**LAPORAN PROYEK PENGOLAHAN DATA BESAR**

**Sentiment Analysis of Music Topic Review in Twitter Application With Text Blob**



**Disusun Oleh:**

|  |  |  |
| --- | --- | --- |
| 1 | 12S18003 | Citra Hutajulu |
| 2 | 12S18004 | Rosalia Pane |
| 3 | 12S18017 | Putri Yohana Panjaitan |
| 4 | 12S18030 | Unedo Ignasius Manalu |

**PROGRAM STUDI SARJANA SISTEM INFORMASI  
FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO  
INSTITUT TEKNOLOGI DEL  
MEI 2022**

# **DAFTAR ISI**

[**DAFTAR ISI** 2](#_Toc105183738)

[**DAFTAR GAMBAR** 3](#_Toc105183739)

[**DAFTAR TABEL** 4](#_Toc105183740)

[**BAB 1** 5](#_Toc105183741)

[**PENDAHULUAN** 5](#_Toc105183742)

[**1.1 Latar Belakang** 5](#_Toc105183743)

[**1.2 Rumusan Masalah** 6](#_Toc105183744)

[**1.3 Tujuan** 6](#_Toc105183745)

[**1.4 Manfaat** 7](#_Toc105183746)

[**1.5 Ruang Lingkup** 7](#_Toc105183747)

[**BAB 2** 8](#_Toc105183748)

[**SOLUSI** 8](#_Toc105183749)

[**2.1 Big Data Architecture** 8](#_Toc105183750)

[**2.2 Machine Learning Pipeline** 9](#_Toc105183751)

[**BAB 3** 11](#_Toc105183752)

[**IMPLEMENTASI** 11](#_Toc105183753)

[**3.1. Stream Data From Twitter** 11](#_Toc105183754)

[**3.2. Sentiment Analysis with Textblob** 13](#_Toc105183755)

[**3.3. Visualisasi Data** 16](#_Toc105183756)

[**BAB IV** 19](#_Toc105183757)

[**HASIL DAN PEMBAHASAN** 19](#_Toc105183758)

[**4.1 Hasil Prediksi** 19](#_Toc105183759)

[**4.2 Hasil Visualisasi** 19](#_Toc105183760)

# **DAFTAR GAMBAR**

[**Gambar 1 Big Data Architecture** 8](#_Toc105185256)

[**Gambar 2 Machine Learning Pipeline** 9](#_Toc105185257)

[**Gambar 3 menginisiasi api dan access token** 11](#_Toc105185258)

[**Gambar 4 Melakukan autentikasi** 11](#_Toc105185259)

[**Gambar 5 Import Library** 12](#_Toc105185260)

[**Gambar 6 Mengambil data dari twitter** 12](#_Toc105185261)

[**Gambar 7 Mengambil data twitter dengan topik trending music** 12](#_Toc105185262)

[**Gambar 8 Membuat Dataframe** 13](#_Toc105185263)

[**Gambar 9 Import library dan pembacaan data\_tweet\_music.csv** 14](#_Toc105185264)

[**Gambar 10 Melakukan remove stopwords** 14](#_Toc105185265)

[**Gambar 11 Membuat spark session** 14](#_Toc105185266)

[**Gambar 12 Melakukan Data Cleaning** 15](#_Toc105185267)

[**Gambar 13 Melakukan text classification dengan fungsi Textblob** 15](#_Toc105185268)

[**Gambar 14 Menjalankan semua fungsi** 15](#_Toc105185269)

[**Gambar 15 Menampilkan hasil eksekusi** 16](#_Toc105185270)

[**Gambar 16 Membuat label positif, negative dan neutral** 16](#_Toc105185271)

[**Gambar 17 Membuat kolom baru dengan nama type polarity** 17](#_Toc105185272)

[**Gambar 18 Memvisualisasikan data dengan pie chart** 17](#_Toc105185273)

[**Gambar 19 Membuat visualisasi dengan label Positive, Negative dan Neutral** 17](#_Toc105185274)

[**Gambar 20 Memvisualisasikan data subjectivity dengan line graph** 18](#_Toc105185275)

[**Gambar 21 Contoh polarity yang tinggi** 19](#_Toc105185276)

[**Gambar 22 Contoh polarity yang rendah** 19](#_Toc105185277)

[**Gambar 23 Tampilan wordcloud pada data twitter topik trending music** 20](#_Toc105185278)

[**Gambar 24 Tampilan visualisasi hasil sentyment analysis** 21](#_Toc105185279)

# **DAFTAR TABEL**

[**Tabel 1 Jumlah data tweet sesuai label prediksi** 19](#_Toc105183661)

# **BAB 1**

# **PENDAHULUAN**

## **1.1 Latar Belakang**

*Twitter* merupakan salah satu sosial media yang sangat populer di masyarakat. Pada *twitter*, setiap pengguna berinteraksi dengan pengguna lainnya melalui sebuah pesan yang biasa disebut *tweet. Twitter* dan masyarakat mengakui bahwa pengaruh yang berkembang pada *twitter* sangatlah besar. Bentuk interaksi pada twitter adalah singkat dan jelas. Hal itu memungkinkan penggunanya lebih mudah untuk mendapatkan dan mengetahui informasi [1]. Twitter juga digunakan untuk mengungkapkan emosi mengenai sesuatu hal, baik untuk memuji ataupun mencela dalam bentuk emosi.

