PROPOSAL PROYEK

11S4037 – Pemrosesan Bahasa Alami

Sentiment Analysis of Covid 19 Tweets

PENGUSUL

12S17015 – Dodi Sanjaya Butarbutar

12S17020 – Jovan Pioma Pakpahan

12S17048 – Rizky Marganda Sitohang

12S17057 – Chatrin

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO

INSTITUT TEKNOLOGI DEL

NOVEMBER 2020

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB PENDAHULUAN

Pada bab ini dijelaskan mengenai latar belakang, tujuan, manfaat dan ruang lingkup disertai penjabaran.

1.1 Latar Belakang

Sentiment Analysis (SA) adalah analisis tentang perbandingan pendapat, sikap, dan emosi seseorang terhadap suatu entitas. Entitas dapat mewakili individu, peristiwa, atau topik. *Sentiment analysis* sering juga disebut dengan *opinion mining* kedua hal tersebut memiliki ekspresi yang dapat ditukarkan, yang dimana baik SA maupun OM mengekspresikan makna timbal balik. Namun, beberapa peneliti mengatakan bahwa SA dan OM memiliki pengertian yang sedikit berbeda. Opinion Mining (OM) mengekstrak dan menganalisis pendapat orang-orang tentang entitas sementara *sentiment analysis* mengidentifikasi sentimen yang diekspresikan dalam teks nya kemudian dianalisis. Oleh karena itu target SA adalah menemukan opini lalu mengidentifikasi sentiment yang diungkapkan dan kemudian mengklasifikasikan polaritasnya. [1]

Sentiment analysis memiliki 3 tingkat dalam melakukan klasifikasi yaitu, *document level*, *sentence level*, dan, *aspect level*. *Document level* bertujuan untuk mengklasifikasikan opini pada sebuah dokumen sebagai pernyataan sentiment positif atau negatif. Hal ini mengasumsikan bahwa seluruh dokumen adalah sebagai unit untuk informasi dasar yang membicarakan satu topik. *Sentence level* bertujuan untuk mengklasifikasikan sentimen yang diekspresikan dalam suatu kalimat. Langkah pertama adalah mengidentifikasi apakah kalimat tersebut subjektif atau objektif. Jika kalimat tersebut subjektif maka *sentence level* pada SA akan menentukan apakah kalimat tersebut menyatakan pendapat positif atau negatif. [1]

Dataset merupakan hal yang penting dalam pemenjalankan proyek ini, dimana dataset yang digunakan adalah opini yang ada di dalam twitter ataupun kometar yang terdapat di dalam twitter terkait *covid 19*. Dengan tinjauan ini maka bisa di dapatkan hasil dari pengklasifikasian komentar di twitter mengenai *covid 19* negatif, positif, atau netral.

1.2 Tujuan

Tujuan proyek ini adalah:

1. Memberikan insight mengenai *covid 19* berdasarkan postingan yang ada pada media sosial twitter
2. Membuat model yang dapat melakukan analisis sentimen terhadap pandangan pengguna sosial media twitter terkait *covid 19*

1.3 Manfaat

Manfaat proyek ini adalah:

1. Memberikan kesimpulan terkait respon pengguna media sosial twitter terhadap *covid 19*
2. Memberikan model analisis sentimen untuk menghasilkan insight dengan memanfaatkan pemrosesan bahasa alami

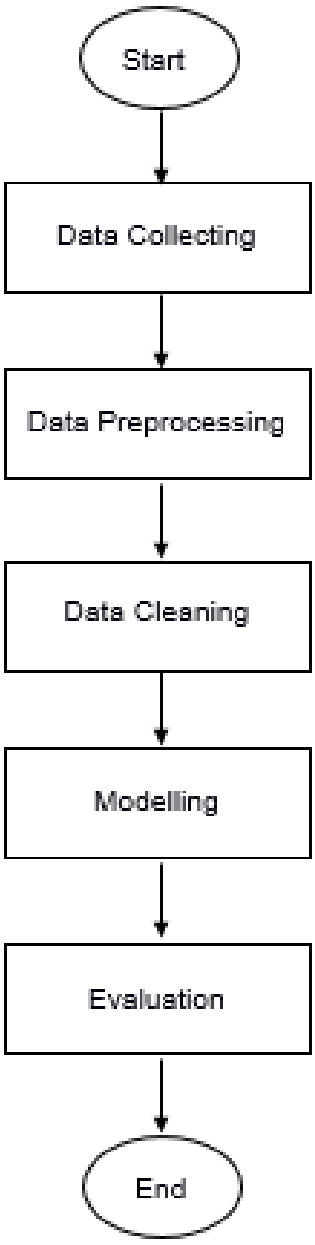
1.4 Ruang Lingkup

Ruang lingkup dari proyek ini adalah:

1. Membangun sistem yang digunakan untuk menganalisis sentimen dari sebuah ungkapan komentar pada twitter terkait *covid* 19.
2. Menggunakan dataset yang berisi kumpulan komentar yang ada pada twitter.

