SENTIMEN ANALISIS PILPRES 2024 PADA MEDIA SOSIAL TWITTER MENGGUNAKAN *NAÏVE BAYES CLASSIFIER*

SKRIPSI

Karya tulis sebagai salah satu syarat
untuk memperoleh gelar Tingkat Sarjana

Oleh

MUCHAMMAD FAHD ISHAMUDDIN
NPM : 411550050180048

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS LANGLANGBUANA
2023

LEMBAR PENGESAHAN

SENTIMEN ANALISIS PILPRES 2024 PADA MEDIA SOSIAL TWITTER MENGGUNAKAN NAÏVE BAYES CLASSIFIER

Oleh

MUCHAMMAD FAHD ISHAMUDDIN
NPM : 411550050180048

untuk memperoleh gelar Tingkat Sarjana dari
Program Studi Teknik Informatika
Universitas Langlangbuana

Menyetujui

Pembimbing 1	Pembimbing 2
(Prof. Dr. Pembimbing 1, M.T.)	(Dr. Pembimbing 2, S.T., M.T.)
NIDN: <NIDN>	NIDN: <NIDN>

Mengetahui

Dekan Fakultas Teknik	Ketua Program Studi Teknik Informatika
(Dr. Dekan Fakultas Teknik, M.T.)	(Dr. Kaprodi TIF, M.Si)
NIDN: <NIDN>	NIDN: <NIDN>

LEMBAR PERNYATAAN KEASLIAN TUGAS AKHIR

Saya yang bertanda tangan dibawah ini :

Nama	:	MUCHAMMAD FAHD ISHAMUDDIN
NPM	:	411550050180048
Judul Tugas Akhir	:	SENTIMEN ANALISIS PILPRES 2024 PADA MEDIA SOSIAL TWITTER MENGGUNAKAN NAÏVE BAYES CLASSIFIER

Menyatakan dengan sebenarnya bahwa penulisan Tugas Akhir ini berdasarkan hasil penelitian, pemikiran dan pemaparan asli dari saya sendiri, baik untuk naskah laporan maupun kegiatan pembangunan aplikasi yang tercantum sebagai bagian dari Tugas Akhir ini. Jika terdapat karya orang lain, saya akan mencantumkan sumber yang jelas.

Demikian pernyataan ini saya buat dengan sesungguhnya dan apabila dikemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh karena karya tulis ini dan sanksi lain sesuai dengan peraturan yang berlaku di Universitas Langlangbuana. Demikian pernyataan ini saya buat dalam keadaan sadar tanpa paksaan dari pihak manapun.

Bandung, 14 April 2023
Yang membuat pernyataan,

MUCHAMMAD FAHD ISHAMUDDIN
NPM : 411550050180048

*Dipersembahkan kepada kedua orang tuaku dan guru-guruku, *
semoga Allah selalu memberkati mereka.

KATA PENGANTAR

Penulis sangat berterima kasih pada Prof. Dr. Marsilam Hutabarat, M.T. Dam Dr. Jaya Sumarna, S.T., M.T. sebagai Pembimbing, atas segala saran, bimbingan dan nasehatnya selama penelitian berlangsung dan selama penulisan Skripsi ini.

Terima kasih disampaikan kepada Kementerian Riset Teknologi dan Pendidikan Tinggi atas bantuan Beasiswa Pendidikan Pascasarjana (BPPs) yang diterima selama pendidikan program Tingkat Sarjana ini.

(dan seterusnya)

ABSTRAK

Pemilihan Umum (PEMILU) merupakan pesta demokrasi rakyat yang diselenggarakan 5 tahun sekali, pada pesta rakyat tersebut salah satunya ada Pemilihan Presiden. PEMILU 2024 sangatlah istimewa yakni terjadi pada era modern dan banyaknya pemilih pemula yang menjadi peserta pada pesta demokrasi tersebut. Data pada tahun 2022 menyatakan bahwa 191 juta jiwa masyarakat Indonesia sudah aktif dalam ber media sosial, seperti yang kita ketahui bahwa *spotlight* seluruh media sekarang berpindah menjadi media digital. media digital pada saat ini menjadi tempat masif-nya para juru kampanye untuk mendongkrak elektabilitas calon yang diusung oleh partainya, mulai dari Prabowo Subianto, Anies Baswedan, Ridwan Kamil dan Ganjar Pranowo. Hal tersebut memantik penulis untuk menganalisa nilai sentiment pada tweet yang terkumpul sejak awal hingga akhir 2022, penulis menganalisa menggunakan model machine learning naive bayes classifier dan melakukan penulisan pada python notebook, setelah melakukan modeling penulis mengevaluasi denan monte-carlo cross validation dan pengaplikasian web sederhana dengan library streamlit.

Kata Kunci: *Machine Learning, Naïve bayes classifier, Pilpres 2024, Prabowo, Ganjar Pranowo, Anies Baswedan, streamlit, monte-carlo cross validation, pyhton notebook.*

ABSTRACT

The General Election (PEMILU) is a people's democratic party that is held every 5 years, one of which is the Presidential Election. The 2024 ELECTION is very special, because it took place in the modern era and many first-time voters participated in this democratic party. Data for 2022 states that 191 million Indonesian people are already active on social media, as we know that the focus of all media has now turned to digital media. Digital media is currently a massive place for campaigners to boost the electability of candidates promoted by their parties, starting from Prabowo Subianto, Anies Baswedan, Ridwan Kamil and Ganjar Pranowo. This sparked the author to analyze the sentiment value in the tweets collected from the beginning to the end of 2022, the authors analyzed using the machine learning naive Bayes classifier model and wrote on a python notebook, after modeling the authors evaluated it with monte-carlo cross validation and simple web applications with streamlit library.

Keywords: *Machine Learning, Naïve bayes classifier, Pilpres 2024, Prabowo, Ganjar Pranowo, Anies Baswedan, streamlit, monte-carlo cross validation, pyhton notebook.*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR SINGKATAN DAN ISTILAH

Singkatan	Kepanjangan
CRISP-DM	*Cross Industry Standard Process for Data Mining*
ML	*Machine Learning*
MCCV	*Monte Carlo Cross validation*
PEMILU	Pemilihan Umum
PILPRES	Pemilihan Presiden

DAFTAR SIMBOL

DAFTAR LAMPIRAN

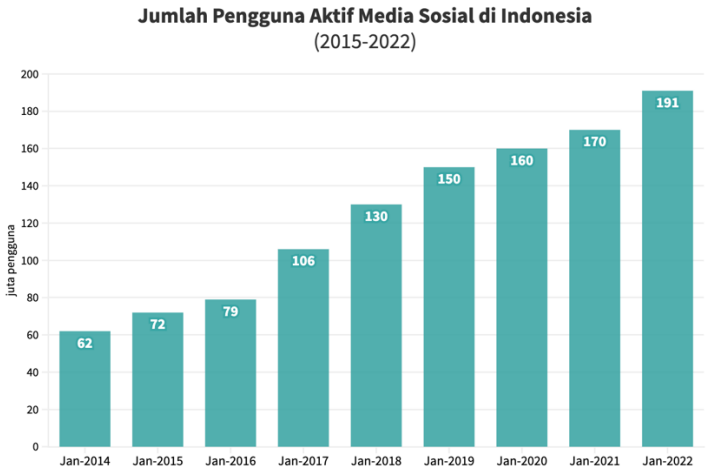
BAB I PENDAHULUAN

I.1 Latar Belakang

Indonesia merupakan negara yang memiliki bentuk pemerintahan presidensial dan demokrasi. Pemerintahan presidensial berarti kepemimpinan pada negara tersebut dipimpin oleh seorang presiden, Demokrasi berarti kekuasaan tertinggi ada di tangan rakyat sehingga yang dapat memilih siapa pemimpin pada negara tersebut. Seperti yang kita ketahui jika pemilihan presiden diadakan dengan Pemilu (Pemilihan Umum) yang bertujuan untuk menentukan eksekutif dan legislatif serta diselenggerakan oleh KPU (Komisi Pemilihan Umum).

Pemilu pertama kali dilaksanakan pada 29 September 1955 untuk memilih anggota DPR dilanjutkan pada 15 Desember 1955 untuk memilih anggota Dewan Konstituante, pada saat ini Indonesia masih dipimpin oleh Ir. Soekarno. Pada tahun 1967 ada SUPERSEMAR (Surat Perjanjian Sebelas Maret) yang menyatakan penyerahan kepemimpinan dari Ir. Soekarno kepada Soeharto. Pemilu pada tahun 1971, 1977, 1997 ketiga pemilu tersebut hanya digunakan untuk memilih DPR hingga akhirnya pada tahun 1999, setelah masa kepemimpinan Ir. B. J. Habibie, FREng. Diadakan pemilihan presiden (pilpres) tetapi melalui sidang paripurna MPR,dengan mencatatkan Abdurrahman Wahid (gusdur) menjadi presiden ke-4 Indonesia dan didampingi Megawati Soekarnoputri sebagai wakil presiden, hingga pada tahun 2004 merupakan pertama kalinya pemilihan presiden dilakukan secara luberjurdil (langsung, umum, bebas, rahasia, jujur dan adil) dengan memenangkan Susilo Bambang Yudhoyono sebagai presiden dan Jusuf Kalla sebagai wakil presiden. Terakhir dilaksanakn pada 2019 yang dimenangkan oleh petahan yakni Ir. H. Joko Widodo.

Hingga sekarang pemilihan presiden masih dilakukan untuk menentukan pemimpin negara Indonesia, tetapi ada yang berbeda antara zaman terdahulu dengan zaman sekarang, pada zaman sekarang kita bisa mengetahui reaksi masyarakat terhadap pemilu dan juga calon presiden yang diusung, apalagi melalui media sosial yang merupakan komponen primer manusia era modern.



