****

**Sentimen Analisis pada Streaming Data Twitter untuk Mengetahui Reaksi Publik terhadap Lagu Cupid**

**Proyek Penambangan Data Besar**

**Dosen Pengampu**

**Humasak Tommy Argo Simanjuntak, ST, M.ISD**

**PENGUSUL**

**KELOMPOK 1**

Oleh:

**12S19013 AGNES SAGITA LUMBANTOBING**

**12S19021 IRFAN JUMADIN SIREGAR**

**12S19034 SAMUEL CHRISTIAN SITORUS**

**12S19035 NIXON DANIEL HUTAHAEAN**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**LAGUBOTI**

**2023**

**DAFTAR ISI**

[DAFTAR GAMBAR 3](#_Toc135600554)

[DAFTAR POTONGAN KODE PROGRAM 4](#_Toc135600555)

[BAB 1 PENDAHULUAN 5](#_Toc135600556)

[1.1 Latar Belakang 5](#_Toc135600557)

[1.2 Rumusan Masalah 6](#_Toc135600558)

[1.3 Tujuan dan Manfaat 6](#_Toc135600559)

[1.4 Ruang Lingkup 6](#_Toc135600560)

[BAB 2 ARSITEKTUR SISTEM 7](#_Toc135600561)

[2.1 Arsitektur Sistem 7](#_Toc135600562)

[2.2 Machine Learning Pipeline 8](#_Toc135600563)

[BAB 3 IMPLEMENTASI 10](#_Toc135600564)

[3.1 Stream Data From Twitter 10](#_Toc135600565)

[3.2 Stream Data From Twitter 12](#_Toc135600566)

[BAB 4 HASIL DAN PEMBAHASAN 15](#_Toc135600567)

[4.1 Hasil Prediksi 15](#_Toc135600568)

[4.2 Hasil Evaluasi 15](#_Toc135600569)

DAFTAR GAMBAR

[Gambar 1. 1 Arsitektur Sistem 7](#_Toc135600392)

[Gambar 1. 2 Machine Learning Pipeline 8](#_Toc135600393)

DAFTAR POTONGAN KODE PROGRAM

[Potongan Kode 1. 1 Inisiasi API dan Access Token 10](#_Toc135600430)

[Potongan Kode 1. 2 Alamat Broker Kafka 10](#_Toc135600431)

[Potongan Kode 1. 3 Kata Kunci Pencarian 10](#_Toc135600432)

[Potongan Kode 1. 4 Mengakses API Twitter 11](#_Toc135600433)

[Potongan Kode 1. 5 Import Library 11](#_Toc135600434)

[Potongan Kode 1. 6 Pengambilan Data Twitter 12](#_Toc135600435)

[Potongan Kode 1. 7 Import Library 12](#_Toc135600436)

[Potongan Kode 1. 8 Membaca Data Streaming 13](#_Toc135600437)

[Potongan Kode 1. 9 Membaca Data dari Kafka 13](#_Toc135600438)

[Potongan Kode 1. 10 Membersihkan Data Teks Tweet 13](#_Toc135600439)

[Potongan Kode 1. 11 Melakukan Klasifikasi Sentimen 14](#_Toc135600440)

**BAB 1  
PENDAHULUAN**

Pada bab ini dijelaskan mengenai latar belakang, pertanyaan penelitian, tujuan penelitian, dan sistematika dari penulisan dokumen tugas akhir.

* 1. **Latar Belakang**

*Sentiment analysis* merupakan kategori khusus penambangan data yang memperkirakan temperamen pendapat orang menggunakan pemrosesan bahasa alami, linguistik tipe komputasi, dan analisis data dalam bentuk teks. *Sentiment analysis* digunakan untuk melihat opini terhadap sebuah masalah menuju ke opini positif atau netral. *Sentiment analysis* digunakan di berbagai media sosial, *Twitter* merupakan salah satu media sosial yang digunakan untuk memberikan opini melalui *tweet*. Pengguna *twitter* akan memberikan opini tentang suatu hal, seperti salah satu lagu yang sedang populer dikalangan masyarakat khususnya muda mudi . Opini dari pengguna tersebut akan sangat bermanfaat bagi pengguna lain dan rumah produksi suatu musik tersebut.

Musik merupakan salah satu topik yang trending pada twitter, banyak orang yang memberikan komentar atau *tweet* terhadap topik tersebut. ada beberapa jenis *tweer* yang dilontarkan atau diberikan terhadap topik musik. *Tweet* yang masih tersusun secara acak menyebabkan kesulitan bagi pengguna dalam mengetahui opini positif, dan netral. Berdasarkan permasalahan tersebut akan dilakukan penelitian tentang *Twitter Analysis Sentiment* untuk mengklasifikasikan *tweet* positif, ataupun netral. Klasifikasi opini diperlukan untuk memudahkan pengguna dalam melihat opini positif, dan netral.

