ANALISIS SENTIMEN PEMILIHAN PADA PEMILU 2024 MELALUI TWITTER: PENDEKATAN TEXT MINING DAN KLASIFIKASI K-MEANS

SKRIPSI

DISUSUN OLEH

AHMAD RAIHAN LUBIS

2009010050

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

MEDAN

2024

ANALISIS SENTIMEN PEMILIHAN PADA PEMILU 2024 MELALUI TWITTER: PENDEKATAN TEXT MINING DAN KLASIFIKASI K-MEANS

SKRIPSI

Disajikan sebagai salah satu syarat untuk meraih gelar Sarjana Ilmu Komputer (S.Kom) pada program Studi Sistem Informasi Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara

AHMAD RAIHAN LUBIS

2009010050

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

MEDAN

2024

LEMBAR PENGESAHAN

Judul Skripsi	:	ANALISIS SENTIMEN PEMILIHAN PADA PEMILU 2024 MELALUI TWITTER: PENDEKATAN TEXT MINING DAN KLASIFIKASI K-MEANS
Nama Mahasiswa	:	Ahmad Raihan Lubis
NPM	:	2009010050
Program Studi	:	Sistem Informasi

Menyetujui

Komisi Pembimbing

(Fatma Sari Hutagalung S,Kom.,M.Kom)

NIDN. 0117019301

Ketua Prodi	Dekan
(Martino,S.Pd.,S.Kom.,M.Kom)	**(Dr. Al – Khowarizmi,S.kom.,M.Kom)**
NIDN. 0128029302	NIDN. 0127099201

PERNYATAAN ORISINALITAS

ANALISIS SENTIMEN PEMILIHAN PADA PEMILU 2024 MELALUI TWITTER: PENDEKATAN TEXT MINING DAN KLASIFIKASI K-MEANS

SKRIPSI

Saya menyatakan bahwa artikel ini adalah karya saya sendiri kecuali beberapa kutipan dan ringkasan yang mengakui masing- masing sumber .

Medan , juni 2024

Yang membuat pernyataan

d Raihan Lubis

NPM.2009010050

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMI

Sebagai sivitas akademika Universitas Muhammiyah Sumatera Utara, saya berlangganan di bawah ini:

Nama	: Ahmad Raihan Lubis
NPM	: 2009010050
Program Studi	: Sistem informasi
Karya Ilmiah	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Muhammadiyah Sumatera Utara Hak Bedas Royalti Non-Eksekutif (Non-Exclusive Royalty free Right) atas penelitian skripsi saya yang berjudul:

ANALISIS SENTIMEN PEMILIHAN PADA PEMILU 2024 MELALUI TWITTER: PENDEKATAN TEXT MINING DAN KLASIFIKASI K-MEANS

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non- Eksekutif ini, Universitas Muhammadiyah Sumatera Utara berhak menyimpan, mengalih media, memformat, mengelola dalam bentuk database, merawat dan mempublikasikan Skripsi saya ini tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemegang dan atau sebagai pemilik hak cipta..
Demikian pernyataan ini dibuat dengan sebenarnya.

Medan, Juni 2024

Penulis

Ahmad Raihan Lubis

RIWAYAT HIDUP

DATA PRIBADI

Nama Lengkap	: Ahmad Raihan Lubis
Tempat dan Tanggal Lahir	: Medan, 24-Agustus-2002
Alamat Rumah	: Merelan Raya,Tanah Enam Ratus No.16
Telefon/hp	: 0822-8459-9591
Email	:ahmadraihan24082@gmail.com
Instansi Tempat Kerja	: -
Alamat Kerja	: -

DATA PENDIDIKAN

SD	: SDN 060877 MEDAN	TAMAT : 2014
SMP	: SMPN 38 MEDAN	TAMAT : 2017
SMA	: SMK TRITECH INFORMATIKA	TAMAT : 2020

KATA PENGANTAR

Assalamualaikum.wr.wb.

Puji dan syukur penulis ucapkan kepada Allah SWT atas limpahan berkat, rahmat,serta kemudahan yang sudah diberikan sehingga penulis dapat menyelesaikan Proposal Penelitian ini yang ialah syarat untuk menerima gelar Sarjana Komputer pada Program Studi Sistem Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara. Tak lupa juga shalawat serta salam kepada Nabi Muhammad SAW yang sudah memberi petunjuk kepada kita ke jalan yang lurus. Dalam kurun waktu pengerjaan Proposal Penelitian ini penulis menyadari bahwasanya sangat banyak pihak yang berjasa turut menolong penulis dalam penyelesaian Proposal Penelitian ini. Dalam kesempatan ini penulis mengucapkan terimakasih kepada:

1. Bapak Prof.Dr.Agussani,M.AP, selaku Rektor Universitas Muhammadiyah Sumatera Utara.
2. Bapak Dr. Al-Khowarizmi,S.Kom.,M.Kom, selaku Dekan Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara.
3. Bapak Halim Maulana,S.T.,M.Kom, selaku Wakil Dekan I Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara.
4. Bapak Lutfi Basit,S.Sos.,M.I.Kom, selaku Wakil Dekan III Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara.
5. Bapak Martiano,S.Kom.,M.Kom,selaku Kepala Program Studi Sistem Informasi Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara.
6. Ibu Yoshida Sary,S.E.,S.Kom.,M.Kom selaku Sekretaris Program Studi Sistem

Informasi Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara.

7. Ibu Fatma Sari Hutagalung, S.kom.,M.Kom. selaku dosen pembimbing yang sudah meluangkan waktu membimbing penulis selama pengerjaan Proposal Penelitian ini.
8. Bapak,Ibu Dosen dan Staff pengajar Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara.
9. Kedua Orangtua, ibu dan ayah saya serta kakak,abang ,adik dan keluarga besar saya yang turut mendukung dan memberikan saya semangat.
10. Staff Biro dan Pegawai Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara.
11. Dan juga teman-teman dan orang-orang di sekeliling saya yang selalu memberikan dukungan dan semangat dalam pengerjaan skripsi ini.
12. Terakhir, teruntuk Mutia Faizah,S.P.si yang sudah menemani dan memberikan semangat serta menolong saya dalam menuntaskan skripsi ini.

Penulis menyadari bahwasanya skripsi ini masih belum sempurna. Oleh sebab itu untuk menyempurnakan skripsi ini, kritik dan saran yang membangun sangat penulis harapkan. Akhir kata penulis berharap semoga skripsi ini dapat bermanfaat.

Medan, Juni 2024

Penulis

Ahmad Raihan Lubis

.

ANALISIS SENTIMEN PEMILIHAN PADA PEMILU 2024 MELALUI TWITTER: PENDEKATAN TEXT MINING DAN KLASIFIKASI K-MEANS

ABSTRAK

Penelitian ini mengkaji opini sentimen terkait pemilih pada pemilu 2024 dengan menggunakan analisis data dari media sosial Twitter. Dengan menggunakan pendekatan text mining dan klasifikasi, penelitian ini mengekstrak informasi berharga dari tweet yang berisi kata kunci terkait pemilu 2024. Proses pengumpulan data dilakukan dengan teknik scraping, dimana tweet dikumpulkan dalam jangka waktu tertentu untuk memastikan representasi yang lengkap.Setelah data terkumpul, dilakukan preprocessing untuk membersihkan dan menyiapkan teks, yang meliputi langkah-langkah seperti tokenaize,stopword dan lebeling. Analisis sentimen kemudian digunakan untuk mengkategorikan tweet menjadi sentimen positif, negatif, atau netral. Algoritme K-Means digunakan untuk mengumpulkan data opini guna membantu mengidentifikasi pola dan tren persepsi publik terhadap kandidat dan isu politik.Hasil analisis menunjukkan adanya distribusi opini yang signifikan di antara kandidat dan isu yang berbeda, sehingga mengungkap dinamika opini publik yang kompleks. Hasil-hasil ini memberikan pemahaman mendalam kepada para pembuat kebijakan, kandidat politik, dan peneliti tentang bagaimana opini publik terbentuk dan bagaimana opini tersebut dapat dipengaruhi selama kampanye pemilu. Selain itu, penelitian ini menyoroti potensi besar penerapan teknologi penambangan teks dan algoritma

Kata Kunci: Analisis Sentimen,Pemilu 2024,Twitter,Text Mining,Klasifikasi,K-means.

ANALISIS SENTIMEN PEMILIHAN PADA PEMILU 2024 MELALUI TWITTER: PENDEKATAN TEXT MINING DAN KLASIFIKASI K-MEANS

ABSTRAK

This research examines opinion sentiment regarding voters in the 2024 election using data analysis from the social media Twitter. By using a text mining approach, this research extracts valuable information from tweets containing keywords related to the 2024 election. The data collection process was carried out using a scraping technique, where tweets were collected over a certain period of time to ensure complete representation After the data is collected, preprocessing is carried out to clean and prepare the text, which includes steps such as tokenaize, stopword and labeling. Sentiment analysis is then used to categorize tweets into positive, negative, or neutral sentiment. The K-Means algorithm is used to collect opinion data to help identify patterns and trends in public perception of political candidates and issues. The results of the analysis show that there is a significant distribution of opinions between different candidates and issues, thus revealing the complex dynamics of public opinion. These results provide policymakers, political candidates, and researchers with a deep understanding of how public opinion is formed and how it can be influenced during election campaigns. Additionally, this research highlights the great potential of applying text mining technologies and algorithms

Keywords:Sentiment Analysis, 2024 Election, Twitter, Text Mining,Klasifikasi,K-means.

MOTTO

ITAMI O KANJIRO!, ITAMI O KANGAERO!, ITAMI O UKETORE!, ITAMI O SHIRE!. ITAMI O SHIRANU MONO NI, HONTOU HO HEIWA WAKARAN!. KOKO YORI SEKAI NI ITAMI O! SHINRA TENSEI !!!"

"Rasakanlah kepedihan!, pikirkanlah kepedihan!, terimalah kepedihan!, ketahuilah kepedihan!, orang yang tidak tahu kepedihan tidak akan mengerti kedamaian yang sbenarnya. Dari sini, dunia harus menerima kepedihan!"

-Pain Akatsuki-

" Orang lain ga akan bisa paham struggle dan masa sulit kita, yang mereka ingin tahu hanya bagian success storiesnya aja, jadi berjuanglah untuk diri sendiri meskipun gak akan ada yang tepuk tangan, kelak diri kita di masa depan akan sangat bangga dengan apa yang kita perjuangkan hari ini "

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1. Latar Belakang

Indonesia merupakan salah satu negara di berbagai penjuru dunia yang menerapkan sistem demokrasi untuk menjalankan negaranya. Sistem pemerintahaan nya di selenggarakan dari rakyat,oleh rakyat , dan untuk rakyat . proses Demokrasi di Indonesia telah mengalami perkembangan yang signifikan sejak reformasi pada tahun 1998, di mana pemerintahan otoriter digantikan dengan sistem demokrasi yang memberikan suara kepada rakyat.

Landasan konsep pada demokrasi di Indonesia adalah pancasila yang secara tekstual terdapat di pembukaan UUD khususnya sila ke-4, yaitu "Kerakyatan yang dipimpin oleh hikmat kebijaksanaan dalam permusyawaratan/perwakilan". Konsep ini menegaskan bahwa kekuasaan dalam negara berada di tangan rakyat, yang diwujudkan melalui proses musyawarah dan perwakilan. Oleh karena itu, sila ke-4 menjadi pondasi yang kuat dalam membangun sistem politik Indonesia yang demokratis, adil, dan berkeadilan. Hal ini menunjukan pentingnya partisipasi aktif rakyat dalam menentukan arah dan kebijakan negara melalui pemilihan umum(pemilu).

