**LAPORAN PROYEK**

**Analysis and Classification Twitter Data Stream of Racism Topic Using Vader Sentiment Analysis**

**12S4058 - PENGOLAHAN DATA BESAR**



**Disusun Oleh:**

1. 12S18036 - Sandraulina Siregar
2. 12S18042 - Indah Oktavia M. Sibarani
3. 12S18050 - Elsa Elfransyah Marbun
4. 12S18053 - Evelin Theresia Panjaitan

**SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2022**

# **DAFTAR ISI**

[DAFTAR ISI 2](#_Toc105328203)

[DAFTAR GAMBAR 3](#_Toc105328204)

[KODE PROGRAM 4](#_Toc105328205)

[BAB 1 PENDAHULUAN 5](#_Toc105328206)

[1.1 Latar Belakang 5](#_Toc105328207)

[1.2 Rumusan Masalah 6](#_Toc105328208)

[1.3 Tujuan 6](#_Toc105328209)

[1.4 Manfaat 6](#_Toc105328210)

[1.5 Ruang Lingkup 7](#_Toc105328211)

[BAB 2 SOLUSI 8](#_Toc105328212)

[2.1 Arsitektur *Big Data* 8](#_Toc105328213)

[2.2 *Machine Learning Pipeline* 9](#_Toc105328214)

[2.2.1 *Twitter Data Stream* 9](#_Toc105328215)

[2.2.2 *Data Preprocessing* 9](#_Toc105328216)

[2.2.3 *Sentiment Analysis* menggunakan Vader 11](#_Toc105328217)

[2.2.4 Visualization 12](#_Toc105328218)

[BAB 3 IMPLEMENTASI 13](#_Toc105328219)

[3.1 *Data Retrieval* 13](#_Toc105328220)

[3.2 *Data Preparation* 15](#_Toc105328221)

[3.3.1 Tokenization 16](#_Toc105328222)

[3.3.2 *Stopword Removal* 18](#_Toc105328223)

[3.3.3 *Stemming* 18](#_Toc105328224)

[3.3.3 *Filtering* 20](#_Toc105328225)

[BAB 4 HASIL & PEMBAHASAN 23](#_Toc105328226)

[4.1 Sentimen Analisis Vader 23](#_Toc105328227)

[4.2 Hasil Visualisasi 25](#_Toc105328228)

[4.3 Kesimpulan 27](#_Toc105328229)

# **DAFTAR GAMBAR**

[Gambar 1 Arsitektur Big Data 6](#_Toc105328120)

[Gambar 2 Machine Learning Pipeline 7](#_Toc105328121)

# **KODE PROGRAM**

[Kode Program 1 Tokenization 10](#_Toc105328151)

[Kode Program 2 Contoh Hasil Tokenization 10](#_Toc105328152)

[Kode Program 3 Stopword Removal 10](#_Toc105328153)

[Kode Program 4 Hasil Stopword Removal 11](#_Toc105328154)

[Kode Program 5 Stemming 11](#_Toc105328155)

[Kode Program 6 Hasil Stemming 11](#_Toc105328156)

[Kode Program 7 Import Library 13](#_Toc105328157)

[Kode Program 8 Twitter API 14](#_Toc105328158)

[Kode Program 9 Autentikasi API 14](#_Toc105328159)

[Kode Program 10 Crawling Data 14](#_Toc105328160)

[Kode Program 11 Read racism.csv 15](#_Toc105328161)

[Kode Program 12 Tampilan Hasil Crawling 15](#_Toc105328162)

[Kode Program 13 Import Library Data Preparation 16](#_Toc105328163)

[Kode Program 14 Cek Data racism.csv 16](#_Toc105328164)

[Kode Program 15 Clean Text - 1 17](#_Toc105328165)

[Kode Program 16 Clean Text - 2 17](#_Toc105328166)

[Kode Program 17 Lihat Hasil Clean Text 18](#_Toc105328167)

[Kode Program 18 Import nltk 18](#_Toc105328168)

[Kode Program 19 Remove Stopwords 18](#_Toc105328169)

[Kode Program 20 Stemming 19](#_Toc105328170)

[Kode Program 21 Menampilkan Hasil Steam Text 19](#_Toc105328171)

[Kode Program 22 Drop Kolom Date 19](#_Toc105328172)

[Kode Program 23 Menampilkan Hasil Clean Data 19](#_Toc105328173)

[Kode Program 24 Filtering - 1 20](#_Toc105328174)

[Kode Program 25 Filtering - 2 20](#_Toc105328175)

[Kode Program 26 Menampilkan Hasil Filtering 21](#_Toc105328176)

[Kode Program 27 Preprocess Data 21](#_Toc105328177)

[Kode Program 28 Hasil Preprocess Data 21](#_Toc105328178)

[Kode Program 29 Import Hasil Cleaning Data 22](#_Toc105328179)

[Kode Program 30 Install Vader Sentiment 23](#_Toc105328180)

[Kode Program 31 Import Library untuk Visualisasi 23](#_Toc105328181)

[Kode Program 32 Read clean\_data 23](#_Toc105328182)

[Kode Program 33 Import SentimentIntensityAnalyzer 24](#_Toc105328183)

[Kode Program 34 polarity\_score 24](#_Toc105328184)

[Kode Program 35 Compound Vader Sentiment 24](#_Toc105328185)

[Kode Program 36 Drop Vader Sentiment Column 25](#_Toc105328186)

[Kode Program 37 Hitung Hasil Vader Counts 25](#_Toc105328187)

[Kode Program 38 Word Cloud 26](#_Toc105328188)

[Kode Program 39 Matplotlib 26](#_Toc105328189)

[Kode Program 40 Grafik 27](#_Toc105328190)

