**LAPORAN PROYEK PENGOLAHAN DATA BESAR**

**Sentiment Analisis pada Streaming Data Twitter untuk Review Produk Miniso**

****

**Disusun oleh:**

**12S18002 – Wiranda Siahaan**

**12S18009 – Novita Enjelia Hutapea**

**12S18041 – Merika H. Manurung**

**12S18046 – Tiara Octavia Situmorang**

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

**MEI 2022**

# 

# 

# **DAFTAR ISI**

# 

# 

# **DAFTAR GAMBAR**

# 

# **DAFTAR TABEL**

# 

# **DAFTAR KODE PROGRAM**

# 

# **Bab 1**

# **Pendahuluan**

## **1.1 Latar Belakang**

Twitter sebagai salah satu media sosial yang cukup populer digemari sebagai salah satu media penyebar informasi yang sangat cepat. Informasi yang beredar juga sangat banyak mulai dari berita, opini, komentar, kritik dan semuanya ada yang bersifat positif, negatif dan netral. Menurut data yang dilansir Statista berdasarkan hasil penelitian PeerReach, Indonesia termasuk pengguna twitter yang aktif ke 3 di dunia dibawah Amerika Serikat dan Jepang. Pada Twitter terdapat istilah bernama tweet yang merupakan sebuah pesan atau status yang dibuat oleh penggunanya. Sebuah tweet dapat mengekspresikan sebuah perasaan atau keadaan dari pengguna Twitter. Tweet dapat mengandung sebuah opini dari penggunanya terhadap kejadian yang dialaminya. Opini tersebut dapat dimanfaatkan sebagai penilaian bagi suatu objek.

Twitter dapat diakses melalui web atau perangkat seluler. Untuk berbagi informasi di Twitter seluas mungkin, Twitter juga menyediakan akses terprogram ke data Twitter kepada perusahaan, pengembang, dan pengguna melalui API (antarmuka pemrograman aplikasi). Pada tingkat tinggi, API adalah cara program komputer "berbicara" satu sama lain sehingga mereka dapat meminta dan mengirimkan informasi. Hal ini dilakukan dengan mengizinkan aplikasi perangkat lunak untuk memanggil apa yang dikenal sebagai titik akhir: alamat yang sesuai dengan jenis informasi tertentu yang diberikan. Twitter mengizinkan akses ke bagian layanan kami melalui API untuk memungkinkan orang membangun perangkat lunak yang terintegrasi dengan Twitter, seperti solusi yang membantu perusahaan menanggapi umpan balik pelanggan di Twitter.

Platform API Twitter menyediakan akses luas ke data Twitter publik yang telah dipilih pengguna untuk dibagikan kepada dunia. Twitter juga mendukung API yang memungkinkan pengguna untuk mengelola informasi Twitter non-publik mereka sendiri (mis., Direct Message) dan memberikan informasi ini kepada pengembang yang telah mereka izinkan untuk melakukannya

Analisis sentimen merupakan bagian dari Natural Language Processing (NLP) dan Machine Learning. Cara kerjanya adalah dengan melakukan klasifikasi terhadap opini positif dan negatif. Analisis sentimen menganalisa pendapat orang, sentimen, evaluasi, penilaian, sikap, dan emosi terhadap entitas seperti produk, jasa, organisasi, individu, masalah, peristiwa atau suatu topik. Sentimen analisis ini bertujuan untuk menentukan sentimen publik terhadap objek tertentu yang disampaikan melalui twitter dengan bahasa indonesia yang dapat membantu usaha untuk melihat opini publik terkait dengan objek tersebut sebagai bahan pertimbangan pengambilan keputusan.

Berdasarkan pemaparan yang telah diuraikan, penelitian ini bermaksud untuk membangun analisis sentimen pada twitter dengan objek review yang diberikan pada produk Miniso.

## **1.2 Rumusan Masalah**

Berdasarkan uraian pada latar belakang dapat dirumuskan permasalahannya yaitu bagaimana menganalisis dan mengklasifikasi sentimen melalui tweet di twitter dengan objek review yang diberikan pada produk Miniso.

## **1.3 Tujuan**

1. Mengetahui pengaruh algoritma terhadap proses analisis sentimen untuk review yang diberikan pada produk Miniso
2. Melakukan tweet preprocessing dan sentiment analysis.
3. Melakukan Klasifikasi pada data Twitter yang diperoleh.