Gambar 1. 1 Grafik pengguna media sosial di Indonesia (dataindonesia.id, 2022)

Pengguna media sosial di Indonesia per tahun 2022 dapat dilihat pada gambar 1.1 di atas dan menunjukkan sudah mencapai 191 juta jiwa dan total masyarakat Indonesia ialah 275 juta jiwa, berarti sudah mencapai 69% jiwa di Indonesia menggunakan sosial media hal ini bisa menjadi alasan media sosial adalah media yang bisa menjadi pertimbangan dalam melihat elektabilitas maupun perspektif masyarakat terhadap Pemilu, hal tersebut merupakan tugas dari juru kampanye atau tim pemenangan dari suatu partai atau calon yang ingin manyalonkan dirinya di Pemilu 2024 untuk menaikkan elektabilitas partai atau calon yang ia dukung. Media sosial yang massif dan menjadi persebaran opini dari masyarakat Indonesia ialah twitter, karena banyaknya pendapat yang disampaikan dari warganet Indonesia akan menghasilkan berbagai macam reaksi, maka dapat dilakukan sentiment analisis untuk mendapatkan nada emosional tweet warganet Indonesia, algoritma yang banyak digunakan untuk mendapatkan sentiment ialah naïve bayes, dengan menghitung probabilitas dengan dasar bayes theorem.

Setelah mendalami permasalahan tersebut, maka penulis tertarik untuk menganalisa data dari twitter tentang Pilpres 2024 dan menuangkannya pada penelitian Skripsi yang berjudul ”**SENTIMEN ANALISIS PILPRES 2024 PADA MEDIA SOSIAL TWITTER MENGGUNAKAN NAÏVE BAYES CLASSIFIER**”

I.2 Rumusan Masalah

Berdasarkan dari latar belakang yang sudah ditulis, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana kita mengetahui akurasi dari pengolahan data menggunakan algoritma naive bayes classifier, sehingga juru kampanye dapat memanfaatkan hasil penelitian ini.
2. Bagaimana juru kampanye dapat melakukan uji sentimen pada data twitter dengan mudah.

I.3 Batasan Masalah

Adapun Batasan masalah yang ada pada penelitian ini adalah sebagai berikut:

1. Data yang digunakan lebih dari 100000 (seratus ribu) data
2. Tweet yang didapat ialah tweet sejak januari 2022 hingga desember 2022
3. Output ialah data sentiment terhadap pilpres 2024 dan web sederhana pengujian sentimen

I.4 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, maka yang menjadi tujuan penelitian yang akan dilakukan antara lain:

1. Menerapkan metode naïve bayes classifier guna melakukan sentiment analisis dari social media twitter agar menjadi pertimbangan kontestan politik.
2. Mengetahui akurasi terbaik dan melakukan sentiment analysis tentang pilpres 2024 guna mengetahui bentuk opini positif, negative dan netral.

I.5 Keluaran Penelitian

Dalam penelitian ini luaran yang akan dihasilkan sebanyak 2(dua) luaran, yakni:

1. Laporan Penelitian, Python Notebook dan visualisasi data
2. Web sederhana penerapan model sentiment analisis

I.6 Sistematika Penulisan

Untuk memberikan gambaran mengenai penelitian ini, maka disusun sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN : membahas secara singkat mengenai latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, keluaran penelitian dan sistematika penulisan.

BAB II LANDASAN TEORI : membahas tentang artikel-artikel jurnal dari karya para peneliti sebelumnya yang berguna dalam proses pembuatan sistem.

BAB III METODOLOGI PENELITIAN : menjelaskan bagaimana urutan langkah penyelesaian masalah berdasarkan rumusan masalah, serta penjelasan metode penelitian yang akan digunakan.

BAB IV HASIL DAN PEMBAHASAN : tentang hasil penelitian dan hasil analisis dari sistem.

BAB V SIMPULAN DAN SARAN : simpulan yang diperoleh dari analisis sistem dan saran yang bermanfaat.

BAB II LANDASAN TEORI

II.1 Teori Terkait Permasalahan

II.1.1 *Sentiment Analysis*

Sentiment Analysis adalah pengumpulan pandangan orang tentang setiap peristiwa yang terjadi dalam kehidupan nyata. Dalam situasi seperti itu di mana dunia sedang melalui, memahami emosi dari orang-orang berdiri sangat penting. Skenario kubur dimana orang tidak bisa keluar dari rumah mereka menuntut eksplorasi-ing apa orang-orang benar-benar berpikir tentang keseluruhan skenario. Oleh karena itu, penulis telah merencanakan pekerjaan ini di bawah menghadapi situasi yang menuntut terutama di media social (Chakraborty, 2020).

Analisis sentimen adalah proses untuk mengidentifikasi dan mengenali atau mengkategorikan emosi pengguna atau pendapat untuk layanan apa pun seperti film, masalah produk, acara, atau setiap atribut adalah positif, negatif atau netral. Sumber untuk analisis ini adalah saluran komunikasi sosial yaitu situs Web yang meliputi *review*, forum diskusi, *blog*, *micro-blog*, *Twitter* dll. Bidang penelitian ini sangat populer saat ini karena data pendapatnya di mana pengguna dapat menemukan ulasannya layanan apa pun yang berguna untuk kehidupan sehari-hari mereka. Besar jumlah data opini disimpan dalam bentuk digital. Untuk topik tertentu atau pendapat analisis sentimen yang menghubungkan penambangan data bekerja dan memberikan output. (Mehta and Pandya, 2020)

II.1.2 *Naïve Bayes Classifier*

Naïve Bayes Classifier adalah metode klasifikasi berdasarkan teorema Bayes. Pengklasifikasi *Naïve Bayes* dikenal lebih baik daripada beberapa metode klasifikasi lainnya. Karena pertama, ciri utama dari *Naïve Bayes* adalah asumsi independensi (naif) yang sangat kuat dari setiap kondisi atau peristiwa. Kedua, modelnya simple dan mudah dibuat. Ketiga, model dapat diimplementasikan untuk set data yang besar. Dasar salah satu teorema *Naïve Bayes* yang digunakan adalah rumus Bayes sebagai berikut: (Han, dkk, 2012).

Metode Naïve bayes classifier berasal dari bayes theorem yang ditemukan oleh Thomas bayes pada tahun 1770. Teorema bayes adalah sebuah teorema dengan

dua penafsiran berbeda. Teorema ini menyatakan seberapa jauh derajad kepercayaan subjektif harus berubah secara rasional ketika diberikan petunjuk baru. Teori ini juga berasal dari penerapan teori probabilitas.

Persamaan 1 adalah teori naïve bayes:

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)} \qquad \textit{Persamaan (1)}$$

P(H|e) = peluang kejadian **H** apabila **e** terjadi

P(e|H) = peluang kejadian **e** apabila **H** terjadi

P(H) = probabilitas kejadian (H)

P(e) = probabilitas (e) atau disebut prior probability. Berlaku jika (e) ≠ 0

II.1.3 *Machine Learning*

Machine learning atau dalam Bahasa Indonesia dikenal dengan pembelajaran mesin adalah aplikasi dari disiplin ilmu kecerdasan buatan (*Artificial Intelligence*). Konsep dari machine learning adalah memberikan kemampuan kepada computer untuk belajar secara mandiri dari sekumpulan data yang sudah diberikan sebelumnya, dengan menggunakan algoritma dan model untuk membuat prediksi. Fokus utama dari machine learning adalah untuk menemukan sebuah pola yang tepat dari sekumpulan data, sehingga dapat menghasilkan suatu model untuk melakukan proses *input-output* tanpa menggunakan kode program secara eksplisit (Tiwari, 2017). *Machine learning* dibagi dalam 3 bentuk, yakni *supervised learning*, *unsupervised learning* dan *generative learning*. Sentiment analisis dengan algoritma naïve bayes menggunakan metode *supervised Learning*.

1. Supervised Learning

Supervised learning adalah bidang pengenalan pola dan statistik dalam ilmu komputer. Ini adalah studi ilmiah tentang algoritma dan model statistik, yang digunakan untuk melakukan tugas tertentu secara efisien, tanpa menggunakan instruksi eksplisit, tetapi mengandalkan model. Algoritme pembelajaran yang

diawasi membangun model matematika dari data sampel untuk membuat prediksi tanpa memerlukan pemrograman eksplisit untuk melakukan tugas. (Yin, Q. 2020).

Supervised Learning adalah suatu metode untuk menciptakan *artificial intelligence* (AI), untuk mengidentifikasi pola dalam kumpulan data yang tidak di klasifikasikan atau tidak di beri label. Algoritma yang bertujuan untuk memperkirakan fungsi pemetaan sehingga ketika ada variabel *input* (X) kita dapat memprediksi variabel *output* (Y). Algoritma supervised learning dapat digunakan untuk memproses berbagai jenis data, mulai data yang terstruktur hingga yang tidak terstruktur. (Altamevia, 2023)

II.2 Teori Pendukung

II.2.1 Python

Python adalah bahasa pemrograman komputer open source untuk tujuan umum. Ini dioptimalkan untuk kualitas perangkat lunak, produktivitas pengembang, portabilitas program, dan integrasi komponen. Python digunakan oleh setidaknya ratusan ribu pengembang dunia di berbagai bidang seperti skrip Internet, pemrograman sistem, antarmuka pengguna, kustomisasi produk, pemrograman numerik, dan banyak lagi. Secara umum dianggap menjadi salah satu dari empat atau lima bahasa pemrograman yang paling banyak digunakan di dunia hari ini.(Lutz, 2011)

Python adalah bahasa pemrograman interpretative yang dianggap mudah dipelajari serta berfokus pada keterbacaan kode, dengan kata lain python diklaim sebagai Bahasa pemrograman yang memiliki kode kode pemrograman yang sangat jelas, lengkap dan mudah untuk dipahami. (JUD, 2019)

II.2.2 *Text Mining*

Text mining adalah salah satu bidang yang sampai saat ini masih berkembang dengan pesat, dengan tugasnya dalam mengekstraksi atau mengumpulkan informasi yang bermakna dari teks alami suatu bahasa. Ini dapat diartikan sebagai proses menganalisis suatu teks untuk kemudian diekstrak informasi-informasi yang berguna dari teks tersebut untuk tujuan terentu. Dalam budaya modern, teks adalah salah satu media dalam pertukaran informasi,

dibanding dengan database, teks tidak terstuktur, memiliki bermacam-macam bentuk, dan lebih sulit ditangani menggunakan algoritma tertentu.(Witten, 2004)

Pada kasus ini, yaitu text mining sumber data yang berupa teks tidak memiliki struktur yang jelas dan memiliki bermacam bentuk, sehingga disebut sebagai unstructured data. Maka dari itu, butuh proses untuk membuat data menjadi lebih terstuktur sehingga ekstraksi informasi dari teks akan lebih mudah, tepat, dan sangat penting dalam proses text mining. Sumber data yang digunakan, yaitu novel berbahasa Indonesia merupakan unstructured data, sehingga butuh proses untuk membuat data menjadi lebih terstuktur. Salah satunya adalah dengan diawali oleh preprocessing, yang mana nanti akan menghasilkan fitur yang lebih represenatif dibanding sumber data novel berbahasa Indonesia yang belum dipersiapkan dan masih tidak berstruktur.

