hilang, seperti `*quote_count*`, `*reply_count*`, `*retweet_count*`, `*favorite_count*`, `*lang*`, `*user_id_str*`, `*conversation_id_str*`, `*username*`, dan `*tweet_url*`. Tipe data yang digunakan meliputi *float64* untuk kolom numerik dan *object* untuk kolom teks. *Dataset* ini menggunakan memori sebesar 832.2+ KB. Sebelum analisis lebih lanjut dilakukan, *preprocessing* data diperlukan untuk

mengatasi nilai-nilai yang hilang dan memastikan data siap untuk analisis seperti analisis sentimen, frekuensi kata, dan interaksi pengguna.

4.2.2 Hasil Memisahkan Kolom Tanggal dan Waktu



user_id_str	conversation_id_str	username	tweet_url	Tanggal	Waktu
6.007472e+07	1.700000e+20	teguhsano	https://twitter.com/teguhsano/status/169703536...	2023-08-30	23:55:35
2.446570e+09	1.700000e+20	jokoignatius1	https://twitter.com/jokoignatius1/status/16970...	2023-08-30	23:53:46
1.450000e+20	1.700000e+20	HennyPujiPurwa1	https://twitter.com/HennyPujiPurwa1/status/169...	2023-08-30	23:53:06
1.450000e+20	1.700000e+20	HennyPujiPurwa1	https://twitter.com/HennyPujiPurwa1/status/169...	2023-08-30	23:52:34
1.290000e+20	1.700000e+20	OmKiki10	https://twitter.com/OmKiki10/status/1697034587...	2023-08-30	23:52:29

Gambar 4.6 Hasil Memisahkan Kolom Tanggal dan Waktu

Gambar 4.6 memperlihatkan proses pemisahan kolom `*created_at*` menjadi kolom `Tanggal` dan `Waktu` menghasilkan *dataset* yang lebih terstruktur dan mudah untuk dianalisis. Kolom `Tanggal` berisi informasi tentang tanggal kapan tweet dibuat dalam format `YYYY-MM-DD`, seperti `2023-08-30` untuk tweet pertama. Kolom `Waktu` berisi informasi tentang waktu spesifik kapan tweet diposting dalam format `HH:MM:SS`, seperti `23:55:35` untuk *tweet* pertama.

Pemisahan ini memungkinkan analisis yang lebih mendetail dan terfokus, seperti mengidentifikasi tren waktu dan frekuensi *tweet* berdasarkan hari, jam, atau periode waktu tertentu. Misalnya, kita bisa mengidentifikasi puncak aktivitas *tweet* atau pola posting harian dengan lebih mudah. Selain itu,analisis temporal dapat dilakukan untuk mengkorelasikan aktivitas *tweet* dengan peristiwa tertentu pada tanggal-tanggal tertentu, memberikan wawasan yang lebih mendalam tentang dinamika penggunaan *twitter* oleh pengguna.

4.2.3 Hasil Penghapusan Data Duplikat

Hasil penghapusan duplikat data menampilkan *dataframe* dengan 10 entri yang berisi kolom-kolom berikut: tanggal, waktu, *username,*

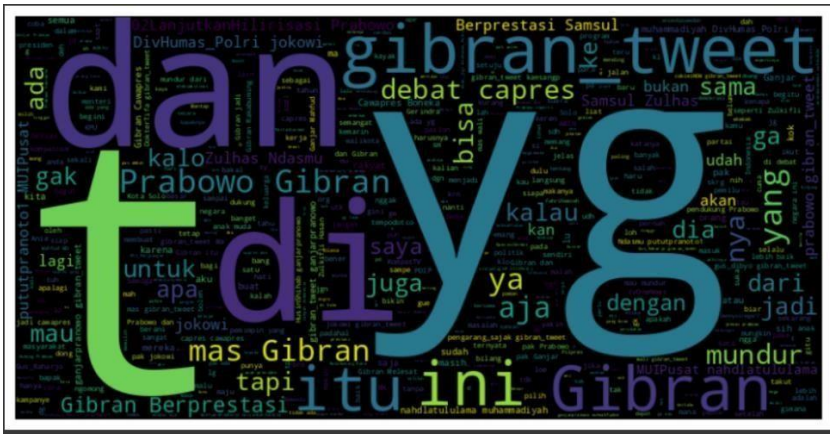
quote_count, reply_count, retweet_count, favorite_count, dan full_text.

	Tanggal	Waktu	username	quote_count	reply_count	retweet_count	favorite_count	full_text
0	2023-08-30	23:55:35	teguhsano	0.0	0.0	0.0	0.0	@gibran_tweet Gak bisa dianeksasi atau diinvasi?
1	2023-08-30	23:53:46	jokoignatius1	0.0	0.0	0.0	0.0	@gibran_tweet tulisan kota solo itu yang bener...
2	2023-08-30	23:53:06	HennyPujiPurwa1	0.0	0.0	0.0	1.0	@gibran_tweet masing-masing dan bisa menjalank...
3	2023-08-30	23:52:34	HennyPujiPurwa1	0.0	0.0	0.0	6.0	@gibran_tweet Nggih leres Bapak. Setiap permas...
4	2023-08-30	23:52:29	OmKiki10	0.0	0.0	0.0	1.0	@FennyAngela6 @gibran_tweet @KemenBUMN @BANKBR...
5	2023-08-30	23:50:54	agent_pin	0.0	0.0	0.0	0.0	@gibran_tweet Kl HULK itu orang HIJAU, kalau o...
6	2023-08-30	23:49:16	dinenggrss	0.0	0.0	0.0	0.0	@gibran_tweet Orang biru kek avatar donk mas
7	2023-08-30	23:46:25	Avario15	0.0	0.0	0.0	0.0	@TaryokoL @NoviAd14 @Arman_topbgt @gibran_twee...
8	2023-08-30	23:42:49	menyamenyanen	0.0	1.0	0.0	0.0	@kukuh_budiarto @gibran_tweet hihi jadi mau ny...
9	2023-08-30	23:42:32	pandusatria_32	0.0	0.0	0.0	0.0	@gibran_tweet Mbok referendum aja mas,