# **BAB 1 PENDAHULUAN**

## **1.1 Latar Belakang**

Media sosial adalah sebuah media *online* dengan para pengguna yang dapat dengan mudah berpartisipasi, berbagi dan menciptakan isi meliputi blog, jejaring sosial, wiki dan dunia virtual. Media sosial atau *social networking* juga merupakan sebuah bentuk layanan internet sebagai komunitas online bagi orang yang memiliki kesamaan aktivitas, ketertarikan pada bidang tertentu, atau kesamaan latar belakang tertentu atau yang biasa disebut sebagai jaringan pertemanan[1]. Adapun beberapa situs jejaring sosial yang populer adalah LinkedIn, Facebook, dan Twitter. Penggunaan media sosial terus berkembang pesat dikarenakan media sosial dilengkapi dengan fitur yang dapat membuat penggunanya saling berkomunikasi dan berinteraksi. Media sosial yang menjadi topik bahasan pada proyek ini adalah Twitter. Dengan Twitter, pengguna dapat saling bertukar opini untuk berbagai isu dan topik ke dalam pesan berbasis teks yang disebut “tweet”. Salah satu topik yang sering menjadi *trending topic* di Twitter adalah rasisme. Rasisme adalah suatu sistem kepercayaan/ doktrin yang menyatakan bahwa perbedaan biologis ras manusia menentukan pencapaian budaya atau individu sehingga suatu ras tertentu merasa lebih superior dan memiliki hak untuk mengatur ras yang lainnya [1].

Data pada Twitter meningkat setiap hari sehingga memerlukan *framework* untuk memproses dan menganalisis *Big Data.* Beberapa *framework* yang dapat mengatasi hal tersebut adalah Hadoop, Apache Spark dan HBase. Pada proyek ini kami menggunakan apache spark dikarenakan apache spark memiliki beberapa komponen dan dukungan dari berbagai bahasa pemrograman seperti Spark Core, Spark SQL, MLlib dan lainnya. Apache Spark melakukan pemrosesan data secara *in-memory* sehingga waktu pemrosesannya menjadi lebih cepat.

Untuk mengenali dan mengekstraksi opini dalam bentuk teks, dapat dilakukan dengan menggunakan salah satu bidang Natural Language Processing (NLP) yaitu *Sentiment analysis*. Teknik dasar *Sentiment analysis* dibagi menjadi dua yaitu *Rule-based* dan *Machine learning-based*. Teknik *Rule-based sentiment analysis* menggunakan kamus kata-kata yang diberi label sentiment tertentu. Sedangkan pada teknik *Machine learning based sentiment analysi*s, fitur-fitur di dalamnya dapat mengenali sentimen (sudut pandang seseorang) dalam sebuah teks. Dengan *Sentiment analysis*, mengekstraksi informasi opini yang terdapat dalam data teks dari platform Twitter dapat dilakukan secara otomatis sehingga informasi yang sebelumnya tidak terstruktur dapat diubah menjadi data yang lebih terstruktur.

Pada dasarnya analisis sentimen bertujuan untuk menemukan pendapat, umpan balik atau ulasan, dan kemudian mengidentifikasi sentimen dan mengklasifikasikan *polarity* yaitu positif, negatif atau netral. Untuk mendapatkan *class* sentimen tersebut, kami menggunakan metode analisis Vader (*Valence Aware Dictionary and Sentiment Reasoner*). Vader menerapkan metode Lexicon. Metode Lexicon adalah metode yang tidak memerlukan data training atau data yang telah dilabeli namun sudah tersedia dalam kamus lengkap dengan kepolaritasan sentimennya. Oleh karena itu, pada proyek ini dilakukan sentiment analysis pada data Twitter dengan topik rasisme menggunakan Vader *sentiment analysis* [2]

## **1.2 Rumusan Masalah**

Berdasarkan latar belakang yang telah dikemukakan pada Bab 1, maka rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana melakukan pengambilan data Twitter dengan menggunakan API (*Application Programming Interface*) yang telah disediakan oleh Twitter.
2. Bagaimana memproses data secara *real-time* dan menerapkan *sentiment analysis* pada *tweet* dengan *keyword* yang telah ditentukan
3. Bagaimana membuat sistem yang mampu melakukan klasifikasi dan analisis sentimen dari topik tertentu pada Twitter menggunakan vader sentiment.
4. Bagaimana menampilkan visualisasi dari hasil data yang telah diperoleh.

## **1.3 Tujuan**

Adapun tujuan dari pengerjaan proyek adalah sebagai berikut.

1. Memperoleh data *trend* Twitter dengan *keyword* #racism
2. Mengolah data *real-time* menggunakan sistem yang mendukung komputasi dan pemrosesan secara *streaming*.
3. Mengklasifikasi dan menganalisis penilaian mengenai persepsi/opini terhadap aksi Rasisme berdasarkan sentimen publik yang ada di Twitter menggunakan metode *Vader sentiment analysis*.
4. Mendapatkan visualisasi dari data hasil yang telah diperoleh.

## **1.4 Manfaat**

Manfaat dari pengerjaan proyek ini adalah:

1. Bagi mahasiswa

Manfaat yang diharapkan akan didapatkan oleh mahasiswa melalui pengerjaan proyek ini adalah pemahaman dan keterampilan bagi seluruh mahasiswa tim proyek untuk mengimplementasikan teknik *machine learning* (*scalable machine learning*) pada *Big data* dengan teknologi pengolahan data besar menggunakan *Apache spark* untuk pemrosesan dan analisis. Selain itu, mahasiswa diharapkan mampu untuk menangani dan mengolah *data streaming* secara *real-time* dengan sistem yang mendukung.

1. Bagi masyarakat

Manfaat yang diharapkan akan didapatkan masyarakat melalui pengerjaan proyek ini adalah mendapatkan gambaran bagaimana pendapat atau opini masyarakat mengenai rasisme yang terjadi.

