**LAPORAN PROYEK PENGOLAHAN DATA BESAR**

**Sentiment Analysis   
Pada Streaming Data Twitter**

****

**Disusun Oleh:**

| **12S18033** | **Cristina Sriwahyuni Hasibuan** |
| --- | --- |
| **12S18038** | **Naomi A. Simatupang** |
| **12S18049** | **Natasya Sitorus** |
| **12S18060** | **Elsaday Bakara** |

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2022**

**DAFTAR GAMBAR**

**DAFTAR TABEL**

1. **PENDAHULUAN**

Pada bab ini membahas mengenai latar belakang, tujuan, dan ruang lingkup dari pengerjaan proyek yang akan dilakukan.

**1.1 Latar belakang**

Analisis sentimen atau opinion mining adalah proses untuk memahami, mengekstrak dan mengolah data

Analisis Sentimen adalah studi komputasional dengan menggunakan opini, sentimen, dan emosi dari objek pengguna (orang) melalui entitas atau atribut yang dimiliki yang diekspresikan dalam bentuk teks. Sentimen analisis yang merupakan pemrosesan bahasa dengan menggunakan pendekatan pembelajaran mesin untuk mendefinisikan apakah suatu penggalan teks dapat dikategorikan pada respon positif atau negatif. Saat ini masyarakat yang sudah lekat dengan teknologi komunikasi sangat mudah dalam menyampaikan tanggapan, baik dalam bentuk ulasan, saran, atau komentar. Tanggapan yang diberikan oleh masyarakat dapat ditemukan dalam berbagai aspek, misalnya pendidikan, ekonomi, hiburan, dan sebagainya.

Dewasa ini produk *skincare* sedang sangat diminati di masyarakat terutama dikalangan remaja menuju dewasa, dapat memberikan tanggapannya langsung terkait kesan yang diterima selama menggunakan produk tersebut. Respon/ulasan masyarakat terhadap sebuah produk menjadi penentu apakah produk tersebut dapat dikategorikan bagus atau tidak. Akibat kemudahan pemberian ulasan oleh pengguna produk dan semakin banyaknya jenis dan brand produk yang beredar di masyarakat, muncullah permasalahan terkait sulitnya menentukan tanggapan pengguna termasuk pada respon positif atau negatif dengan data yang sangat banyak. Untuk mempermudah menentukan review produk Klasifikasi sentimen bertujuan untuk mengatasi masalah ini dengan secara otomatis mengelompokkan review pengguna menjadi opini positif atau negatif. Untuk itu perlunya pengkajian ulang tentang review produk kosmetik dengan cara pengklasifikasian review tersebut kedalam class positif dan negatif agar pada akhirnya konsumen baru dapat mengetahui tanggapan konsumen lain tentang produk tersebut secara cepat dan tepat, berdasarkan *review* yang sudah diberikan oleh pengguna sebelumnya. Setiap review akan diproses sehingga menghasilkan klasifikasi sentimen yang positif dan negatif. Data yang diberikan masih berupa review dari berbagai kalangan pengguna dengan bahasa dan tanda baca yang bebas, oleh karena itu data yang diperoleh terlebih dahulu di *preprocessing* agar mudah untuk diklasifikasikan.

Algoritma Naive Bayes adalah merupakan salah satu algoritma klasifikasi yang banyak digunakan pada *data mining* ataupun *text mining*. Algoritma Naive Bayes berdasarkan teorema Bayes, bahwa semua kegiatan memberikan sebuah kontribusi yang sama penting atau saling bebas pada pemilihan kelas tertentu. Salah satu metode klasifikasi untuk menentukan gambaran persepsi masyarakat (respon, opini, review, dll) di dalam Text Mining adalah metode Naïve Bayes yang sering disebut dengan Naive Bayes Classifier. Penelitian yang dilakukan adalah tentang sentimen analisis untuk mengklasifikasikan Tweet Review Scarlett whitening. Penelitian ini termasuk ke dalam sebuah penelitian yang bernama *fined grained* *sentiment analysis* yaitu analisis pada suatu kalimat komentar. Data dari Twitter tersebut akan dapat diproses menggunakan *text mining*, kemudian dilanjutkan dengan mengklasifikasikan Tweet ke dalam dua kelas, yaitu positive dan negative.