Meningkatnya jumlah interaksi sosial publik terjadi melalui komputer, dimana semuanya dibagikan mulai dari lelucon, cerita, berita dan terutama dunia industri musik. Narasi media mengenai industri musik banyak menuai komentar negatif dan positif. Dalam perkembangan teknologi saat ini adalah sebuah hal yang luar biasa terhadap musisi untuk melakukan promosi lagu, berkomunikasi dengan penggemar secara langsung atau tidak langsung. Media sosial telah mengubah cara penggemar berinteraksi dengan selebritas favorit mereka dan juga diantara mereka satu sama lain. Salah satu media sosial yang menyediakan komunikasi seperti adalah *twitter* [2].

*Twitter* juga menyediakan api *twitter* yang bertujuan agar para pengembang dapat dengan mudah untuk melakukan analisis terhadap data pada *twitter*. Analisis yang pada umumnya digunakan adalah sentimen analisis. Sentimen Analisis *(opinion mining*) merupakan suatu implementasi yang diperoleh ketika penambangan teks ( *text mining* ) digunakan untuk mengklasifikasikan informasi berupa teks berdasarkan sentimen/pandangan subjektif pada teks tersebut. Analisis sentimen dilakukan secara otomatis menggunakan algoritma pembelajaran mesin (*machine learning*). Algoritma ini akan mendeteksi kecenderungan opini yang terkandung dalam suatu teks dan akan diklasifikasikan menjadi kelas positif, negatif, ataupun netral.

Tujuan utama dilakukan sentimen analisis adalah untuk mengklasifikasikan polaritas informasi dalam dokumen baik kalimat maupun kata-kata. Analisis sentimen bekerja secara sistematis untuk mengidentifikasi, mengekstrak, dan mempelajari kondisi dan informasi subjektif yang secara luas diterapkan pada analisis pendapat konsumen, ulasan, tanggapan survey, dan media sosial. Data yang digunakan pada sentimen analisis adalah data yang di streaming lewat *twitter* menggunakan *apache spark*. *Apache spark* berguna untuk membangun jalur pemrosesan data dengan mudah dan dapat mengakses data dari berbagai sumber. *Apache spark* dapat juga menyediakan data yang kemudian akan dilakukan text mining dan klasifikasi terhadap *tweet* atas data pada *twitter* sehingga mendapat informasi baik buruknya *tweet* tersebut.

Musik merupakan salah satu topik yang *trending* pada *twitter,* banyak orang yang memberikan komentar atau *tweet* terhadap topik tersebut. Ada beberapa jenis *tweet* yang dilontarkan atau diberikan terhadap topik musik. Mulai dari komentar yang positif, negatif, subjektif dan objektif. Oleh karena itu perlu dilakukan analisis sentimen terhadap komentar atau *tweet* pada topik musik, supaya mendapatkan informasi terkait berapa banyak *tweet* yang positif, negatif, subjektif, dan objektif [3].

*Textblob* merupakan salah satu metode NLP (*Natural Language Processing*) untuk mengklasifikasikan kalimat berdasarkan polarity dan subjektif. Tipe data polarity dan subjektif itu adalah *float* dimana *range* dari data polarity adalah (-1) - 1 dan range data subjektif adalah 0-1.

Oleh karena itu akan dilakukan analisis sentimen menggunakan *textblob* terhadap data *twitter* yakni topik musik untuk mendapatkan polarity dan subjektif terhadap data *twitter* tersebut.

## **1.2 Rumusan Masalah**

Adapun rumusan masalah yang dilakukan pada proyek ini adalah:

1. Bagaimana cara memperoleh data pada *streaming* data pada *twitter*
2. Bagaimana cara untuk mengklasifikasikan informasi *tweet* musik pada *twitter* untuk mengetahui baik dan buruknya sebuah *tweet*

## **1.3 Tujuan**

Adapun tujuan dilakukannya proyek pengolahan data besar ini yaitu:

1. Menerapkan *sentiment analysis* pada *streaming* data *twitter* untuk *tweet* yang berisi topik musik.
2. Mengklasifikasikan informasi mengenai *tweet* musik pada *twitter* dengan menggunakan *Text Blob* untuk mengetahui baik buruknya sebuah *tweet*.