## **1.4 Manfaat**

Manfaat dari pengerjaan proyek :

Proyek ini diharapkan nantinya dapat memberikan pemahaman dan keterampilan bagi tim proyek untuk mengimplementasikan teknik machine learning (scalable machine learning) pada data besar khususnya pada salah satu metode dari pengklasifikasian yaitu Logistic Regression dengan teknologi pengolahan data besar menggunakan Spark Mlib API untuk memproses dan menganalisis data besar.

## **1.5 Ruang Lingkup**

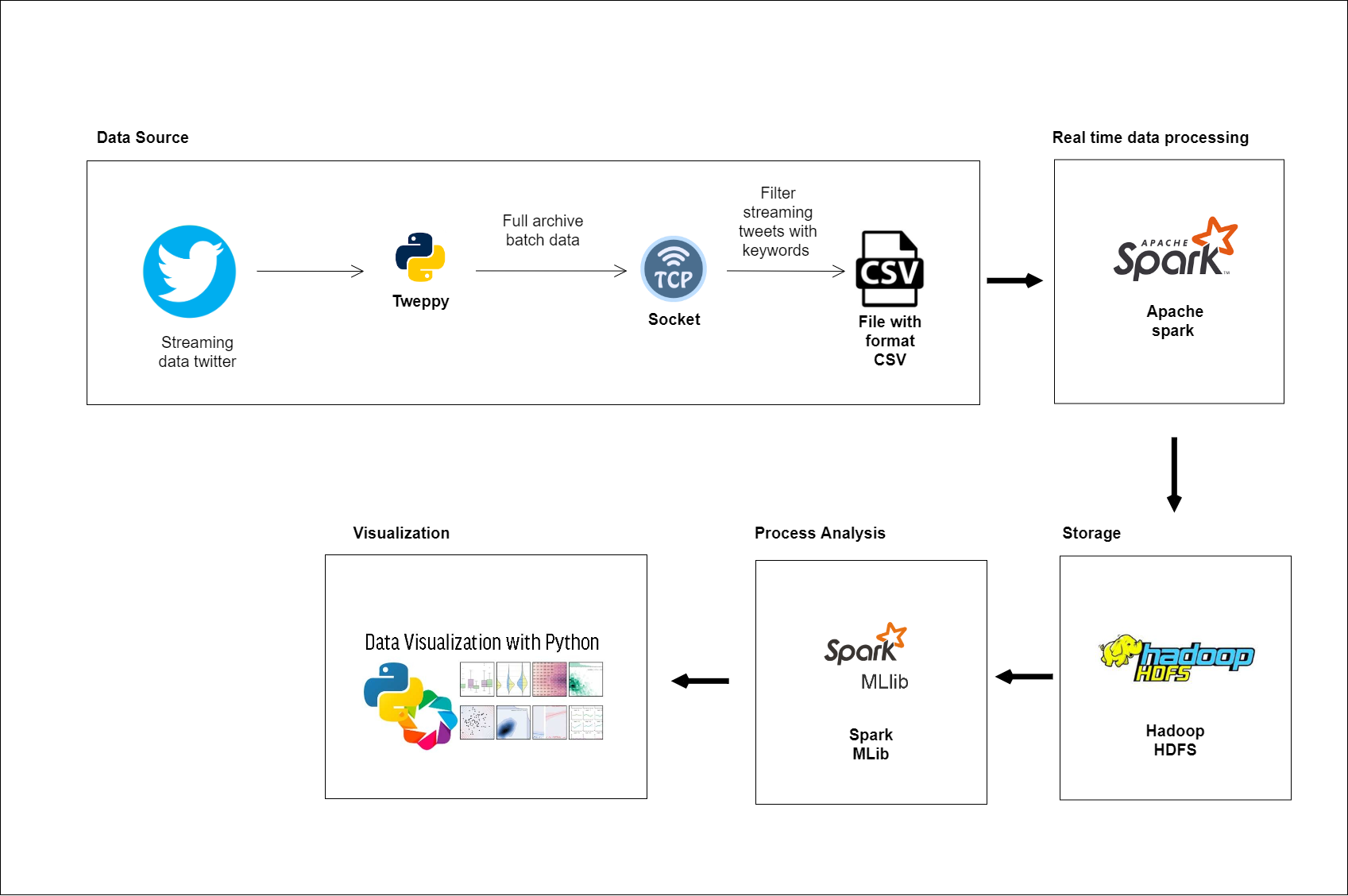
Ruang lingkup dalam pengerjaan proyek ini adalah

1. Melakukan pengolahan data menggunakan salah satu metode yang ada dari *Classification.*
2. Penelitian ini hanya didasarkan pada tanggapan yang diberikan pada platform Twitter tentang review yang diberikan pada produk Miniso.