## **1.5 Ruang Lingkup**

Ruang lingkup dari pengerjaan proyek ini adalah sebagai berikut:

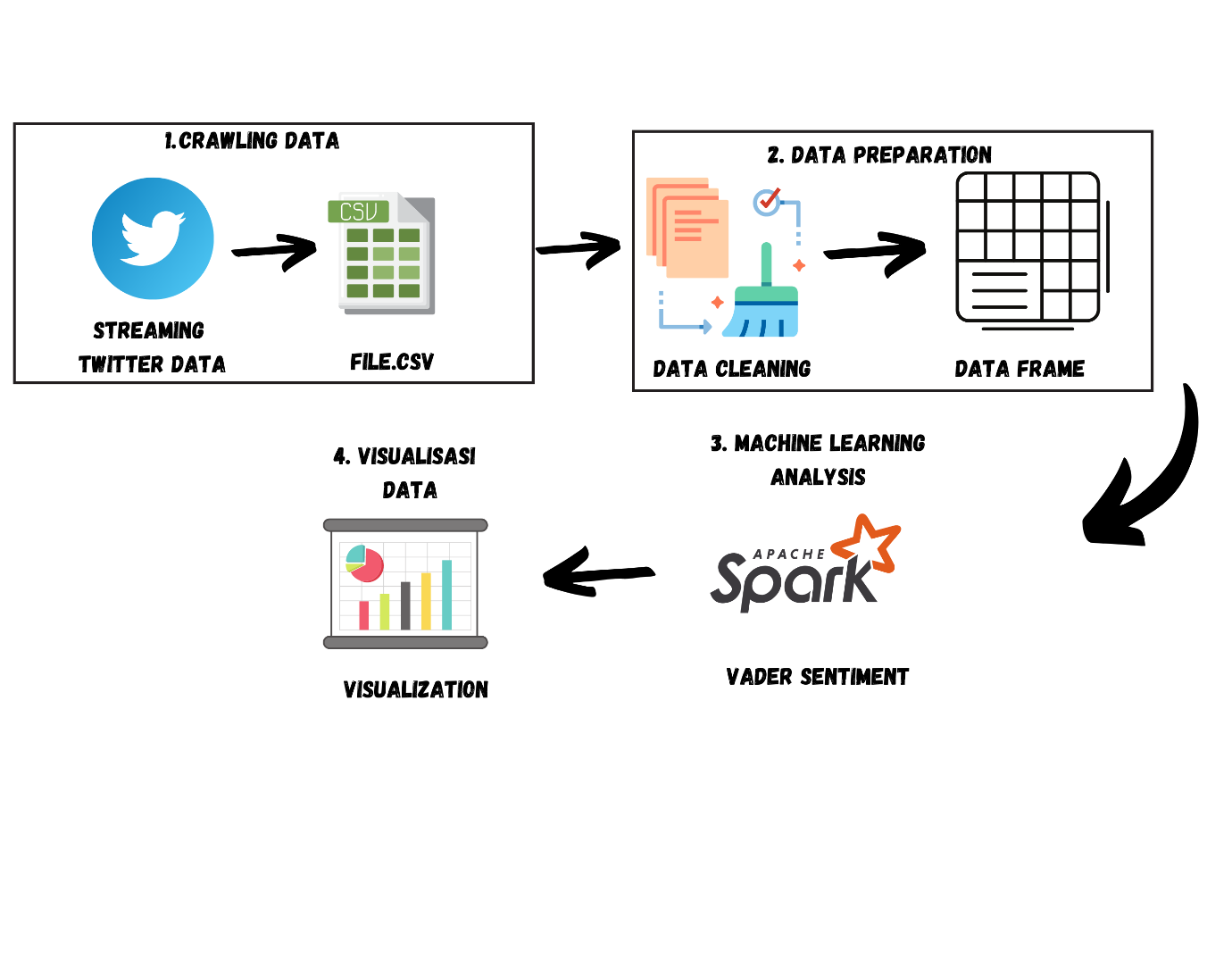
1. Dataset yang digunakan adalah data Twitter dengan topik rasisme dan menggunakan *keyword* #racism.
2. Sistem yang digunakan untuk mengenali dan mengekstraksi data adalah Vader *Sentiment analysis*, salah satu bidang dari *Natural Language Processing (NLP)*.
3. Pada proyek ini, mahasiswa mengintegrasikan dan mengimplementasikan teknologi yang digunakan dalam pengolahan data besar, menggunakan Spark.

# 

# **BAB 2 SOLUSI**

## **2.1 Arsitektur *Big Data***

Berikut ini merupakan arsitektur big data yang menjadi solusi terhadap trend topik Rasisme pada Twitter.



Gambar 1 Arsitektur Big Data

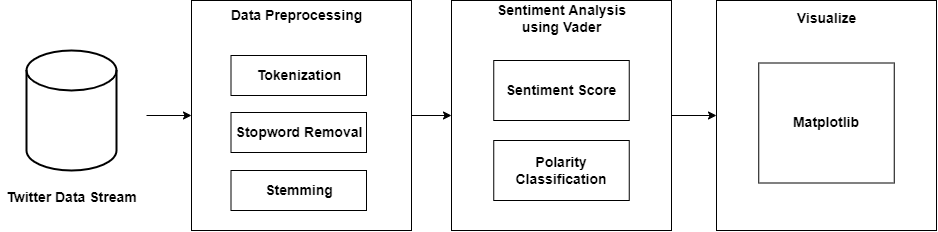
Pada pengerjaan proyek, data yang digunakan adalah data teks Twitter khususnya topik rasisme dengan *keyword* #racism. Data disiapkan terlebih dahulu dengan melakukan *crawling* data dari Twitter menggunakan API *key* Twitter. Proses pengambilan data Twitter dibantu dengan bahasa pemrograman Python dengan menggunakan library tweepy. API Key Twitter adalah Application Programming Interface (API), di dalamnya terdapat suatu layanan berisi sekumpulan perintah, fungsi, komponen dan juga protokol yang disediakan untuk mempermudah penggunanya pada saat membangun suatu sistem perangkat lunak. API Key Twitter memiliki consumer key, consumer secret, access key, dan access secret yang digunakan untuk mengakses data Twitter yang dibutuhkan. Setelah data berhasil di *crawling*, maka akan dihasilkan file dengan format .csv.

Setelah mendapatkan data, akan dilakukan *Data preparation* untuk mengubah *raw data* menjadi informasi yang lebih bersih untuk dapat digunakan pada pemrosesan selanjutnya yaitu analisis. Proses ini dilakukan untuk menangani kumpulan data yang masih memiliki kesalahan seperti data yang tidak konsisten, nilai yang hilang, format yang tidak sesuai dan sebagainya. Selanjutnya, sentimen yang diungkapkan dalam *tweet* akan diklasifikasikan dengan menerapkan *Vader Sentiment Analyzer* ke kumpulan data.