II.2.3 Jupyter *Notebook*

Jupyter adalah organisasi non-profit untuk mengembangkan software interaktif dalam berbagai bahasa pemrograman. Notebook adalah satu software buatan Jupyter, adalah aplikasi web open-source yang memungkinkan Anda membuat dan berbagi dokumen interaktif yang berisi kode live, persamaan, visualisasi, dan teks naratif yang kaya.(Prijono, 2019)

Pada penelitian ini, Jupyter Notebook digunakan sebagai salah satu text editor untuk menuliskan kode-kode program Python serta memvisualisasikan data hasil olah pada sistem ini.

II.2.4 *Preprocessing*

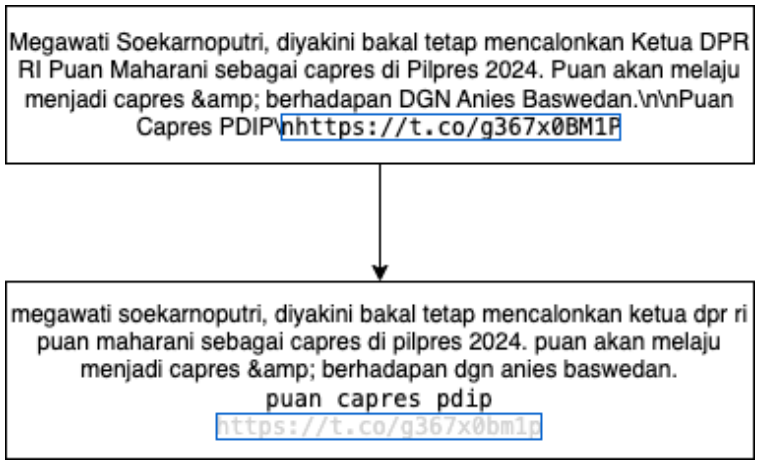
Proses preprocessing merupakan tahap dimana sumber data input diolah Kembali sebelum kemudian diproses lebih lanjut dan dijadikan bahan data pada machine leardning. Pada kasus teks tweet, text mining masih memliliki struktur yang bercampur dimana pada tweet masih ada mention dan link, serta data yang NaN pada data hasil scrapping, sehingga dibutuhkan proses yang merubah bentuknya menjadi data yang terstruktur. Proses ini akan melakukan penyeragaman case tweet yang merubah semua tweet menjadi lowercase, menghilangkan tanda mention serta username termention, menghilangkan tautan pada tweet, kemudian

membuat token dari data input, sehingga data lebih bersih, terstruktur dan dapat diolah lebih lanjut.

Pada penelitian ini, tahap preprocessing yang diterapkan adalah *Case Folding, Lemmatization, Stopword Removal dan Tokenizing.*

1. *Case Folding*

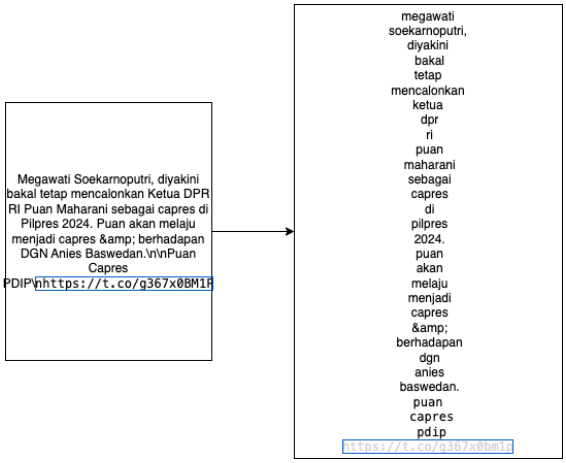
   Case folding digunakan untuk menyeragamkan seluruh teks dalam case yang seragam, abik menjadi huruf kecil (*lowercase*) ataupun huruf kapital (*uppercase*). Case folding digunakan pada penelitian ini adalah penyeragaman menjadi lowercase. Contoh dari penerapan case folding bisa dilihat pada gambar dibawah ini:



Gambar 2. 1 Contoh case folding

2. *Tokenizing*

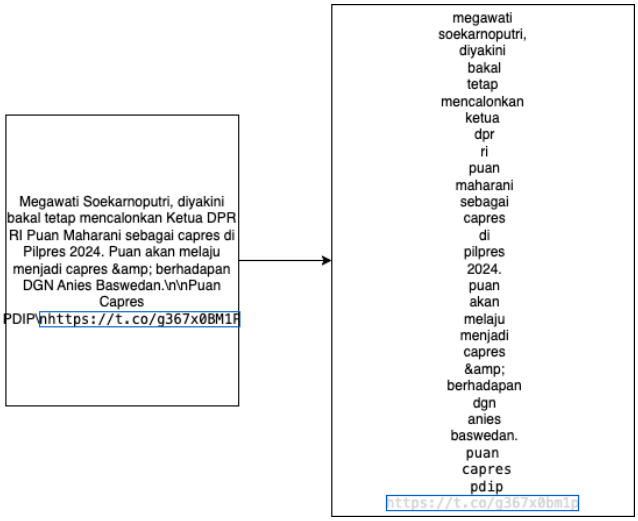
   Proses *Tokenizing* adalah proses dimana string dipotong menjadi beberapa bagian dengan melihat delimiternya, seperti tipe kapitalisasi, keberadaan digit, tanda baca, karakter special dan sebagainya. Pemecahan dokumen menjadi kata – kata tunggal silakukan dengan caara men-scan dokumen dan setiap kata akan teridentifikasi atau terpisahkan dengan kalimatnya oleh delimiter. Tokenizing adalah proses diamana data input dibagi menjadi beberapa token sesuai dengan jumlah kalimat menggunakan delimiter '.' Pada teks input contoh tokenizing ada pada gambar dibawah.



Gambar 2. 2 Contoh Tokenizing

3. *Stemming*

Stemming adalah teknik pada natural language processing yang digunakan untuk mengembalikan kata kepada kata dasarnya yang disesuaikan dengan kamus Bahasa Indonesia, proses stemming dilakukan dengan menggunakan library sastrawi. Stemming digunakan pada kebutuhan yang berhubungan dengan text mining seperti information retrieval yang dilakukan pada tahap preprocessing.



Gambar 2. 3 Contoh stemming

4. *Stopword Removal*

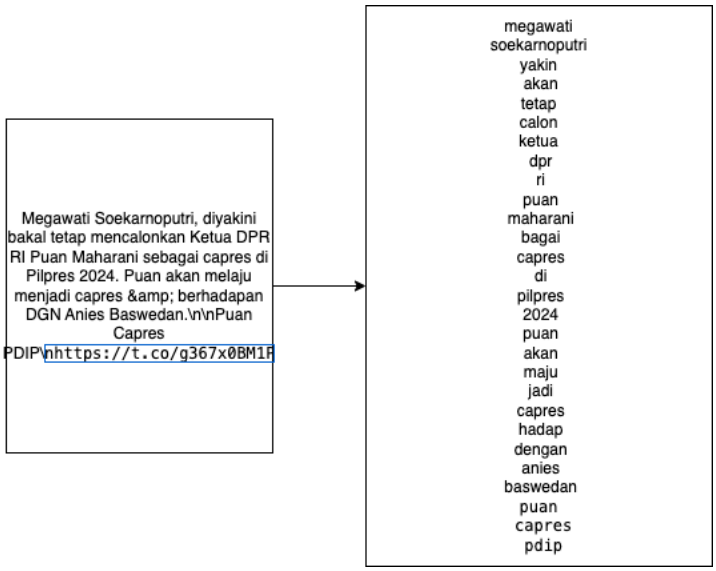
Stopword Removal adalah tahap pemilihan kata-kata yang dianggap penting. Terdapat dua metode yang dapat digunakan dalam tahap stopword removal, yakni:

a) ***Stoplist***

Pada metode ini, kita menyuaokan kumpulan kata yang tidak deskriptif/tidak pendting yang disebut stoplist. Kata yang termasuk ke dalam stoplist akan dibuang dan tidak digunakan pada proses selanjutnya.

b) ***Wordlist***

Wordlist merupakan kebalikan dari stoplist, pada metode ini kita menyiapkan kumpulan kata yang deskriptif yang disebut wordlist. Hanya kata yang termasuk ke dalam wordlist yang akan digunakan pada proses selanjutnya, sementara kata lainnya akan dibuang