*Analysis Sentiment* akan mengidentifikasi sentimen pada sebuah *tweet* paling relevan pada sebuah dokumen tunggal maupun banyak dokumen kemudian membuat ringkasan yang mewakili, sehingga *aspect oriented* berbasis sentimen dilakukan dengan cara melakukan klasifikasi polaritas kalimat kedalam kelompok positif, dan netral, kemudian menentukan aspek mana yang paling penting untuk dijadikan ringkasan final, karena banyaknya aspek yang terdapat pada setiap *tweet* sehingga pihak terkait kesulitan dalam melakukan *tweet* akan dirangkum berdasarkan sentimen pada aspek yang ada pada *tweet*.

Maka akan dilakukan analisis sentimen terhadap data *twitter* yakni topik musik untuk mendapatkan *polarity* dan subjektif terhadap data twitter tersebut.

* 1. **Rumusan Masalah**

Berdasarkan uraian pada latar belakang diatas maka rumusan masalah dalam penyelesaian penelitian ini adalah Bagaimana reaksi publik terhadap lagu cupid yang sedang dibahas menggunakan *Analysis Sentiment*?

* 1. **Tujuan dan Manfaat**

Tujuan dari penelitian ini adalah Mengetahui reaksi publik terhadap lagu cupid yang sedang dibahas menggunakan *Analisis Sentimen*.

* 1. **Ruang Lingkup**

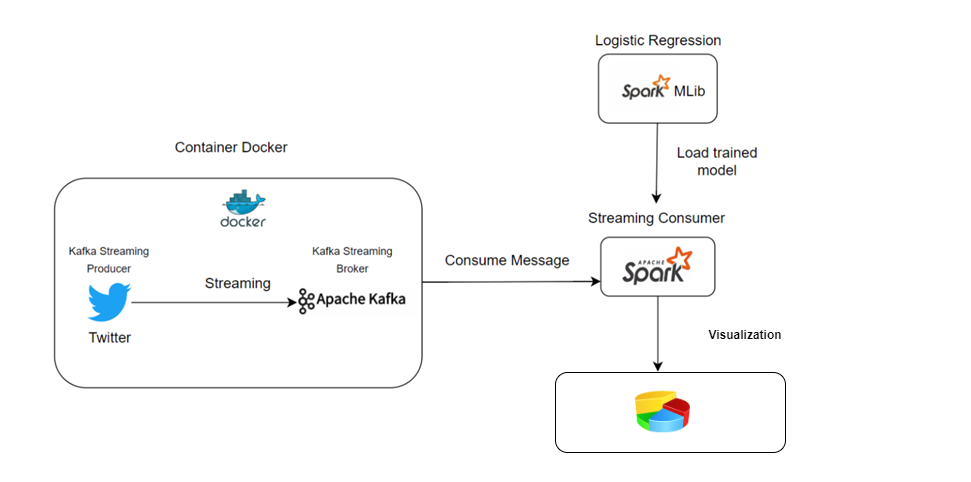
Ruang lingkup yang akan dilakukan dalam penelitian ini adalah Data yang digunakan adalah *tweet* positif dan netral pada lagu cupid yang ada pada *twitter* dengan melakukan API *streaming.*

**BAB 2**  **ARSITEKTUR SISTEM**

Bab ini menjelaskan arsitektur sistem yang dapat digunakan sebagai solusi dalam melakukan analisis sentimen Live Streaming Data Twitter menggunakan keyword “twitter” untuk mengetahui pandangan dan reaksi masyarakat melalui diskusi pada twitter terhadap film terbaru. Analisis sentimen dilakukan untuk mengekstrak sentimen dari teks dalam bahasa Inggris pada tweet yang terkait dengan musik terbaru dari data streaming Twitter. Hasil analisis sentimen akan memberikan informasi tentang apakah tweet tersebut memiliki sentimen positif, maupun netral. Metode yang digunakan dalam analisis sentimen Arsitektur Big Data yang digunakan dalam penelitian ini mencakup struktur keseluruhan dari sistem *logical* dan *physical* yang digunakan untuk menyimpan, mengakses, dan memproses big data.

1. **Arsitektur Sistem**

Berikut ini merupakan arsitektur sistem pada analisis sentimen pada streaming data twitter untuk reaksi terhadap lagu cupid.