## **1.4 Manfaat**

Manfaat dari pengerjaan proyek adalah sebagai berikut:

1. Mengetahui cara dan proses analisis menggunakan *sentiment analysis* pada *streaming* data *twitter.*
2. Mengetahui seberapa banyak *tweet* positif dan negatif dari *tweet* musik *twitter* berdasarkan klasifikasi menggunakan *Textblob.*

## **1.5 Ruang Lingkup**

Ruang lingkup dalam pengerjaan proyek ini yaitu:

1. Algoritma yang digunakan dalam pengerjaan proyek ini adalah *textblob* untuk mengklasifikasikan data.
2. Data yang digunakan adalah *tweet* positif dan negatif pada topik music yang ada pada *twitter* dengan melakukan API streaming.

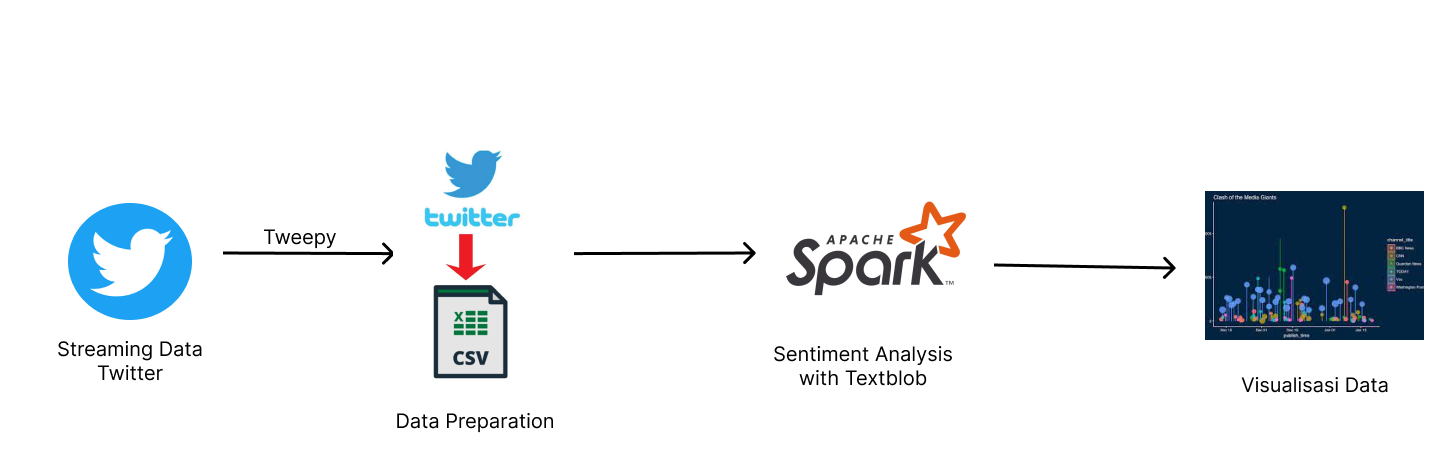
# 

# **BAB 2**

# **SOLUSI**

## **2.1 Big Data Architecture**

Dalam melakukan analisis sentimen terhadap masalah pada topik music pada twitter, dibutuhkan sebuah solusi untuk mendapatkan analisis sentimen. Berikut big data architecture design sebagai solusi pada analisis ini:



**Gambar 1 Big Data Architecture**

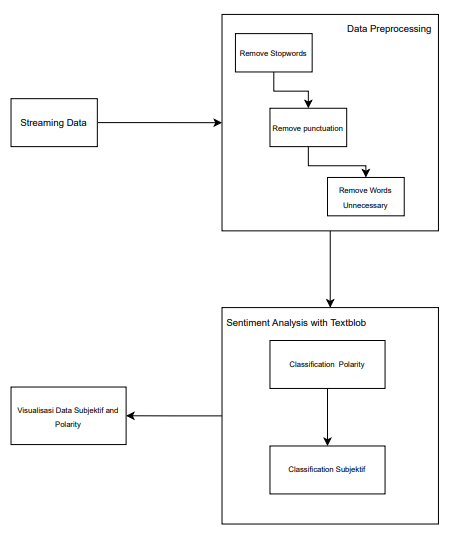
Data yang akan diproses menggunakan apache spark dan machine learning yakni text blob berasal dari api key. Untuk mendapatkan data tersebut, data akan diambil dengan menggunakan library python yakni tweepy. Pada library tweepy, data akan di filter menggunakan fungsi tweepy.filter dengan keywords yang ditentukan. Kemudian setiap data yang di filter, data tweet dan data user yang melakukan tweet akan disimpan pada file csv. File csv tersebut yang akan diproses pada spark streaming.