Pemilu menjadi mekanisme utama di mana rakyat secara langsung maupun melalui perwakilan mereka memilih para pemimpin dan wakil-wakilnya. Dengan demikian, proses pemilu menjadi sarana utama untuk mewujudkan prinsip kerakyatan yang diamanatkan dalam sila keempat Pancasila. Pemilu di Indonesia mencerminkan keberagaman politik, sosial, dan budaya yang kaya dalam konteks demokrasi yang dinamis. Sebagai negara dengan populasi yang besar dan ragam suku, agama, dan kepentingan politik, pemilu di Indonesia menjadi panggung bagi beragam partai politik dan calon untuk bersaing memperoleh dukungan dari masyarakat. yang telah menghasilkan sistem pemilu yang semakin terbuka dan inklusif.

Keberadaan pemilu berpengaruh pada kemunculan berita yang ada pada media masa.

Isu terkait hal tersebut menjadi pembicaraan dan headline yang populer. Pada penelitian ini analisa yang dilakukan pada berita yang dibuat dan diunggah oleh salah satu kelompok pers yang terkenal yaitu Detik.com. data yang diambil melalui scraping twitter dari akun resmi Detik.com. Data tersebut akan dianalisa untuk melihat trend dan pola yang ada sehingga informasi yang berguna bisa didapatkan dari hal tersebut. Proses analisa akan mengacu pada text mining untuk memproses opini yang terjad. Dan sementara , mangacu analisis kasifikasi berita ketiga calon presiden dan wakil presiden dan .

Text mining dalam proses Analisis Opini Pemilu 2024 dengan Twitter adalah tentang memproses teks tidak terstruktur yang terdapat dalam tweet dan mengekstraksi informasi berharga darinya. Proses penambangan teks ini melibatkan beberapa langkah seperti pembersihan data, penandaan, ekstraksi dan pembobotan kata menggunakan teknik seperti TF-IDF (Term Frekuensi-Invers Dokumen Frekuensi). Dengan bantuan penambangan teks, data mentah tweet diubah menjadi format yang dapat dianalisis secara kuantitatif, yang memungkinkan untuk mengenali pola, tren, dan emosi yang tersembunyi di dalam teks. Hasil text mining menyediakan fitur-fitur penting yang digunakan untuk menentukan sentimen suatu tweet pada langkah klasifikasi.

Klasifikasi Dalam proses analisis sentimen pemilih pada pemilu 2024 melalui Twitter, algoritma K-means bertujuan untuk mengelompokkan tweet berdasarkan sentimen yang dikandungnya. Fungsi utama klasifikasi ini adalah untuk membagi tweet ke dalam kategori sentimen berbeda seperti positif, negatif, dan netral, sehingga memungkinkan analisis lebih mendalam terhadap opini publik yang muncul. Dengan mengidentifikasi opini dari tweet, peneliti dapat memahami preferensi pemilih, tren opini, dan tanggapan terhadap kandidat atau isu yang berbeda, yang dapat menjadi informasi berharga untuk strategi kampanye dan pengambilan keputusan kebijakan.

Dalam hal ini keberadaan pemilu sebagai salah satu ajang kontestasi terbesar dalam pemilihan pemimpin tentu menarik minat untuk dibahas. Kemunculan berita terhadap pasangan calon tentu mempengaruhi jalannya kontestasi dan sangat menarik untuk diteliti. Pasangan calon merupakan tokoh publik yang mendapar sorotan sangat besar pasa kontestasi kali ini. Sebagai tokoh publik yang mendapat sorotan besar oleh masa sudah tentu berita dan kabar dari mereka sangat ditunggu dan tentu adanya kemunculan berita tersebut akan memberikan dampak yang signifikan terhadap proses-prose tertentu.

Analisis sentimen pada pemilihan pemilu 2024 melalui Twitter menggunakan pendekatan Text Mining dan klasifikasi Algoritma K-Means bertujuan untuk mengelompokkan tweet berdasarkan pola sentimen yang terkandung di dalamnya. K-Means digunakan sebagai metode clustering untuk memisahkan tweet-tweet ke dalam sebuah kelompok-kelompok yang memiliki sebuah karakteristik sentimen yang sama, memungkinkan pemahaman yang lebih mendalam terhadap pandangan dan opini pemilih terkait kandidat dan isu-isu yang relevan dalam pemilihan tersebut.

Untuk menganalisa sentiment tersebut melalui pendekatan text mining dan klasifikasi mengunakan algoritma K-Means. Dalam pendekatan sentiman masyarakat terkait isu-isu pemilu yang dapat digunakan untuk menganalisis sentimen masyarakat. Diantaranya ada bebrapa latar belakang Dalam penelitian ini , Mailoa (2019). dengan judul Analisis Sentimen Data Twitter Mengunakan Metode *Text Mining* Tentang Masalah Obesitas di Indonesia menunjujan bahwa metode *text mining* lebih di dominasi oleh sentiman positif dengan nilai akurasi algoritma *Naive Nayes Classifier* berada dalam kategori “Excellent” atau baik. (Anni, 2020) dengan berjudul Analisis Sentiman Pandemi Covid-19 Pada *Streaming* Twitter Pada *Text Mining Pyton* menunjukan bahwa sentimen netral paling tinggi di bandingkan dengan sentimen negatif atau positif. (Dianati, 2022) dengan judul “Analisis Sentimen

Kinerja Dewan Perwakilan Rakyat(DPR) Pada Twitter Mengunakan Metode *Haive Bayes Classifer*"dalam penelitian ini analisis sentimen masyarakat terhadap kinerja dewan perwakilan rakyat. (Kahfi, 2017), dengan judul "Analisis Sentimen Komentar Kebijakan *Full Day School* (Fds) Dari *Facebook Page* Kemendikbud RI Menggunakan Algoritma *Naïve Bayes Classifie".*

Berdasarkan latar belakang di atas penulis memutuskan untuk melakukan penelitian dengan menggunakan judul "**Analisis Sentimen Pemilih pada Pemilu 2024 melalui Twitter: Pendekatan Text Mining dan Algoritma K-Means**". Dalam konteks ini, penelitian yang berfokus pada analisis sentimen terhadap pemilu yang di lakukan di indonesia pada 2024 memiliki potensi untuk memberikan wawasan yang berharga bagi masyarkat terkait terjadinya pemilu di Indonesia

1.2. Permasalahan Peneltian

Permasalahan utama yang ingin coba diselesaikan dalam proses penelitian ini adalah bagaimana melakukan analisis sentimen terhadap pemilihan pada Pemilu 2024 melalui Twitter dengan menggunakan pendekatan text mining dan klasifikasi beralgortima K-Means.

1.3. Batasan Masalah

Adapun batasan masalah dari penelitian ini di antaranya:

1. Data yang digunakan berasal dari platform Twitter dari akun Detik.com. Penelitian ini memanfaatkan tweet-tweet yang relevan dengan Pemilu 2024.
2. Fokus penelitian adalah pada penggunaan pendekatan text mining dan algoritma K-Means untuk menganalisis sentimen berita terkait dengan Pemilu 2024 di Twitter dari akun Detik.com.

1.4. Tujuan Penelitian

Adapun sebuah tujuan yang akan dicapai dalam proses penelitian ini adalah:

1. penelitian ini bertujuan untuk mengembangkan metode analisis sentimen yang dapat mengidentifikasi dan mengelompokkan berita yang muncul dari tweet-tweet yang terkait dengan Pemilu 2024 di Twitter dari akun Detik.com.
2. Mengimplementasikan algoritma K-Means untuk mengelompokkan sentimen-sentimen tersebut menjadi kategori yang berbeda.
3. Menerapkan peoses text mining untuk memproses dan menganalisis data tweet terkait Pemilu 2024.

1.5. Manfaat Penelitian

Adapun manfaat yang hasilkan dari penelitian ini di antaranya:

1.Bagi Mahasiswa

a. Mahasiswa akan memperoleh pengetahuan tentang metode analisis sentimen, text mining, dan algortima K-Means yang penting dalam pengolahan data teks dan analisis sentimen.

b. memperoleh pengalaman praktis dalam merancang,bmelaksanakan, dan menganalisis penelitian, yang merupakan keterampilan berharga untuk masa depan akademik atau profesional.

2.Bagi Masyarakat

a. Hasil dari penelitian dikiranya dapat memberikan informasi berharga kepada pemangku kepentingan, seperti Badan Pengawas Pemilu (Bawaslu) atau penyelenggara pemilu, untuk memahami dinamika opini dan sentimen pemilih di media sosial.

b. Analisis sentimen yang komprehensif memberikan landasan untuk membuat sebuah keputusan yang lebih baik lagi dalam merancang strategi kampanye politik, memahami isu-isu yang signifikan bagi pemilih, serta merumuskan kebijakan yang responsif terhadap aspirasi publik.

c. Sebagai sumber referensi pada penelitian selanjutnya yang berhubungan dengan Text mining dan Klasifikasi dari Algoritma K-Means

BAB II
LANDASAN TEORI

2.1. Analisis Sentimen

Sentimen analisis adalah proses penggunaan text analitic untuk mendapatkan berbagai sumber data dari internet dan beragam platfrom media sosial. Tujuannya adalah untuk memperoleh opini dari pengguna yang terdapat pada platform tersebut.

Menurut (Liu ,2011) analisis sentimen memproses pada bidang luas pemrosesan bahasa alami, linguistik komputasi, dan text teks yang bertujuan untuk menganalisis opini, perasaan, penilaian, sikap, penilaian, dan perasaan seseorang tentang topik, produk, layanan, organisasi, atau individu tertentu. atau kinerja. Sedangkan menurut (Oktinas dan Willa, 2017). Analisis sentimen adalah suatu teknik atau metode yang digunakan untuk mengidentifikasi bagaimana suatu emosi diungkapkan melalui teks dan bagaimana emosi tersebut dapat diklasifikasikan sebagai emosi positif atau negatif. menurut (Saifuidin dan Irawan,2018) ,Analisis Sentimen adalah mengelompokan polaritas teks pada tingkat dokumen, kalimat atau fitur/aspek dan menentukan pendapat mana yang diungkapkan dalam dokumen. bersifat positif,negatif dan netral. Sedangkan menurut (Ghozal dan Sugiharto 2022), analisis sentimen adalah proses menentukan polaritas teks dalam suatu dokumen atau kalimat dan mengelompokkannya sehingga dapat didefinisikan kategori positif, negatif, atau perasaan netral. Saat ini, analisis sentimen banyak digunakan oleh para peneliti sebagai bagian dari penelitian ilmu komputer. Jajak pendapat juga dapat dibandingkan dengan penelitian opini karena berfokus pada opini positif atau negatif. Dalam analisis sentimen, terdapat penambangan data dilakukan untuk menganalisis, mengolah, dan mengekstrak data tekstual yang terkandung dalam suatu entitas seperti layanan, produk, orang, fenomena, atau topik tertentu. Proses analisis dapat mencakup peninjauan teks, forum, tweet atau blog, dan pre-prosessing data, termasuk proses

tokenization, stopword, penghapusan, stemming, deteksi sentimen, dan klasifikasi sentimen. (Laurensz & Sediyono, 2021).

sehingga sentiemen analysis dapat menyimpulkan emosional sedih,gembira,atau marah Dasar analisis sentimen adalah mengelompokkan sebuah teks menjadi kalimat dan dokumen kemudian menentukan apakah pendapat yang diungkapkan dalam kalimat atau dokumen tersebut positif, negatif atau netral. (Dehaff,2010).