Gambar 4.7 Hasil Penghapussan Data Duplikat

Gambar 4.7 adalah sebuah *dataframe* yang berisi 10 entri unik, yang memastikan bahwa tidak ada data yang berulang. Data ini mencakup berbagai interaksi *tweet*, termasuk balasan, kutipan, *retweet*, dan favorit yang menunjukkan keterlibatan pengguna dengan *tweet* tersebut. Dengan membersihkan data dari duplikat, analisis lebih lanjut dapat dilakukan dengan akurasi yang lebih tinggi, tanpa risiko bias dari data yang berulang.

4.2.4 Hasil *WORDCLOUD*



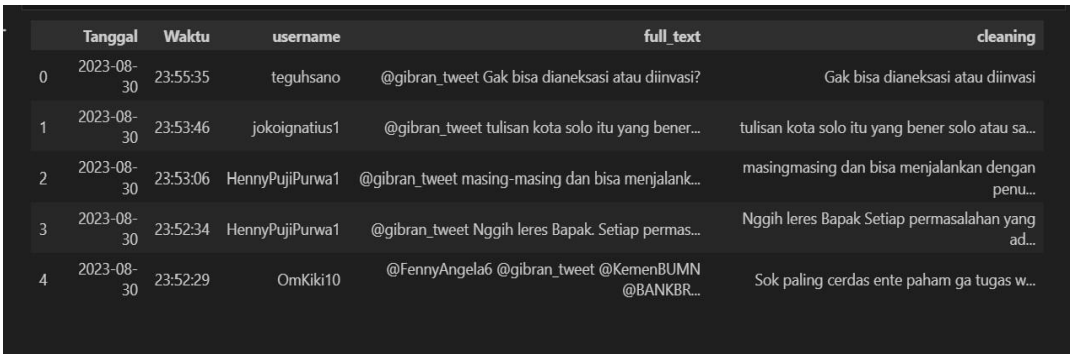
Gambar 4.8 Hasil *WORDCLOUD*

Dari data kata pada gambar 4.8 yang disajikan, dapat dilihat bahwa kata-kata seperti "gibran_*tweet*", "Gibran", "dan", "yg", "di", "yang", "Prabowo", "ini", "itu", dan "mas" mendominasi teks yang dianalisis. Frekuensi tinggi kata "@gibran_*tweet*" dan "Gibran" menunjukkan bahwa diskusi terpusat pada topik yang berkaitan dengan akun *twitter* @gibran_*tweet* dan individu bernama "Gibran". Sementara itu, kata-kata umum seperti "dan","yg", "di", dan "yang" mencerminkan struktur bahasa Indonesia yang

digunakan dalam kalimat untuk menghubungkan atau merujuk pada objek tertentu.

4.3 Hasil Data *Cleaning*

Hasil dari penghapusan *URL, HTML, Emoji,, Simbol, number, username* dapat di liahat pada gambar 4.9.



	Tanggal	Waktu	username	full_text	cleaning
0	2023-08-30	23:55:35	teguhsano	@gibran_tweet Gak bisa dianeksasi atau diinvasi?	Gak bisa dianeksasi atau diinvasi
1	2023-08-30	23:53:46	jokoignatius1	@gibran_tweet tulisan kota solo itu yang bener...	tulisan kota solo itu yang bener solo atau sa...
2	2023-08-30	23:53:06	HennyPujiPurwa1	@gibran_tweet masing-masing dan bisa menjalank...	masingmasing dan bisa menjalankan dengan penu...
3	2023-08-30	23:52:34	HennyPujiPurwa1	@gibran_tweet Nggih leres Bapak. Setiap permas...	Nggih leres Bapak Setiap permasalahan yang ad...
4	2023-08-30	23:52:29	OmKiki10	@FennyAngela6 @gibran_tweet @KemenBUMN @BANKBR...	Sok paling cerdas ente paham ga tugas w...

Gambar 4.9 Hasil Data *Cleaning*

Pada tahap cleaning yang terdapat pada gambar 4.9 terlihat perubahan teks dari teks awal " @gibran_tweet Gak bisa dianeksasi atau diinvasi?" menjadi "Gak bisa dianeksasi atau diinvasi"

4.3.1 Case Folding

Case folding yang telah diterapkan hasilnya dapat dilihat dalam gambar 4.10 di bawah ini:

	Tanggal	Waktu	username	full_text	cleaning	case_folding
0	2023-08-30	23:55:35	teguhsano	@gibran_tweet Gak bisa dianeksasi atau diinvasi?	Gak bisa dianeksasi atau diinvasi	gak bisa dianeksasi atau diinvasi
1	2023-08-30	23:53:46	jokoignatius1	@gibran_tweet tulisan kota solo itu yang bener...	tulisan kota solo itu yang bener solo atau sa...	tulisan kota solo itu yang bener solo atau sa...
2	2023-08-30	23:53:06	HennyPujiPurwa1	@gibran_tweet masing-masing dan bisa menjalank...	masingmasing dan bisa menjalankan dengan penu...	masingmasing dan bisa menjalankan dengan penu...
3	2023-08-30	23:52:34	HennyPujiPurwa1	@gibran_tweet Nggih leres Bapak. Setiap permas...	Nggih leres Bapak Setiap permasalahan yang ad...	nggih leres bapak setiap permasalahan yang ad...
4	2023-08-30	23:52:29	OmKiki10	@FennyAngela6 @gibran_tweet @KemenBUMN @BANKBR...	Sok paling cerdas ente paham ga tugas w...	sok paling cerdas ente paham ga tugas w...