Gambar 1. 1 Arsitektur Sistem

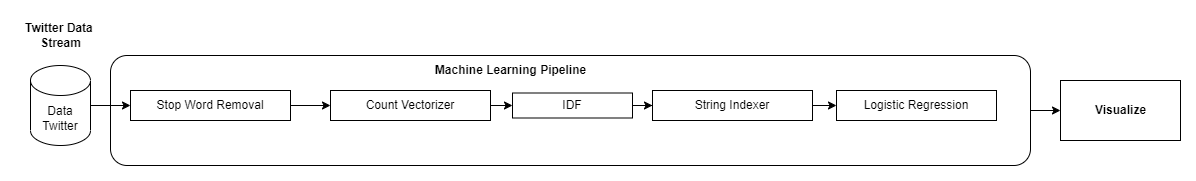
Pada gambar tersebut merupakan alur arsitektur sistem yang akan digunakan dalam melakukan analisis sentimen pada Live Streaming Twitter dengan menggunakan kata kunci “cupid”. Arsitektur sistem dibentuk agar dapat mengolah data yang terstruktur dan tidak terstruktur. Berikut adalah penjelasan setiap tahapan yang dilakukan pada arsitektur sistem. Pada arsitektur sistem tersebut, akan dilakukan ekstrak data dari Twitter API, membuat aliran data menggunakan Apache Kafka dan mengubah data untuk melakukan analisis sentimen waktu nyata dengan Spark Streaming dan menyimpan hasilnya ke Delta Lake atau MongoDB.

Pada apache streaming, dimulai dengan mengimpor pustaka yang diperlukan, API key dan membuat produser kosong. Produser kemudian dihubungkan ke aliran data API Twitter dan mengirimkan data ke topik twitter. Kemudian menghubungkan konsumen ke aliran produser dan mengubah data.

Selanjutnya membuat entitas menggunakan streaming data, dengan mengimport library dan membuat objek Spark, jika ingin menggunakan model lain harus mengubah nilai path\_to\_model. Setelah terhubung ke topik Twitter, mengambil langsung data terakhir dan memasukkan nilai keluaran sebagai string (teks tweet). Dibuat fungsi preprocessing dan memuat model pre train sebagai mengubah data sesuai pipa yang ditentukan selama train (ekstraksi fitur, dll) dan memilih kolom yang akan diambil. Selanjutnya hasil data tersebut dapat disimpan secara lokal dan divisualisasikan dengan menggunakan pie chart.

1. **Machine Learning Pipeline**

Machine learning pipeline adalah salah satu jenis pipeline yang terus mengolah dan mempersiapkan data untuk keperluan pelatihan model machine learning.



Gambar 1. 2 Machine Learning Pipeline

Berdasarkan proses Machine Learning Pipeline, data Twitter diambil secara real-time menggunakan Tweepy dan difilter menggunakan fungsi Tweepy untuk mendapatkan data yang relevan. Data streaming ini merupakan sumber data untuk melatih model machine learning untuk analisis sentimen terhadap suatu topik tertentu. Pada data preprocessing dilakukan penghilangan kata-kata yang termasuk stopwords dan menghapus tanda baca pada kalimat dan menghapus kata-kata yang tidak perlu dalam melakukan sentiment analysis. Kemudian data akan diklasifikasikan dengan polarity dan subjektif, dengan melakukan klasifikasi label positive, netral. Setelah dilakukan klasifikasi data hasil klasifikasi tersebut akan divisualisasikan dengan menggunakan grafik.

**BAB 3** **IMPLEMENTASI**

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 mendapat token akses *twitter* tersebut, akan dibuat program untuk mendapatkan data *twitter* secara *streaming.* Pertama yaitu menginisiasi api dan *access token* untuk mengautentikasi. Berikut merupakan potongan codenya:

|  |
| --- |
| 1. """API ACCESS KEYS""" 2. consumerKey = 'QnupptSEsqShuukhaPGsk0Svf' 3. consumerSecret = "HuD4bnPtEZcE1IVFDcluzuYcngbSMMYxKyfYA8MTXPUYZvn4N0" 4. accessToken = '1390698546353999878-KExzyOWL2qm1zfmxZdrYmcnkaVbjue' 5. accessTokenSecret = 'dkEkoVa4QwW3Z8Rzvau6r37C7DhnUPU2zu91P6GZmm7Ec' 6. bearerToken = 'AAAAAAAAAAAAAAAAAAAAALP5cAEAAAAACkdJeE7i23YH3L4qeYWpvoeqZAA%3DoOefkiPYcftGrIzdm9YgPi7hnYrI6ymuojs6031Dd5K5U7CWul' |