Kita dapat mencari opini tentang produk, merek, atau orang dan menentukan apakah opini tersebut dipandang positif atau negatif:

a. Presepsi prodak baru
b. Presepsi merek
c. Menejemen reputasi

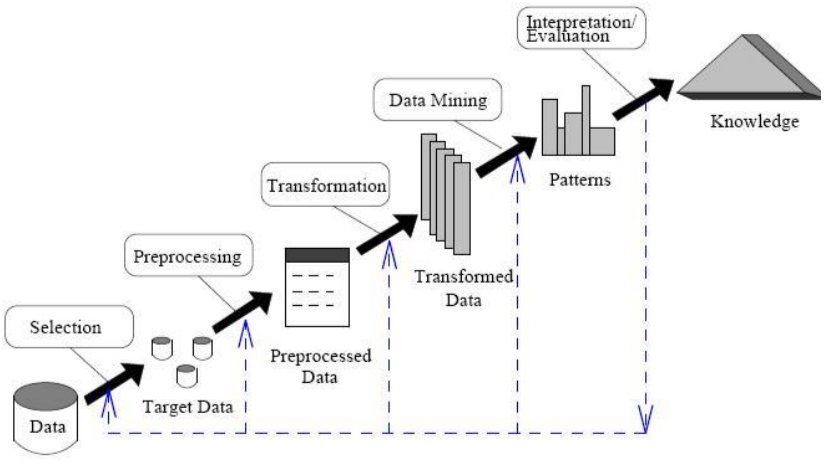
Ekspresi mengacu pada fokus topik tertentu, pernyataan terkait suatu topik dapat memiliki arti berbeda untuk pernyataan yang sama tentang berbeda. mata pelajaran Perilaku dapat merujuk pada alasan, pendapat atau penilaian, keadaan yang merugikan (bagaimana penulis ingin mempengaruhi pembaca). Sehingga dapat disimpulkan tujuan dari sebuah analisis sentimen adalah untuk menentukan apakah perilaku atau pendapat seorang peneliti yang berfokus pada topik tertentu.

2.2 Data Mining

Secara umum data mining terbagi menjadi dua (dua) kata: Data, yaitu sekumpulan fakta yang terekam atau suatu entitas yang tidak mempunyai arti dan saat ini diabaikan; dan mining adalah suatu proses penambangan, maka data mining dapat diartikan sebagai proses penambangan data yang menghasilkan output berupa data (Tan,P.N, dkk, 2015). Data mining merupakan kombinasi beberapa disiplin ilmu yang menggabungkan teknik dari pembelajaran mesin, pengenalan pola, statistik, database dan visualisasi untuk memecahkan masalah

penggalian informasi dari database yang besar (Mardi,2017). Data mining merupakan proses memperoleh informasi yang berguna dari suatu database yang besar untuk diekstraksi sehingga dapat diperoleh informasi baru darinya dan dapat membantu dalam pengambilan keputusan. Data mining adalah proses menganalisis data dari berbagai sumber dan merangkumnya menjadi informasi atau data atau pola yang berkaitan dengan peningkatan keuntungan, pengurangan biaya atau bahkan keduanya (Suntoro, 2019).Penambangan data adalah proses pencarian informasi berguna secara otomatis dari repositori besar. Teknik data mining digunakan untuk mengeksplorasi database besar sebagai sarana untuk menemukan pola baru dan berguna. Namun, tidak semua pekerjaan pencarian informasi dapat digambarkan sebagai penambangan informasi. (Witten & Frank, 2011).

Data mining adalah proses mencari informasi menarik dari data yang tersimpan di database, data berkuruan besar , warehouse atau lokasi penyimpanan data lainnya (Han dan Kamber, 2001) Sistem data mining disajikan dalam tabel ini:



Gambar 2.1. Sistem Data Mining

Data mining dapat diterapkan). untuk berbagai jenis database, seperti database relasional, gudang data, database peristiwa, database terkait objek dan objek, database spasial, data deret waktu dan data temporal, database teks dan multimedia, database heterogen dan warisan, dan WWW.

Knowledge mining mempunyai istilah lain yang memiliki arti serupa dengan Knowledge Mining, yaitu Knowledge Discovery in Database (KDD). Data mining dan KDD memiliki tujuan yang sama, yaitu menggunakan data yang ada di database kemudian mengolah data tersebut untuk mendapatkan informasi baru yang berguna. Selain itu, masih banyak istilah lain yang memiliki arti serupa dengan penambangan pengetahuan, seperti penemuan pengetahuan dalam database, penemuan pengetahuan dalam database, pengetahuan Pursua, analisis pola/data, arkeologi data, dan penambangan data. Banyak orang menganggap data mining sebagai sinonim untuk istilah lain yang umum digunakan, penemuan data pengetahuan atau KDD, sementara yang lain menganggap data mining hanya sebagai langkah penting dalam proses pencarian informasi (J. Tema, M.Kamber dan J.Pei 2012)

2.2.1 Tujuan dan Peran Data Mining

Tujuan dan peran data mining secara umum dapat dibagi menjadi dua kategori utama, yaitu

1. Prediktif. Tujuannya adalah untuk memprediksi atau memperkirakan nilai suatu atribut tertentu berdasarkan nilai atribut lainnya. Atribut yang diprediksi biasanya disebut dengan variabel target atau variabel terikat, sedangkan atribut yang digunakan untuk melakukan prediksi disebut dengan variabel penjelas atau variabel bebas. .
2. Deskriptif. Tujuannya adalah untuk mendapatkan pola (korelasi, tren, cluster, lintasan dan anomali) yang merangkum hubungan mendasar dalam data. Tugas penambangan data deskriptif sering kali bersifat eksploratif dan seringkali memerlukan teknik pasca-pemrosesan untuk mengonfirmasi dan memperjelas hasilnya. (Han dan Kamber, 2012).

2.2.2 Metode Data Mining

Menurut (Rerung, 2018) Berdasarkan tujuan dan Peran data mining dalam proses deskripsi data, tugas data mining dapat dibagi menjadi beberapa metode, antara lain:

a. Deskripsi, mendeskripsikan pola dan tren data

b. Estimasi, target bersifat numerik, bukan kategoris. Model dibangun dengan menggunakan kumpulan data (catatan) lengkap yang memberikan nilai variabel target sebagai nilai prediksi.

c. Prediksi. Nilai hasilnya ada di masa depan.

d. Klasifikasi. Yaitu variabel kategori sasaran.

e. Klaster,Array entri yang mirip satu sama lain tanpa variabel target.

f. Asosiasi Mencari atribut yang terlihat secara bersamaan.

2.3 Text Mining

Text mining adalah proses mengekstraksi informasi dari sebuah teks. Informasi biasanya diperoleh melalui prediksi model dan tren pembelajaran dari model statistik. Analisis penambangan teks, menambahkan fitur linguistik yang disimpulkan dan menghapus beberapa.lalu menambahkannya ke database, menentukan pola dalam data terstruktur, dan terakhir mengevaluasi dan menafsirkan keluarannya. Penambangan teks biasanya berarti kombinasi relevansi, kebaruan, dan minat (Oktinas & Willa 2017). Text mining adalah teknik yang digunakan untuk menyelesaikan masalah klasifikasi, pengelompokan, pelacakan informasi, dan pencarian informasi (Berry, M.W., & Kogan, J. 2010). Text mining adalah proses menganalisis data berupa teks. Text mining adalah teknik yang digunakan untuk membuat analisis data tidak terstruktur yang berbentuk teks. Ada dua langkah utama dalam analisis text mining, yaitu. pra-pemrosesan dan

integrasi data tidak terstruktur dan analisis statistik dari data tersebut, yang menurut Fransis dan Flynn telah diproses sebelumnya untuk mengekstrak konten dari teks. Sedangkan menurut buku Shallow Wiess, text mining adalah pengubahan data teks menjadi data numerik, atau dengan kata lain text mining mengubah data tidak terstruktur menjadi data terstruktur (Zaki Hariansyah, 2022).

Text mining adalah bidang penelitian baru dan menarik yang berupaya memecahkan masalah kelebihan informasi menggunakan penambangan data, pembelajaran mesin, pemrosesan bahasa alami (NLP), pengambilan informasi (IR), dan teknik manajemen informasi. Text mining mencakup tahapan pra-prosesing kumpulan dokumen, seperti klasifikasi teks, ekstraksi data, ekstraksi istilah. (Hasan & Wahyudi, 2018). Penambangan teks mengacu pada ekstraksi informasi dan pola tidak implisit, sebelumnya tidak diketahui, dan berpotensi berharga secara otomatis atau semi-otomatis dari data teks tidak terstruktur yang sangat besar, seperti teks bahasa alami (Hassani et al., 2020).

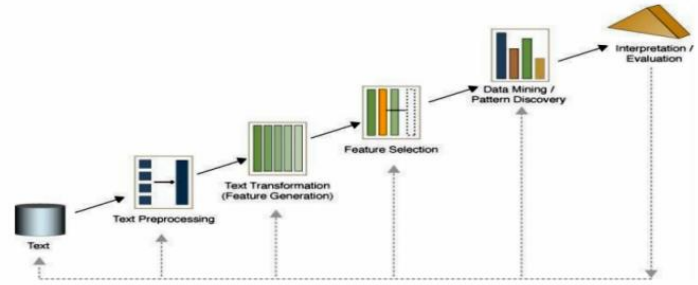
2.3.1 Sistem Text Klasifikasi

Dalam penambangan teks, sistem klasifikasi teks adalah proses peninjauan dan evaluasi yang dipantau secara ketat untuk memastikan keakuratan dan relevansi hasil. Sebuah proses Sistem klasifikasi teks yang merupakan suatu proses penentuan yang bertujuan untuk memisahkan dan mengelompokkan teks dapat dikategorikan baik secara langsung atau otomatis berdasarkan klasifikasi konten sistem.Proses ini melibatkan beberapa langkah, dimulai dengan pengumpulan data teks mentah, yang kemudian melewati langkah-langkah pra-pemrosesan seperti pembuatan token, pembuatan prospek, dan penghapusan ekor untuk mempersiapkan teks untuk analisis lebih lanjut. Setelah langkah prapemrosesan, teks yang dibersihkan dan disiapkan diekstraksi fiturnya,

biasanya sebagai vektor yang mewakili fitur penting teks.

Dalam penambangan teks, sistem klasifikasi teks adalah proses peninjauan dan evaluasi yang terkontrol. Sistem klasifikasi teks adalah suatu proses yang menentukan pemisahan dan pengelompokan teks secara langsung atau otomatis sesuai dengan klasifikasi isi teks yang disediakan oleh sistem. Berikut gambaran proses text mining (Firdaus dan Firdaus, 2021

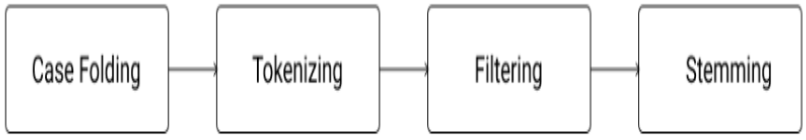
Algoritme ini dilatih menggunakan kumpulan data yang telah diberi label sebelumnya untuk mempelajari pola karakteristik dan fitur setiap kelas. Setelah model klasifikasi dilatih, model tersebut dapat diterapkan pada data teks baru dan mengklasifikasikan teks tersebut ke dalam kategori yang sesuai. Proses ini memungkinkan pengelompokan teks secara efisien dan akurat, baik tentang analisis opini, topik diskusi, atau kategori konten lainnya. Dengan sistem klasifikasi teks yang andal, penambangan teks dapat memberikan informasi berharga dan memfasilitasi pengambilan keputusan berdasarkan data di berbagai bidang, termasuk pemasaran, analisis media sosial, dan penelitian akademis.. Berikut ini adalah sebuah gambar proses dari perjalanan teks mining(ali firdaus,2021)



Gambar 2.2 Text mining proses

2.3.2 Text Preprocessing

Preprocessing teks merupakan sekumpulan proses yang harus dilakukan untuk mengolah dataset dalam suatu kumpulan sebagai data masukan. Beberapa proses yang dilakukan dalam text preprocessing yaitu tagging, stopword removal, dan detangling (Enda esyuda dan Bambang,2015).Bagian penambangan teks menggunakan prapemrosesan data untuk mengekstrak informasi menarik dan relevan dari data teks tidak terstruktur. Information retrieval (IR) digunakan untuk memutuskan dokumen mana dalam suatu koleksi yang harus dicari untuk memenuhi kebutuhan informasi pengguna (Gurusamy,V.,2014), sehingga keputusan pencarian diproses dengan membandingkan istilah pencarian dengan istilah indeks..