Gambar 2. 4 Contoh stopword removal

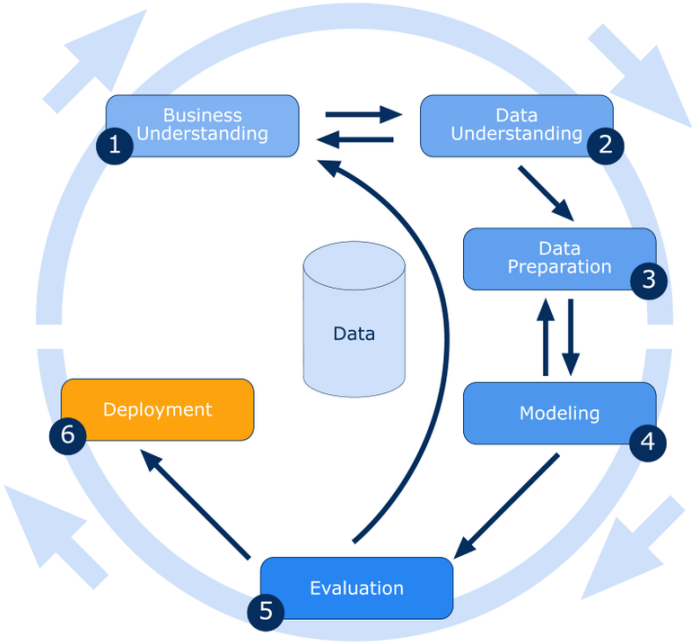
II.2.5 *Monte Carlo Cross Validation*

Monte Carlo Cross Validation (MCCV), suatu bentuk validasi model, pertama kali diperkenalkan oleh Picard dan Cook (1984). Shao (1993) membuktikan bahwa metode ini konsisten secara asimtotik dan memiliki peluang lebih besar daripada LOO untuk memilih model terbaik dengan kemampuan prediksi yang lebih akurat. MCCV meninggalkan bagian penting dari sampel pada

suatu waktu selama pembuatan model dan validasi dan mengulangi prosedur berkali-kali. Jika dibandingkan dengan metode biasa untuk memilih variabel prediktor terbaik (yaitu regresi bertahap dan menggunakan statistik seperti Mallows Cp atau hipotesis P-value), MCCV mungkin lebih diinginkan karena mengevaluasi model yang berbeda sesuai dengan kemampuan prediksi mereka menggunakan banyak model yang berbeda. kombinasi set data validasi. Menariknya, MCCV belum diuji dalam analisis regresi hidrologi, di mana orang sering berurusan dengan kumpulan data teramati yang sangat terbatas dan langka. (Haddad, 2013)

II.2.6 Cross Industry Standard Process for Data Mining (CRISP-DM)

Cross Industry Standard Process for Data Mining atau yang biasa kita sebut dengan CRISP-DM merupakan model proses independen industri untuk data mining. CRISP-DM sendiri terdiri dari enam fase iteratif dari *business understanding* hingga deployment(Felix. K, 2021)



Gambar 2. 5 Alur CRISP-DM

point point pada gambar 2. 5 akan dijelaskan dibawah ini:

1. *Business Understanding*

Bisnis harus dinilai untuk mendapatkan gambaran tentang sumber daya yang tersedia dan dibutuhkan. penentuan tinjauan data mining merupakan salah satu

aspek terpenting dalam fase ini. Jenis data mininig harus dijelaskan dan kriteria keberhasilan data mining. Rencana proyek wajib dibuat.

2. Data Understanding

Mengumpulkan data dari sumber data, mengeksplorasi dan mendeskripsikannya serta memeriksa kualitas data adalah tugas penting dalam fase ini.

3. Data Preparation

Pemilihan data harus dilakukan dengan menentukan kriteria inklusi dan eksklusi. kualitas data yang buruk dapat ditangani dengan cleaning data.

4. Modeling

Tahap permodelan terdiri dari pemilihan teknik permodelan, membangun kasus uji dan model. Pada tahap ini dilakukan metode statistika dan Machine Learning untuk penentuan terhadap teknik data mining, alat bantu data mining, dan algoritma data mining yang akan diterapkan. Semua teknik data mining dapat digunakan.

5. Evaluation

fase evaluasi hasilnya diperiksa terhadap tujuan bisnis yang ditetapkan. oleh karena itu, hasilnya harus ditafsirkan dan tindakan lebih lanjut harus ditentukan.

6. Deployment

fase deployment atau rencana penggunaan model adalah tahap yang paling dihargai dari proses CRISP-DM. Perencanaan untuk Deployment dimulai selama Business Understanding dan harus menggabungkan tidak hanya bagaimana untuk menghasilkan nilai model, juga bagaimana mengkonversi skor keputusan, dan bagaimana untuk menggabungkan keputusan dalam sistem operasional.

II.3 Penelitian-penelitian Terdahulu

II.3.1 Implementasi Metode Naïve Bayes untuk Analisis Sentimen Warga Jakarta Terhadap Kehadiran Mass Rapid Transit

Sarika A, Helena N, Noor F, Ika N.(2019), melakukan penelitian menggunakan data dari sosial media yaitu Twitter dengan keyword "MRTJakarta" yang dilakukan selama masa uji coba public MRT yaitu dari tanggal 5 – 23 maret 2019. Tweet yang diambil sebanyak 1000 tweet (800 tweet untuk training dan 200 tweet untuk testing). Dalam peneliltian ini naive bayes dapat memprediksi sentimen dari tweet yang sudah dikumpulkan terkait animo masyarakat terhadap MRTJakarta dengan akurasi sebesar 75%.

II.3.2 Algoritma Naïve Bayes Classifier Untuk Analisis Sentiment Pengguna Twitter Terhadap Provider By.u

Ike V, Bagas S.(2022), melakukan penelitian ini dapat diambil kesimpulan bahwa Algoritma Naïve Bayes Classifier dapat melakukan analisis sentimen dengan benar dan melakukan klasifikasi secara otomatis setelah melalui tahapan- tahapan proses, yaitu Preprocessing data, pembobotan kata, membuat model untuk klasifikasi otomatis dan dibuatnya data training untuk melatih klasifikasi pada data testing. Tahapan proses tersebut dapat berjalan dengan baik dan mengklasifikasikan data dengan parameter positif dan negatif. Setelah dilakukan 3 kali pengujian didapatkan hasil akurasi 80%, 80%, dan 85%. Didapatkan hasil akurasi paling tinggi pada pengujian terakhir yakni sebesar 85%. Dengan pengujian menggunakan 3 dataset yang memiliki jumlah data yang berbeda, dan setelah mendapatkan hasil tingkat akurasi dari proses analisis sentimen dapat disimpulkan bahwa jumlah dataset dalam pengujian sangat berpengaruh terhadap tingkat akurasi Algoritma Naïve Bayes Classifier. Hal ini ditunjukan oleh hasil tingkat akurasi pada pengujian ketiga dengan 3000 dataset mendapatkan nilai akurasi 85%, lebih besar daripada pengujian pertama dengan 1000 dataset yang hanya memiliki akurasi sebesar 80%.

II.3.3 Sentiment Analysis Menggunakan Naïve Bayes Classifier pada Tweet Tentang Zakat

Adhyaksa H (2020), hasil klasifikasi sentiment dari 50 tweet data uji menggunakan algoritma naïve bayes dengan seleksi fitur Term-Frequency serta metode lexicon Based, didapatkan jumlah sentiment positif yang lebih dominan dibandingkan sentiment negative maupun netral dikarenakan pada pengujian dengan metode lexicon based terdapat lebih banyak tweet yang mengandung kata dalam kamus lexicon positif disbanding kata dalam kamus lexicon negative. Selanjutnya, pada pengujian dengan masing-masing seleksi fitur, sentiment positif lebih dominan dikarenakan tidak keseimbangan jumlah sentiment positif, negative dan netral dalam klasifikasi data latih menggunakan metode lexicon based dimana sentiment positif lebih besar sehingga system lebih condong dalam mengklasifikasi sentiment positif

II.4 State Of Art

State of the art adalah hal yang cukup penting bagi penelitian, bermanfaat untuk mengetahui bagaimana berkembangnya ilmu pada bidang masalah general yang sedang diteliti sampai penulis menemukan masalah penelitian yang dapat memberikan kontribusi. State of art dapat diketahuli lewat penelitian terdahulu.

Perbandingan penelitian terdahulu dapat dilihat dari segi persamaan penelitian atau perbedaan yang ada pada penelitian sebelumnya. Penelitian terdahulu juga ditujukan untuk membantu menemukan inspirasi bagi penelitian selanjutnya. Penelitian terdahulu yang dapat menjadi acuan penelitian ini dikemas dalam bentuk tabel dan deskripsi agar dapat mempermudah perbandingan antar satu penelitian dan penelitian lainnnya. Tabel state of art dapat dilihat pada tabel 2.1

Tabel 2. 1 Tabel *State Of Art*

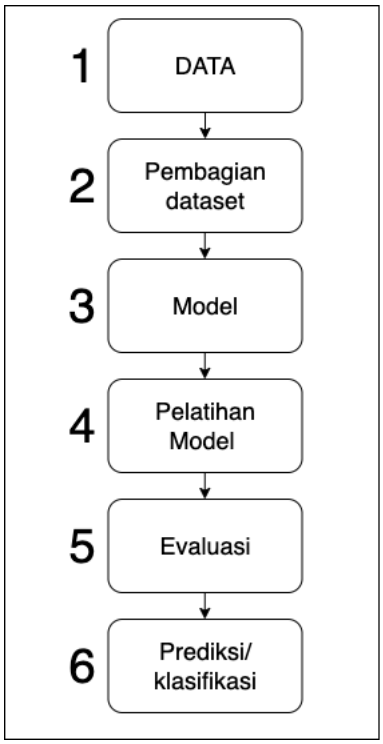
No.	Judul Penelitian	Penulis	Tahun Penelitian	Kata Kunci	Jangka Waktu Data	Banyaknya *tweet*	Pembagian dataset	akurasi
1.	Implementasi Metode Naïve Bayes untuk Analisis Sentimen Warga Jakarta Terhadap Kehadiran Mass Rapid Transit	Sarika A, Helena N, Noor F, Ika N.	2019	"MRTJakarta"	5 - 23 Maret 2019	1000 *tweet*	800 *Training* 200 *Test* (80:20)	75%
2.	Algoritma Naïve Bayes Classifier Untuk Analisis Sentiment	Ike V, Bagas S	2022	"By.U"	23 - 26 Mei 2021	1000,2000,3000 *tweet* 3x pengujian	-	85%

	Pengguna Twitter Terhadap Provider By.u							
3.	Sentiment Analysis Menggunakan Naïve Bayes Classifier pada Tweet Tentang Zakat	Adhyaksa, H.	2020	"@baznasindonesia"	4 Juni 2019	1000 *tweet*	950 *Training* 50 *Test* (95:5)	74%

BAB III METODOLOGI PENELITIAN

III.1 Metode Penelitian

Metode penelitian yang digunakan ialah pendekatan supervised learning. metode ini memiliki tahapan tahapan yaitu pengumpulan data, preprocessing data, pelabelan data, pemilihan model dan prediksi, bisa dilihat pada gambar 3.1.