**1.2 Tujuan**

Adapun tujuan dari penelitian ini adalah untuk melakukan analisis sentimen pada *streaming* data twitter mengenai perkembangan covid

**1.3 Manfaat**

Manfaat dari pengerjaan proyek ini yaitu:

1. Bagi Mahasiswa

Proyek ini diharapkan nantinya dapat memberikan pemahaman dan keterampilan bagi tim proyek untuk mengimplementasikan, mengintegrasikan, dan mengimplementasikan teknologi yang digunakan dalam pengolahan data besar.

**1.4 Ruang Lingkup**

Ruang Lingkup dalam pengerjaan proyek ini adalah

1. Menggunakan data dari API twitter.
2. Menggunakan Jupyter Notebook sebagai *worksheet.*
3. Analisis sentimen akan dilakukan dengan klasifikasi review produk dengan menggunakan SVM(Support Vector Machine).

1. **PERMASALAHAN**

Pada bab ini dijelaskan masalah yang akan diselesaikan dalam proyek Pengenalan Big Data. Permasalahan proyek pengenalan Big Data yang akan diselesaikan adalah melakukan klasifikasi analisis sentimen pada perkembangan covid di indonesia. arsitektur sistem akan didesain dapat mengolah data terstruktur dan data tidak terstruktur. hal tersebut berguna untuk mengetahui cara klasifikasi yang dapat digunakan apabila data yang digunakan merupakan data streaming. Berdasarkan kebutuhan diatas maka permasalahan yang akan dirumuskan adalah bagaimana suatu sistem dapat mengklasifikasikan perkembangan covid yang diberikan pengguna sebagai ulasan yang bernilai positif atau ulasan yang negtif menggunakan algoritma SVM.

1. **SOLUSI**

**2.1 Big Data Architecture**

Berikut merupakan arsitektur big data yang menjadi solusi terhadap permasalahan untuk mengklasifikasi setiap review sehingga menghasilkan klasifikasi sentimen yang positif dan negatif.



Gambar 1. Architecture

Arsitektur sistem dibentuk agar dapat mengolah data yang terstruktur dan tidak terstruktur kemudian menggunakan spark untuk melakukan implementasi analysis Machine Learning, berikut tahapan yang akan digunakan:

1. Menyediakan data Hal yang pertama yang harus dipersiapkan adalah data. Data terdiri dari dua jenis yaitu terstruktur dan tidak terstruktur. Data yang akan digunakan yaitu data terstruktur dan data tidak terstruktur. Data yang terstruktur dapat diketahui apabila data tersebut telah tersimpan dalam sebuah format file. Data yang tidak terstruktur atau yang disebut dengan unstructured data merupakan data yang tidak mengikuti suatu susunan format tertentu sebagai contoh data yang berasal dari sebuah media sosial atau website.

2. ETL Process Setelah mengetahui jenis data yang digunakan, kemudian akan dilanjutkan dengan tahapan ETL Process. Extraction, Transfer, Loading atau dikenal dengan ETL adalah sebuah proses sebuah fase pemrosesan data dari sumber data ke dalam satu penyimpanan yang konsisten dan dimuat ke dalam gudang data. Data terstruktur akan diproses dengan menggunakan Apache Sqoop, sedangkan semi struktur, terstruktur dan tidak terstruktur seperti data streaming akan diproses dengan menggunakan Apache Flume.