BAB ISI

Pada bab ini dijelaskan tahapan pemrosesan bahasa alami yang akan diterapkan dalam bentuk diagram alir disertai penjabaran. Berikut merupakan diagram alir dari tahapan pengerjaan proyek twitter sentiment analysis.



Gambar 1. ***Flow Chart Project***

2.1 Data Collecting

Tahapan pertama yang dilakukan untuk melakukan pemrosesan bahasa alami pada proyek ini adalah pengumpulan data. Adapun data yang kami gunakan untuk proyek ini adalah dataset yang diperoleh dari kaggle.com.

Data tweet merupakan data ungkapan pada twitter terkait covid 19. Tweets telah ditarik dari Twitter dan manual tagging telah dilakukan kemudian. Names dan usernames telah diberi kode untuk menghindari masalah privasi.

Berikut merupakan kolom yang terdapat pada dataset:

1. Username
2. Screen name
3. Location
4. Tweet at
5. Original tweet
6. Label

2.2 Data Preprocessing

Seperti preprocessing lain dalam analisis teks, kami tidak diharuskan untuk menghapus emoji, *slang* (singkatan), emotikon, *punctuations* (tanda baca), dan sebagainya. Hal ini dikarenakan Vader dapat menghasilkan skor berdasarkan ini. Beberapa bagian yang penting adalah Huruf Besar (Kapitalisasi), *punctuation*, *conjunctions*, *degree modifiers, use of emoticons, use of emojis, use of slang,use of negation*, dan *slang word as modifiers*. Pengusul proyek belum menghapus *stopwords* dari teks, karena kata kunci seperti "*not*" memiliki arti penting dalam analisis sentimental.

Untuk melakukan *data preprocessing* ada beberapa tahapan yang dilakukan yaitu:

1. Importing data & libraries

   Dimulai dengan mengimpor data dan semua libraries yang diperlukan.

2. Tokenization

   Data dalam bentuk text, contohnya tweet yang masuk/di-input akan dipisah-pisah terlebih dahulu. Proses tersebut disebut juga sebagai tokenisasi. Tokenisasi dilakukan untuk mempermudah proses analisa dari sebuah kalimat text. Tokenizer digunakan karena tidak dapat menganalisis seluruh kalimat, untuk itu digunakan regex untuk menandai kalimat menjadi daftar kata.

   Ketika kita membagi deskripsi menjadi kata-kata individu, kita harus membuat kosakata dan tambahan kita dapat menambahkan fitur baru yaitu panjang deskripsi.

   Setelah itu, dilakukan tokenisasi dengan 1- gram tokenizer dan 2-gram tokenizer.

2.3 Data Cleaning

Data pada umumnya sangat kotor sehingga perlu diolah terlebih dahulu sebelum dianalisis. Data cleaning ada suatu proses untuk mengelolah data agar data tidak lagi memiliki incomplete

data. Incomplete data ini seperti adanya data yang duplikat, data yang hilang dan juga data yang valuenya tidak diberikan sehingga menimbulkan misunderstanding. Pada project ini penulis menggunakan data Corona_tweets dimana pada data ini tidak perlu dilakukan *data cleaning* karena data ini sudah cukup lengkap untuk kemudian dianalisis.

2.4 Modelling

Sebelum masuk ke dalam proses modelling umumnya akan ditentukan terlebih dahulu prosedur estimasi kinerja. Hal yang perlu diperhatikan dalam tahapan modelling adalah mengetahui cara kerja (algortima) setiap model. Ketidaktahuan terhadap algortima metode yang digunakan akan berdampak besar jika menghadapi kasus ketika hasil tidak berjalan dengan baik. Parameter, optimasi, kemampuan generalisasi dan regularisasi perlu menjadi perhatian dalam proses modelling. Pada *project* ini penulis menggunakan algoritma K-Means. Algoritma K-Means adalah salah satu "*unsupervised machine learning algorithms*" yang paling sederhana dan populer. Tujuan dari algoritma ini adalah untuk menemukan grup dalam data, dengan jumlah grup yang diwakili oleh variabel K. Variabel K sendiri adalah jumlah *cluster* yang kita inginkan.

2.5 Evaluation

Metrik untuk mengevaluasi hasil trainning model, dapat digunakan metrik *accuracy*, *precission*, *recall*, *F1 Score*, dan lain-lain.

BAB RENCANA KERJA

Pada bab ini akan dijelaskan jadwal kegiatan dalam bentuk *Gantt Chart* dan pembagian tugas.

3.1 Jadwal Kegiatan

Berikut ini adalah informasi jadwal kegiatan.

Tabel 1. Jadwal penelitian

No.	Kegiatan	Minggu																										
		Minggu 1							Minggu 2							Minggu 3							Minggu 4						
		1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1.	*Data collecting*																												
2.	*Data preprocessing*																												
3.	*Data cleaning*																												
4.	*Modelling*																												
5.	Evaluation																												

3.2 Pembagian Tugas

Pembagian tugas akan dilampirkan pada laporan akhir.

DAFTAR PUSTAKA

[1] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.