Gambar 3. 1 Alur pendekatan *Supervised Learning*

III.2 Metodologi Pengembangan Sistem

Metode pengembangan sistem adalah pendekatan atau prosedur yang digunakan untuk merencanakan, merancang, mengembangkan, mengimplementasikan, dan memelihara sistem perangkat lunak atau sistem informasi. Metode ini menyediakan panduan dan kerangka kerja untuk mengatur langkah-langkah dalam pengembangan sistem, memastikan bahwa proyek berjalan secara terstruktur, efisien, dan efektif.

III.2.1 Cross Industry Standard Process for Data Mining

penelitian ini menggunakan Cross industry standard process for data mining (CRISP-DM) sebagai metode pengembangan sistem, CRISP-DM memiliki 6

tahapan yakni, business understanding, data understanding, data preparation, modelling, evaluation, dan deployment. Alasan menggunakan CRISP-DM ialah karena memiliki tahapan deploying yang mana tahapan tersebut sesuai dengan luaran dari penelitian ini.

III.2.2 Scrapping

Metode pengambilan data pada penelitian ini menggunakan scrapping, dengan menggunakan library snscrape yang menggunakan bahasa pemrograman python, data yang direquest memiliki query " pilpres 2024 OR prabowo OR ganjar OR anies OR ahy OR pdip OR gerindra OR pdip OR megawati OR puan OR jokowi OR pan OR pkb OR nasdem until:2022-12-31 since:2022-01-01" yang berarti data yang diambil adalah data yang ada sejak awal januari hingga akhir desember, data yang berhasil didapatkan ialah sebanyak 112844.

III.2.3 Pendekatan Supervised Learning

Penelitian ini menggunakan pendekatan supervised learning dengan pelaksanaan dimulai dari data, pembagian dataset, model, pelatihan model, evaluasi dan prediksi/klasifikasi. pengambilan pendekatan supervised learning sebagai metode penelitian karena model naive bayes classifier merupakan jenis machine learning dengan metode supervised learning dan memiliki output klasifikasi.

III.2.4 Modeling

Penelitian ini menggunakan modeling naive bayes classifier, model ini diambil dari bayes theorem, model ini adalah salah satu model untuk melakukan klasifikasi, yang mana pada penelitian ini penulis ingin melatih model yang dapat melakukan klasifikasi positif, negatif dan netral.

III.2.5 Monte Carlo Cross Validation

Monte Carlo Cross Validation merupakan metode yang digunakan untuk evaluasi, cara kerja MCCV ialah pengetesan secara acak pada data test dan training sehingga dapat menghasilkan representasi yang baik pada model hasil test dan train.

III.3 Tahapan Penelitian

Berikut adalah tahapan penelitian dengan metode penelitian pendekatan supervised learning yang ada pada gambar 3.1.

1. Data

Aktivitas yang dilakukan dalam proses machine learning ialah pengumpulan data, pengumpulan data dilakukan dengan metode scraping yang sudah dijelaskan pada point III.2.2, dari proses ini didapatkan data kotor sebanyak 112844 data kotor, sehingga harus dibersihkan dahulu agar bisa menghilangkan nilai bias yang ada pada data dan dapat mempengaruhi output dari machine learninng.

2. Pembagian dataset

Setelah mendapatkan data, dan sebelum melakukan pelatihan pada model machine learning yang akan digunakan, peneliti harus melakukan pembagian dataset, pada pmbagian dataset memiliki aturan, yakni data test jangan sampai melebihi data training, karena akan menimbulkan banyak bias. Penulis mengambil perbandingan 80:20 yang berarti 80% data menjadi data training dan 20% menjadi data uji.

3. Model

Model yang dipilih penulis sesuai dengan judul penelitian, yakni Naive bayes classifier, dimana model ini merupakan model jenis klasifikasi. Klasifikasi merupakan bentuk output yang diinginkan dimana model dapat melakukan klasifikasi pada naskah baru. Klasifikasi yang dicari ialah apakah naskah tweet ini positif, negatif, dan netral.

4. Pelatihan model

Model dilatih dengan menggunakan dataset pelatihan. Selama pelatihan, model mencoba menemukan pola atau relasi yang ada antara fitur-fitur data dan label yang sesuai. Tujuannya adalah untuk membuat model yang dapat mempelajari dan mewakili hubungan yang ada dalam data.

5. Evaluasi

Pada tahap ini penulis menggunakan metode Monte Carlo Cross Validation (MCCV) pada tahap evaluasi. MCCV melakukan evaluasi dengan cara mengacak pengambilan sample uji, serta nilai yang dicari ialah akurasi, presisi, recall, F1-score.

6. Prediksi/Klasifikasi

Setelah model dilatih dan dievaluasi, model dapat digunakan untuk melakukan prediksi atau klasifikasi pada data baru yang belum diketahui. Model ini memanfaatkan pola dan relasi yang telah dipelajari dari dataset pelatihan untuk menghasilkan prediksi atau klasifikasi yang sesuai. serta hasil model diexport dengan format pickle.

BAB IV HASIL DAN PEMBAHASAN

Metodologi pengembangan sistem yang digunakan pada penelitian ini ialah Cross Industry Standard Process for Data mining. Metode ini memilik 6 langkah yaitu, business understanding, data understanding, data preparation, modeling, evaluation dan deployment.

IV.1 Business Understanding

Pemilihan umum merupakan acara demokrasi bagi warga Indonesia dan hanya diselenggarakan setiap 5 tahun sekali, dalam 5 tahun tersebut pemimpin yang terpilih wajib melakukan tanggung jawab sebagai pemimpin negara yang telah dipilih oleh masyarakat Indonesia, setelah waktu berlalu 5 tahun masyarakat dapat memilih pemimpin baru kembali. Pemilihan umum pada era digital cukup menarik, karena semua elemen masyarakat dapat berdiskusi perihal calon pemimpin negara lewat sosial media, lebih spesifik yakni pada media sosial twitter. melalui satu cuitan dalam media sosial dapat mempengaruhi opini masyarakat terhadap calon tersebut, belum lagi pada tahun 2024 nanti banyak pemilih pemula dan pertama kalinya mengikuti pesta demokrasi 5 tahunan ini.

IV.1.1 Goal

Tujuan dari analisis sentimen ini ialah:

1. Mengetahui akurasi dari dataset yang ada terhadap sentimen
2. Membuat web sederhana yang menggunakan model machine learning hasil analisis sentimen

IV.2 Data Understanding

Data yang digunakan dalam melakukan analisis ini merupakan data dari twitter yang didapat dengan cara crawling menggunakan library python snscrape.

penulis menggunakan query "pilpres 2024 OR prabowo OR anies OR ganjar since:2022-01-01 until:2022-12-31" query tersebut berarti mencari tweet yang memiliki keyword pilpres 2024, prabowo, anies, ganjar dan data tersebut di twit pada awal januari hingga akhir desember tahun 2022.

```
import pandas as pd
from tqdm.notebook import tqdm
import snscrape.modules.twitter as sntwitter


import datetime

scraper = sntwitter.TwitterSearchScraper('pilpres 2024 OR prabowo OR anies OR ganjar since:2023-01-01 until:2023-05-31')
```

Gambar 4. 1 Query Scrapping

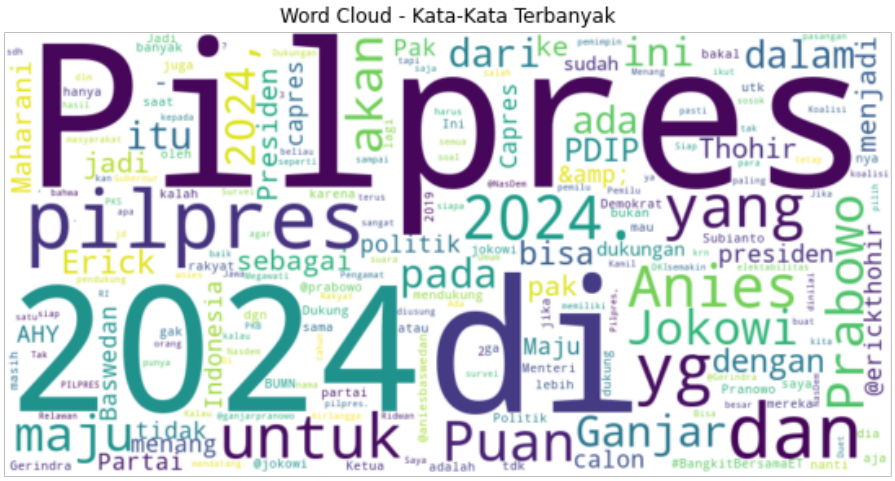
```
tweets = []
n_tweet = 100000
for i, tweet in tqdm(enumerate(scraper.get_items()),total=n_tweet):
    data = [tweet.date, tweet.renderedContent, tweet.user.username]
    tweets.append(data)
    if i > n_tweet:
        break
```