Gambar 2.3 Text preprosesing

Teks preprocessing adalah pembersihan, dan penyederhanaan potongan teks agar dapat diproses lebih lanjut.Berikut adalah tahapan dalam proses ini.menurut (sitti dan novi,2019).

1. Case Folding huruf artinya mengubah huruf kapital pada seluruh tinjauan dokumen data praktikum dan mengubah data ujian menjadi huruf kecil.

2. Tokenisasi berarti memecah kata kata demi kata menjadi satu kesatuan.

3. filtering adalah proses pembersihan dokumen dari kata-kata yang tidak diperlukan. Kata yang dibersihkan adalah kata yang mengandung entitas tweet seperti mention, retweet, hashtag dan link URL, serta simbol atau kode karakter numerik (noise text),

misalnya: ('>and!([1-5]+);')

4 .Stemming adalah proses dalam sistem IR yang mengubah kata dalam dokumen menjadi kata dasar dengan menggunakan langkah-langkah berikut (pratama dan enda, 2015)

1) Periksa apakah kata tersebut ada di kamus, jika ya kata tersebut ditemukan. Namun jika belum, lanjutkan ke langkah berikutnya.

2) Hilangkan akhiran kapital yaitu: "-lah", "-kah", "-ku", "-mu" atau "-nya".

3) Cari awalan dan akhiran yang tidak diperbolehkan, yaitu: ("esti-" dan "-i"), ("di-" dan "-an"), ("ke-" dan "-i, -kan"), ("mi" dan "-an"), ("se-" dan "-i, -kan").

4) Menghilangkan kata-kata seperti sufiks turunan yaitu: "-i", "-an", "-kan".

5) Hilangkan kata-kata seperti awalan awalan yaitu: "di-", "ke-", "se-", "te-". "me untuk menjadi saya-".

2.4. Klasifikasi

Klasifikasi adalah evaluasi dan penempatan objek data ke dalam kategori tertentu dari antara kategori yang tersedia. Klasifikasi membangun model berdasarkan data pelatihan yang ada dan kemudian menggunakan model tersebut untuk mengklasifikasikan data baru. Klasifikasi dapat didefinisikan sebagai tugas yang melakukan pelatihan/pembelajaran berdasarkan fungsi tujuan yang memetakan setiap kumpulan atribut (fitur) ke sekumpulan label kelas yang tersedia. Suatu sistem klasifikasi diharapkan dapat mengklasifikasikan seluruh kumpulan data dengan benar, namun tidak dapat dipungkiri bahwa kinerja sistem dapat 100% benar, sehingga sistem klasifikasi juga harus mengukur kinerjanya sendiri. Secara umum kinerja klasifikasi diukur menggunakan matriks konfusi. (Utomo.2020).

Klasifikasi merupakan Istilah yang mengacu pada suatu metode penyusunan secara sistematis atau menurut aturan atau praktik tertentu yang telah ditentukan sebelumnya.

Secara harfiah, klasifikasi dapat berarti membagi sesuatu ke dalam kategori-kategori. Menurut ilmu pengetahuan, klasifikasi adalah proses pengelompokan benda berdasarkan persamaan dan perbedaannya.(Irma devi.2016)Klasifikasi adalah proses menugaskan setiap item data ke kategori atau kategori yang telah ditetapkan sebelumnya. Dua proses penting yang dilakukan dalam proses klasifikasi yaitu pembelajaran (training) dan pengujian. Pada proses pembelajaran, model dilatih dengan cara menjalankan proses pembelajaran menggunakan data yang telah mempunyai kelas/pengidentifikasi, dan proses yang kedua adalah proses pengujian untuk menguji/memvalidasi model menggunakan data uji.(putra sugiarto,2020).Klasifikasi Proses penemuan pola (atau fitur) yang menggambarkan dan membedakan suatu kelas data atau konsep yang bertujuan untuk memprediksi kelas objek yang pengidentifikasi kelasnya tidak diketahui (Annur, 2018).

2.5 Clustering

Clustering bagain dari sebuah klasifikasi yang tidak terkendali (unsupervised classification).. Pengertian clustering adalah proses pengelompokan atau pengklasifikasian objek berdasarkan informasi dari data yang menjelaskan hubungan antar objek dengan prinsip memaksimalkan kemiripan antar anggota suatu kelas dan meminimalkan kemiripan antar kelas/cluster. Grup data mining berguna ketika Anda mencari pola distribusi dalam suatu kumpulan data yang berguna dalam proses analisis data. Kesamaan objek biasanya bermula dari kedekatan nilai atribut yang menggambarkan objek data, sedangkan objek data biasanya direpresentasikan sebagai titik-titik dalam ruang multidimensi.(rahman,2017).

Clustering disebut sebagai segmention. Metode ini mengidetifikasi kelompok dalam sebuah kasus yang didasarkan pada kelompok atribut yang memiliki kemiripan. Cara kerja clustering memisahkan sejumlah kelompok data berdasarkan ciri masing-masing, dimana

objeknya dapat berupa orang, peristiwa dan lainya yang didistribusikan ke dalam kelompok sehingga terdapat beberapa tingkatan yang saling berhubungan antar cluster, kuat dan lemahnya antar anggota dari cluster yang berbeda terlihat pada anggota cluster yang sama(N. Jannah & T. Yulianto,2016)

Menurut (Widodo 2013clustering, atau klasifikasi, adalah metode membagi suatu kumpulan data menjadi beberapa kelompok berdasarkan kesamaan yang telah ditentukan. Cluster adalah kumpulan atau kumpulan objek data yang sejenis satu sama lain dalam satu cluster yang sama dan berbeda dengan objek dalam cluster yang berbeda. Objek-objek dikelompokkan menjadi satu atau lebih cluster sehingga objek-objek dalam satu cluster mempunyai banyak kesamaan dengan yang lain. Objek-objek dikelompokkan berdasarkan prinsip memaksimalkan kemiripan objek-objek dalam satu cluster yang sama dan memaksimalkan perbedaan antar cluster yang berbeda. Kemiripan objek biasanya diperoleh dari nilai-nilai atribut yang menggambarkan objek data tersebut, sehingga objek data biasanya direpresentasikan sebagai. titik dalam ruang multidimensi. Pengelompokan ini memungkinkan kita untuk mengklasifikasikan wilayah padat. , menangani data yang berisik dan mudah diterjemahkan

2.6 Algoritma K-Means Clustering

K-Means merupakan salah satu metode pengelompokan data non-hierarki (partisi) yang dapat membagi data menjadi dua kelompok atau lebih. Metode ini membagi data menjadi satu bagian, dimana data yang sebuah mempunyai sebuah karakteristik yang sama ditempatkan pada sebuah kelompok yang sama dan data yang memiliki karakteristik berbeda pada kelompok yang lain untuk meminimalkan perbedaan intra kelompok dan memaksimalkan perbedaan antar kelompok (Adiya dan Desnelita, 2019). Algoritma K-Means merupakan salah

satu algoritma clustering yang termasuk dalam kelompok Unsupervised learning yang digunakan untuk mengelompokkan data menjadi beberapa kelompok dengan menggunakan sistem (A. Wanto Et Al. 2020). Algoritma K-Means merupakan salah satu teknik clustering berbasis jarak yang membagi data menjadi beberapa cluster, dan algoritma ini hanya bekerja pada atribut numerik atau numerik (J. Han, M.Kamber, dan J.Oei, 2010).

Menurutu Yulia & agus (2016) K-Means adalah metode pengelompokan data non-hierarki yang berupaya membagi data yang ada menjadi satu atau lebih cluster atau kelompok sehingga data dengan karakteristik yang sama dikelompokkan ke dalam satu cluster dan data dengan karakteristik berbeda ke dalam kelompok lain (pengelompokan berbasis jarak). metode yang membagi data menjadi beberapa cluster dan algoritma ini hanya bekerja dengan atribut numerik. Algoritma K-Means sangat terkenal dengan kesederhanaannya dan kemampuannya dalam mengumpulkan data berukuran besar dan outlier dengan sangat cepat. pada satu tahap pemrosesan cluster tertentu, ia berpindah ke cluster lain.

Algoritma k-means itu sendiri algoritma yang membagi data menjadi beberapa cluster sehingga satu cluster berisi data yang serupa dan cluster lainnya berisi data yang berbeda. Sarwono menjelaskan lebih lanjut algoritma pengelompokan K-Means (Rohmawati W,Defiyanti, & Jajuli, 2015):

1.Tetapkan jumlah kelompok yang akan dibentuk menjadi k.

2. Hasilkan k nilai acak untuk centroid awal (pusat) cluster.

3. Hitung jarak setiap data masukan dari setiap centroid dengan menggunakan rumus Euclidean distance hingga ditemukan jarak terdekat setiap data ke centroid. Berikut persamaan jarak Euclidian Distance:

$$d(xi, \mu j)$$

data kriteria

: centroid pada cluster ke-j

4. Klasifikasikan semua data berdasarkan kedekatannya dengan centroid (jarak terkecil)

5. Update nilai centroid. Nilai centroid baru diperoleh dari mean cluster yang bersangkutan dengan menggunakan rumus

sebagai parameter untuk menentukan klasifikasi data.

$$\mu j(t+1) = \frac{1}{} \sum_{j \in Nsj} \qquad Sj\ xj$$

Keterangan:

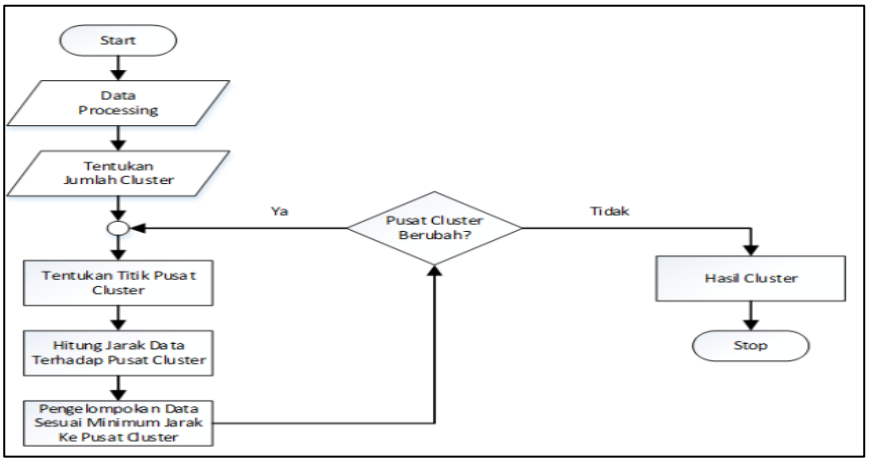
$(t + 1)$: *centroid* baru pada iterasi ke (t+1) Nsj :

banyak data pada *cluster* Sj.

