KLASIFIKASI TWEET MENGENAI COVID-19

Disusun untuk memenuhi tugas besar
pada Mata Kuliah Pengolahan Bahasa Alami Semester Enam
yang diampu oleh Dr. Retno Kusumaningrum, S.Si, M.Kom.

Disusun oleh:

1.	**Aufarizq M Niza Bayzoni**	**24060118130075**
2.	**Novian Fifin Ristianto**	**24060118120054**
3.	**Linggar Maretva Cendani**	**24060117120031**
4.	**Fetty Krisnaeni**	**24060117120032**
5.	**Muhammad Azzam Hanif**	**24060117140096**

PROGRAM STUDI STRATA 1 INFORMATIKA
DEPARTEMEN ILMU KOMPUTER/INFORMATIKA
FAKULTAS SAINS DAN MATEMATIKA
UNIVERSITAS DIPONEGORO
SEMARANG
2021

ABSTRAK

Covid-19 merupakan wabah virus yang saat ini sedang melanda di Indonesia bahkan seluruh penjuru dunia. Gejala yang ditimbulkan oleh virus ini diantaranya yaitu demam, batuk, kelelahan, dan pada kasus yang serius dapat menimbulkan hilangnya indera perasa dan penciuman, sesak nafas, hingga hilangnya kemampuan berbicara atau bergerak. Dengan adanya pandemi Covid-19 ini sangat mempengaruhi kehidupan masyarakat terutama bidang ekonomi dan pendidikan. Berbagai berita dan perbincangan mengenai Covid-19 juga semakin ramai setiap harinya di beberapa situs berita maupun media sosial, salah satunya yaitu Twitter. Di Twitter sudah tersedia berbagai macam informasi dan bahasan terbaru mengenai Covid-19 yang sudah dituliskan oleh penggunanya melalui pesan *tweet* atau cuitan yang telah dibuat. Pada penelitian ini akan dibuat sebuah sistem yang dapat menggolongkan atau mengklasifikasi perbincangan tentang Covid-19 yang terdapat di Twitter menggunakan model Naive Bayes Multinomial.

Kata kunci: Covid-19, Klasifikasi Tweet, Naive Bayes Multinomial.

KATA PENGANTAR

Puji dan syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa, karena berkat rahmat dan karunia-Nya kami dapat menyusun dan menyelesaikan laporan Klasifikasi Tweet Mengenai Covid-19 dengan baik dan tepat pada waktunya.

Laporan ini dibuat untuk memenuhi tugas besar mata kuliah Pengolahan Bahasa Alami untuk mahasiswa S1 Informatika Universitas Diponegoro. Laporan ini dibuat dengan bantuan, bimbingan, dan arahan dari Ibu Dr. Retno Kusumaningrum, S.Si, M.Kom. selaku dosen pengampu mata kuliah Pengolahan Bahasa Alami, teman-teman mahasiswa yang mengambil mata kuliah Pengolahan Bahasa Alami, serta pihak-pihak lain yang terlibat baik secara langsung maupun tidak langsung dalam pembuatan laporan ini. Oleh karenanya, kami mengucapkan terima kasih kepada seluruh pihak yang terlibat.

Pembuatan laporan ini tidak luput dari kesalahan dan kekurangan, oleh karena itu kami memohon maaf dan mengundang para pembaca untuk memberikan kritik dan saran yang membangun. Semoga laporan ini dapat bermanfaat bagi penulis pada khususnya dan pembaca pada umumnya.

Semarang, 20 Juni 2021

Tim Penulis

DAFTAR ISI

BAB I
PENDAHULUAN

A. Latar Belakang

Covid-19 adalah sebuah wabah virus yang saat ini sedang melanda di seluruh penjuru dunia. Virus ini mengakibatkan penderitanya menderita demam, batuk, kelelahan, dan pada kasus yang serius dapat menimbulkan hilangnya indera perasa dan penciuman, sesak nafas, hingga hilangnya kemampuan berbicara atau bergerak. Penyebaran virus Covid-19 sangat cepat karena melalui *droplet* yang berasal dari orang yang terjangkit, sehingga harus dilakukan penanganan serius seperti pemberlakuan penguncian antar daerah dan karantina orang yang terjangkit.

Pandemi Covid-19 sudah sangat mempengaruhi kehidupan bermasyarakat. Perbincangan tentang permasalahan ini pun sangat banyak baik mengenai sosialisasi protokol kesehatan, sosialisasi gejala terjangkit Covid-19, laporan harian kasus terjangkit Covid-19, dan pengembangan vaksin Covid-19, hingga ke permasalahan seperti dampak pandemi terhadap kehidupan ekonomi dan sosial masyarakat. Perbincangan tersebut dapat dijumpai di banyak tempat, salah satunya adalah media sosial seperti Twitter.