Vader adalah alat analisis sentimen berbasis aturan dan leksikon yang digunakan untuk mengekspresikan sentimen di media sosial. Pertama, kami menganalisis intensitas sentimen untuk mengkategorikan kumpulan data. Kemudian metode skor polaritas digunakan untuk menentukan sentimen. Vader *Sentiment Analyzer* digunakan untuk mengklasifikasikan tweet yang telah diproses sebelumnya menjadi positif, negatif, atau netral. Nilai gabungan adalah metrik yang berguna untuk mengukur sentimen dalam tweet tertentu. Dalam metode yang diusulkan, nilai ambang batas digunakan untuk mengkategorikan tweet sebagai positif, negatif, atau netral. Tahapan terakhir yang dilakukan yaitu membuat visualisasi data. Hasil yang diperoleh dari proses sentimen analisis yang telah dilakukan kemudian akan disimpan untuk selanjutnya divisualisasikan.

## **2.2 *Machine Learning Pipeline***

*Machine learning pipeline* merupakan salah satu jenis *pipeline* data yang terus menghasilkan dan menyiapkan data untuk pelatihan data.



Gambar 2 Machine Learning Pipeline

### **2.2.1 *Twitter Data Stream***

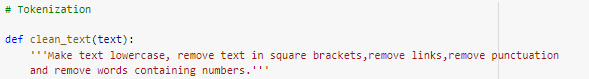
Twitter menyediakan layanan API yang dapat digunakan untuk diintegrasikan dengan aplikasi lain sehingga mempermudah dalam proses akuisisi data Twitter, seperti Search API, Streaming API, dan Twitter Firehose. Twitter streaming API digunakan untuk memperoleh data yang real time. API Twitter hanya bisa diakses melalui permintaan otentikasi. Twitter menggunakan Open Authentication (OAuth) dan setiap permintaan harus dilakukan oleh pengguna Twitter.

### **2.2.2 *Data Preprocessing***

*Data preprocessing* adalah proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan untuk proses *mining* selanjutnya seperti proses *sentiment analysis* dan lainnya. Berikut langkah-langkah yang dilakukan dalam data preprocessing.

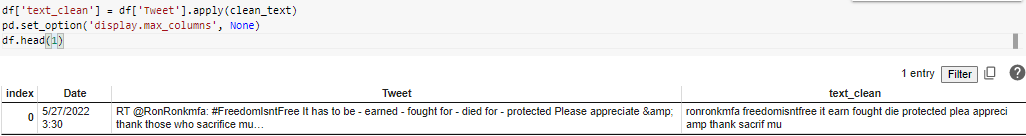
#### **2.2.2.1 *Tokenization***

Tokenisasi adalah proses penggantian elemen data sensitif yang sebenarnya dengan elemen data non-sensitif yang tidak memiliki nilai yang dapat dieksploitasi untuk tujuan keamanan data.



Kode Program 1 Tokenization

Pada data, tokenisasi dilakukan dengan mengubah kumpulan *term* dengan cara menghapus semua karakter tanda baca yang ada pada token. Sebelum melakukan pemisahan kata pada kalimat, terlebih dahulu dilakukan pembersihan tanda baca, tag html dan angka.

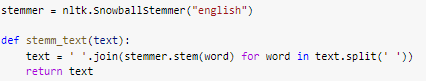


Kode Program 2 Contoh Hasil Tokenization

Pada tokenisasi akan dilakukan pemotongan string input berdasarkan tiap kata yang menyusunnya. Pada tahapan ini yang diperoleh hanya kumpulan kata-kata dari suatu teks/dokumen.

#### **2.2.2.2 *Stopword Removal***

Stopword Removal dilakukan untuk mengetahui suatu kata masuk ke dalam stopword atau tidak. Proses stopword adalah proses pembuangan kata yang tidak memiliki arti ataupun tidak relevan. Term yang diperoleh pada proses tokenisasi akan dicek pada stopword, apabila suatu kata termasuk dalam stopword maka kata tersebut tidak akan diproses. Namun jika suatu kata tidak termasuk dalam stopword maka akan masuk ke proses selanjutnya. Term yang diperoleh dari tahap tokenisasi dicek dalam suatu daftar *stopword*, apabila sebuah kata masuk di dalam daftar *stopword* maka kata tersebut tidak akan diproses lebih lanjut. Sebaliknya apabila sebuah kata tidak termasuk di dalam daftar stopword maka kata tersebut akan masuk ke proses berikutnya.



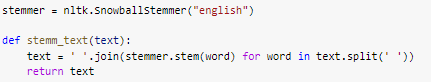
Kode Program 3 Stopword Removal



Kode Program 4 Hasil Stopword Removal

#### **2.2.2.3 *Stemming***

Stemming adalah proses untuk membentuk kata dasar. Term yang diperoleh pada proses *stopword* akan di *stemming*. *Stemming* digunakan untuk mereduksi bentuk term dalam menghindari ketidakcocokan yang dapat mengurangi recall, dimana term-term yang berbeda namun memiliki makna dasar yang sama akan direduksi menjadi satu bentuk.



Kode Program 5 Stemming



Kode Program 6 Hasil Stemming

### **2.2.3 *Sentiment Analysis* menggunakan Vader**

Vader digunakan sebagai model dalam melakukan analisis sentimen dan mampu menentukan keragaman data melalui intensitas kekuatan emosional yang ada sesuai dengan kamus data Lexicon. Vader menggabungkan analisis kualitatif dan validasi empiris menggunakan kebijaksanaan dan penilaian manusia. Vader mampu memberikan polaritas yang berbeda. Vader adalah contoh dari metode Lexicon. Pendekatan Lexicon bertujuan untuk memetakkan kata menjadi sentimen dengan membangun Lexicon atau ‘kamus sentimen’. Keuntungan dari penggunaan Vader adalah tersedia kamus yang berisi nilai dari setiap kata. Proses penentuan polaritas kalimat didapatkan dari penyatuan attribut ‘compound’ dari setiap kata yang tersedia. Kriteria pengelompokan positif, netral dan negatif. Hasil *compound* lebih dari 0,05 untuk kategori positif, hasil *compound* -0,05 dan 0,05 termasuk kategori netral dan untuk *compound* dibawah -0,05 termasuk kategori negatif.