Gambar 4.10 Hasil *Floding*

Kolom '*full_text'* menampilkan teks asli *tweet*: "@gibran_tweet Gak bisa dianeksasi atau diinvasi?". Setelah proses pembersihan dalam kolom '*cleaning*', teks menjadi "Gak bisa dianeksasi atau diinvasi". Kemudian, dalam kolom *'case_folding'*, semua huruf dalam teks diubah menjadi huruf kecil, menghasilkan teks "gak bisa dianeksasi atau diinvasi.". Hal ini bertujuan untuk membuat teks lebih seragam dan konsisten, mempermudahanalisis lanjutan

4.1 Hasil Penguji Perangkat Lunak
Penelitian ini akan menggunakan perangkat lunak berbasis Python dalam melabeli data. Berikut adalah langkah-langkahnya:

seperti klasifikasi sentimen atau pemrosesan teks lainnya.

4.3.2 Hasil Tokenization

Hasil *Tokenization* yang telah di lakukan dapat dilihat pada gambar 4.11

full_text	cleaning	case_folding	tokenize
@gibran_tweet Gak bisa dianeksasi atau diinvasi?	Gak bisa dianeksasi atau diinvasi	gak bisa dianeksasi atau diinvasi	[gak, bisa, dianeksasi, atau, diinvasi]
@gibran_tweet tulisan kota solo itu yang bener...	tulisan kota solo itu yang bener solo atau sa...	tulisan kota solo itu yang bener solo atau sa...	[tulisan, kota, solo, itu, yang, bener, solo, ...
@gibran_tweet masing-masing dan bisa menjalank...	masingmasing dan bisa menjalankan dengan penu...	masingmasing dan bisa menjalankan dengan penu...	[masingmasing, dan, bisa, menjalankan, dengan,...
@gibran_tweet Nggih leres Bapak. Setiap permas...	Nggih leres Bapak Setiap permasalahan yang ad...	nggih leres bapak setiap permasalahan yang ad...	[nggih, leres, bapak, setiap, permasalahan, ya...
@FennyAngela6 @gibran_tweet @KemenBUMN @BANKBR...	Sok paling cerdas ente paham ga tugas w...	sok paling cerdas ente paham ga tugas w...	[sok, paling, cerdas, ente, paham, ga, tugas, ...

Gambar 4.11 Hasil *Tokenization*

Pada tahap tokenisasi, teks "gak bisa dianeksasi atau diinvasi" diubah menjadi daftar kata-kata terpisah: [gak, bisa, dianeksasi, atau, diinvasi]. Proses ini memecah teks menjadi unit-unit yang lebih kecil berupa kata-kataindividual, yang nantinya dapat diolah secara terpisah dalam analisis teks lebih lanjut. Dengan tokenisasi, teks dapat dipecah menjadi unit-unit yang lebih kecil untuk mempermudah analisis dan pemrosesan teks lebih lanjut.

4.3.3 Stopword Removal

Setelah penghapusan *stopwords*, kata-kata tersebut diubah menjadi: [gak, dianeksasi, diinvasi] seperti yang terlihat pada gambar 4.12.

full_text	cleaning	case_folding	tokenize	stopword removal
@gibran_tweet Gak bisa dianeksasi atau diinvasi?	Gak bisa dianeksasi atau diinvasi	gak bisa dianeksasi atau diinvasi	[gak, bisa, dianeksasi, atau, diinvasi]	[gak, dianeksasi, diinvasi]
@gibran_tweet tulisan kota solo itu yang bener...	tulisan kota solo itu yang bener solo atau sa...	tulisan kota solo itu yang bener solo atau sa...	[tulisan, kota, solo, itu, yang, bener, solo, ...	[tulisan, kota, solo, bener, solo, sala, mas]
@gibran_tweet masing-masing dan bisa menjalank...	masingmasing dan bisa menjalankan dengan penu...	masingmasing dan bisa menjalankan dengan penu...	[masingmasing, dan, bisa, menjalankan, dengan,...	[masingmasing, menjalankan, penuh, tanggung, t...
@gibran_tweet Nggih leres Bapak. Setiap permas...	Nggih leres Bapak Setiap permasalahan yang ad...	nggih leres bapak setiap permasalahan yang ad...	[nggih, leres, bapak, setiap, permasalahan, ya...	[nggih, leres, permasalahan, mohon, diadukan, ...
@FennyAngela6 @gibran_tweet	Sok paling cerdas	sok paling cerdas	[sok, paling, cerdas,	[sok, cerdas, ente,

Gambar 4.12 Hasil *Stopword Removal*

Proses ini melibatkan penghapusan kata-kata umum atau stopwords dari teks, seperti "bisa" dan "atau", yang biasanya tidak memberikan kontribusi

signifikan terhadap pemahaman makna dari teks. Dengan menghapus *stopwords*, fokus analisis dapat diperkuat pada kata-kata yang lebih berartiatau relevan dalam konteks tertentu, sehingga meningkatkan akurasi dan efisiensi dalam proses analisis teks.