6. Ulangi langkah 2-5 hingga tidak ada anggota cluster yang berubah.

Ketika langkah 6 selesai, nilai pusat cluster (μj) dari iterasi terakhir digunakan Jika langkah 6 telah terpenuhi, maka nilai pusat *cluster* (μj) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukanklasifikasi data.

Dengan implementasi data dengan metode K-Means Clustering diharapkan dapat membantu ngelompokan sentimen masyarakat terhadapa pemilu yang terjadi pada tahun 2024, sehingga masyarakat atau pihak-pihak lain yang melihat penelitian ini dapat lebih terbantu untuk meliahat sentimen masarakat terhadap pemilu terhadap analisis dan mengambil keputusan berdasarkan hasil pengelompokan tersebut, Berikut ini adalah Flowchart proses algoritma k-means Clustering :



Gambar 2.4 Flowchart algoritma K-means Clustering

2.7 Pemilu (Pemilihan Umum)

Pemilihan Umum Parlemen (Pemilu) sebagai sebuah proses politik yang dinamis hanya dapat berjalan sesuai keinginan dan lancar jika setiap pemilih mengikuti aturan main yang telah disepakati. Pemilu merupakan sarana persaingan jabatan politik dalam pemerintahan, berdasarkan seleksi formal terhadap warga negara yang berhak (Susanto, Asy‟ari, & Sri Hardjanto, 2016).

Pemilu merupakan sarana dimana masyarakat dapat berpartisipasi dalam menentukan arah penyelenggaraan pemerintahan. Ada pemilu yang pada akhirnya menjadi wahana legitimasi politik pemerintah yang berkuasa, karena melalui pemilu gagasan bahwa pemerintah memerlukan persetujuan dari yang diperintah dapat dikonsolidasikan. Tanpa pemilihan parlemen untuk memilih anggota parlemen, patut dipertanyakan apakah pemerintah menyatakan dirinya sebagai pemerintahan rakyat, meskipun pembentukannya tidak berdasarkan hasil pemilihan parlemen. Oleh karena itu penyelenggaraan pemilu sangatlah penting (John Morgan &Felix V´ardy.2012). Pemilu menurut Pasal 1 ayat (1) Undang-Undang Penyelenggara Pemilu Nomor 15 Tahun 2011 merupakan sarana pelaksanaan hak

rakyat untuk menentukan nasib sendiri dan harus dilaksanakan secara langsung, terbuka, bebas, rahasia, jujur, dan adil dalam Negara Kesatuan. Republik Indonesia (selanjutnya). disebut Negara Kesatuan Republik Indonesia) berdasarkan Pancas dan Undang-Undang Dasar Negara Republik Indonesia Tahun 1945 (selanjutnya disebut Undang-Undang Dasar Negara Republik Indonesia Tahun 1945).

2.8 Twitter

Twitter merupakan layanan media sosial yang sangat populer di kalangan penggiat media sosial. Twitter juga berguna sebagai alat untuk menyampaikan pendapat (seperti kritik), untuk kampanye politik, alat pembelajaran, alat komunikasi dan hiburan (S. R. Elisabet,2021)

Twitter adalah sebuah platform media sosial yang memungkinkan pengguna untuk membagi pemikiran, berita, gambar , video dan pesan dalam bentuk pendek memerlukan tweet. Meningkatnya penggunaan media sosial membuka sebuah peluang baru untuk menganalisis beberapa aspek dan model komunikasi. Misalnya, data media sosial dapat dianalisis untuk mendapatkan wawasan mengenai isu, tren, partai politik, dan jenis informasi lainnya. Twitter adalah alat yang sangat ampuh untuk berita, aktivisme, jejaring sosial, dan komunikasi online. Ini telah digunakan dalam banyak konteks, mulai dari kampanye politik hingga gerakan sosial, dan telah menjadi platform penting dalam ekosistem media sosial (A. Rivaldy,2021)Berikut adalah beberapa fitur yang ada pada *twitter:*

1. *Tranding topic* adalah fitur yang menampilkan topik atau pembahasan teratas berupa *hastag* yang banyak dibicarakan pengguna *twitter*.
2. *Hastag* adalah fitur yang dapat mengelompokkan *tweet* atau pesan.
3. *Retweet* adalah fitur untuk membagikan *tweet* dari pengguna lain.

4. *Following* adalah fitur untuk menghubungkan antar pengguna atau sering disebut teman.

Data twitter dapat diambil menggunakan aplikasi atau API yang disediakan oleh *twitter*. Jika dibandingkan dengan media sosial lainnya, tidak mudah untuk mengumpulkan data secara terbuka. Media sosial lainya tidak mengizinkan data akses karena kebijakan keamanan yang berbeda-beda.

twitter juga mempunyai beberapa kecocokan dengan data mining, sebagai berikut (Wandani, 2021):

1. Format data *twitter* yang cocok dan nyaman bagi peneliti untuk dianalisis
2. Peraturan *twitter* untuk data relatif fleksible jika dibandingkan dengan API lainnya.

Twitter mempunyai desain yang *user friendly* atau mudah diakses bagipenggunanya

2.9 Google Colab

Collaborative, atau lebih dikenal Colab adalah produk riset Google yang memungkinkan pengguna menulis dan menjalankan kode Python langsung melalui browser. Colab berbagai tujuan, mulai dari pembelajaran mesin, analisis data hingga pelatihan. Dengan Colab, pengguna dapat memanfaatkan kekuatan komputasi server Google tanpa harus menginstal perangkat lunak tambahan di komputer mereka. Selain itu, Colab mendukung penggunaan GPU secara gratis, yang sangat berguna untuk mempercepat pelatihan model pembelajaran mesin. Fitur seperti integrasi dengan Google Drive, kemampuan berbagi dan kolaborasi secara real-time, serta akses ke beberapa pustaka Python menjadikan Colab alat yang sangat canggih dan praktis bagi peneliti, ilmuwan data, dan pelajar yang ingin berinteraksi dan menguji kode mereka. cara yang terhubung dengan baik. Secara teknis, Colab

adalah layanan desktop Jupyter yang dihosting yang dapat digunakan tanpa instalasi dan menyediakan akses gratis ke sumber daya komputasi, termasuk GPU. Sumber daya Colab tidak dijamin dan bersifat terbatas, dan batas penggunaan dapat bervariasi. Hal ini diperlukan agar Colab dapat menyediakan sumber daya secara gratis. Pengguna yang menginginkan akses lebih andal ke sumber daya yang lebih baik dapat menggunakan Colab Pro. Pengenalan Colab Pro adalah langkah pertama yang diambil Google untuk melayani pengguna yang ingin melakukan lebih banyak hal dengan Colab. Sasaran jangka panjang Google adalah terus menawarkan Colab versi gratis sambil terus berkembang untuk memenuhi kebutuhan pengguna Google.

Salah satu fitur utama Google Colab adalah kemampuan berkolaborasi secara real-time. Mirip dengan berbagi dokumen di Google Docs, pengguna dapat berbagi buku catatan mereka dengan orang lain. Menurut revou.co(2020), berikut beberapa keuntungan utama menggunakan Google Colab:

Akses mudah ke sumber daya komputasi: Google Colab menyediakan akses gratis ke perangkat keras seperti GPU dan TPU. Pengguna dapat melakukan fungsi kerja yang memerlukan daya komputasi tinggi tanpa investasi infrastruktur yang mahal.

Tidak diperlukan konfigurasi: Google Colab dapat digunakan untuk menulis dan menjalankan kode Python langsung di notebook yang disertakan. Pengguna tidak perlu menyiapkan lingkungan pengembangannya sendiri, yang seringkali memakan banyak waktu.

Kolaborasi waktu nyata: Google Colab mendukung kolaborasi waktu nyata. Artinya, anggota tim dapat mengerjakan buku catatan yang sama pada waktu yang sama, sehingga memfasilitasi pertukaran ide dan kolaborasi proyek.

Integrasi dengan Google Drive dan GitHub: Google Colab terintegrasi dengan Google Drive dan GitHub. Pengguna dapat menyimpan pekerjaan mereka langsung ke cloud dan

mengakses proyek dari mana saja

Akses ke perpustakaan pembelajaran mesin: Google Colab memiliki banyak perpustakaan pembelajaran mesin seperti TensorFlow dan PyTorch. Pengguna dapat segera memulai proyek pembelajaran mesin mereka tanpa harus melalui proses instalasi dan konfigurasi perpustakaan.

2.9 Penelitian Terdahulu

Berikut adalah tabel penelitian terdahulu yang mendukung kerangka teoritis pada penelitian ini..

Tabel 2. 1 Tabel Penelitian Terdahulu

No	REFRENSI	Objek	Metode	Hasil Penelitisn
1	Sentimen Analisis pada Data Tweet Pengguna Twitter Terhadap Produk Penjualan Toko Online Menggunakan Metode K-Means	Penjualan toko online Terhadap penguna twitter	Metode k-means	Clustering berdasarkan kata kunci "mulaijadulu" dan "toppers" merupakan kata yang paling sering muncul dalam tweet sedangkan "gajian" dan "gratis ongkir" merupakan kata
2	Sentimen Analisis pada Data Tweet Pengguna Twitter Terhadap Produk Penjualan Toko Online Menggunakan Metode K-Means	Mengetahui Kalimat Positif maupun Negatif pada Buletin APTIKOM	Metode k-means	- Hasil Penelitian ini adalah text mining \ n dokumen negatif APTIKOM untuk menentukan analisis negatif. Kalimat positif dan negatif ditunjukkan pada Gambar 3. Proses terakhir adalah algoritma k-means yang menghasilkan pusat massajumlah kalimat netral dan jumlah kalimat positif dan negatif paling sedikit seperti yang terlihat pada Gambar 4 dan SSE sebesar 75.0 %
3	Automasi Penentuan Tren Topik Skripsi	Automasi Menentukan topik tren	k-means clustreing	Kesimpulan penelitian ini adalah sebagai berikut: aplikasi yang diusulkan dapat bekerja dengan baik

	Menggunakan Algoritma K-Means Clustering. Fuadi, Wahyu(2022)			dengan akurasi 84% pada 70 data uji.
4	Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi : (Sebuah Ulasan). Ali Firdaus(2021)	Dalam masalah inoformasi	*Text mining* *Pola algoritm a*	Kemudian beberapa model umum dijelaskan untuk memahami evolusi penambangan teks secara komprehensif. Dan terakhir, penulis mengklasifikasikan penambangan teks ke dalam analisis tren melalui klasifikasi teks, pengelompokan teks, aturan ekstraksi hubungan, dan
5	Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM) Tinegas,Rian (2020)	Analisis sentimen Layanan indihome	Klasifikasi Dengan support vector machine	Hasil pengujian memberikan keakuratan analisis pengujian sebesar 88,89% dengan 536 data set, 439 data positif dan 97 data negatif. Hal ini menunjukkan bahwa metode Multinominal Naive Bayes dapat digunakan untuk analisis opini data audit.
6	Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. Utomo D (2020)	Anlisis pada reduksi artibt oenyakit jantung.	Klasifikasi dan Principal Component naive baye-	Hasil proses klasifikasi dapat digunakan untuk membuat analisis algoritma C5.0. Tidak terdapat perubahan tingkat presisi dalam hal recall, presisi dan akurasi pada dataset penyakit jantung yang diperoleh sebelum reduksi atribut dan setelah reduksi atribut. Sedangkan akurasi yang diperoleh pada algoritma C5.0 sebesar 93,38%. Untuk algoritma Naive Bayes Classifier (NBC), selisih tingkat akurasi sebelum reduksi sebesar 99,01%, sedangkan

				setelah reduksi sebesar 98,53%. Hasil pengujian yang dilakukan menunjukkan bahwa algoritma Naive Bayesian Classifier (NBC) dapat menangani proses klasifikasi dataset penyakit jantung dengan lebih baik
7	Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia. Fridom Mailo (2019)	Masalah obesitas di indonesia Melalui twitter	Analisis sentime n tetx mining	Berdasarkan hasil analisis sentimen tweet menggunakan text mining, ditemukan bahwa sentimen positif lebih dominan sebanyak 22.246 (51,2%) tweet, disusul negatif sebanyak 12.015 (27,7%) tweet dan netral. opini 9174 (21,1%) dari total 43.435 tweet. Nilai akurasi algoritma pengklasifikasi Naive Bayes sebesar 94% dan masuk dalam kategori "Klasifikasi Sangat Baik" yang berarti algoritma pengklasifikasi Naive Bayes berhasil. memprediksi dengan baik kategori suasana hati penelitian ini..