### **2.2.4 Visualization**

*Data visualization* merupakan tampilan berupa grafis atau visual dari informasi dan data. *Visualization* juga berarti mengubah kumpulan data menjadi elemen visual agar lebih sederhana untuk ditampilkan. Salah satu *library* Python yang dapat digunakan untuk data *visualization* adalah matplotlib. Dengan menggunakan matplotlib, data dapat diubah menjadi plot, histogram, diagram batang atau jenis lainnya dengan sintaks yang telah diterapkan.

# 

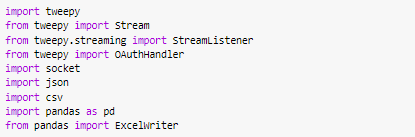
# **BAB 3 IMPLEMENTASI**

## **3.1 *Data Retrieval***

Data yang akan digunakan dalam penelitian ini adalah data tweet dari aplikasi twitter dengan keyword “*racism*”. Data yang dikumpulkan berupa teks yang diambil menggunakan teknik crawling. Tahap awal untuk melakukan proses pengambilan data dari Twitter, yaitu harus memiliki key number dan secret number dari API Key Twitter. Untuk mendapatkan API Key Twitter terlebih dahulu harus mendaftarkan atau melakukan pengajuan terhadap pihak developer atau pengembang Twitter melalui <https://developer.twitter.com/en>. Berikut merupakan *key and tokens* yang akan digunakan:

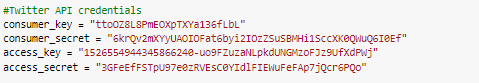
|  |
| --- |
| [CONSUMER KEYS]  Api Key:  ttoOZ8L8PmEOXpTXYa136fLbL  API Key Secret:  6krQv2mXYyUAOIOFat6byi2IOzZSuSBMHi1SccXK0QWuQ6I0Ef  [AUTHENTICATION TOKENS]  Bearer Token: AAAAAAAAAAAAAAAAAAAAADrHcgEAAAAAdqE%2FYrPHDFy5YsxamJzLUoPztnc%3D5GdoOIjruy0GFBAddabHS5fT5WNLrDZFuDcN803UDGJHqfEI5w  Access Token:  1526554944345866240-uo9FZuzaNLpkdUNGMzoFJz9UfXdPWj  Access Token Secret:  3GFeEfFSTpU97e0zRVEsC0YIdlFIEWuFeFAp7jQcr6PQo |

Setelah berhasil mendapatkan key dan token API Twitter, proses diawali dengan import beberapa library yang digunakan untuk terhubung dengan API Twitter.



Kode Program 7 Import Library

Selanjutnya, tambahkan Twitter API credentials yang sudah diperoleh sebelumnya melalui Twitter Developer. Adapun *Key* dan *token* yang digunakan adalah API Key, API Key Secret, Access Token dan Access Token Secret.



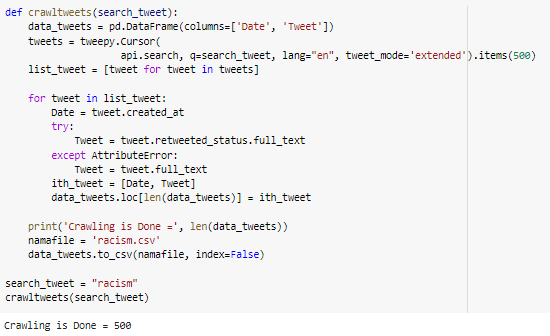
Kode Program 8 Twitter API

Setelah library yang digunakan berhasil di import, selanjutnya dilakukan autentikasi koneksi API yang digunakan.



Kode Program 9 Autentikasi API

Berikut adalah tahapan yang menunjukkan proses *crawling data*. Data akan disimpan ke dalam bentuk dataframe dengan nama kolom “date” dan “tweet”. Tweet yang akan diambil berjumlah 500 dengan bahasa inggris. Untuk keyword yang digunakan pada proyek ini adalah “racism”. Ketika proses crawling sudah selesai dilakukan, maka akan menampilkan “Crawling is Done = 500”.



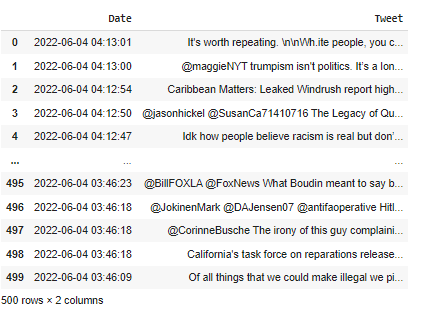
Kode Program 10 Crawling Data

Untuk memastikan data yang telah berhasil di crawling telah sesuai, maka dijalankan sintaks berikut untuk membaca data racism.csv



Kode Program 11 Read racism.csv

Berikut merupakan tampilan file racism.csv yang telah berhasil didapatkan melalui proses *crawling data* dari Twitter menggunakan Twitter API.

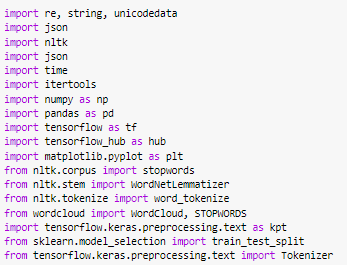


Kode Program 12 Tampilan Hasil Crawling

Dari tampilan dataset racism.csv yang telah didapatkan di atas, dapat dilihat bahwa dataset sudah memiliki identitas pada masing-masing kolom dan jumlah data sudah sesuai dengan jumlah yang telah dideklarasikan sebelumnya, yaitu 500.