4.3.4 Steming Data

Proses *steming data* terdapat pada gambar 4.13

full_text	cleaning	case_folding	tokenize	stopword removal	steming_data
@gibran_tweet Gak bisa dianeksasi atau diinvasi?	Gak bisa dianeksasi atau diinvasi	gak bisa dianeksasi atau diinvasi	[gak, bisa, dianeksasi, atau, diinvasi]	[gak, dianeksasi, diinvasi]	gak aneksasi invasi
@gibran_tweet tulisan kota solo itu yang bener...	tulisan kota solo itu yang bener solo atau sa...	tulisan kota solo itu yang bener solo atau sa...	[tulisan, kota, solo, itu, yang, bener, solo, ...	[tulisan, kota, solo, bener, solo, sala, mas]	tulis kota solo bener solo sala mas
@gibran_tweet masing-masing dan bisa menjalank...	masingmasing dan bisa menjalankan dengan penu...	masingmasing dan bisa menjalankan dengan penu...	[masingmasing, dan, bisa, menjalankan, dengan,...	[masingmasing, menjalankan, penuh, tanggung, t...	masingmasing jalan penuh tanggung terimakasih
@gibran_tweet Nggih leres Bapak. Setiap permas...	Nggih leres Bapak Setiap permasalahan yang ad...	nggih leres bapak setiap permasalahan yang ad...	[nggih, leres, bapak, setiap, permasalahan, ya...	[nggih, leres, permasalahan, mohon, diadukan, ...	nggih les masalah mohon adu kepala wilayah mas...
@FennyAngela6 @gibran_tweet @KemenBUMN @BANKBR	Sok paling cerdas ente paham ga tugas w	sok paling cerdas ente paham ga tugas w	[sok, paling, cerdas, ente, paham, ga, tugas	[sok, cerdas, ente, paham, ga, tugas, walikota...	sok cerdas ente paham ga tugas walikota bank b

Gambar 4.13 Hasil Steming Data

Gambar 4.13 memperlihatkan kata-kata seperti "dianeksasi" dan "diinvasi" menjadi "aneksasi" dan "invasi". Hal ini membantu dalam mengurangi variasi kata yang memiliki akar kata yang sama, sehingga fokus dapat lebih ditempatkan pada makna dasar dari setiap kata dalam teks. Proses ini tidak hanya menyederhanakan teks, tetapi juga membantu dalam klasifikasi dan pemrosesan lanjutan seperti penghitungan frekuensi kata atau identifikasi pola-pola tertentu dalam data yang dikumpulkan dari media sosial *twitter*.