Gambar 4. 2 Get data

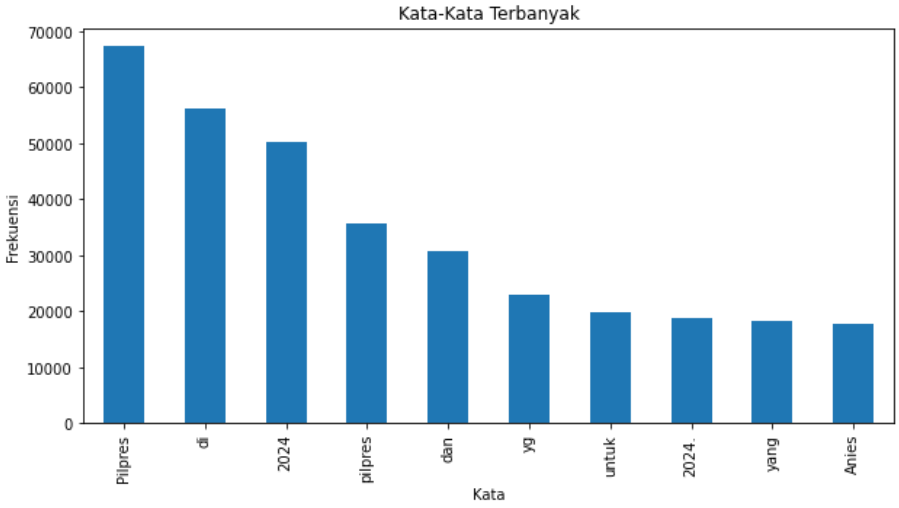
Data disimpan menggunakan format CSV(comma separated view), agar memudahkan penulis meneliti menggunakan bahasa pemrograman python yang dibantu dengan library pandas. Data tersebut memiliki kolom yakni:

1. Date : tanggal dimana tweet tersebut dibuat
2. Tweet: isi dari tweet tersebut
3. Username: pengguna twitter yang melakukan tweet

Setelah melakukan visualisasi data, terhadap dataset yang digunakan maka menghasilkan wordcloud seperti gambar 4.1 dan bar plot pada gambar 4.2



Gambar 4. 3 Word Cloud kata terbanyak



Gambar 4. 4 Bar plot kata terbanyak

IV.3 Data Preparation

Sebelum dilakukan pelatihan pada machine learning, data wajib dibersihkan dahulu, fase ini disebut dengan preprocessing yang mana output dari preprocessing ini ialah membuat data yang sebelumnya kotor menjadi bersih, sehingga bisa merendahkan bias yang ada pada dataset, hal ini sangatlah vital dalam pembuatan machine learning agar data yang digunakan bisa mencapai akurasi yang diinginkan dan rendah akan bias. tahapan preprocessing pada penelitian ini ada 4 yaitu:

1. Case Folding dan Filtering

   Case folding dan filtering merupakan tahap awal dari preprocessing, output dari tahap ini ialah penyamarataan huruf menjadi lowercase, serta menghapus username twitter, retweet dan hashtag pada data tweet. Code yang akan digunakan ada pada gambar 4. 5 pertama, kita harus menulis import library regex yakni 'import re' sudah ditulis pada line 1, lalu menulis fungsi preprocess_tweet2 dengan parameter tweet. Filtering dilakukan dengan mendeklarasi emoji pattern dengan ascii dari emoji, kemudian menghapus angka, melakukan case folding, menghapus tautan yang ada pada tweet, menghapus username dan hashtag yang ada pada tweet, lalu menghapus "\n" yang ada pada setiap tweet, serta remove punctuations seperti koma, titik dan lainnya. contoh tweet awal ada pada gambar 4.6

```
def preprocess_tweet2(tweet):
    EMOJI_PATTERN = re.compile(
    "(["
    "\U0001F1E0-\U0001F1FF"  # flags (iOS)
    "\U0001F300-\U0001F5FF"  # symbols & pictographs
    "\U0001F600-\U0001F64F"  # emoticons
    "\U0001F680-\U0001F6FF"  # transport & map symbols
    "\U0001F700-\U0001F77F"  # alchemical symbols
    "\U0001F780-\U0001F7FF"  # Geometric Shapes Extended
    "\U0001F800-\U0001F8FF"  # Supplemental Arrows-C
    "\U0001F900-\U0001F9FF"  # Supplemental Symbols and Pictographs
    "\U0001FA00-\U0001FA6F"  # Chess Symbols
    "\U0001FA70-\U0001FAFF"  # Symbols and Pictographs Extended-A
    "\U00002702-\U000027B0"  # Dingbats
    "])")
    tweet = re.sub(r'[0-9]+','', str(tweet))
    tweet = tweet.lower()  # convert to lower case
    tweet = re.sub(r"http\S+|www\S+|https\S+", '', tweet, flags=re.MULTILINE)  # remove URLs
    tweet = re.sub(r'\@\w+|\#\w+', '', tweet)  # remove mentions and hashtags
    tweet = re.sub(r'\d+', '', tweet) # remove numbers
    tweet = re.sub(r'\.',' ', tweet) #
    tweet = tweet.translate(str.maketrans("", "", string.punctuation))  # remove punctuations
    tweet = tweet.strip()
    tweet = re.sub(EMOJI_PATTERN, r'', tweet)
    tweet = re.sub(r'\n+', '', tweet)
    tweet = re.sub(r'^\s+', '', tweet)
    return tweet
data['tweet']=data['tweet'].apply(preprocess_tweet2)
```

Gambar 4. 5 Fungsi case folding dan filtering

'Berkat kinerja di BUMN, serta kedekatan dengan pemerintah, membuat elektabilitas terus meningkat. Atas dasar itu, dinilai jadi Cawapres terkuat di Pilpres 2024.\n\n@erickthohir\n#BangkitBersamaET https://t.co/1kxqJpwT66'

Gambar 4. 6 Tweet semula

Setelah kita mengamati apa yang ada pada gambar 4.6, kita lakukan proses yang ada pada gambar 4.5 maka hasilnya dapat dilihat pada gambar 4.7.

berkat kinerja di bumn serta kedekatan dengan pemerintah membuat elektabilitas terus meningkat atas dasar itu dinilai jadi cawapres terkuat di pilpres

Gambar 4. 7 Tweet setelah casefolding dan filtering

2. Tokenisasi

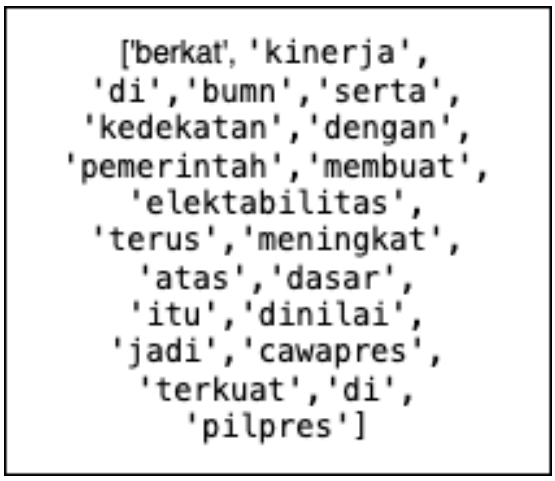
Tokenisasi merupakan tahap ketika setiap kalimat/tweet dipisah menjadi per kata agar mudah melakukan tahap selanjutnya yakni Stemming dan stopword removal, fungsi untuk tokenisasi dapat dilihat pada gambar 4.8.

```python
from nltk.tokenize import word_tokenize

def tokenize_column(text):
    if isinstance(text, str):  # Memastikan bahwa text adalah string
        return word_tokenize(text)
    else:
        return []  # Mengembalikan list kosong jika text bukan string

# Contoh penggunaan:
data['tweet'] = data['tweet'].apply(tokenize_column)
```

✓ 16.9s

Gambar 4. 8 Fungsi Tokenisasi

output yang dihasilkan dari fungsi tokenisasi pada gambar 4.8 dapat dilihat pada gambar 4.9.



Gambar 4. 9 Output Tokenisasi

bisa dilihat, data yang asalnya seperti gambar 4.8, kini berubah menjadi pengutipan tiap kata, yang menjadi data tersebut ter-tokenisasi.

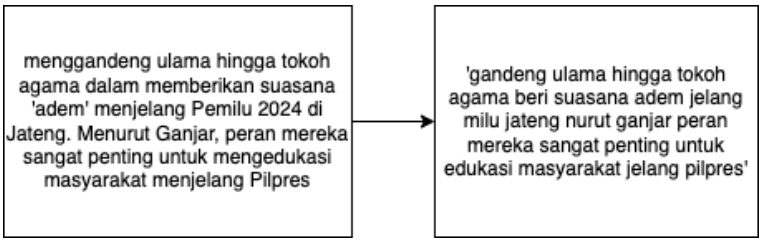
3. Stemming/Lemmatization

Proses ini, adalah proses pelemasan kata, atau membuat kata menjadi kembali ke kata asal. Seperti kalimat 'menjadi', setelah di stemisasi akan menjadi kalimat 'jadi' dan kalimat yang lainnya pun menjadi kembali pada kalimat asal. Proses ini termasuk proses penting pada preprocessing. Fungsi untuk melakukan stemming dapat dilihat pada gambar 4.10.

```
stemmer = StemmerFactory().create_stemmer()
def stemming(batch):
    # Menerapkan stemming pada setiap teks dalam batch
    stemmed_batch = [stemmer.stem(text) for text in batch]

    # Melakukan penghapusan stopwords pada setiap teks dalam batch

    return stemmed_batch

# Menambahkan hasil preprocessing ke dalam DataFrame
data['tweet'] = data['tweet'].apply(stemming)

# Hasil preprocessing data
print(data)
```
✓ 48.4s

Gambar 4. 10 Fungsi Stemming

setelah kalimat dimasukan pada fungsi stemming, yang mana data stemming bahasa indonesia didapat oleh library sastrawi. kita harus mendeklarasikan 'from Sastrawi.Stemmer.Stemmerfactory import StemmerFactory' lalu deklarasi variabel baru yakni stemmer, seperti pada gambar 4.10 dimana variabel ini memanggil fungsi dari StemmerFactory. untuk hasil outputnya dapat dilihat pada gambar 4.11.