BAB III
METODOLOGI PENELITIAN

3.1 Pendekatan Penelitian

Pada tahap ini Penelitian ini dilakukan untuk mencari tau hasil analisa sentiment dari pengguna *twitter* terkait fenomena pemilu (pemilihan umum) yang terjadi di indonesia selama 5 tahuan sekali yang diklasifikasikan berdasarkan sentiment positif dan negatif. Sehingga dari hal tersebut dapat diketahui apakah keberadaan pemilihan umum (pemilu) ini berdampak baik atau buruk bagi masyarakat.Teknik text mining digunakan untuk mengekstrak fitur dari teks, seperti frekuensi kata , untuk mendapatkan gambaran awal pola dan topik tweet. Selanjutnya algoritma K-Means digunakan untuk mengklasifikasikan tweet menjadi beberapa cluster berdasarkan kesamaan sentimen. Proses ini melibatkan pemilihan jumlah cluster yang optimal dan menghitung pusat massa setiap cluster. Hasil pemeringkatan ini dianalisis untuk mengidentifikasi sentimen dan tren dominan terhadap kandidat atau isu tertentu pada yang nantinya akan dianalisis guna mengetahui sentiment masyarakat terkait pemilu 2024 .

3.2 Teknik Pengumpulan Data

Dalam penelitian ini penulis mengumpulkan informasi dan pengetahuan yang mendukung proses penelitian, dimana proses pengumpulan datanya adalah sebagai berikut:

3.2.1 Studi pustaka

Pada tahap ini pengumpulan data tinjauan pustaka ini, penulis mengumpulkan berbagai teori terkait skripsi sebagai bahan untuk melengkapi penelitian ini. Sumber teori berasal dari berbagai referensi yang memberikan latar belakang konsep yang diperlukan, hasil penelitian terdahulu seperti jurnal dan tesis yang berkaitan dengan topik penelitian, serta artikel yang

mengkaji aspek-aspek tertentu dari penelitian yang dilakukan. Selain itu, penulis juga mengunjungi beberapa website yang menyediakan informasi dan metode penerapan menggunakan text mining, klasifikasi teks dan algoritma k-means clustering. Melalui tinjauan literatur ini, penulis bertujuan untuk memperoleh pemahaman yang luas dan komprehensif mengenai teknik dan metode yang digunakan dalam penelitian, memastikan bahwa pendekatan tersebut didasarkan pada praktik terbaik dan temuan terkini di lapangan. Proses pengumpulan data ini tidak hanya membantu mengidentifikasi kerangka teori yang kuat, namun juga memberikan wawasan praktis mengenai implementasi dan penerapan metode text mining dan klasifikasi beralgoritma k-means clustering untuk menganalisis opini masyarakat pada pemilu 2024 untuk melakukan penelitian. yang mempunyai dasar ilmiah yang kuat dan penting.

3.2.2 Scraping

Tujuan dari penelitian ini adalah untuk mengumpulkan data dari sumber media sosial Twitter untuk menjaring tanggapan dan komentar masyarakat terhadap hasil pemilu tahun 2024. Data dikumpulkan pada tanggal 1 Oktober 2023 hingga 30 Desember 2023. Sebanyak 9.192 data terkumpul selama ini. periode. berbagai topik dari "pilihan". Proses pendataan ini terfokus pada media sosial Twitter, karena pengguna platform ini dinilai aktif menyampaikan pandangan dan pendapatnya terhadap berbagai isu politik, termasuk hasil pemilu. Data yang dikumpulkan mencakup beragam pendapat, baik positif, negatif, maupun netral, serta memberikan gambaran menyeluruh mengenai persepsi masyarakat terhadap hasil pemilu 2024. Analisis terhadap data ini diharapkan bisa memberikan sebuah pemahaman yang lebih mendalam mancakup sentimen publik dan bagaimana isu pemilu dibahas di media sosial, yang pada akhirnya dapat menjadi masukan bagi penelitian lebih lanjut atau pengambilan kebijakan yang lebih sesuai dengan opini publik

3.3 Preprocessing

Preprocessing bertujuan untuk proses perubahan atau pembersihan data yang mentah menjadi sebuah data , yang di mana data tersebut masih memiliki noise di dalamnya. Maka dari itu, proses dari preprosesing ini untuk menghilangkan sebuah noise yang ada pada data. Dan juga ,proses pembersihan data dari preprosesing dengan Dokumen-dokumen yang beisi kebanyakan tidak banyak memiliki struktur target sehingga informasi yang dikandungnya tidak dapat diekstraksi secara langsung. Preprocessing sangat diperlukan ketika memilih pengolah kata sebagai indeks. Kata-kata yang mewakili dokumen yang nantinya digunakan untuk membuat model pengambilan informasi dan aplikasi text mining lainnya. Tahap Preprocessing pada penelitian ini terdiri dari beberapa bagian tahap di antara nya:

1. Case Folding

   Di sini, casefolding berfungsi untuk mengubah huruf besar menjadi huruf kecil. Proses ini diterapkan sedemikian rupa sehingga semua data yang diproses memiliki token yang seragam dan juga memudahkan langkah selanjutnya selama pemrosesan

2. Tokenizing

   Proses tokenisasi bekerja dengan cara memisahkan kalimat-kalimat menjadi satu kesatuan data. Pada langkah ini tokenisasi dilakukan untuk memudahkan peralihan ke langkah transformasi, karena kalimat yang tersedia tidak diproses, melainkan kalimat diproses kata demi kata.

3. Filtering

   Pengolahan data sudah selesai tahap Tokenizing, tahap selanjutnya adalah tahap filtering. Di sini, pemfilteran adalah proses menghilangkan kata-kata yang tidak berpengaruh. Kata-kata umum yang tidak sering muncul dan tidak mempunyai arti

disebut kata-kata stopword. Proses ini menggunakan dokumen yang berisi stopword yang digunakan dalam proses klasifikasi. Contoh kata yang tidak diperlukan adalah: apa, di dalam dan, dari, yang lain. dari.

4. Steaming

Pada tahap ini steaming berkerja untuk membersihkan imbuhan dalam sebuah kata yang terdapat pada awal, akhir ataupun kombinasi dari keduanya.

3.4 Lebeling

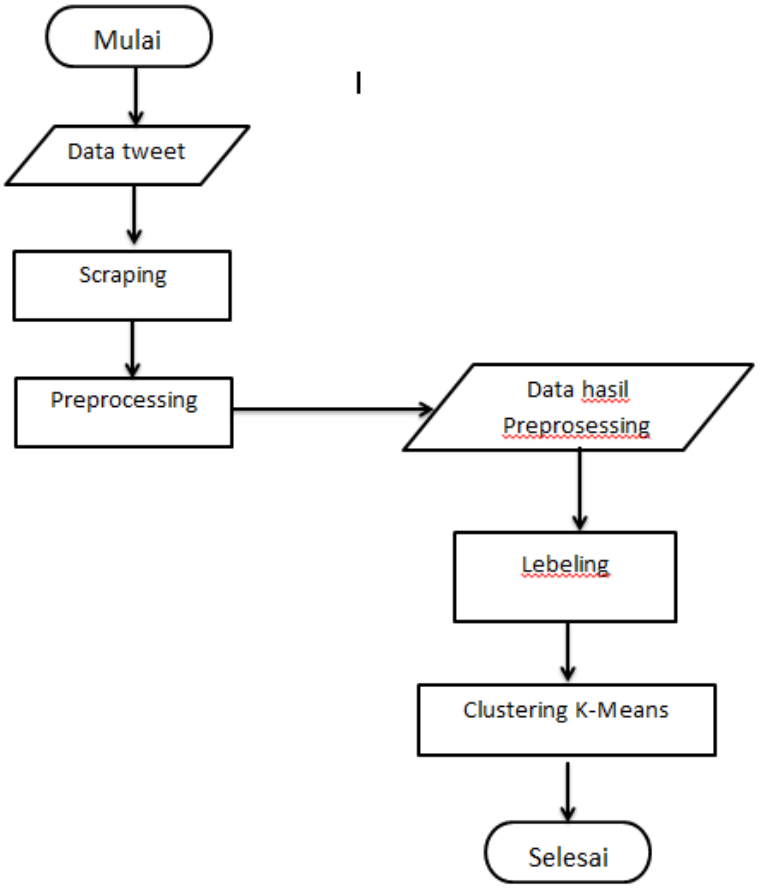
Pada langkah ini, data yang dikumpulkan oleh Twitter dibersihkan dan anotasi sentimen ditambahkan untuk memfasilitasi analisis lebih lanjut. Proses pembersihan data yang melibatkan penghapusan data duplikat, spam, dan pucat sehingga hanya data yang benar-benar relevan dengan pemilu 2024 yang digunakan. Setelah dibersihkan, setiap data diberi label opini, yang mencakup kategori seperti positif, negatif, dan netral. Penanda sentimen ini sangat penting karena membantu mengidentifikasi pola dan kecenderungan emosi pemilih. Dengan menggunakan tag, analisis dapat lebih mudah mengidentifikasi topik-topik penting, sering muncul, dan relevan dalam pemilu 2024, seperti kepercayaan terhadap proses pemilu, persepsi terhadap kandidat, dan isu-isu kontroversial yang menjadi perhatian publik.Selain itu, penandaan emosi membantu mengurangi kesalahan klasifikasi dengan menetapkan kelas unik untuk setiap bagian data. Hal ini membuat analisis sentimen menjadi lebih akurat dan andal, karena setiap kumpulan data diperiksa sepenuhnya untuk memastikan keakuratan label yang diberikan. Proses penandaan juga meningkatkan konsistensi data dan mengurangi inkonsistensi analitis dengan memastikan bahwa semua data serupa memiliki tag yang sama. Hal ini sangat penting untuk memperoleh hasil analisis yang benar dan dapat diandalkan. Menelaah data berlabel permanen mungkin lebih efektif dibandingkan persepsi masyarakat terhadap hasil pemilu 2024. Pada akhirnya, analisis yang lebih akurat dan

konsisten ini dapat berkontribusi secara signifikan terhadap pemahaman yang lebih baik mengenai dinamika opini publik dan membantu menciptakan kebijakan yang lebih responsif dan tepat sasaran.