4.3.5 Hasil Remove Nilai Null

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8875 entries, 0 to 8875
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Tanggal           8875 non-null   object
 1   Waktu             8875 non-null   object
 2   username          8875 non-null   object
 3   quote_count       8875 non-null   float64
 4   reply_count       8875 non-null   float64
 5   retweet_count     8875 non-null   float64
 6   favorite_count    8875 non-null   float64
 7   full_text         8875 non-null   object
 8   cleaning          8875 non-null   object
 9   case_folding      8875 non-null   object
 10  tokenize          8875 non-null   object
 11  stopword removal  8875 non-null   object
 12  steming_data      8875 non-null   object
dtypes: float64(4), object(9)
memory usage: 970.7+ KB
```

Gambar 4.14 Hasil *Remove Nilai Null*

Hasil pada gambar 4.14 menunjukkan informasi tentang *DataFrame* setelah dilakukan penghapusan nilai null. *DataFrame* ini memiliki total8875 baris dan 13 kolom. Setiap kolom memiliki jumlah non-null yangsama dengan total baris, yaitu 8875, yang menandakan bahwa tidak ada nilai null di dalamnya.

4.3.6 Hasil Normalization

Hasil normalisasi teks "tidak aneksasi invasi" mungkin bergantung pada konteks dan aturan normalisasi yang digunakan. Dalam beberapa kasus, normalisasi mungkin tidak mengubah teks tersebut karena teks tersebut sudah dalam bentuk yang cukup normal. Namun, jika aturan normalisasi termasuk mengubah kata "tidak" menjadi bentuk yang lebih umum, maka hasilnya mungkin menjadi "tidak aneksi invasi" seperti pada gambar 4.15.

Normalisasi teks bertujuan untuk mengubah teks ke dalam bentuk yang lebih standar atau terstruktur, sehingga mempermudah proses analisis atau pemrosesan selanjutnya. Dalam hal ini, normalisasi dapat mencakup pengubahan kata-kata menjadi bentuk dasarnya,penggabungan kata-kata yang

serupa, atau penghapusan unsur-unsur yang tidak relevan.

full_text	cleaning	case_folding	tokenize	stopword removal	steming_data	hasil_normalisasi
@gibran_tweet Gak bisa dianeksasi atau diinvasi?	Gak bisa dianeksasi atau diinvasi	gak bisa dianeksasi atau diinvasi	['gak', 'bisa', 'dianeksasi', 'atau', 'diinvasi']	['gak', 'dianeksasi', 'diinvasi']	gak aneksasi invasi	tidak aneksasi invasi
@gibran_tweet tulisan kota solo itu yang bener...	tulisan kota solo itu yang bener solo atau sa...	tulisan kota solo itu yang bener solo atau sa...	['tulisan', 'kota', 'solo', 'itu', 'yang', 'be...	['tulisan', 'kota', 'solo', 'bener', 'solo', '...	tulis kota solo bener solo sala mas	tulis kota solo benar solo sala mas
@gibran_tweet masing-masing dan bisa menjalank...	masingmasing dan bisa menjalankan dengan penu...	masingmasing dan bisa menjalankan dengan penu...	['masingmasing', 'dan', 'bisa', 'menjalankan',...	['masingmasing', 'menjalankan', 'penuh', 'tang...	masingmasing jalan penuh tanggung terimakasih	masingmasing jalan penuh tanggung terimakasih
@gibran_tweet Nggih leres Bapak. Setiap permas...	Nggih leres Bapak Setiap permasalahan yang ad...	nggih leres bapak setiap permasalahan yang ad...	['nggih', 'leres', 'bapak', 'setiap', 'permasa...	['nggih', 'leres', 'permasalahan', 'mohon', 'd...	nggih les masalah mohon adu kepala wilayah mas...	nggih les masalah mohon adu kepala wilayah mas...
FennyAngela6 @gibran_tweet @KemenBUMN @BANKBR...	Sok paling cerdas ente paham ga tugas w...	sok paling cerdas ente paham ga tugas w...	['sok', 'paling', 'cerdas', 'ente', 'paham', '...	['sok', 'cerdas', 'ente', 'paham', 'ga', 'tuga...	sok cerdas ente paham ga tugas walikota bank b...	sok cerdas ente paham tidak tugas walikota ban...

Gambar 4.15 Hasil Normalisasi

4.4 Hasil Labeling Data

	Tanggal	Waktu	username	hasil_normalisasi	Sentiments
0	2023-08-30	23:55:35	teguhsano	tidak aneksasi invasi	Tidak
1	2023-08-30	23:53:46	jokoignatius1	tulis kota solo benar solo sala mas	Tidak
2	2023-08-30	23:53:06	HennyPujiPurwa1	masing masing jalan penuh tanggung jawab terim...	Tidak
3	2023-08-30	23:52:34	HennyPujiPurwa1	nggih les masalah mohon adu kepala wilayah mas...	Tidak
4	2023-08-30	23:52:29	OmKiki10	sok cerdas ente paham tidak tugas walikota ban...	Iya

Gambar 4.16 Hasil Labeling Data Perkalimat

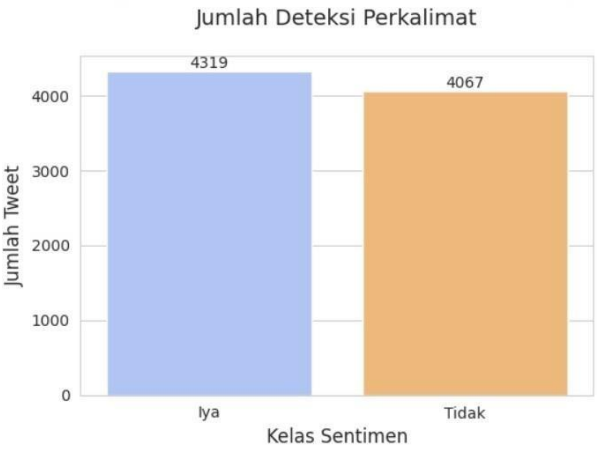
Gambar 4.16 Merupakan hasil lableing perkalimat dari analisis sentimen yang dilakukan terhadap teks *tweet*, *tweet* yang dianalisis telah diklasifikasikan ke dalam dua kategori sentimen, yaitu "Tidak" dan "Iya", menunjukkan apakah *tweet* tersebutdianggap memiliki sentimen negatif atau positif.

Sebagai contoh, baris pertama menunjukkan bahwa tweet yang diposting oleh pengguna "teguhsano" pada tanggal 30 Agustus 2023 jam 23:55:35 memiliki hasil normalisasi "tidak aneksasi invasi" dan diklasifikasikan sebagai sentimen "Tidak". Ini menunjukkan bahwa *tweet* tersebut dianggap memiliki sentimen negatif berdasarkan analisisnya setelah proses normalisasi dilakukan.

	Tanggal	Waktu	username	hasil_normalisasi	Sentiments
0	2023-08-30	23:55:35	teguhsano	tidak aneksasi invasi	Tidak
1	2023-08-30	23:53:46	jokoignatius1	tulis kota solo benar solo sala mas	Tidak
2	2023-08-30	23:53:06	HennyPujiPurwa1	masing masing jalan penuh tanggung jawab terim...	Tidak
3	2023-08-30	23:52:34	HennyPujiPurwa1	nggih les masalah mohon adu kepala wilayah mas...	Iya
4	2023-08-30	23:52:29	OmKiki10	sok cerdas ente paham tidak tugas walikota ban...	Iya

Gambar 4.17 Hasil Labeling Data Perkata

Dalam gambar 4.17 meskipun ada kata "tidak" dalam kalimat pada baris 3, sentimen perkalimatnya tetap dianggap "Tidak", sedangkan dalam kasus baris 4, sentimen perkata menunjukkan "Iya". Hal ini menunjukkan bahwa penilaian sentimen secara perkalimat dapat menghasilkan interpretasi yang lebih tepat tergantung pada konteks kalimat secara keseluruhan, sementara sentimen berdasarkan kata-kata mungkin tidak selalu mencerminkan keseluruhan makna kalimat.



Gambar 4.18 Hasil Jumlah Deteksi Perkalimat

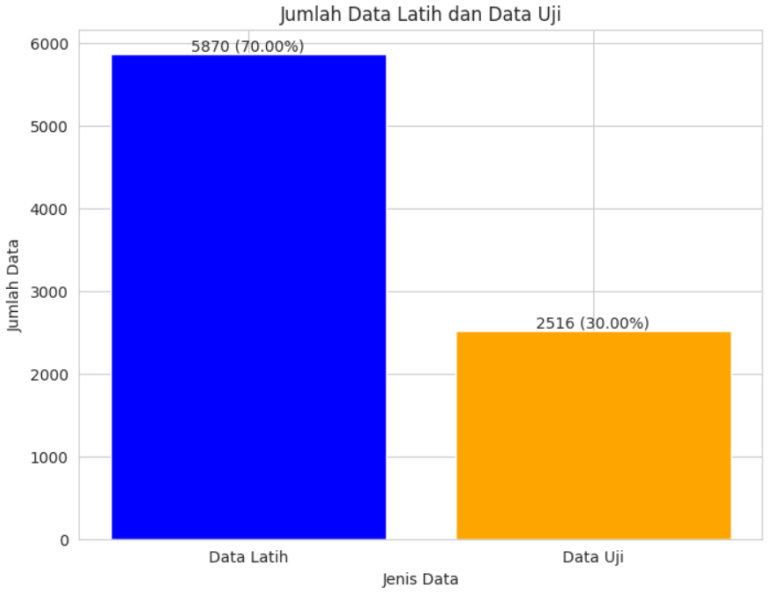
Gambar 4.19 Hasil Jumlah Deteksi Perkata

Diagram pada gambar 4.18 dan gambar 4.19 tersebut memberikan gambaran visual tentang distribusi jumlah data deteksi perkalimat dan per kata berdasarkan kelas sentimen "iya" dan "tidak". Data perkalimat menunjukkan bahwa dari total 4000 *tweet*, sebanyak 4319 *tweet* diklasifikasikan sebagai "iya", sementara 4067 *tweet* diklasifikasikan sebagai "tidak".

Sementara itu, data perkata menunjukkan bahwa dari total 5000 *tweet*, sebanyak 5016 *tweet* diklasifikasikan sebagai "iya" dan 3370 *tweet* diklasifikasikan sebagai "tidak". Dari diagram tersebut, dapat disimpulkan bahwa jumlah data yang diklasifikasikan sebagai "iya" lebih tinggi daripada yang diklasifikasikan sebagai "tidak" baik pada data perkalimat maupun per kata.

4.5 Hasil *Feature Engineering*

Feature engineering, yang bertujuan untuk menguji kinerja fitur dengan model yang dibuat dan meningkatkan fitur untuk direpresentasikan dalam bentuk data numerik atau matriks data. Sebelum menerapkan feature engineering, dataset dibagi menjadi data training dan data testing dengan rasio 70:30. Hasil dari pembagian tersebut adalah 5.870 data untuk data training dan 2.516 data untuk data testing.



Gambar 4.20 Hasil *Feature Engineering*

4.6 Hasil Model Evaluation *Algoritma K-Nearest Neighbor*