Gambar 4. 11 Output dari proses Stemming

Dapat dilihat kalimat menggandeng menjadi gandeng, memberikan menjadi beri, inilah hasil dari proses stemming menggunakan library sastrawi.

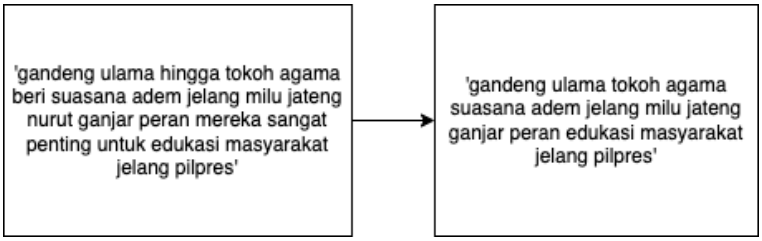
4. Stopword Removal

Proses *stopword removal* tak kalah penting dengan proses stemming, proses ini berfungsi untuk menghilangkan kalimat kalimat yang sering muncul pada kalimat, jika kita ambil contoh bahasa inggris, kalimat yang akan hilang yakni, *i, me, you*, dll. Kalimat yang sangat sering muncul dan tidak mempengaruhi dari nilai sentimen tersebut. Fungsi Stopword Removal ada pada gambar 4.12.

```
stopword_remover = StopWordRemoverFactory().create_stop_word_remover()
def stopword(batch):
    # Melakukan penghapusan stopwords pada setiap teks dalam batch
    cleaned_batch = [stopword_remover.remove(text) for text in batch]
    cleaned_batch = [text for text in cleaned_batch if 'yg' not in text.lower()]

    return cleaned_batch


# Menambahkan hasil preprocessing ke dalam DataFrame
data['tweet'] = data['tweet'].apply(stopword)

# Hasil preprocessing data
print(data)
```
✓ 2.8s

Gambar 4. 12 Fungsi Stopword Removal

Sama seperti stemming, stopword removal juga menggunakan library dari sastrawi dengan fungsi StopWordRemoverFactory dan dipanggil dengan variabel 'stopword_remover'. *Output* dari *stopword removal* ada pada gambar 4.13.



Gambar 4. 13 Output Stopword Removal

jika dilihat secara seksama kalimat hingga, beri, nurut hilang, karena kalimat tersebut ada pada kamus stopword removal atau kalimat itu termasuk pada kalimat stopword, sehingga kalimat tersebut hilang.

Setelah melakukan preprocessing, kita melakukan labeling yakni memberi label kepada setiap tweet yang ada pada dataset, apakah ini positif, negatif atau netral. labeling sangatlah krusial karena label adalah target yang akan dilatih sebagai hasil dari latihan tersebut. Fungsi labelling ada pada gambar 4.14.

```
from nltk.tokenize import word_tokenize
def labelling(tweet):
    # tweet = preprocess_tweet(tweet)
    clean_tokens = word_tokenize(tweet) #tokenize
    # clean_tokens = [stemmer.stem(word) for word in tweet_tokens if word not in
    positive_words = open("positive.txt").read().splitlines()
    negative_words = open("negative.txt").read().splitlines()
    positive_count = sum([1 for word in clean_tokens if word in positive_words])
    negative_count = sum([1 for word in clean_tokens if word in negative_words])
    if positive_count > negative_count:
        return 'Positive'
    elif positive_count < negative_count:
        return 'Negative'
    else:
        return 'Neutral'
data['sentiment'] = data['tweet'].apply(labelling)
✓ 2m 27.3s
```

Gambar 4. 14 Fungsi Labelling

Setelah melakukan labelling maka akan menambah kolom baru lagi yakni kolom sentimen, karena pada gambar 4. 14 jelas, bahwa kolom sentimen dihasilkan dari fungsi labelling, dan menghasilkan data neutral sebanyak 41061 (empat puluh satu ribu enam puluh satu), data positif sebanyak 38145 (tiga puluh delapan ribu seratus empat puluh lima), dan data negatif sebanyak 19340 (sembilan belas ribu tiga ratus empat puluh), dapat dilihat pada gambar 4.15.

```
data.sentiment.value_counts()
✓ 0.0s

Neutral     41061
Positive    38145
Negative    19340
Name: sentiment, dtype: int64
```

Gambar 4. 15 Hasil Labelling

IV.4 Modeling

Model yang akan digunakan yakni Naive Bayes Classifier, yang dasarnya diambil dari bayes theorem, rumusnya dapat dilihat pada persamaan(1). Model ini sangatlah terkenal pada proses klasifikasi, selain sentimen analisis, model ini biasa digunakan pada forecasting cuaca. Proses Modelling dapat dilihat pada gambar 4.16

```python
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score, f1_score, precision_score, recall_score,
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import Pipeline
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

vectorizer = CountVectorizer()

# Melakukan transformasi teks menjadi vektor fitur
X = vectorizer.fit_transform(data['tweet'].astype(str))
# Y = vectorizer.fit_transform(data['sentiment'])

epoch = 1
accuracy_values = np.zeros(epoch)
# Membagi data menjadi data latih dan data uji
for epochs in range(epoch):
    X_train, X_test, y_train, y_test = train_test_split(X, data['sentiment'], test_size=0.2, random_state=42)
    naive_bayes = MultinomialNB()
    training = naive_bayes.fit(X_train, y_train)
    prediction =  naive_bayes.predict(X_test)

    accuracy = accuracy_score(y_test, prediction)
    accuracy_values[epochs] = accuracy

    print("epoch:", epochs + 1, "acc:", accuracy)
plt.plot(range(1,epoch + 1), accuracy_values, marker='o')
plt.xlabel('epoch')
plt.ylabel('Accuracy')
plt.title('ACC per Epoch')
plt.show()



y_pred = naive_bayes.predict(X_test)

# Menghitung akurasi
accuracy = accuracy_score(y_test, y_pred)
print("Akurasi: {:.2f}%".format(accuracy * 100))
```

Gambar 4. 16 Modelling

Pada proses ini penulis membuat data latih sebanyak 80% dari banyaknya dataset yang ada, yang diambil secara acak menggunakan parameter random_state. Maka X_train sebanyak 80% label X dan 20% sisanya menjadi X_test, hal ini juga berlaku pada y_train dan y_test. Output dari training data uji menghasilkan akurasi sebesar 70.38%. Dapat dilihat pada gambar 4.17 yang menunukan hasil pelatihan dataset dengan model Naive Bayes Classifier.

```
epoch: 1 acc: 0.7038140020898641
```

Gambar 4. 17 Hasil Training

IV.5 Evaluation

Monte-Carlo Cross Validation digunakan sebagai evaluasi dari dataset, MCCV (Monte-Carlo Cross Validation) merupakan pengujian setiap data uji dan data latih yang diambil secara acak dan dicari akurasinya, untuk code MCCV dapat dilihat pada gambar 4.18.

```
from sklearn.model_selection import ShuffleSplit
from sklearn.model_selection import cross_val_score
shuffle_split = ShuffleSplit(n_splits=150, test_size=0.2, random_state=42)

# Melakukan cross-validation dengan ShuffleSplit
scores = cross_val_score(naive_bayes, X, data['sentiment'], cv=shuffle_split, scoring='accuracy')
x = []
# Menampilkan skor akurasi untuk setiap iterasi cross-validation
for i, score in enumerate(scores):
    print(f"Iterasi {i+1}: {score}")
    x.append(score)

# Menampilkan rata-rata skor akurasi dari cross-validation
print(len(x))
print("Rata-rata skor akurasi: {:.2f} %".format(scores.mean()*100))
print("Skor tertinggi: {:.2f} %".format(scores.max()*100))
plt.plot(range(len(x)), x, marker='o')
```

Gambar 4. 18 Monte-Carlo Cross Validation

dapat dilihat pada variabel shullfe_split saya mendeklarasikan sebanyak 150 kali acak dengan pengambilan data uji sebanyak 20% dan random pengambilan sebanyak 42. Variabel score memanggil cross_val_score yang mana memanggil cv=shuffle_split yakni metode cross validation menggunakan shuffle_split. Hasil dari evaluasi ini dapat dilihat pada gambar 4.19.

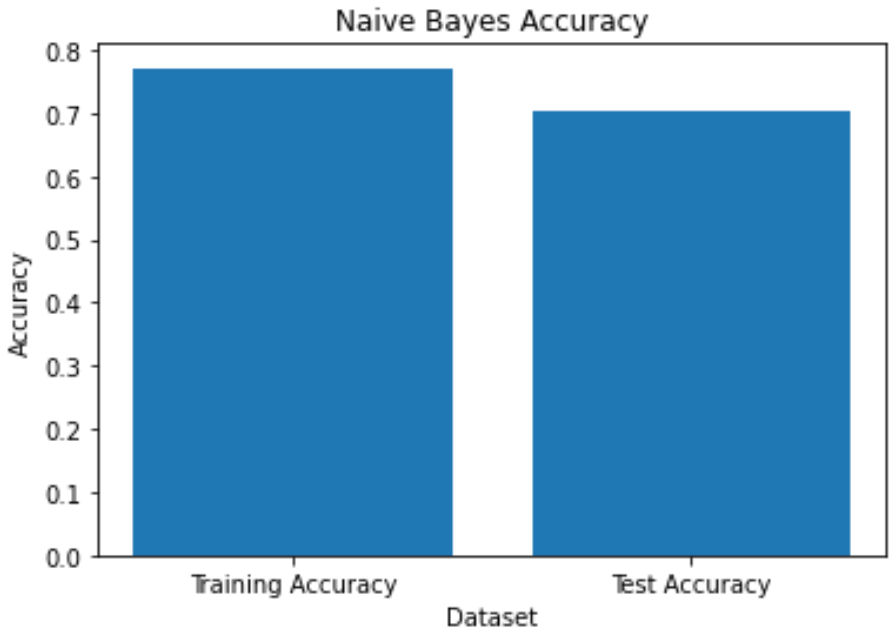
```
Iterasi 149: 0.6966562173458725
Iterasi 150: 0.6955067920585162
150
Rata-rata skor akurasi: 69.71 %
Skor tertinggi: 70.41 %
```