Potongan Kode 1. 1 Inisiasi API dan Access Token

Setelah mendefinisikan beberapa API keys yang diperlukan untuk mengakses API twitter. Selanjutnya mengatur tingkat logging ke INFO yang akan memungkinkan program mencetak pesan logging ke konsol. Dan membuat objek produsen Kafka daan menentukan alamat broker Kafka.

|  |
| --- |
| 1. logging.basicConfig(level=logging.INFO) 2. producer = KafkaProducer(bootstrap\_servers='localhost:9092') |

Potongan Kode 1. 2 Alamat Broker Kafka

Selanjutnya mendefinisikan kata kunci pencarian untuk mencari tweet dan nama topik Kafka yang akan digunakan untuk mengirimkan data tweet.

|  |
| --- |
| 1. search\_term = 'cupid' 2. topic\_name = 'twitter' |

Potongan Kode 1. 3 Kata Kunci Pencarian

Setelah melakukan inisiasi pada api dan access token akan dilakukan autentikasi pengguna dengan kredensial API twitter dan mengembalikan objek ‘api’ yang digunakan untuk mengakses API twitter, berikut potongan kodenya:

|  |
| --- |
| 1. def twitterAuth(): 2. # create the authentication object 3. authenticate = tweepy.OAuthHandler(consumerKey, consumerSecret) 4. # set the access token and the access token secret 5. authenticate.set\_access\_token(accessToken, accessTokenSecret) 6. authenticate.secure = True 7. # create the API object 8. api = tweepy.API(authenticate, wait\_on\_rate\_limit=True) 9. return api |

Potongan Kode 1. 4 Mengakses API Twitter

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

|  |
| --- |
| 1. import tweepy 2. from kafka import KafkaProducer 3. import logging 4. import json |

Potongan Kode 1. 5 Import Library

Library tweepy digunakan untuk melakukan streaming data pada twitter. Library logging untuk melakukan logging atau pencatatan pesan dari sebuah program seperti pesan error, informasi, peringatan dan lainnya. Library json digunakan untuk membantu pengolahan data dalam format JSON. Dan untuk mengimpor kelas “Kafkaproducer” digunakan dari modul kafka.

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

|  |
| --- |
| 1. class TweetListener(tweepy.StreamingClient): 2. def on\_data(self, raw\_data): 3. logging.info(raw\_data) 4. tweet = json.loads(raw\_data) 5. if tweet['data']: 6. data = { 7. 'message': tweet['data']['text'].replace(',', '') 8. } 9. producer.send(topic\_name, value=json.dumps(data).encode('utf-8')) 10. return True 11. @staticmethod 12. def on\_error(status\_code): 13. if status\_code == 420: 14. # returning False in on\_data disconnects the stream 15. return False 16. def start\_streaming\_tweets(self, search\_term): 17. self.add\_rules(tweepy.StreamRule(search\_term)) 18. self.filter() |

Potongan Kode 1. 6 Pengambilan Data Twitter

Class TweetListener digunakan untuk mengambil data user twitter dan text tweet, kemudian data tersebut disimpan ke dalam variabel tweets. Kelas ini digunakan untuk memproses data tweet yang diterima dari Twitter API dan mengirimkan data tersebut ke Kafka broker.

1. **Stream Data From Twitter**

*Sentiment analysis* bertujuan untuk mengklasifikasikan terkait identifikasi terhadap informasi subjektif dan ulasan pada media sosial. Pada sub bab ini akan menjelaskan kode yang berfungsi untuk membaca data real-time dari Kafka, melakukan preprocessing pada data, dan menerapkan model untuk membuat prediksi sentimen dari teks dalam data. Selanjutnya, prediksi akan ditampilkan di konsol sebagai output. Berikut adalah penjelasan dari setiap bagian kode:

Dalam melakukan analisis sentimen, yang dilakukan pertama adalah import library.

|  |
| --- |
| 1. import re 2. import findspark 3. from pyspark.ml import PipelineModel 4. from pyspark.sql import SparkSession 5. from pyspark.sql.functions import from\_json, udf 6. from pyspark.sql.types import StructType, StructField, StringType, ArrayType |

Potongan Kode 1. 7 Import Library

Untuk melakukan inisialisasi SparkSession dan Spark Context untuk membaca data streaming dari Kafka dan melakukan analisis menggunakan PySpark dan menentukan path\_to\_model yang menunjukkan ke model Machine Learning yang telah dilatih sebelumnya.