```
Accuracy: 0.75
===============
              precision    recall  f1-score   support

         Iya       0.75      0.76      0.75      1274
       Tidak       0.75      0.74      0.75      1242

    accuracy                           0.75      2516
   macro avg       0.75      0.75      0.75      2516
weighted avg       0.75      0.75      0.75      2516
```

Gambar 4.21 Hasil Model *Evaluasi KNN Perkalimat*

Hasil evaluasi model perkalimat pada gambar 4.21 tersebut menunjukkan akurasi sebesar 0.75, yang mengindikasikan bahwa model memiliki tingkat keakuratan sebesar 75%. Selain itu, dilaporkan juga hasil metrik *precision*, *recall*, dan *f1-score* untuk setiap kelas (Iya dan Tidak), serta untuk keseluruhan model.

Untuk kelas "Iya", *precision* adalah 0.75, yang berarti 75% dari prediksi yang diklasifikasikan sebagai "Iya" adalah benar. *Recall* sebesar 0.76 menunjukkan bahwa 76% dari keseluruhan data "Iya" berhasil diprediksi dengan benar oleh model. Nilai *f1-score* untuk kelas "Iya" adalah 0.75.

Sementara untuk kelas "Tidak", *precision* sebesar 0.75 menunjukkan bahwa75% dari prediksi yang diklasifikasikan sebagai "Tidak" adalah benar. *Recall* sebesar 0.74 menunjukkan bahwa 74% dari keseluruhan data "Tidak" berhasildiprediksi dengan benar oleh model. *Nilai f1-score* untuk kelas "Tidak" jugaadalah 0.75.

Kesimpulannya, hasil evaluasi menunjukkan bahwa model memiliki kinerjayang seimbang antara kelas "Iya" dan "Tidak", dengan akurasi total sebesar 75%.

