**LAPORAN PROYEK**

**Analysis and Classification Twitter Data Stream of Pemilu Topic Using Spark Mlib**

**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**

# 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, forum dan dunia virtual. Andreas Kaplan dan Michael Haenlein mendefinisikan media sosial sebagai “Sebuah kelompok aplikasi berbasis internet yang membangun di atas dasar ideologi dan teknologi Web 2.0, dan yang memungkinkan penciptaan dan pertukaran “*user-generated content*”. Profesor J.A. Barnes pada tahun 1954 memperkenalkan jejaring sosial (*Social Network*), “Jejaring sosial adalah struktur sosial yang terdiri dari elemen-elemen individu atau organisasi. Jejaring menunjukan jalan dimana mereka berhubungan karena kesamaan sosialitas, mulai dari mereka yang dikenal sehari-hari sampai dengan keluarga”. *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, lazim disebut sebagai jaringan pertemanan. Adapun beberapa situs jejaring sosial populer adalah LinkedIn, Facebook dan Twitter.

Perkembangan pengguna jejaring sosial saat ini terus berkembang pesat, sehingga setiap hari terdapat sejumlah komentar, pendapat dan artikel yang dibuat oleh pengguna pada media sosial. Layanan jejaring sosial biasanya berbasis web, dilengkapi dengan berbagai fitur bagi penggunanya agar dapat saling berkomunikasi dan berinteraksi. Media sosial yang dijadikan pembahasan dalam proyek ini adalah Twitter. Twitter menjadi salah satu media sarana bertukar opini untuk berbagai *issue* dan topik terkini dari berbagai bidang dan dalam kehidupan sehari-hari. Pengguna twitter yang sudah terdaftar dapat mengirim dan membaca pesan berbasis teks hingga 140 karakter yang disebut *tweet.* Sedangkan yang belum terdaftar hanya bisa membaca *tweet.*

Sejumlah postingan teks yang terdapat pada Twitter meningkat setiap harinya sehingga membuat *database* yang ada tidak dapat menangani ekstensif jumlah data dalam waktu singkat. *Database* dirancang untuk memproses data terstruktur, namun terdapat batasan dengan data besar. Beberapa *framework* yang dapat mengatasi hal tersebut adalah *Hadoop, Apache Spark,* dan *HBase*. Data yang ada pada Twitter termasuk ke dalam data besar, menurut website resmi *Apache*: *Apache Spark* adalah framework yang digunakan untuk memproses, menanyakan, dan menganalisis Big Data. *Apache Spark* melakukan pemrosesan data melalui *in-memory*, sehingga waktu pemrosesan lebih cepat daripada *framework* sejenis. *Apache Spark* memiliki beberapa komponen dan dukungan dari berbagai bahasa pemrograman, seperti *Spark Core, Spark SQL, MLlib* dan lainnya. Selain itu, *framework* ini juga menyediakan berbagai *library* yang dapat memenuhi kebutuhan pemrosesan data. *Hadoop* merupakan *open source framework* di bawah lisensi Apache, yaitu platform perangkat lunak yang digunakan sebagai *analytic engine* yang memungkinkan penulisan perintah (*write*) dan menjalankan (*run*) aplikasi yang memproses data dalam jumlah besar. Terdapat 2 komponen utama dari hadoop, yaitu *Hadoop Distributed File System (HDFS)* yang mendistribusikan data ke *cluster* sehingga dapat diukur dan cepat. Selain itu, terdapat komponen *MapReduce* untuk memproses data dengan cara yang handal dan juga cepat.

Untuk mengenali dan mengekstraksi opini dalam bentuk teks, dapat 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 sentimen tertentu. Sedangkan pada teknik *machine learning based sentiment analysis,* fitur-fitur di dalamnya dapat mengenali sentimen (sudut pandang seseorang) dalam sebuah teks. Mengekstraksi informasi opini yang terdapat dalam data teks dari platform Twitter secara otomatis dengan pendekatan komputasional penting untuk dilakukan. Informasi berbentuk teks banyak terdapat di internet dalam format forum, blog, media sosial, serta situs berisi review. Dengan bantuan sentiment analysis, informasi yang tadinya tidak terstruktur dapat diubah menjadi data yang lebih terstruktur. Pengerjaan proyek ini akan melakukan eksperimen dengan menggunakan teknik sentiment analysis pada dataset Twitter menggunakan framework Apache spark. Karakteristik real-time pada data Twitter membutuhkan sistem yang mendukung pemrosesan secara streaming. Oleh karena itu, pengerjaan proyek akan dilakukan menggunakan framework Apache spark. Eksekusi dari aplikasi yang dibangun menggunakan Spark dapat mendukung pemrosesan data secara real-time, sehingga dapat digunakan untuk membangun pipa pemrosesan Big data dari berbagai sumber data, menuju penyimpanan data secara terus menerus.