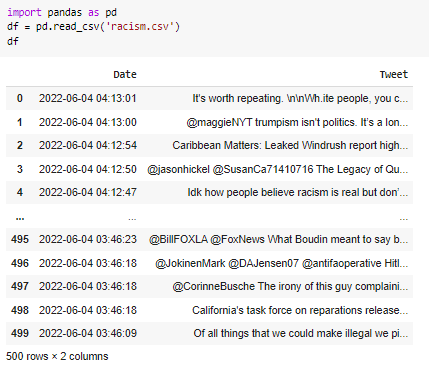
## **3.2 *Data Preparation***

Setelah menyelesaikan tahapan Data Retrieval, dimana dataset racism.csv telah didapatkan, maka akan dilakukan proses *data preparation.* Untuk memulai proses Data Preparation, dilakukan import library yang diperlukan terlebih dahulu.



Kode Program 13 Import Library Data Preparation

Setelah import library berhasil dilakukan, selanjutnya data racism.csv akan diread untuk memastikan bahwa data yang akan di cleaning sudah sesuai. Dapat dilihat bahwa data sudah sesuai sehingga cleaning data sudah dapat dilakukan.

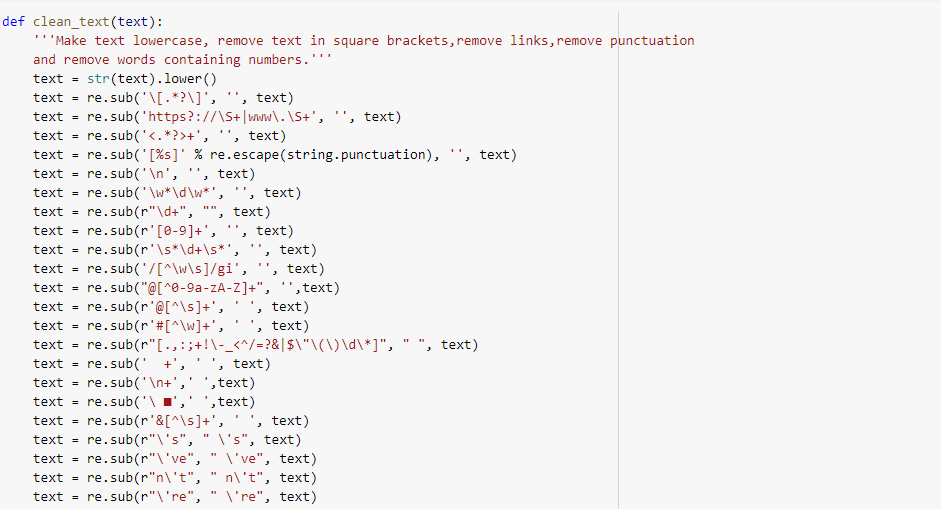


Kode Program 14 Cek Data racism.csv

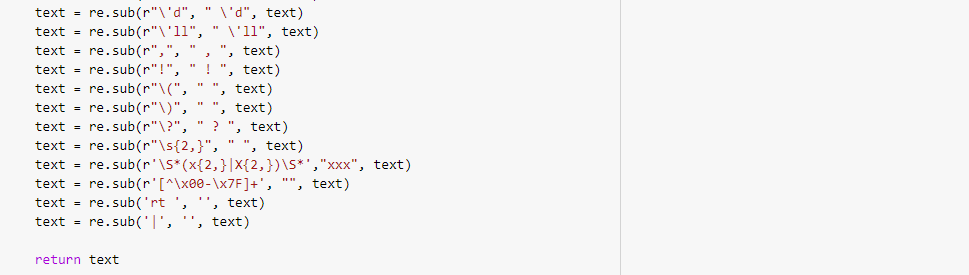
### **3.3.1 Tokenization**

Tokenisasi adalah proses penggantian elemen data sensitif yang sebenarnya dengan elemen data non-sensitif yang tidak memiliki nilai yang dapat dieksploitasi untuk tujuan keamanan data. Pada data, tokenisasi dilakukan dengan mengubah kumpulan *term* dengan cara menghapus semua karakter tanda baca yang ada pada token. Sebelum melakukan pemisahan kata pada kalimat, terlebih dahulu dilakukan pembersihan tanda baca, tag html dan angka.

Pada tokenisasi akan dilakukan pemotongan string input berdasarkan tiap kata yang menyusunnya. Pada tahapan ini yang diperoleh hanya kumpulan kata-kata dari suatu teks/dokumen.

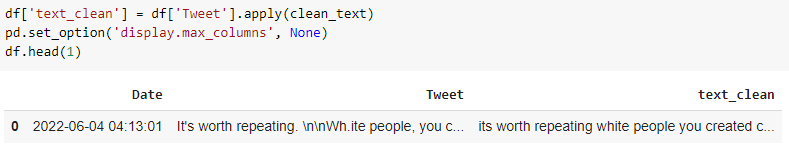


Kode Program 15 Clean Text - 1



Kode Program 16 Clean Text - 2

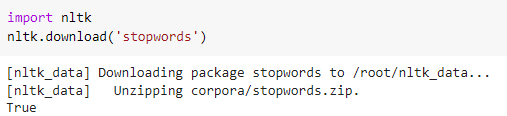
Kemudian jalankan sintaks berikut untuk menampilkan hasil data yang telah berhasil di tokenisasi. Kami mengatur ketentuan data yang ditampilkan hanya satu yaitu data pada kolom yang pertama. Untuk tweet yang sudah berhasil di tokenisasi akan ditampilkan ke dalam kolom baru yaitu “text\_clean”.



Kode Program 17 Lihat Hasil Clean Text

### **3.3.2 *Stopword Removal***

Pertama dilakukan import library nltk untuk membantu bekerja dengan pemodelan teks.



Kode Program 18 Import nltk

Stopwords adalah kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Pada pengerjaan proyek ini stopwords yang akan dihapus dalam bahasa inggris yakni ‘u’, ‘im’, ‘c’.

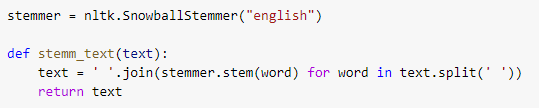
****

Kode Program 19 Remove Stopwords

Dapat dilihat pada kolom text\_clean, bahwa kata pada tweet yang memiliki stopword telah dihapus.