3.5 Metode K-means

Metode K-means menganalisis hasilkan opini masyarakat terhadap objek penelitian terkait komentar masyarakat Indonesia terhadap pemilu (Pemilihan Umum) tahun 2024. K-means clustering merupakan algoritma yang populer dalam data mining dan machine learning, yang digunakan untuk mengelompokkan data ke dalam beberapa cluster atau kelompok yang telah ditentukan. Dalam konteks analisis sentimen, k-means membantu mengelompokkan komentar-komentar yang memiliki kesamaan pendapat, baik positif, negatif, atau netral. Proses ini dimulai dengan memilih sejumlah cluster yang berpusat pada k secara acak, kemudian melakukan iterasi untuk meminimalkan jarak antara titik data dan pusat cluster hingga terbentuk cluster yang optimal.Penggunaan k-means clustering untuk analisis opini pemilu 2024 melibatkan beberapa langkah penting. Pertama, data komentar publik yang dikumpulkan dengan mengikis media sosial Twitter dibersihkan untuk menghilangkan gangguan seperti spam, iklan, dan duplikat. Setelah proses pembersihan, data teks diproses pada tahap preprocessing yang meliputi penandaan, decoding dan penghapusan stopword untuk memudahkan normalisasi kata. Normalisasi ini penting untuk memastikan algoritma mengidentifikasi varian kata yang memiliki arti sama dengan benar. Selanjutnya, fitur-fitur penting diekstraksi dari data teks dan diubah menjadi vektor yang dapat digunakan dengan algoritma k-means. Vektor-vektor ini kemudian digunakan dalam proses pengelompokan untuk mengelompokkan komentar berdasarkan kesamaan sentimentalnya.Hasil dari proses clustering K-means memberikan wawasan mendalam

tentang bagaimana reaksi masyarakat Indonesia terhadap pemilu 2024. Dengan mengelompokkan komentar berdasarkan opini, peneliti dapat mengidentifikasi pola-pola tertentu yang mungkin tidak terlihat melalui analisis manual. Misalnya, peneliti dapat menemukan kelompok besar komentar positif terkait keberhasilan atau kebijakan kandidat tertentu, atau sebaliknya kelompok komentar negatif yang mengkritik aspek tertentu dalam pemilu. Analisis lebih lanjut terhadap klaster ini dapat membantu memahami kekhawatiran masyarakat dan membantu kandidat atau partai politik merespons kekhawatiran dan aspirasi masyarakat secara lebih efektif. Dengan demikian, penggunaan k-means clustering dalam analisis sentimen tidak hanya memberikan gambaran persepsi masyarakat yang lebih jelas, namun juga dapat menjadi alat strategis dalam pengambilan keputusan politik.



Gambar 3.1 kerangkah Penelitian

3.5 Waktu dan Tempat Penelitian

3.5.1 Waktu Penelitian

Tabel 3. 1 Tabel Waktu Penelitian

No.	Tahapan	Januari 2024	Febuari 2024	Maret 2024	April 2024	Mei 2024
1.	Pengajuan Judul					
2.	Pengumpulan data					
3.	Penyusunan Proposal					
4.	Seminar Proposal					
5.	Analisis Data dan Penelitian					
6.	Penyusunan Sripsi					
7.	Sidang Meja Hijau					
8.	Penyempurnaan Skripsi dan Penulisan Artikel Jurnal					

3.5.2 Tempat Penelitian

Penelitian ini dilakukan pada sentimen yang ada pada media sosial *twitter* tentang isu pinjaman online. Pengambilan dan pengolahan data menggunakan *tools Google*

Colab dengan dibantu *library Pyhon Pandas* serta API *twitter* untuk scraping data.

3.6 Perangkat Penelitian

Perangkat atau alat baik berupa perangkat keras (*Hardware*) ataupun perangkat lunak (*Software*) yang digunakan dalam penelitian ini dipilih sesuai dengan kebutuhan penelitian. Keduanya digunakan berdasarkan kapasitas dan kemampuan dari tiap perangkat. Perangkat keras dan pengkat lunak digunakan dengan maksimal sehingga penelitan ini dapat berjalan dengan baik. Berikut adalah deksripsi tiap perangkat:

Tabel 3. 2 Kebutuhan Perangkat Keras

No	Nama Perangkat	Deskripsi
1	Laptop	*Asus VivoBook Max*
2	Processor	*AMDA A9*
3	RAM	4GB
4	Penyimpanan	1TB HDD

Tabel 3. 3 Kebutuhan Perangkat Lunak

No.	Nama	Deskripsi
1.	*Windows 10 64-bit*	*Operating System*
2.	*Python 3.10.11 dan Googlr colab*	Tools untuk membangun dan melatih (training) model
3.	*Pandas*	Pustaka *Python* untuk memanipulasi dan menganalisis data
4.	*Scikit-learn*	Pustaka *Python* yang menyediakan algoritma dan fungsi untuk analisis data, termasuk pengelompokan data (*klasifikasi*).
5.	*Matplotlib dan Seaborn*	Pustaka *Python* untuk visualisasi data.
6.	*Microsoft Office 2010*	*Tools* untuk membuat hasil laporan penelitian

BAB IV
HASIL DAN PEMBAHASAN

4.1 Data Selection

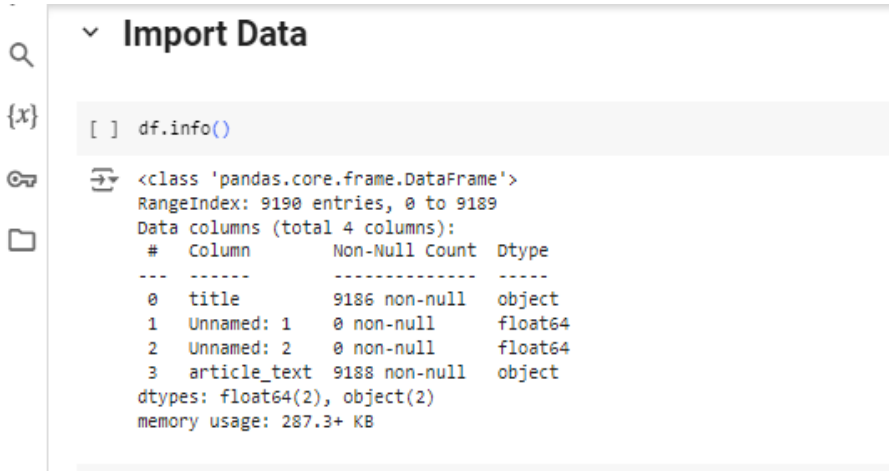
Pada tahap ini data selection akan melakukan proses pencarian data dan pengambilan data yang akan di gunakan adalah data berita yang diambil dari akun twitter Detik.com, data yang diambil adalah data yang berkaitan dengan isu pemilu dari mulai 01-10-2023 sampai dengan 30-12-2023 data yang diperoleh sebanyak 9.192 data. Data yang diperoleh memiliki kolom ‘title’, ‘author’, ‘publish_date’, ‘article_text’, ‘url’, ‘main_image’ dan ‘tag’. Untuk keperluan penelitian data akan dilakukan tahapan preprocesing dan nantinya akan mengambil kolom yang hanya diperlukan saja.

	title	author	publish_date	article_text	url	main_image	tag
2	Hasto Ingin Debat Pilpres Perta	detikNews	2023-12-10 23:31:00+	Sekjen PDIP sekaligus Sekretaris	https://news.de	https://awsimages.deti	['hasto kristiyanto', 'hasto', 'sekjen pdip', 'tpn ganjar-mahfud', 'pemilu', 'pc
3	Terima Kunjungan Gibran di Po	detikNews	2023-12-10 23:09:00+	Mantan Ketua Umum Pengurus	https://news.detik.com/pemilu/d-7082	['said aqil siradj', 'gibran rakabuming', 'pemilu', 'pilpres', 'politik']	
4	Contoh Format dan Isi Surat Pe	detikNews	2023-12-10 22:38:00+	Surat pernyataan KPPS Pemilu	https://news.de	https://awsimages.deti	['surat pernyataan kpps pemilu 2024', 'pendaftaran kpps pemilu 2024', 'kpp
5	9 Ribu Personel Gabungan Dike	detikSumut	2023-12-10 22:23:00+	Sekitar 9.000 personel	https://www.de	https://awsimages.deti	['natal', 'tahun baru', 'nataru', 'pengamanan nataru', 'sumut', 'polda sumut'
6	Amarah Ukraina Usai Rusia Bak	detikNews	2023-12-10 22:12:00+	Ukraina marah usai Rusia	https://news.de	https://awsimages.deti	['ukraina', 'rusia', 'pemilu rusia', 'round-up']
7	TKN Prabowo-Gibran Buka Pint	detikNews	2023-12-10 21:59:00+	Sekretaris Tim Kampanye	https://news.de	https://awsimages.deti	['tkn', 'prabowo-gibran', 'dudung abdurachman']
8	Gibran Dadakan Makan Gultik d	detikNews	2023-12-10 21:51:00+	Cawapres nomor urut 2 Gibran	https://news.de	https://awsimages.deti	['gibran rakabuming raka', 'gultik blok m', 'blok m', 'gibran rakabuming', 'gil
9	TPN Ganjar-Mahfud Anggap Du	detikNews	2023-12-10 21:45:00+	Sekretaris Tim Pemenangan	https://news.de	https://awsimages.deti	['abuya muhtadi', 'hasto kristiyanto', 'hasto', 'tpn ganjar-mahfud', 'pemilu',
10	TKN: Prabowo ke Marapi Sebag	detikNews	2023-12-10 21:32:00+	Tim Kampanye Nasional (TKN)	https://news.de	https://awsimages.deti	['tkn', 'prabowo subianto', 'erupsi gunung marapi']
11	Prabowo-Gibran Siap Hadapi D	detikNews	2023-12-10 21:31:00+	Pasangan capres dan cawapres	https://news.de	https://awsimages.deti	['debat pilpres', 'prabowo-gibran', 'prabowo', 'gibran', 'pemilu', 'debat', 'pc
12	Contoh Format dan Isi Surat Pe	detikNews	2023-12-10 21:26:00+	Surat pendaftaran KPPS Pemilu	https://news.de	https://awsimages.deti	['surat pendaftaran kpps pemilu 2024', 'pendaftaran kpps pemilu 2024', 'kp
13	Debat Perdana Pilpres, TPD Jab	detikJabar	2023-12-10 21:23:00+	Debat perdana Pilpres 2024 bakal	https://www.de	https://awsimages.deti	['debat pilpres 2024', 'debat capres cawapres', 'ganjar mahfud', 'jawa barat'
14	Gerindra Rekom Khofifah Cagu	detikJatim	2023-12-10 21:08:00+	Partai Gerindra resmi	https://www.de	https://awsimages.deti	['pemilu', 'pemilu 2024', 'pilgub jatim 2024', 'khofifah indar parawansa', 'en
15	Bamsoet Harap Saksi TPS Purba	detikNews	2023-12-10 20:48:00+	Wakil Ketua Umum Partai Golkar	https://news.de	https://awsimages.deti	['mpr', 'bamsoet']
16	Gibran Sambangi Ponpes Al-Ts	detikNews	2023-12-10 20:41:00+	Cawapres nomor urut 2 Gibran	https://news.detik.com/pemilu/d-7082	['gibran rakabuming raka', 'gibran', 'said aqil', 'pemilu', 'politik']	
17	Dosen USU Jadi Panelis Debat F	detikSumut	2023-12-10 20:33:00+	Komisi Pemilihan Umum (KPU) RI	https://www.de	https://awsimages.deti	['debat capres', 'debat capres cawapres', 'ahmad taufan damanik']
18	Zulhas Sebut Gibran Siap Debat	detikBali	2023-12-10 20:32:00+	Wakil Ketua Dewan	https://www.de	https://awsimages.deti	['zulkifli hasan', 'kampanye zulkifli hasan di ntb', 'pan', 'prabowo-gibran', 'g
19	Yusril Jadi Kuasa Hukum Hadap	detikNews	2023-12-10 20:31:00+	Pasangan calon nomor urut 2	https://news.de	https://awsimages.deti	['prabowo-gibran', 'yusril ihza mahendra', 'yusril', 'pn jakpus', 'pemilu']
20	Cara Cek NIK Terdaftar Parpol F	detikNews	2023-12-10 20:23:00+	Komisi Pemilihan Umum (KPU)	https://news.de	https://awsimages.deti	['cek anggota parpol', 'cara cek nik anggota parpol', 'pemilu 2024', 'parpol 2
21	2 Kecamatan di Wajo Masuk Da	detikSulsel	2023-12-10 20:23:00+	Polres Wajo, Sulawesi Selatan	https://www.de	https://awsimages.deti	['politik sulsel', 'daerah rawan pemilu', 'pilkada 2024', 'pemilu 2024', 'wajo'
22	Kampanye di Lombok, Zulhas S	detikNews	2023-12-10 20:17:00+	Ketua Umum PAN Zulkifli Hasan	https://news.de	https://awsimages.deti	['pan', 'zulkifli hasan']
23	Bamsoet Ajak Partisipasi Aktif I	detikNews	2023-12-10 20:13:00+	Ketua MPR RI Bambang Soesatyo	https://news.de	https://awsimages.deti	['mpr', 'bamsoet']