Pada apache streaming akan dilakukan pembuatan session untuk melakukan proses analisis terhadap file csv tersebut. File csv tersebut akan dibaca dengan function spark.read\_file\_csv(). Kemudian akan dilakukan data cleaning untuk menghilangkan hal hal yang tidak diperlukan pada data tersebut. Setelah dilakukan proses data cleaning, data tersebut akan di analisis menggunakan textblob untuk mendapatkan data mengenai angka polarity dan subjective. Polarity berarti positive dan negative kalimat pada data tersebut sedangkan subjective adalah apakah kalimat tersebut bersifat subjective atau objective. Tipe data pada polarity dan subjective adalah float.

Setelah didapat polarity dan subjective, akan dilakukan visualisasi data menggunakan grafik. Dari grafik tersebut didapat informasi mengenai seberapa banyak kalimat yang positif, seberapa banyak kalimat yang negatif, seberapa banyak kalimat yang merupakan kalimat subjektif dan seberapa banyak kalimat yang merupakan kalimat objektif. Setelah itu, dilakukan end session pada spark session yang telah dibuat di awal.

## **2.2 Machine Learning Pipeline**

Berikut ini merupakan *machine learning pipeline* yang menjadi solusi terhadap permasalahan untuk memprediksi apakah tweet untuk topik musik tersebut memiliki makna positif, negatif, atau netral.

****

**Gambar 2 Machine Learning Pipeline**

Pada arsitektur, data akan dilakukan streaming pada spark dengan cara mengload data yang sudah di stream lewat twitter. Dalam stream tersebut dilakuakn data preprocessing. Pada data preprocessing dilakukan penghilangan kata kata yang termasuk stop words. Kemudian menghapus tanda-baca pada kalimat dan menghapus kata kata yang tidak perlu dalam melakukan sentiment analysis. Setelah dilakukan data preprocessing, data tersebut akan diklasifikasi dengan polarity dan subjektif dengan menggunakan textblob. Pada klasifikasi polarity akan dilakukan klasifikasi dengan label positive, neutral dan negative. Pada klasifikasi subjektif akan dilakukan pengklasifikasi apakah kalimat tweet tersebut objektif atau subjektif. Setelah dilakukan klasifikasi, data hasil klasifikasi tersebut akan divisualisaikan dengan menggunakan grafik.

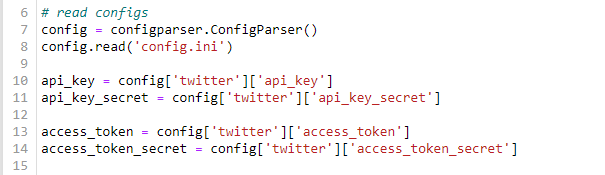
# **BAB 3**

# **IMPLEMENTASI**

## **3.1. Stream Data From Twitter**

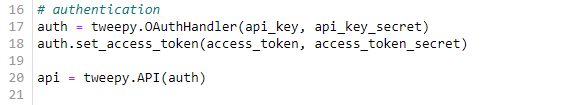
Data yang diperlukan pada penelitian ini merupakan data yang berasal dari *twitter*. Aplikasi *Twitter* menyediakan api untuk mendapatkan data yang dibutuhkan oleh peneliti. Untuk mendapatkan api tersebut, peneliti harus melakukan pendaftaran ke dalam portal *developer twitter*. Setelah mendaftarkan ke dalam portal *developer twitter*, peneliti akan me*request* api token dengan mengisi beberapa form yang disediakan pada portal tersebut. Setelah mengisi form tersebut, peneliti akan menunggu *approve* dari pihak *twitter* agar mendapatkan api token untuk akses data *twitter*.

Setelah mendapat token akses *twitter* tersebut, akan dibuat program untuk mendapatkan data *twitter* secara *streaming*. Hal pertama yang dilakukan adalah menginisiasi api dan *access token* untuk mengauthentikasi. Berikut potongan codenya:



**Gambar 3 menginisiasi api dan access token**

Setelah melakukan inisiasi pada api dan *access token* akan dilakukan autentikasi, berikut potongan codenya:



**Gambar 4 Melakukan autentikasi**

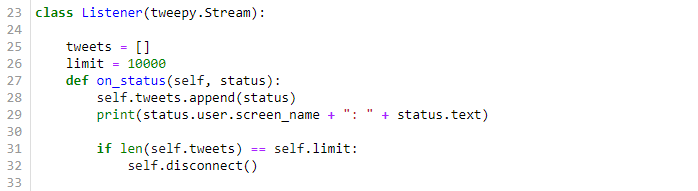
Kemudian lakukan *import library* yang dibutuhkan untuk melakukan *streaming* data pada *twitter*.