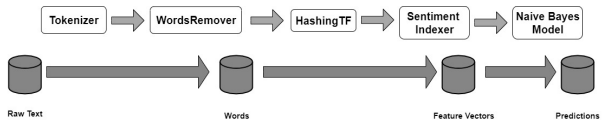
3. Proses penyimpanan (storage) pada database Setelah data diproses kemudian data akan disimpan dalam HBase. HBase adalah database terdistribusi yang berorientasi pada kolom. HBase adalah program yang berjalan diatas Hadoop Distributed File System (HDFS) yang mampu memproses data dalam skala besar secara interaktif. HBase baik digunakan karena memiliki sifat fault tolerant, artinya HBase mampu menangani keutuhan data meskipun terjadi sebuah kegagalan pada sistem yang digunakan, dengan cara mengolah data kembali berdasarkan historical pengolahan data tersebut.

4. Implementasi analisis Machine Learning Pada tahap ini, data yang disimpan akan dilakukan pemrosesan data dimana berguna untuk menghilangkan data yang noise, kemudian data yang telah dilakukan preprocessing, akan dipakai kembali untuk melakukan klasifikasi analisis sentimen pada review Scarlett Whitening.

5. Memvisualisasikan data Pada tahapan ini, hasil klasifikasi sentimen analisis pada review Scarlett Whitening akan diolah kembali untuk divisualisasikan berdasarkan hasil klasifikasi sentimen analisis untuk mempermudah dalam memahami hasil pengolahan data.

1. **MACHINE LEARNING PIPELINE**

Pada bab ini akan dijelaskan Machine Learning Pipeline yang akan digunakan dalam sentiment analysis - Tweet Covid19.



Berikut tahapan Machine Learning Pipeline yang akan digunakan sebagai berikut:

**1. Tokenizer**

Tokenizer merupakan proses pemisahan teks menjadi kata, frasa, simbol atau elemen bermakna lainnya yang disebut dengan token. Tujuannya adalah mengeksplorasi kata kata dalam sebuah kalimat. Daftar token menjadi masukan untuk diproses lebih lanjut. Tokenisasi adalah proses untuk memotong dokumen menjadi pecahan kecil yang dapat berupa bab, sub-bab, paragraf, kalimat, dan kata (token). Pada proses ini akan menghilangkan whitespace .

**2. Stop Words Removal**

Banyak kata dalam dokumen yang sangat sering muncul namun pada dasarnya tidak memiliki arti karena digunakan untuk menggabungkan kata dalam kalimat. Stop Word Removal merupakan penghapusan kata-kata pada dokumen yang tidak memiliki arti tersebut .

**3. Hashing TF Tahap**

HashingTF merupakan proses untuk melakukan transformer yang nantinya akan mengambil kata dan mengubah kata tersebut menjadi suatu vektor dengan panjang yang tepat.

**4. String Indexer**

String Indexer merupakan sebuah proses yang dilakukan untuk meng-encode kolom dengan label string menjadi label indeks. Indeks dibuat dalam bentuk [0, nomor label] 7 yang diurutkan berdasarkan frekuensi label, dimana indeks 0 menunjukkan bahwa label tersebut sering muncul.

**5. Naive Bayes**

Model Naive Bayes model dibangun dengan menggunakan algoritma Naive Bayes. Naive Bayes merupakan algoritma yang digunakan untuk klasifikasi multi kelas. Naive Bayes menerapkan fungsi statistik sederhana berdasarkan teorema bayes dengan asumsi keberadaan dari suatu fitur tertentu terhadap suatu kelas yang tidak berhubungan dengan fitur lainnya .

Naive Bayes classifier menggunakan prior probability (yaitu nilai probabilitas yang diyakini benar sebelum melakukan eksperimen) dari setiap label yang merupakan frekuensi masing-masing label pada training set dan kontribusi dari masing-masing fitur. Klasifikasi Naive Bayes dapat dilatih dengan sangat efisien dalam bentuk supervised learning. Spark.mllib mendukung Naive Bayes Multinomial dan Naive Bayes Bernoulli. Model Naive Bayes ini digunakan untuk mengklasifikasikan dokumen.