|  |
| --- |
| 1. if \_\_name\_\_ == "\_\_main\_\_": 2. findspark.init() 3. # Path to the pre-trained model 4. path\_to\_model = r'' 5. # Config 6. spark = SparkSession \ 7. .builder \ 8. .master("local[\*]") \ 9. .appName("TwitterSentimentAnalysis") \ 10. .config("spark.jars.packages", "org.apache.spark:spark-sql-kafka-0-10\_2.12:3.3.0") \ 11. .getOrCreate() |

Potongan Kode 1. 8 Membaca Data Streaming

Selanjutnya membuat schema StructType yang digunakan untuk membaca data dari Kafka. Membaca data streaming dari Kafka dengan konfigurasi tertentu seperti menggunakan format kafka, subscribe ke topik ‘twitter’ dan menggunakan nilai terbaru.

|  |
| --- |
| 1. # Schema for the incoming data 2. schema = StructType([StructField("message", StringType())]) 3. # Read the data from kafka 4. df = spark \ 5. .readStream \ 6. .format("kafka") \ 7. .option("kafka.bootstrap.servers", "localhost:9092") \ 8. .option("subscribe", "twitter") \ 9. .option("startingOffsets", "latest") \ 10. .option("header", "true") \ 11. .load() \ 12. .selectExpr("CAST(value AS STRING) as message") 13. df = df \ 14. .withColumn("value", from\_json("message", schema)) |

Potongan Kode 1. 9 Membaca Data dari Kafka

Membersihkan data dengan melakukan pre-processing pada data tweet dengan menghapus tanda baca dan URL, kemudian menggunakan lambda function untuk membuat fungsi udf (user-defined function) yang digunakan untuk membersihkan data teks tweet.

|  |
| --- |
| 1. # Pre-processing the data 2. pre\_process = udf( 3. lambda x: re.sub(r'[^A-Za-z\n ]|(http\S+)|(www.\S+)', '', x.lower().strip()).split(), ArrayType(StringType()) 4. df = df.withColumn("cleaned\_data", pre\_process(df.message)).dropna() |

Potongan Kode 1. 10 Membersihkan Data Teks Tweet

Melakukan klasifikasi sentimen pada data yang telah dibersihkan dan memilih kolom-kolom yang relevan untuk diprediksi, yaitu teks tweet dan prediksi sentimen. Selanjutnya akan menampilkan hasil prediksi pada konsol dan memulai streaming dan menunggu proses stream untuk selesai.

|  |
| --- |
| 1. # Load the pre-trained model 2. pipeline\_model = PipelineModel.load(path\_to\_model) 3. # Make predictions 4. prediction = pipeline\_model.transform(df) 5. # Select the columns of interest 6. prediction = prediction.select(prediction.message, prediction.prediction) 7. # Print prediction in console 8. prediction \ 9. .writeStream \ 10. .format("console") \ 11. .outputMode("update") \ 12. .start() \ 13. .awaitTermination() |

Potongan Kode 1. 11 Melakukan Klasifikasi Sentimen

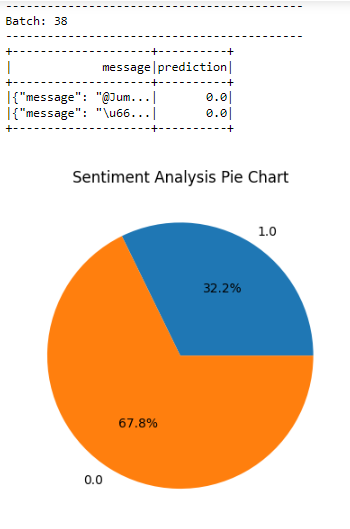
BAB 4  
HASIL DAN PEMBAHASAN

* 1. Hasil Prediksi

Pada sentimen analisis, akan dilakukan prediksi terhadap data tweet pada twitter mengenai polarity dan subjectivity. Pada perlakukan prediksi polarity, terdapat dua label yakni, positive, dan neutral. Untuk hasil prediksi dari data tweet ditampilkan secara real time sehingga tidak mencantumkan total persentase prediksi label positif dan netral. Selain itu, diperlukan *code* tambahan untuk melakukan visualisasi data agar dapat dilakukan analisis total persentase prediksi label positif dan netral.

* 1. Hasil Evaluasi

Setelah data tweet tersebut sudah dilakukan sentiment analysis, hasil dari sentiment analysis tersebut akan di visualisasikan menggunakan pie chart. Pie chart tersebut digunakan untuk memberikan informasi terkait total persentase prediksi label positif dan netral. Berikut adalah gambar dari hasil visualisasi yang telah dilakukan:



Gambar 4. 1 Visualisasi hasil prediksi