Gambar 4. 19 Hasil Evaluasi

Hasil dari evaluasi ini cukup baik, karena tidak terlalu jauh dengan akurasi pada modelling, jika hasil jauh, seperti 60% maka bisa dicurigai data tersebut underfitting.

selain melakukan evaluasi dengan MCCV, kita juga harus membandingkan hasil prediksi data latih dan data uji, hasil dari pengujian data latih mengasilkan

akurasi sebesar 77% tidak jauh dengan data uji yakni 70,38%, data dinyatakan overfit jika data latih dan data uji memiliki jarak yang relatif jauh bisa hingga selisih 20%. gambar barplot perbedaan data uji dan data latih ada pada gambar 4.20



Gambar 4. 20 Perbadingan data uji dan data latih

IV.6 Deployment

Deployment pada penelitian ini yakni menggunakan streamlit sebagai library untuk membuat web sederhana dengan python yang bermodalkan format pickel dari data latih. code pembuatan model dapat dilihat pada gambar 4.21.

```
model_data = {
    'model': naive_bayes,
    'vectorizer': vectorizer
}

with open('trained_model1.pkl', 'wb') as file:
    pickle.dump(model_data, file)
```

Gambar 4. 21 Save model format Pickle

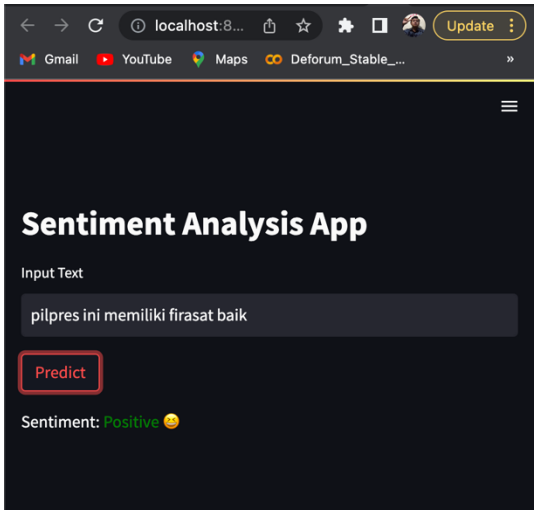
Variabel yang ada pada model_data dapat dilihat pada gambar 4.16 dimana naive_bayes ialah MultinomialNB pada fungsi modelling dan vectorizer pada model menggunakan CountVectorizer.

Selanjutnya ialah membuat code streamlit yang ada pada gambar 4.12.

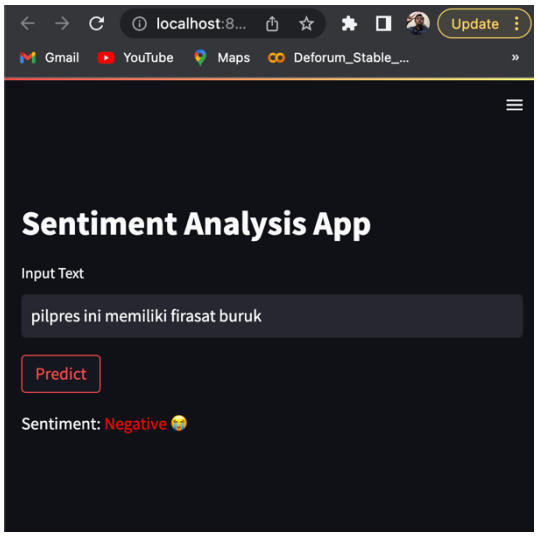
```python
import streamlit as st
import pickle
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory #stopword remover
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory #stemming


# Load model dan objek CountVectorizer dari file pickle
with open('trained_model.pkl', 'rb') as file:
    model_data = pickle.load(file)

#

model = model_data['model']
vectorizer = model_data['vectorizer']
stemmer = StemmerFactory().create_stemmer()
stopword_remover = StopWordRemoverFactory().create_stop_word_remover()
# Fungsi untuk melakukan prediksi sentimen
def predict_sentiment(text):
    stem_text = stemmer.stem(text)
    clean_text = stopword_remover.remove(stem_text)
    text_vectorized = vectorizer.transform([clean_text])
    prediction = model.predict(text_vectorized)
    return prediction[0]

# Tampilan aplikasi dengan Streamlit
st.title("Sentiment Analysis App")

# Input teks
text = st.text_input("Input Text")

# Tombol untuk melakukan prediksi
if st.button("Predict"):
    if text:
        sentiment = predict_sentiment(text)
        if sentiment == 'Positive':
            st.write("Sentiment:", "<span style='color:green;'>Positive 😃</span>", unsafe_allow_html=True)
        elif sentiment == 'Negative':
            st.write("Sentiment:", "<span style='color:red;'>Negative 😡</span>", unsafe_allow_html=True)
        else:
            st.write("Sentiment:", "<span style='color:white;'>Neutral 😅</span>", unsafe_allow_html=True)
    else:
        st.write("Please input text.")
```

Gambar 4. 22 Code streamlit

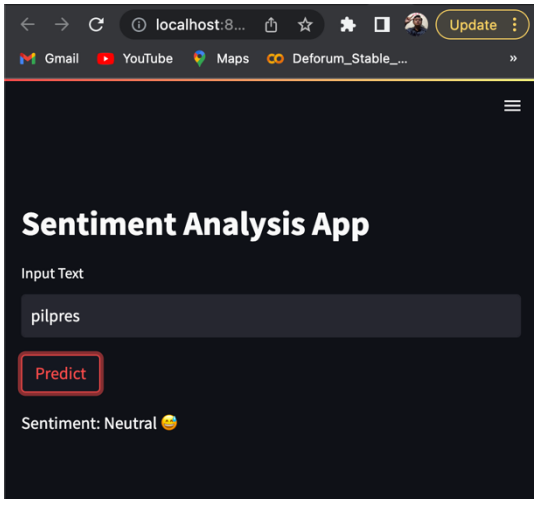
Gambar diatas menunjukkan model di import, lalu load value yang ada pada model tersebut yakni model dan vectorizer. Gambaran *output* dari code streamlit dapat dilihat pada gambar 4.23 yang menjelaskan gambar positif, 4.24 menjelaskan gambar negatif dan 4.25 menjelaskan gambar netral.



Gambar 4. 23 Hasil positif



Gambar 4. 24 Hasil negatif



Gambar 4. 25 Hasil netral

BAB V SIMPULAN DAN SARAN

V.1 Simpulan

<Sampaikan simpulan sesuai dengan point-point tujuan penelitian>

V.2 Saran

<Sampaikan saran-saran yang bisa dilakukan oleh pengguna atau peneliti selanjutnya untuk menyempurnakan penelitian ini>

DAFTAR PUSTAKA

PROSIDING

Albarda (2010): Portal Kantor Digital, *Proceeding Seminar TIK Nasional ke-6*., STEI Institut Teknologi Bandung.

JURNAL

Ali, N., Babar, M.A. (2009):Modeling Service Oriented Architectures of Mobile Applications by Extending SoaML with Ambients, *IEEE Computer Society vol. 978-0-7695-3784-9/09*

LINK

Budi, Setiawan, "*Panduan Penggunaan Notasi UML dalam Pengembangan Aplikasi GIS*", http://setiawanbudi.blogspot.com/artikel?17, diambil pada tanggal 17 Agustus 2018 pukul 19.00 WIB

BUKU

Creswell, J.W. (2003): *Research Design Qualitative, Quantitive, and Mixed Methods Approaches*, Sage Publications, Second Edition, London

JURNAL

Dörner, D., Draxler,S., Pipek, V., and Wulf, V.(2009): End Users at the Bazaar: Designing Next-Generation Enterprise Resource Planning Systems, *IEEE Software vol. 0740-7459/09*

BUKU

Erl, Thomas. (2005): *Service Orienterd Architecture,* Prentice Hall, New Jersey.

JURNAL

Fei, H.R., Xiang, W.H., Qian, X., Pei, J.X., Hui, L. (2009): Service oriented decentralized access control for military systems in Net-Centric Environment, *IEEE Computer Society vol. 978-0-7695-3643-9*

BUKU + EDITOR

Sugiman, A., Djubiantono, T., Aziz, F., Subandi, J.S., Wagiman, R. A., Aceng, D.C., Hardi, W.E., Elan, J.M., Darsono, A.C, dan Agus (2015) : Pengukuran Kadar Gangguang Sinyal GPS Menggunakan Model Bernoulli, *Jurnal Telematika,* 48, 661-667.

DISERTASI/ TESIS/ SKRIPSI

Wijaya, R. (1996) : *Diagnosis Penyakit Tipus dengan Metode PCR,* Disertasi Program Doktor, Universitas Langlangbuana, 25 – 29.

LINK

Zuckerberg, Mark, "How to Obtain Accurate GPS Location With Android OS", *IEEE Online Journal, Edition IX/18, 2018*, http://ejournal.ieee.org/artickel?1123, diambil pada tanggal 20 Semptember 2018 pukul 19.00 WIB

Sumber YouTube

Galih, Sandika, "*Belajar NodeJS | 1. Apa Itu NodeJS*?", Tutorial Web Programming UNPAS, https://youtu.be/sSLJx5t4OJ4?t=687, diambil pada tanggal 20 Februari 2021 pukul 19.00 WIB

Sumber Wikipedia

gcuomo (24 October 2013). "*JavaScript Everywhere and the Three Amigos* (Into the wild BLUE yonder!)". www.ibm.com, https://en.wikipedia.org/wiki/Node.js#cite_note-6, diambil pada tanggal 20 Februari 2021 pukul 19.00 WIB

LAMPIRAN A: PYTHON NOTEBOOK SCRAPPING

LAMPIRAN B: PYTHON NOTEBOOK CLEANING DATA

LAMPIRAN C: PYTHON NOTEBOOK SENTIMEN ANALISIS