## **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 metode *Naive Bayes Classifier*
4. Bagaimana menampilkan visualisasi dari hasil data yang telah diperoleh

## **1.3 Tujuan**

Adapun tujuan dari pengerjaan proyek adalah:

1. Memperoleh data trend Twitter dengan keyword #pemilu2024.
2. Melakukan pengolahan data *real-time* menggunakan sistem yang mendukung komputasi dan pemrosesan secara *streaming*.
3. Mengklasifikasi dan menganalisis penilaian mengenai persepsi/opini masyarakat terhadap pemilu 2024 mendatang berdasarkan sentimen publik yang ada di Twitter menggunakan metode *Naive Bayes*.
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

## **1.5 Ruang Lingkup**

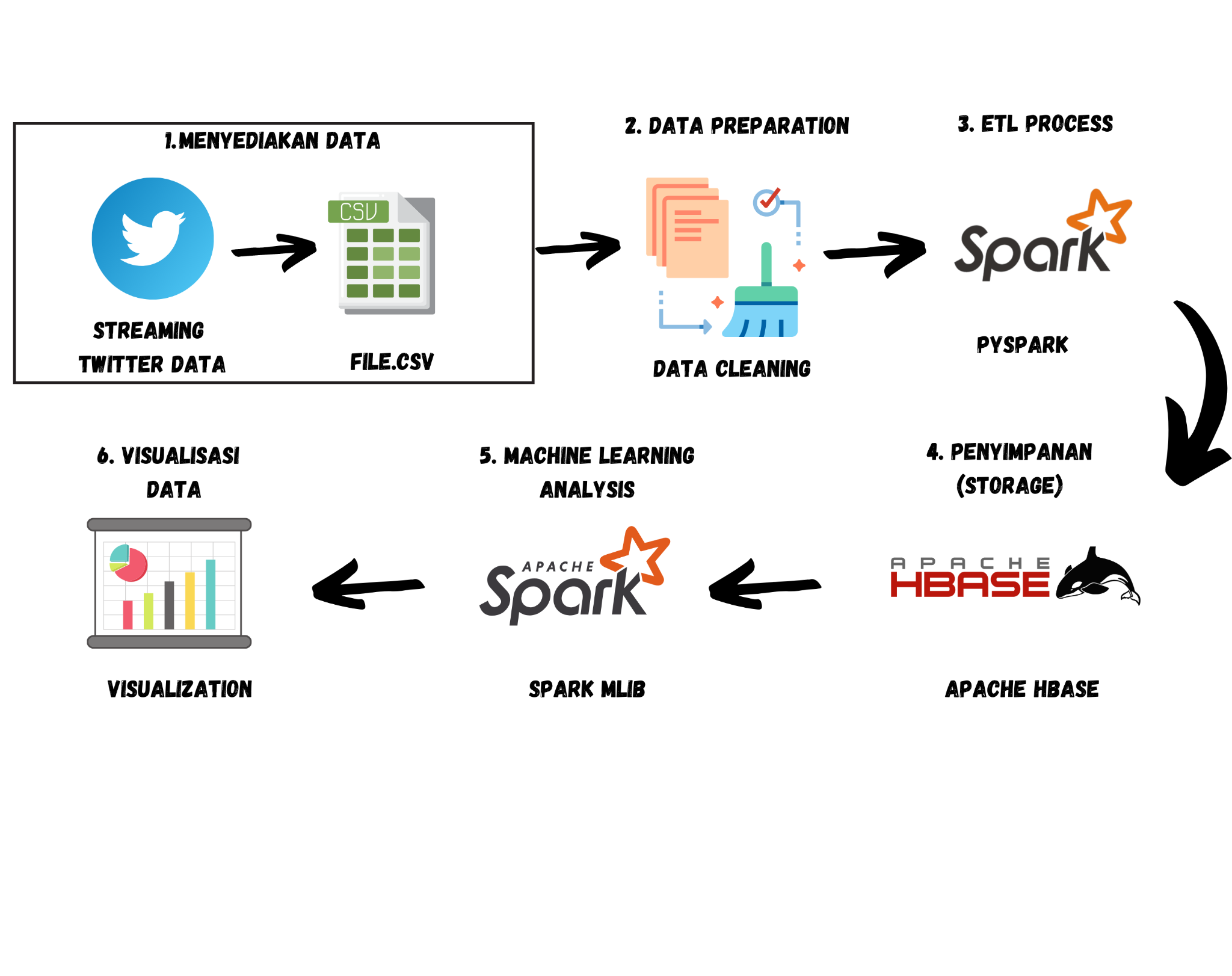
Ruang lingkup dari pengerjaan proyek ini adalah sebagai berikut:

1. Dataset yang digunakan adalah data twitter, dengan topik pemilu dan menggunakan keyword #pemilu2024.
2. Sistem yang digunakan untuk mengenali dan mengekstraksi data adalah *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 Hadoop dan Apache Spark.

# BAB 2 SOLUSI

**2.1 Arsitektur *Big Data***

Berikut ini merupakan arsitektur big data yang menjadi solusi terhadap trend pemilu 2024 pada Twitter.

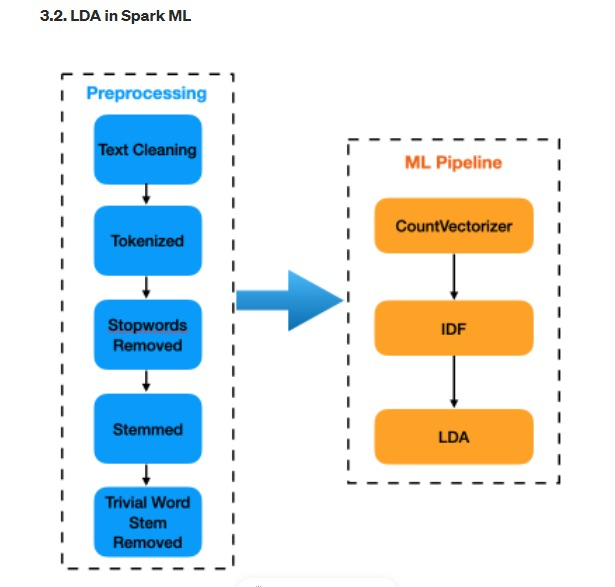


Pada pengerjaan proyek, data yang digunakan adalah data teks twitter khususnya topik pemilu dengan keyword #pemilu2024. Pertama data disediakan terlebih dahulu dengan cara melakukan crawling data dari twitter menggunakan API key twitter. Proses pengambilan data Twitter dibantu dengan bahasa pemrograman 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 pengguna nya pada saat membangun suatu sistem perangkat lunak. API Key Twitter memiliki consumer keys, consumer secret, access key, dan accsess secret yang digunakan untuk mengakses data Twitter yang dibutuhkan. Setelah data berhasil dicrawling, maka akan menghasilkan file berbentuk csv.

Kedua, dilakukan data preparation untuk mengubah data mentah menjadi lebih efisien. Ketiga melakukan ETL process menggunakan pyspark. Keempat data disimpan di storage. Adapun storage yang kami gunakan adalah apache HBase. Kelima, dilakukan machine learning analysis menggunakan apache spark mlib. Terakhir, hasil yang diperoleh dari data analytics yang telah dilakukan kemudian akan disimpan untuk selanjutnya divisualisasikan. Visualisasi tersebut dapat memberikan pemahaman mengenai pola yang ditemukan dari hasil analisis data yang telah diperoleh dan memperoleh informasi dengan mudah.

**2.2 *Machine Learning Pipeline***

Berikut merupakan *machine learning pipeline* yang menjadi solusi untuk permasalahan analisis sentimen *keyword #*pemilu2024 pada Twitter.



**2.2.1**

# BAB 3 IMPLEMENTASI

# 

# **BAB 4 HASIL & PEMBAHASAN**

# **Buat visualisasi**

# Referensi

<https://github.com/stamatelou/twitter_sentiment_analysis>

<https://github.com/Upasna22/Twitter-Sentiment-Analysis-using-Apache-Spark->

<https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api>

<https://towardsdatascience.com/sentiment-analysis-on-streaming-twitter-data-using-spark-structured-streaming-python-fc873684bfe3>

<https://medium.com/analytics-vidhya/exploring-twitter-streaming-data-using-python-and-spark-3f4f189ec660>