```
Accuracy: 0.7087928464977645
===============
              precision    recall  f1-score   support

         Iya       0.77      0.73      0.75      1993
       Tidak       0.63      0.69      0.66      1362

    accuracy                           0.71      3355
   macro avg       0.70      0.71      0.70      3355
weighted avg       0.71      0.71      0.71      3355
```

Gambar 4.22 Hasil Model *Evaluasi KNN Perkata*

Hasil evaluasi model berdasarkan perkata pada gambar 4.22 akurasi model tercatat sebesar 70.88%, yang menunjukkan bahwa model mampu mengklasifikasikan dengan benar sekitar 70.88% dari total data yang diuji. Presisi untuk kelas "Iya" adalah 77%, yang berarti dari semua prediksi "Iya" yang dibuat oleh model, 77% di antaranya benar-benar "Iya". Sebaliknya, presisiuntuk kelas "Tidak" adalah 63%, menunjukkan bahwa dari semua prediksi "Tidak", hanya 63% yang benar-benar "Tidak". *Recall* untuk kelas "Iya" adalah 73%, menunjukkan bahwa dari semua data aktual yang benar-benar "Iya", model berhasil mengidentifikasi 73% di antaranya dengan benar. Untuk kelas "Tidak", *recall* adalah 69%, yang berarti model berhasil mengidentifikasi 69% dari semua data aktual yang benar-benar "Tidak". *F1-Score*, yang menggabungkan presisi dan *recall*, adalah 0.75 untuk kelas "Iya" dan 0.66 untuk kelas "Tidak". Dengan *support* masing-masing 1993 untuk kelas "Iya" dan 1362 untuk kelas "Tidak", *macro average* dan *weighted average* untuk presisi, *recall*, dan *F1-Score* semuanya berkisar antara 0.70 hingga 0.71. Dari hasil ini, dapat disimpulkan bahwa model lebih baik dalam mengidentifikasi *tweet* yang termasuk dalam kategori "Iya" dibandingkan "Tidak", namun masih ada ruang untuk perbaikan dalam meningkatkan presisi dan *recall*, terutama untuk kelas "Tidak".

4.7 Analisa Pendapat Ahli Bahasa Indonesia (Ibu Hasnawati Nasution)

Setelah mengevaluasi performa model klasifikasi, langkah berikutnya adalah membandingkan hasil tersebut dengan penilaian dari ahli Bahasa Indonesia. Sampel data yang digunakan untuk evaluasi dapat ditemukan pada lampiran, yaitu "Tabel Hasil Evaluasi Berdasarkan Pendapat Ahli Bahasa Indonesia". Hasil analisisnya disajikan pada tabel 4.1 berikut:

Tabel 4.1 Hasil Analisa Pendapat Ahli Bahasa Indonesia

No	Analisa	Hasil Klasifikasi	Konteks Tambahan
1	Tidak aneksasi invasi	Tidak	Tidak ada target spesifik, tidak ada konotasi negatif.
2	Tulis kota solo benar solo sala mas	Tidak	Tidak ada target spesifik, tidak ada konotasi negatif.
3	Masing-masing jalan penuh tanggung jawab terima kasih	Tidak	Tidak ada target spesifik, tidak ada konotasi negatif.
4	Nggih les masalah mohon adu kepala wilayah masing-masing kepala wilayah tingkat rendah tingkat tingkat teratas begitu makan gaji buta tupoksinya	Tidak	Tidak ada target spesifik, tidak ada konotasi negatif.
5	Sok cerdas ente paham tidak tugas walikota bank bumn luar ranah walikota	Iya	Mengandung konotasi negatif, ada target spesifik (walikota).
6	Hahaha kalau bawaslu tidak mengasih sangai bobby gibran ya meniru	Tidak	Tidak ada konotasi negatif.
7	Sudah maju ya mas dari target selesai tapi bisa maju september tahu tutup	Tidak	Tidak ada target spesifik, tidak ada konotasi negatif.
8	Gibran bawaslu dan berani	Tidak	Tidak ada konotasi negatif.
9	Cangkeme mudeng e curhat mulane ra nde utek	Tidak	Tidak ada target spesifik, tidak ada konotasi

			negatif.
10	Gibran anjing kayak tahi lo bangkai	Iya	Mengandung konotasi negatif, menyebut nama spesifik (Gibran), disebarkan di media sosial.

Tabel 4.1 di atas menggambarkan bagaimana berbagai kalimat dianalisis berdasarkan konteks tambahan yang diperlukan menurut pendapat ahli bahasa untuk menentukan apakah kalimat tersebut termasuk dalam kategori *cyberbullying*.

Menurut Ibu Hasnawati Nasution yang menjabat sebagai widyabasa ahli muda pada instansi kantor bahasa provinsi lampung. untuk menentukan apakah suatu kalimat termasuk dalam kategori *cyberbullying*, tidak cukup hanya melihat kalimat tersebut secara terpisah. Ada beberapa konteks tambahan yang perlu diperhatikan agar analisis ini akurat dan komprehensif. Ahli bahasa menjelaskan bahwa konsep dan definisi *cyberbullying* melibatkan penggunaan teknologi digital untuk mengirim, memposting, atau membagikan konten negatif, berbahaya, atau tidak benar tentang seseorang yang ditargetkan secara spesifik. Oleh karena itu, hanya mengandalkan satu kalimat tanpa konteks tambahan tidak cukup untuk mengidentifikasinya sebagai *cyberbullying*.

Konteks kalimat juga sangat penting. Sebuah kalimat tidak bisa dianggap sebagai *cyberbullying* hanya berdasarkan kata-kata yang digunakan. Pentinguntuk melihat konteks yang lebih luas, termasuk apakah kalimat tersebut ditujukan kepada seseorang yang spesifik, dan jika ada elemen tambahan seperti nama, foto, atau informasi pribadi lainnya yang disebutkan. Misalnya, kalimat yang menyebut nama seperti "Gibran" harus dianalisis lebih lanjut untuk menentukan konteks penggunaannya.