# **Bab 2**

# **Solusi**

## **2.1 Big Data Arsitektur**



**Data Source,** data diambil menggunakan *developer credentials* untuk dapat melakukan autentikasi, dan terhubung dengan API dari twitter. Dengan memanfaatkan pustaka *tweepy* data akan diekstrak dari versi API dengan arsip lengkap, yang kemudian akan di filter sesuai dengan *keyword* yang akan digunakan pada pemodelan, data yang akan diberikan disimpan pada file dengan format CSV. Pada proses pengambilan data, penggunaan *socket* TCP antara *tweepy* dengan spark dimana *socket* TCP akan membantu proses pengiriman tweet ke proses streaming spark.

**Real time data processing,** dalam pengolahan data real time, akan digunakan apache spark sehingga data kemudian akan disimpan pada hadoop HDFS atau Hadoop Distributed File System dalam bentuk database.

**Process analysis,** kemudian dilanjutkan dengan proses analisis menggunakan Machine Learning dengan Spark MLib yang mendukung berbagai penggunaan algoritma.

**Visualization**, data yang telah berhasil didapatkan dengan melalui proses tersebut kemudian akan divisualisasikan dengan memanfaatkan beberapa pustaka yang telah disediakan oleh Python.

### **2.2.1 ETL Process**

Proses ETL merupakan proses integrasi data dengan menggabungkan data dari berbagai sumber ke dalam suatu penyimpanan dan dimuat ke dalam warehouse atau sistem lainnya.

#### **2.2.1.1 Extract**

Hal pertama yang dilakukan adalah mengekstraksi atau mengidentifikasi semua data dari sumbernya sebelum dipindahkan ke tempat penyimpanan yang lain

#### **2.2.1.2 Transform**

Setelah tahap ekstraksi selesai, data perlu dipindahkan ke sistem target/ perantara untuk diproses lebih lanjut. Transform merupakan proses membersihkan dan mempersiapkan agregasi untuk analisis. Hal ini dilakukan untuk membantu memastikan data yang akan diolah siap dan kompatibel. Hal-hal yang dilakukan pada saat transform adalah menghapus data yang tidak konsisten dan data yang sama, menghapus data yang tidak dapat digunakan, mengatur data menurut jenisnya serta menambahkan aturan tertentu yang dapat meningkatkan kualitas data.

#### **2.2.1.3 Load**

Load merupakan tahap terakhir dari proses ETL. Load merupakan proses untuk membuat atau memindahkan data yang telah ditransformasi pada proses sebelumnya ke data warehouse. Data dapat dimuat sekaligus maupun secara terjadwal.

### **2.2.2 Data Preparation**

*Data Preparation* merupakan tahap dimana data dipersiapkan dan dianalisis dengan cara melakukan eksplorasi data sehingga dapat diketahui model atau algoritma yang cocok digunakan untuk mengolah data. Adapun hal yang dilakukan pada saat data preparation adalah EDA (Exploratory Data Analysis) yaitu dengan mengekstrak variabel atau fitur yang penting, menghapus variabel yang tidak diperlukan, memeriksa pencilan (outliers) dan nilai yang hilang (missing values).

### **2.2.3 Data Segregation**

Data *segregation* adalah proses membagi (*split*) dataset menjadi dua bagian yaitu *train data* dan *test data*. *Train data* merupakan kumpulan data yang digunakan untuk melatih algoritma sementara test data merupakan kumpulan data yang digunakan untuk mengetahui performa algoritma yang sudah dilatih sebelumnya dengan model tertentu, sehingga dapat diketahui nilai keakuratannya. Pembagian dataset pada proyek ini menggunakan komposisi perbandingan 70% untuk *train data* dan 30% untuk *test data*. Antar *train data* dan *test data* tidak boleh overlap, karena jika overlap akan merusak proses training model. Sehingga komposisi untuk train data harus jauh lebih besar dari test data.

### **2.2.4 Model Training**

### **2.2.5 Evaluation**

### **2.2.6 Tuning Model**

# 

# **Bab 3**

# **Implementasi**

## **3.1 ETL**

## **3.2 Data Preparation**

## **3.3 Load Data**

## **3.4 Model training**

## **3.5 Evaluation**

## **3.6 Tuning Model**

# **Bab 4**

# **Hasil dan Pembahasan**

## 

## **4.1 Hasil Prediksi**

## **4.2 Hasil Tuning Model**

## **4.3 Hasil Visualisasi**

## **4.4 Kesimpulan**

# 

# 

# **DAFTAR PUSTAKA**