**Gambar 5 Import Library**

Library *tweepy* digunakan untuk melakukan *streaming* data pada *twitter*. Library *configparser* digunakan untuk memudahkan melakukan *configuration* api token dan *access token twitter*. Library pandas digunakan untuk melakukan penyimpanan data ke csv.

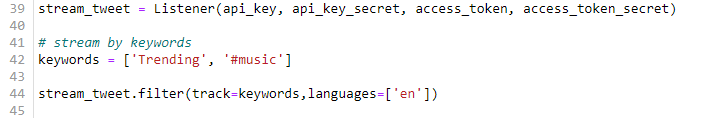
Setelah melakukan import library, akan dilakukan pengambilan data dari *twitter*. Berikut potongan codenya:



**Gambar 6 Mengambil data dari twitter**

*Class listener* digunakan untuk mengambil data *user twitter* dan *text tweet* pada *twitter* dengan batasan data adalah 10.000, kemudian data tersebut disimpan ke dalam variabel *tweets*. Ketika proses sudah mencapai 10.000 maka proses akan dihentikan.

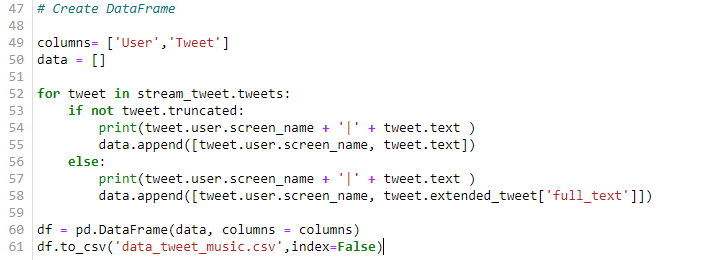
Pada proyek ini tidak semua data *twitter* diambil sebagai data masukan untuk *sentiment analysis*, melainkan data *twitter* yang termasuk ke dalam topik *trending music*. Berikut potongan code:



**Gambar 7 Mengambil data twitter dengan topik trending music**

Variabel *stream\_tweet* akan dilakukan pemanggilan *class listener* dengan parameter api token dan *acces\_token*.

Setelah melakukan pemanggilan class listener, selanjutnya adalah membuat dataframe. Berikut potongan codenya:



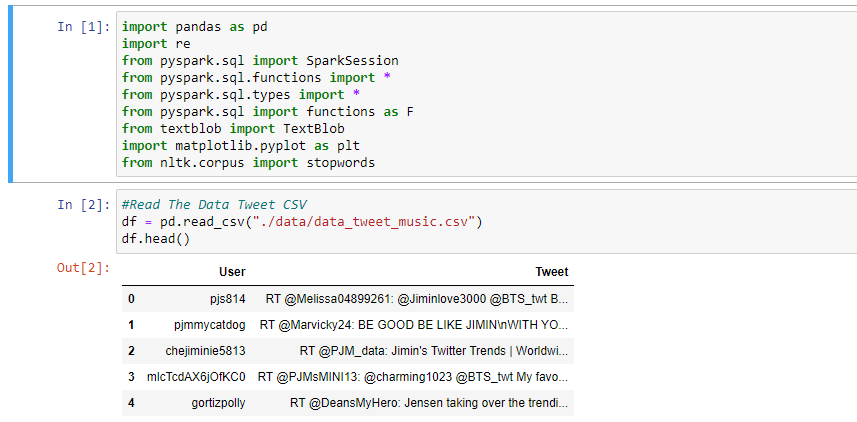
**Gambar 8 Membuat Dataframe**

Dalam membuat dataframe yang pertama dilakukan adalah membuat dataframe dengan nama kolom user dan tweet, kemudian disimpan disimpan ke file dengan name data\_tweet\_music.csv.

## **3.2. Sentiment Analysis with Textblob**

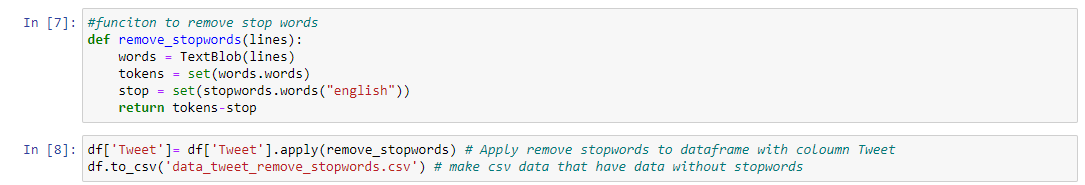
*Sentiment analysis* digunakan untuk membuat klasifikasi informasi terkait identifikasi terhadap informasi subjektif dan ulasan pada media sosial. Dalam melakukan analisis sentimen menggunakan *textblob*, dimana *textblob* digunakan untuk mendapatkan data mengenai angka polarity dan subjective.