Twitter merupakan sebuah media sosial dimana penggunanya dapat menuliskan pesan di lamannya yang dapat dibaca oleh publik yang dinamakan *tweet* atau cuitan. Twitter merupakan sebuah media sosial yang tergolong besar, sehingga perbincangan yang terjadi disana pun juga sangat banyak. Oleh karenanya, sebuah sistem untuk melakukan klasifikasi terkait perbincangan yang terjadi diperlukan untuk dapat mempelajari tentang hal apa yang sedang ramai diperbincangkan terkait dengan masalah pandemi Covid-19.

B. Perumusan Masalah

Berdasarkan latar belakang yang sudah dipaparkan, maka rumusan masalah dalam penelitian ini adalah bagaimana membentuk sebuah sistem yang dapat menggolongkan atau mengklasifikasi perbincangan tentang Covid-19 yang terdapat di Twitter.

BAB II
DASAR TEORI

a. TFIDF Classifier

TFIDF adalah sebuah pengklasifikasi yang didasarkan pada algoritma umpan balik relevansi Rocchio dan menggunakan berat kata TFIDF (TFIDF *word weights*). Dasar dari algoritma ini adalah merepresentasikan setiap dokumen *d* menjadi sebuah vektor $\vec{d} = (d^{(1)},..., d^{(|F|)})$ dalam ruang vektor sehingga dokumen yang memiliki kemiripan konten akan memiliki vektor yang mirip. Setiap dimensi ruang vektor merepresentasikan sebuah kata yang dipilih berdasarkan proses tersebut. Nilai dari elemen vektor$d^{(i)}$ untuk sebuah dokumen *d* akan dikalkulasikan sebagai kombinasi statistik TF(*w, d*), yaitu jumlah kemunculan kata *w* dalam dokumen *d*, dan DF(*w*), yaitu jumlah dokumen yang terdapat kemunculan kata *w* di dalamnya sebanyak minimal satu kali. Algoritma TFIDF kemudian akan mempelajari sebuah kelas model dengan mengkombinasikan kumpulan vektor yang telah dibentuk menjadi sebuah vektor prototype. (Joachims, 1996).

b. Multinomial Naive Bayes Classifier

Naïve Bayes merupakan salah metode pembelajaran mesin probabilistik. Seperti namanya, metode ini mengasumsikan bahwa setiap atribut dari data tidak bergantung satu sama lain. Pada dasarnya, asumsi bahwa setiap kata tidak bergantung satu dengan yang lain pada metode Naïve Bayes ini berlawanan dengan keadaan sebenarnya. Hal ini dikarenakan suatu dokumen atau teks perlu memiliki kata yang saling berhubungan agar dokumen tersebut memiliki makna. Akan tetapi, metode ini terbukti mampu memberikan hasil yang cukup memuaskan apabila diterapkan di bidang klasifikasi teks (Jo, 2019).

Salah satu model dari Naïve Bayes yang sering digunakan dalam klasifikasi teks adalah multinomial Naïve Bayes (Manning, 2008). Multinomial Naïve Bayes merupakan metode supervised learning, sehingga setiap data perlu diberikan label sebelum dilakukan training.

BAB III
METODOLOGI

Metodologi yang dilakukan pada tugas ini dapat dijabarkan menjadi point sebagai berikut:

1. Dataset

   Dataset yang digunakan pada tugas ini adalah data cuitan pengguna Twitter yang diambil berdasarkan kata kunci vaksin pencegahan atau pengobatan, dan perkembangan Covid-19 yang dijadikan sebagai fitur kelas pada dataset.

2. Data *Preprocessing*

   Data cuitan yang didapat akan di-*preprocessing*. Proses *preprocessing* yang dilakukan adalah pembuangan nama pengguna, tagar, dan tautan pada cuitan, melatih pembuat vektor TF-IDF, membentuk pembuat kode label, dan membagi dataset menjadi data latih dan data uji.

3. Naive Bayes

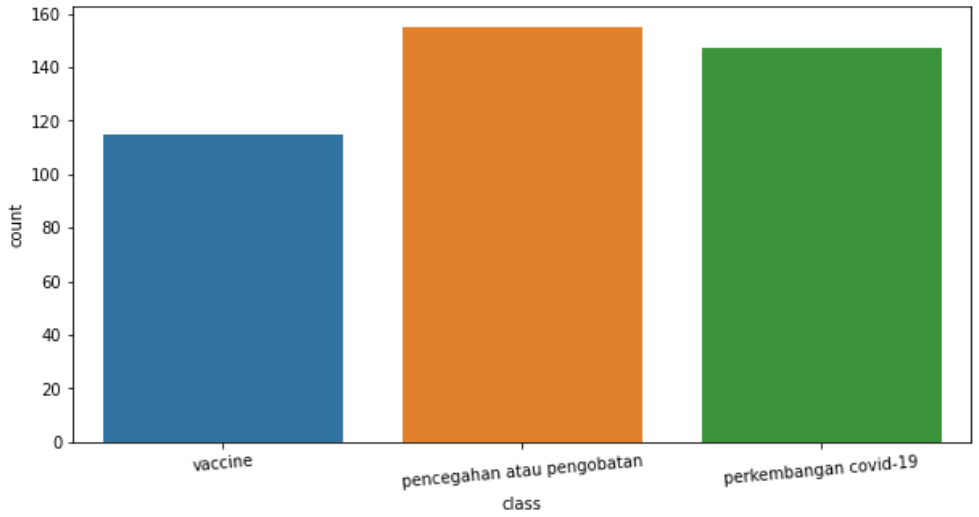
   Data yang telah di-*preprocessing* akan dikonversikan menjadi vektor kemudian model Naive Bayes Multinomial dilatih dengan menggunakan vektor data latih. Model yang telah dilatih diuji akurasinya dengan menggunakan vektor data uji kemudian disimpan.