Selain itu, kata-kata yang digunakan dalam kalimat tersebut harus dianalisis apakah mengandung konotasi negatif atau ofensif menurut Kamus Besar Bahasa Indonesia (KBBI) atau penggunaan umum di media sosial. Kata-kata seperti "goblok," "sampah," atau "monyet" dapat dianggap sebagai penghinaan dan

memiliki konotasi negatif, terutama jika ditujukan kepada seseorang secara spesifik. Penyebaran di media sosial juga memainkan peran penting. Kalimat yang diposting atau disebarkan di platform media sosial dapat memperbesar dampak negatifnya terhadap target. Misalnya, kalimat yang mengandung konotasi negatif dan menyebut nama spesifik, seperti "Gibran anjing kayak tahi lo bangkai," memiliki potensi besar untuk dianggap sebagai *cyberbullying* jika diposting di media sosial.

4.8 Perbandingan Hasil Pendapat Ahli Bahasa Dengan Model Sistem

Berikut adalah tabel perbandingan hasil analisis kalimat terkait *cyberbullying* oleh sistem dan ahli bahasa:

Tabel 4.2 Perbandingan Hasil Pendapat Ahli Bahasa Dengan Model Sistem

No	Kalimat	Label Sistem	Label Ahli Bahasa
1	Tidak aneksasi invasi	Tidak	Tidak
2	Tulis kota solo benar solo sala mas	Tidak	Tidak
3	Masing-masing jalan penuh tanggung jawab terima kasih	Tidak	Tidak
4	Nggih les masalah mohon adu kepala wilayah masing-masing kepala wilayah tingkat rendah tingkat tingkat teratas begitu makan gaji buta tupoksinya	Tidak	Tidak
5	Sok cerdas ente paham tidak tugas walikota bank bumn luar ranah walikota	Iya	Iya
6	Hahaha kalau bawaslu tidak mengasih sangai bobby gibran ya meniru	Iya	Tidak
7	Sudah maju ya mas dari target selesai tapi bisa maju september tahu tutup	Tidak	Tidak
8	Gibran bawaslu dan berani	Iya	Tidak
9	Cangkeme mudeng e curhat mulane ra nde utek	Iya	Tidak
10	Gibran anjing kayak tahi lo bangkai	Iya	Iya

Tabel 4.2 di atas menggambarkan perbandingan hasil analisis kalimat terkait *cyberbullying* oleh sistem otomatis berbasis *machine learning* dan oleh ahli bahasa. Sistem otomatis cenderung menggunakan algoritma untuk mengidentifikasi konotasi negatif dalam kalimat. Pendekatan ini seringkali cepat tetapi bisa kurang akurat tanpa konteks tambahan. Di sisi lain, ahli bahasa mempertimbangkan konteks yang lebih luas, termasuk target spesifik, konotasi negatif, dan penyebaran di media sosial.

4.9 Kelebihan dan Kekurangan Sistem

4.9.1 Kelebihan

1. Sistem otomatis dapat memproses dan menganalisis teks dalam jumlah besar dengan sangat cepat, jauh lebih cepat daripada manusia.
2. Mampu menangani volume data yang sangat besar, termasuk ribuan hingga jutaan tweet, tanpa kesulitan.
3. Algoritma machine learning memberikan hasil yang konsisten karena tidak terpengaruh oleh faktor-faktor seperti kelelahan atau perubahanmood yang bisa mempengaruhi manusia.
4. Setelah sistem dikembangkan dan dilatih, biaya operasional untuk analisis data dalam skala besar cenderung lebih rendah dibandingkan dengan mempekerjakan banyak ahli bahasa.

4.9.2 Kekurangan

1. Sistem otomatis seringkali kurang mampu memahami konteks yang lebih luas atau nuansa dalam bahasa yang bisa mempengaruhi maknasebenarnya dari teks.
2. Kualitas hasil sangat tergantung pada data latih. Jika data latih tidak representatif atau berkualitas rendah, hasil analisis juga akan kurang akurat.
3. Algoritma mungkin mengalami kesulitan dalam menangani kalimat-kalimat yang ambigu atau memiliki makna ganda yang mudah dipahami oleh manusia.
4. Mesin sering kesulitan mendeteksi sarkasme atau ironi, yang bisa menyebabkan kesalahan dalam klasifikasi.

4.10 Kelebihan dan Kekurangan Ahli Bahasa

4.10.1 Kelebihan

1. Ahli bahasa dapat memahami konteks yang lebih luas, nuansa, dan makna implisit dalam teks yang mungkin terlewat oleh sistem otomatis.
2. Manusia bisa menyesuaikan analisis mereka berdasarkan perkembangan atau perubahan dalam bahasa dan budaya.
3. Ahli bahasa lebih baik dalam menangani kalimat yang ambigu atau memiliki makna ganda karena mereka bisa mempertimbangkan berbagai faktor tambahan.
4. Manusia lebih mampu mendeteksi sarkasme atau ironi yang mungkin tidak terdeteksi oleh sistem otomatis.

4.10.2 Kekurangan

1. Analisis oleh manusia jauh lebih lambat dibandingkan dengan sistem otomatis, terutama untuk volume data yang besar.
2. Menggunakan tenaga ahli bahasa dalam jumlah besar bisa menjadi sangat mahal, terutama untuk proyek yang melibatkan data dalam skala besar.
3. Hasil analisis oleh manusia bisa bervariasi karena faktor-faktor seperti kelelahan, bias pribadi, atau perubahan mood.
4. Kemampuan manusia untuk memproses data terbatas dan tidak bisa ditingkatkan dengan mudah seperti sistem otomatis.