Dalam melakukan analisis sentimen, yang dilakukan pertama adalah *import library*, kemudian dilakukan pembacaan data pada data\_tweet\_music.csv. Berikut potongan codenya:



**Gambar 9 Import library dan pembacaan data\_tweet\_music.csv**

Setelah melakukan pembacaan data, kemudian melakukan preprocessing data. Pada tahap preprocessing akan dilakukan *remove stopwords.* Berikut potongan codenya:



**Gambar 10 Melakukan remove stopwords**

*Remove stopword* dilakukan untuk menghapus kata-kata yang termasuk dalam jenis kata sambung, imbuhan dan sebagainya, sehingga kata-kata yang akan digunakan adalah kata dasar. Data yang telah melalui tahap remove stopwords akan disimpan ke dalam file csv dengan nama file data\_tweet\_remove\_stopwords.csv. Data csv tersebut akan diproses pada spark streaming.

Untuk melakukan anilis sentiment pada spark streaming, perlu dilakukan pembuatan spark session. Berikut potongan codenya:



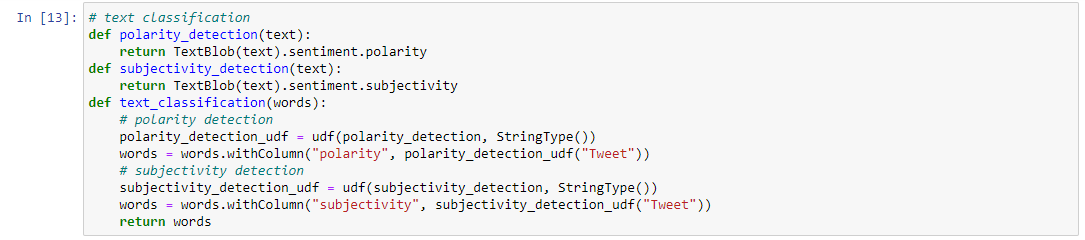
**Gambar 11 Membuat spark session**

Spark session dibuat dengan nama ‘sentiment analysis’. Setelah spark session dibuat, maka akan dibuat satu variable untuk membaca data yang telah bersih dari stopwords. Setelah berhasil dibaca, maka data tersebut akan masuk ke dalam proses data preprocessing. Pada data preprocessing, akan dilakukan data cleaning pada column Tweet yang ada pada variable. Berikut potongan code untuk data preprocessing:



**Gambar 12 Melakukan Data Cleaning**

Kemudian akan dilakukan text classification dengan menggunakan fungsi pada *Textblob* yakni sentiment polarity dan sentiment subjecity. Berikut potongan code pada setiap fungsi teks classification:



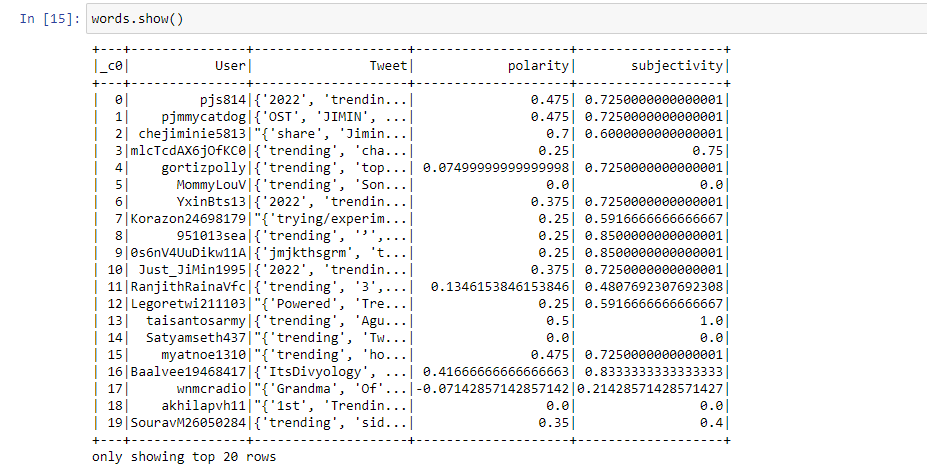
**Gambar 13 Melakukan text classification dengan fungsi Textblob**

Kemudian semua fungsi tersebut dilakukan seperti potongan code di bawah ini:



**Gambar 14 Menjalankan semua fungsi**

Setelah code tersebut dieksekusi, berikut hasil keluarannya:

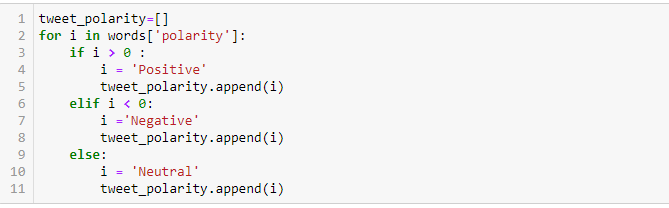


**Gambar 15 Menampilkan hasil eksekusi**

Pada gambar di atas, terdapat penambahan dua kolom yakni polarity dan subjetivity. Pada polarity, terdapat data float dimana ketika hasilnya > 0 maka akan positif, =0 maka netral, dan < 0 maka akan negatif. Pada kolom subjectivity, terdapat data seberapa subjectif kata tersebut. Semakin menuju ke angka 1, maka semakin subjektif kata tersebut.