4. Penilaian Kinerja

   Hasil klasifikasi dari model yang telah dilatih kemudian digunakan untuk membuat sebuah Confusion Matrix dengan sumbu-x sebagai prediksi kelas dan sumbu-y sebagai kelas sesungguhnya. Akurasi model dihitung berdasarkan hasil prediksi kelas dari data uji yang sesuai dengan kelas sesungguhnya.

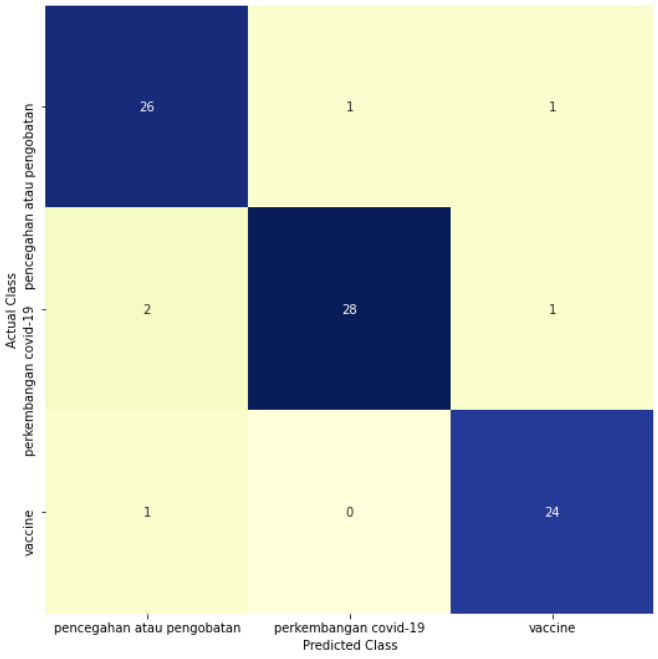
BAB IV
HASIL DAN PEMBAHASAN

Eksperimen dilakukan dengan menggunakan dataset cuitan terkait dengan permasalahan vaksin, pencegahan atau pengobatan, dan perkembangan Covid-19 dengan sebaran kelas seperti yang dapa dilihat pada gambar 1.



Gambar 1. Sebaran kelas dataset yang digunakan

Dataset tersebut kemudian digunakan untuk melatih dan menguji sebuah model Naive Bayes Multinomial yang kemudian memberikan prediksi yang perbandingan antara kelas prediksi dan kelas sesungguhnya dapat dilihat dari Confusion Matrix yang dapat dilihat pada gambar 2.



Gambar 2 Visualisasi *Cluster* Dua Dimensi

Model Naive Bayes Multinomial yang sudah dilatih mempunyai nilai akurasi sebesar 93% dengan detail laporan klasifikasi yang spesifik untuk tiap kelas pada data uji dapat dilihat pada tabel 1.

Kelas	precision	recall	f1-score	support
Pencegahan atau pengobatan	0.90	0.93	0.91	28
Perkembangan covid-19	0.97	0.90	0.93	31
vaccine	0.92	0.96	0.94	25

Tabel 1. Laporan klasifikasi data uji

BAB V
KESIMPULAN

Klasifikasi adalah pengelompokan suatu data berdasarkan kemiripan atribut-atributnya. Klasifikasi data cuitan yang pada tugas besar ini dilakukan berdasarkan kata kunci atau jenis permasalahan yang diperbincangkan di dalam cuitan pengguna Twitter dengan menggunakan model Naive Bayes Multinomial. Berdasarkan hasil uji yang dilakukan, dapat terlihat bahwa model yang dibentuk memiliki akurasi yang cukup tinggi yaitu dengan nilai 93%, sehingga dapat disimpulkan bahwa model Naive Bayes Multinomial merupakan salah satu jenis model yang cocok digunakan untuk melakukan klasifikasi kelas untuk cuitan tentang Covid-19 pada media sosial Twitter.

DAFTAR PUSTAKA

Joachims, Thorsten. 1996. "*A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*". Carnegie Mellon University: Pittaburgh.

T. Jo. 2019. "*Text Mining : Concepts, Implementation, and Big Data Challenge vol. 45*". Springer International Publishing AG : Cham.

C. D and P. R. H. S. Manning. 2008 "*Introduction to Information Retrieval*". Cambridge University Press : New York.

LAMPIRAN

Soure Code:
https://colab.research.google.com/drive/1HvRndXHG_PuDhIpq4kO_Kked2mLfUQjq?usp=sharing