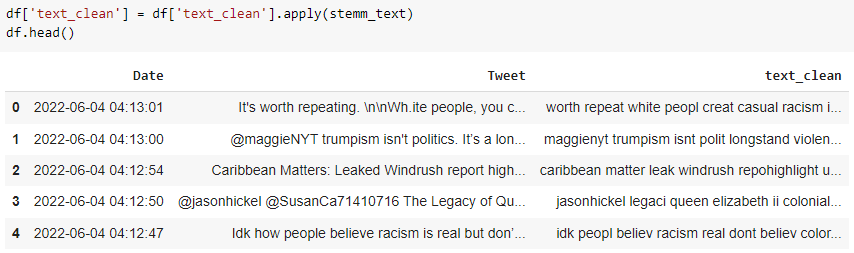
### **3.3.3 *Stemming***

Stemming merupakan proses untuk menghilangkan infleksi kata kedalam bentuk dasar, namun bentuk dasar tersebut tidak berarti sama dengan akar kata (*root word*).

****

Kode Program 20 Stemming

Hasil dari *steam text* ditampilkan menggunakan sintaks berikut.

****

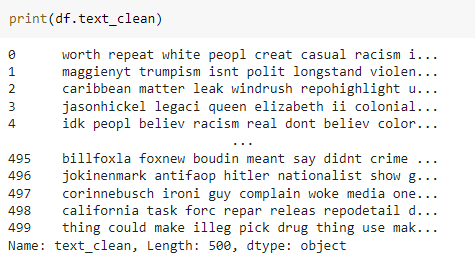
Kode Program 21 Menampilkan Hasil Steam Text

Selanjutnya dilakukan drop kolom date karena data waktu tweet tidak diperlukan.

****

Kode Program 22 Drop Kolom Date

Selanjutnya dilakukan print data untuk memastikan bahwa kolom “date” telah berhasil di drop.

****

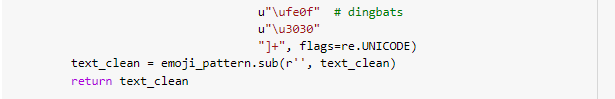
Kode Program 23 Menampilkan Hasil Clean Data

### **3.3.3 *Filtering***

Dikarenakan data yang sudah di cleaning masih mengandung emoji dan spasi, maka dilakukan *filtering* untuk menghapus emoji, spasi atau hal lainnya yang masih tertinggal, misalnya *link, username,* dan *hashtag*.

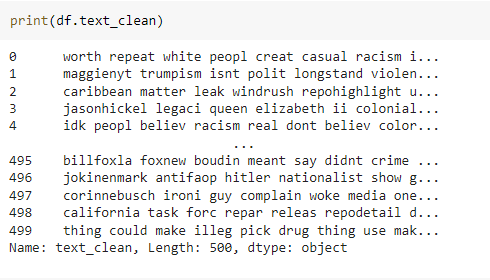
****

Kode Program 24 Filtering - 1

****

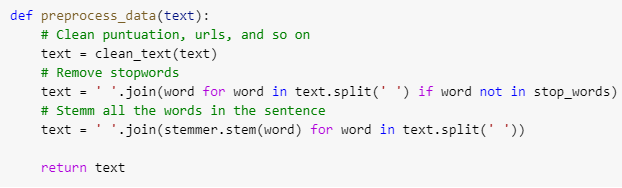
Kode Program 25 Filtering - 2

Selanjutnya dilakukan print data untuk memastikan bahwa proses *filtering* sudah berhasil.

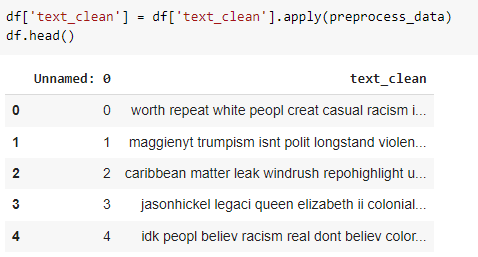
****

Kode Program 26 Menampilkan Hasil Filtering

Untuk memastikan data preprocessing sudah berhasil, kami melakukan preprocess data lagi dan menampilkan hasilnya dalam bentuk data frame.

****

Kode Program 27 Preprocess Data

****

Kode Program 28 Hasil Preprocess Data

Karena proses *cleaning data* sudah selesai, maka dataset akan diunduh dan disimpan dengan format csv. File disimpan dengan nama clean\_data.csv.

****

Kode Program 29 Import Hasil Cleaning Data

# **BAB 4 HASIL & PEMBAHASAN**

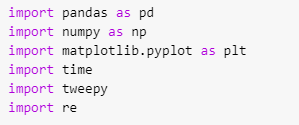
## **4.1 Sentimen Analisis Vader**

Proses sentimen analisis dilakukan setelah menyelesaikan tahapan *data preparation*. Implementasi akan dilakukan menggunakan Vader *sentiment analysis*. Oleh karena itu, perlu terlebih dahulu melakukan instalasi sentimen Vader dengan sintaks berikut ini.



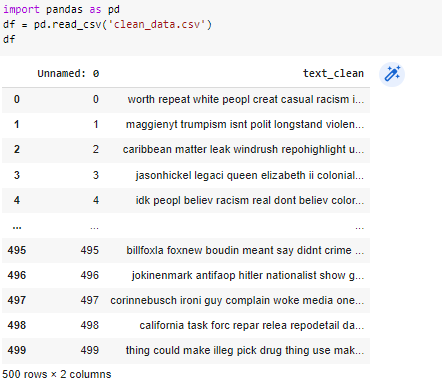
Kode Program 30 Install Vader Sentiment

Setelah instalasi berhasil dilakukan, kemudian dilakukan import library yang diperlukan untuk sentimen analisis dan visualisasi.



Kode Program 31 Import Library untuk Visualisasi

Kemudian dilakukan import file clean\_data.csv. Dan read file untuk memastikan data yang diimpor telah sesuai.



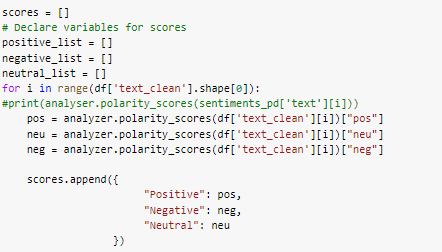
Kode Program 32 Read clean\_data

Setelah library yang dibutuhkan berhasil di import, dilakukan pemanggilan fungsi SentimentIntensityAnalyzer yang merupakan library dari vader lexicon.