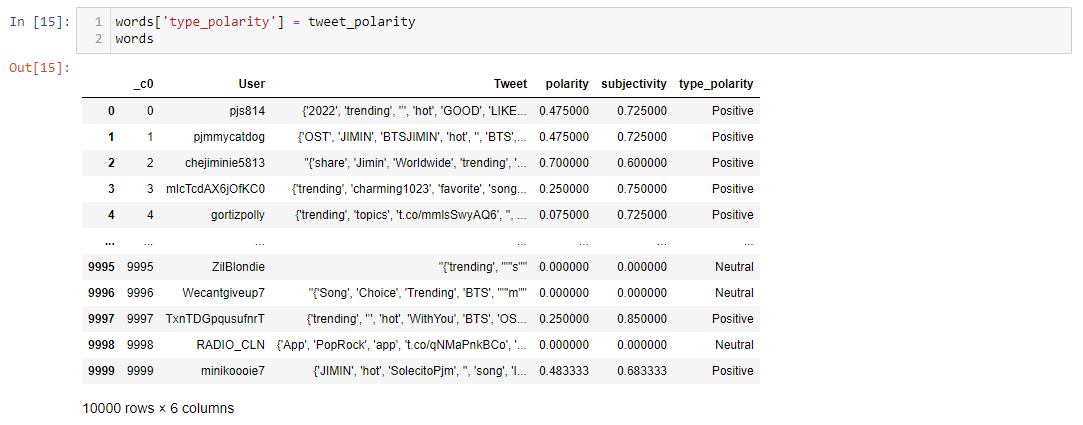
## **3.3. Visualisasi Data**

Setelah dilakukan sentiment analisis, akan dilakukan visualisasi data pada data polarity dan subjectivity. Pada polarity, setiap data pada kolom polarity akan dibuat label positif, negative dan neutral. Untuk data dengan label positif merupakan nilai data yang lebih besar dari 0. Untuk data yang netral merupakan data yang sama dengan 0 dan untuk data dengan label negative merupakan nilai lebih kecil dari 0. Berikut potongan codenya:



**Gambar 16 Membuat label positif, negative dan neutral**

Setelah code tersebut di eksekusi, data yang ada pada tweet\_polarity akan diinput ke dalam dataframe dan membuat kolom baru dengan nama type polarity seperti potongan code di bawah ini:



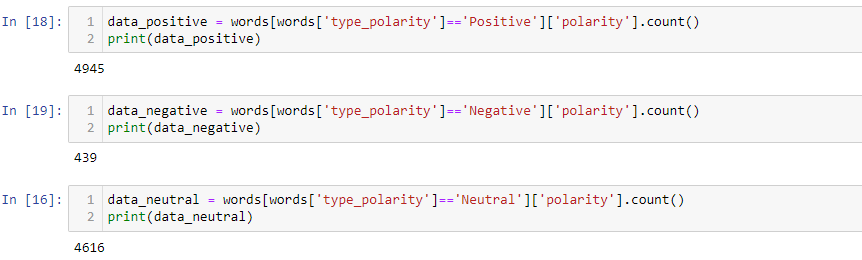
**Gambar 17 Membuat kolom baru dengan nama type polarity**

Kemudian data polarity dengan label akan divisualisasi dengan menggunakan pie chart. Berikut potongan codenya:



**Gambar 18 Memvisualisasikan data dengan pie chart**

Pada code di atas, akan dilakukan pembuatan pie chart dengan label Positive, Negative dan Neutral. Untuk mencari jumlah dari setiap label, digunakan fungsi count() terhadap jumlah label pada kolom type\_polarity. Berikut potongan codenya:



**Gambar 19 Membuat visualisasi dengan label Positive, Negative dan Neutral**

Untuk data subjectivity, akan dilakukan visualisasi dengan line graph. Berikut potongan codenya:



**Gambar 20 Memvisualisasikan data subjectivity dengan line graph**

# **BAB IV**

# **HASIL DAN PEMBAHASAN**

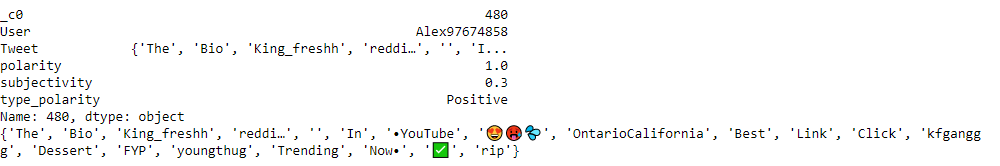
## **4.1 Hasil Prediksi**

Pada sentiment analisis menggunakan textblob, akan dilakukan prediksi terhadap data tweet pada twitter mengenai polarity dan subjectivity. Pada perlakuan prediksi polarity, terdapat tiga label yakni, positive, negative, dan neutral. Berikut jumlah data tweet sesuai label prediksi:

|  |  |  |
| --- | --- | --- |
| No | Label | Jumlah Data |
| 1 | Positive | 4945 |
| 2 | Neutral | 4616 |
| 3 | Negative | 439 |

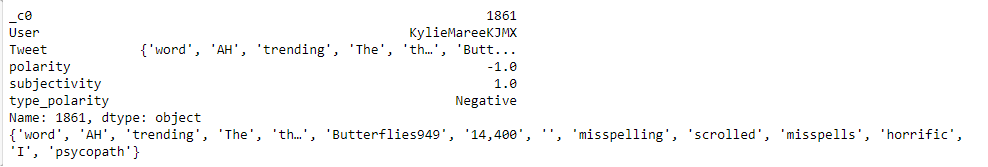
**Tabel 1 Jumlah data tweet sesuai label prediksi**

Dari table di atas, dapat disimpulkan bahwa pada data tweet dengan keyword trending music terdapat lebih banyak tweet yang bersifat positive dan bersifat neutral dan lebih sedikit label negative. Untuk nilai polarity yang paling tinggi adalah 1.0 dan yang paling rendah adalah -1.0. Berikut contoh data dengan polarity yang tinggi:



**Gambar 21 Contoh polarity yang tinggi**

Berikut contoh data dengan polarity yang rendah:



**Gambar 22 Contoh polarity yang rendah**

Untuk rata-rata nilai polarity adalah “0.14650838284169534”. Hal ini menunjukkan bahwa tweet pada topik trending music adalah positive.

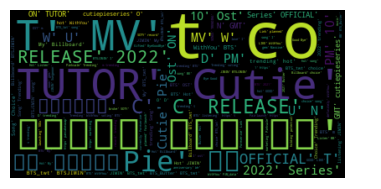
Pada klasifikasi subjectivity, rata-rata nilai dari subjectivity adalah “0.2647827061910537”. Hal ini menunjuukan bahwa setiap tweet pada topik trending music tidak subjectivity karena memiliki nilai rata-rata yang mendekati 0.

## **4.2 Hasil Visualisasi**

Setelah data tweet tersebut sudah dilakukan sentiment analysis, hasil dari sentiment analysis tersebut akan divisualisasikan. Ada beberapa visualisasi yang dilakukan, yakni:

* WordCloud

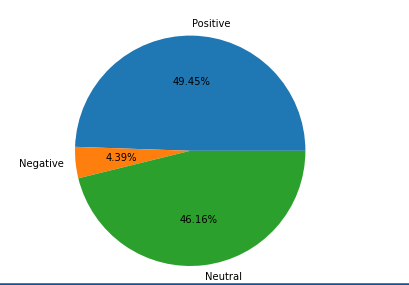
Visualisasi WordCloud digunakan untuk menampilkan kata kata apa yang ada pada data tweet yang telah di lakuka preprocessing. Berikut gambar wordcloud pada data twitter pada topik trending music:



**Gambar 23 Tampilan wordcloud pada data twitter topik trending music**

* Pie Chart

Visualisai pie chart digunakan untuk menvisualisasikan data polarity hasil sentiment analysis. Berikut gambar pie chart yang telah dilakukan:



**Gambar 24 Tampilan visualisasi hasil sentyment analysis**

Pada chart di atas, informasi yang didapar adalah berapa persen jumlah data pada setiap label. Pada chart tersebut dapat terlihat bahwa data tweet yang berlabel positive lebih banyak dibandingkan denga label yang lain. Maka dari itu dapat disimpulkan bahwa data yang digunakan bersifat positive.

# **References**

|  |  |
| --- | --- |
| [1] | E. W. Arief Wibowo, "Paper Review: Data Mining Twitter," *Paper Review: Data Mining Twitter,* pp. 2-10, 2014. |
| [2] | K. Wallach, "A Content Analysis of Twitter Use: Factors That Might Increase Music Sales During an Award Show?," pp. 1-3, 2014. |
| [3] | A. B. J. B. M. M. a. C. Katrin Weller, "Twitter and Society," *Twitter and Society,* pp. 1-4, 2014. |