Kode Program 33 Import SentimentIntensityAnalyzer

Kemudian dilakukan analisis menggunakan metode polarity\_score untuk mengkalkulasi *sentiment score* pada dataset clean\_data.csv.

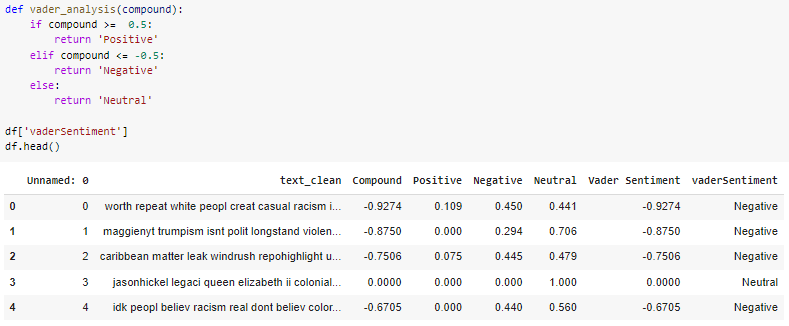


Kode Program 34 polarity\_score

Selanjutnya, menggunakan metrik compound score yaitu metrik yang telah dinormalisasi antara -1 sebagai skor negatif dan +1 sebagai skor positif. Penjelasan lebih detailnya dapat dilihat pada metodologi compound score metodologi berikut:

* Positive sentiment: compound score >= 0.05
* Neutral sentiment: (compound score > - 0.05) *and* (compound score < 0.05)
* Negative sentiment: compound score <= -0.05

Hasil *compound Vader sentiment* dapat dilihat pada kolom vaderSentiment pada Gambar di bawah.



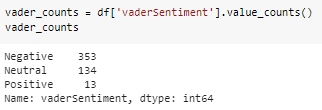
Kode Program 35 Compound Vader Sentiment

Dikarenakan kolom compound dan vader sentiment memiliki data yang sama, maka dilakukan drop kolom vader sentiment agar data yang diperoleh tidak duplikat.



Kode Program 36 Drop Vader Sentiment Column

Kemudian dilakukan penghitungan berdasarkan kolom “vaderSentiment” untuk mengetahui total data yang negatif, netral dan positif.



Kode Program 37 Hitung Hasil Vader Counts

## **4.2 Hasil Visualisasi**

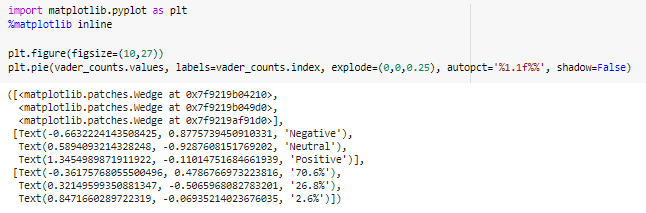
Setelah semua tahapan diselesaikan, maka dilakukan visualisasi untuk mengetahui hasil akurasi persentase dari hasil sentimen analisis yang dilakukan. Pada tahap ini akan menggunakan word cloud dan library matplotlib.

Word cloud merupakan bentuk representasi visual dari dari data teks untuk menggambarkan kumpulan keyword berdasarkan jumlah kata yang paling sering disebut. Semakin besar ukuran font, maka semakin sering kata tersebut muncul.



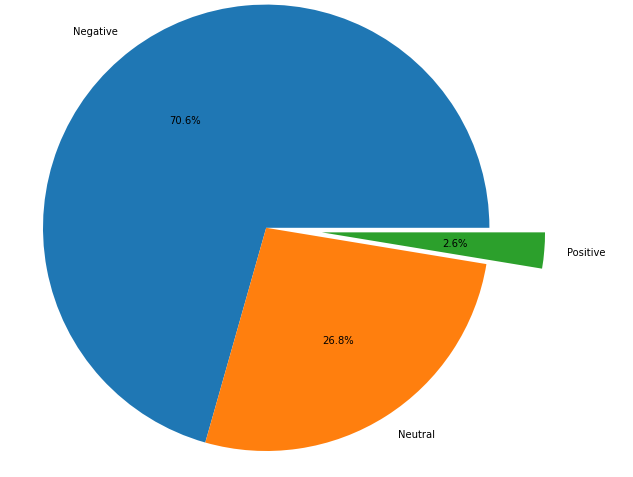
Kode Program 38 Word Cloud

Selain itu kami juga menggunakan library matplotlib. Library matplotlib merupakan library yang digunakan untuk melakukan visualisasi data. Berikut merupakan sintaks untuk menggunakan library matplotlib.



Kode Program 39 Matplotlib

Berikut merupakan hasil implementasi menggunakan library matplotlib. Visualisasi yang dihasilkan ditampilkan menggunakan diagram pie. Dapat dilihat diperoleh nilai persentase dari data yang negatif, positif dan netral. Untuk data negatif diperoleh dengan nilai 70,6%, data positif 2,6% dan netral senilai 26,8%.



Kode Program 40 Grafik

## **4.3 Kesimpulan**

Penggunaan *vader sentiment* dapat membantu untuk mengetahui klasifikasi keyword “racism” pada twitter dari kategori positif, negatif dan netral sesuai dengan skor sentimen yang telah dirumuskan. Berdasarkan hasil visualisasi dapat dilihat bahwa persentase data dengan kategori “negatif” memiliki persentase jumlah paling banyak. Maka dapat ditarik kesimpulan bahwa keyword racism bersifat negatif.

# **REFERENSI PAPER**

[1] S. Legianto, “Implementasi Text Mining Untuk Mendeteksi Hate Speech Pada Twitter,” 2019.

[2] T. Mustaqim, *SENTIMENT ANALYSIS OPINI PELANTIKAN KABINET PEMERINTAH INDONESIA TAHUN 2019 MENGGUNAKAN VADER DAN RANDOM FOREST*. 2020.