Gambar 4.1 Data Selection Scraping

Setelah menampilkan data selection proses scraping selanjutnya memasuki tahapan *explore* adalah tahapan melakukan eksplorasi data diperoleh dari df dataframe memiliki 9190 record dengan 4 kolom. Kolom-kolom tersebut adalah judul, Anonim: 1, Anonim: 2 dan teks_artikel. Dari data tersebut terlihat bahwa kolom judul memiliki 9186 nilai bukan nol, sedangkan kolom artikel memiliki 9188 nilai bukan nol, yang menunjukkan bahwa kedua kolom tersebut memiliki data yang hilang. Pada saat yang sama, kolom Anonymous: 1 dan Anonymous: 2

tidak memiliki nilai bukan nol sama sekali (angka bukan nol adalah 0), dan keduanya memiliki tipe data float64, yang mungkin merupakan kolom minor atau kolom kesalahan dalam proses pengambilan data. Sebaliknya, kolom teks header dan artikel bertipe data objek, yang menunjukkan bahwa kolom tersebut berisi data teks..



Gambar 4.2 Explore Data

4.2 Preprocessing

Pada tahap ini, tujuan dari preprocessing adalah mengubah atau membersihkan data mentah menjadi data yang lebih bersih dan dapat digunakan. Data mentah sering kali mengandung noise yang mewakili gangguan atau informasi tidak relevan yang dapat memengaruhi analisis data. Oleh karena itu, proses pra-pemrosesan ini sangat penting untuk menghilangkan noise pada data.Selain itu, proses pembersihan data pada tahap prapemrosesan juga mencakup beberapa langkah penting lainnya. Salah satunya adalah duplikasi data. Duplikasi data dapat terjadi bila data yang sama dicatat lebih dari satu kali, sehingga dapat menimbulkan bias dan ketidakakuratan dalam analisis. Menghilangkan duplikasi ini membantu memastikan bahwa

analisis didasarkan pada data yang unik dan akurat.Tujuan dari proses preprocessing juga untuk memperbaiki kesalahan data. Kesalahan data dapat disebabkan oleh beberapa hal, seperti kesalahan dalam entri data, kesalahan ketik, atau kesalahan pengambilan sampel. Memperbaiki kesalahan ini sangat penting untuk meningkatkan kualitas data dan memastikan bahwa analisis selanjutnya didasarkan pada data yang benar.Selain deduplikasi dan koreksi kesalahan, proses prapemrosesan juga melibatkan pemeriksaan dan penyesuaian data yang sering berubah atau bertentangan. Data yang bertentangan dapat mempersulit analisis dan mengurangi keandalan hasil yang diperoleh. Oleh karena itu, penting untuk memeriksa informasi dan melakukan perubahan yang diperlukan agar informasi tersebut lebih konsisten dan dapat diandalkan.

Singkatnya, prapemrosesan adalah langkah yang sangat penting dalam siklus pemrosesan data. Proses ini tidak hanya membantu menghilangkan gangguan, duplikasi, dan kesalahan, namun juga memastikan bahwa data yang digunakan untuk analisis bersih, akurat, dan konsisten. Hasil analisis yang diperoleh dengan cara ini lebih dapat diandalkan dan memberikan sebuah hasil gambaran yang lebih bersih, akurat, dan konsisten.

```
# FUNCTION PREPROCESSING

def preprocess_text(kalimat):
    # Mengubah kalimat menjadi huruf kecil
    #lower_case = kalimat.lower()
    if isinstance(kalimat, str):
        # Mengubah kalimat menjadi huruf kecil
        lower_case = kalimat.lower()
        # Menghapus angka dari kalimat
        #hasil = re.sub(r"\d+", "", lower_case)
        hasil = re.sub(r'[,()\d]+', '', lower_case)

        return hasil
    else:
        return kalimat

    # Menghapus hashtag
    hasil = re.sub(r'#\w+', '', hasil)
    hasil = re.sub(r'@\w+', '', hasil)
    hasil = re.sub(r"(?:\@|http?\://|https?\://|www)\S+", "", hasil)

    # Menghapus tanda baca dari kalimat
    hasil = hasil.translate(str.maketrans("", "", string.punctuation))

    # Menghapus spasi pada awal dan akhir kalimat
    hasil = hasil.strip()
```

```
# Menghilangkan Tanda Baca
hasil = hasil.translate(str.maketrans('', '', string.punctuation))

# Mengganti karakter HTML dengan tanda petik
hasil = re.sub('<.*?>', ' ', hasil)

# Mempertimbangkan huruf dan angka
hasil = re.sub('[^a-zA-Z0-9]', ' ', hasil)

# Mengganti line baru dengan spasi
hasil = re.sub("\n", " ", hasil)

# Menghapus single char
hasil = re.sub(r"\b[a-zA-Z]\b", " ", hasil)

# Memisahkan dan menggabungkan kata
hasil = ' '.join(hasil.split())

return hasil
```

Gambar 4.3 Proses Preprosessing

Fungsi preprocess_text dimaksudkan untuk menghilangkan berbagai elemen dalam teks yang tidak diperlukan untuk analisis teks. Pertama, fungsi ini memeriksa apakah inputnya berupa string atau bukan. Jika ada, teks akan diubah menjadi huruf kecil untuk menghindari perbedaan antara huruf besar dan kecil. Angka, tanda kurung, dan koma dihilangkan menggunakan ekspresi reguler. Kemudian hal-hal seperti hashtag, mention, dan URL juga dihapus menggunakan ekspresi reguler lainnya. Fungsi ini juga menghilangkan semua tanda baca menggunakan metode terjemahan dan modul string, serta menghilangkan spasi di awal dan akhir teks. Selain itu, tag HTML diganti dengan spasi dan hanya huruf dan angka yang dipertahankan. Baris baru diubah menjadi spasi dan kata berkarakter tunggal dihapus. Terakhir, kata-kata yang dipisahkan digabungkan hanya dengan satu spasi di antaranya, memastikan tidak ada spasi yang tidak diperlukan. Kode ini membantu mempersiapkan teks untuk analisis lebih lanjut dengan menghapus elemen asing dan memastikan format teks

Tahapan persiapan data melipitu pembersihan data. Hanya mengambil kolom yang diperlukan saya yaitu kolom ‘title dan ‘article_text’, isi dari kedua kolom tersebut berupa teks kotor yang akan dibersihkan. Proses pembersihan data meliputi penghapusan tanda baca, menghapus hastg, menghapus spasi atau withspacemenghilangkan karakter atau link, menghilangkan angka serta mennghapus single char. Untuk menjadi data yang utuh dan siap dikelolah data pada kolom ‘title’ dan kolom ‘article_text” akan di satukan menjadi kolom ‘merge_text’.

```
# MERGE JUDUL & ISI BERITA
df['merged_text'] = df['title'] + ' ' + df['article_text']

# DO PREPROCESSING
df['merged_text_pre'] = df['merged_text'].apply(preprocess_text)
df.head()
```

Gambar 4.4 Proses Marge Text

Kode ini bertujuan untuk menggabungkan kolom 'title' dan 'article_text' di pandas DataFrame menjadi kolom baru bernama 'merge_title_article', lalu membersihkan teks gabungan tersebut menggunakan fungsi preprocess_text. Pertama, setiap nilai di kolom “title” dan “article_text” pada DataFrame digabungkan menjadi satu string yang dipisahkan spasi dan disimpan di kolom “merge_title_article”. Kolom baru ini kemudian diproses dengan preprocess_text untuk menghapus teks dari elemen yang tidak diinginkan, dan hasilnya disimpan di kolom baru bernama “berita”. Kolom sementara “merge_title_article” kemudian dihapus dari DataFrame menggunakan drop(), sehingga DataFrame hanya menyimpan kolom yang sudah dibersihkan di kolom “berita”. Terakhir, head () digunakan untuk menampilkan lima baris DataFrame yang diproses ini, memberikanKolom merge_text selanjutnya dimasukan fungsi preoricesing_text yang telah dibuat diatas untuk membersihkan dah menghasilkan data yang siap diolah. Data tersebut kemudian dimasukan kedalam kolom merge_text_pre.

4.3 Lebeling

Pada tahap ini Data yang telah dibersihkan ditandai dengan lebel sentimen untuk memudahkan analisis lebih lanjut. Hal ini membantu mengidentifikasi pola dan tren emosional di kalangan pemilih.Dengan memberi label pada data dapat mengidentifikasi tema-tema penting yang sering terjadi dan relevan dengan pemilu 2024. Pelabelan membantu mengurangi kesalahan klasifikasi dengan menetapkan kategori unik pada setiap bagian data. Hal ini membuat analisis sentimen lebih akurat dan dapat diandalkan. Proses pelabelan juga meningkatkan konsistensi data dan mengurangi inkonsistensi dalam analisis dengan memastikan bahwa semua data serupa memiliki label yang sama.

Labeling pada tahap ini merujup pada tiga label berita yaitu label ‘’Paslon 1’, ‘Paslon 2’ dan ‘Paslon 3’. Labeling menggunakan file label_map sebagai acuan untuk mengidentifikasi kata yang ada pada data sehingga sistem dapat mengelompokan tiap data pada label yang sesuai.

	label	keyword
2	Paslon 1	amin
3	Paslon 1	anies
4	Paslon 1	anies baswedan
5	Paslon 1	anis muhaimin
6	Paslon 1	cak imin
7	Paslon 1	muhaimin
8	Paslon 1	muhaimin iskandar
9	Paslon 1	baswedan
10	Paslon 1	iskandar
11	Paslon 2	prabowo
12	Paslon 2	subianto
13	Paslon 2	prabowo subianto
14	Paslon 2	gibran
15	Paslon 2	prabowo gibran
16	Paslon 2	rakabuming
17	Paslon 3	ganjar
18	Paslon 3	pranowo
19	Paslon 3	ganjar pranowo
20	Paslon 3	mahfud
21	Paslon 3	ganjar mahfud
22		

Gambar 4.5 Mengidentifikasi Penglebelan

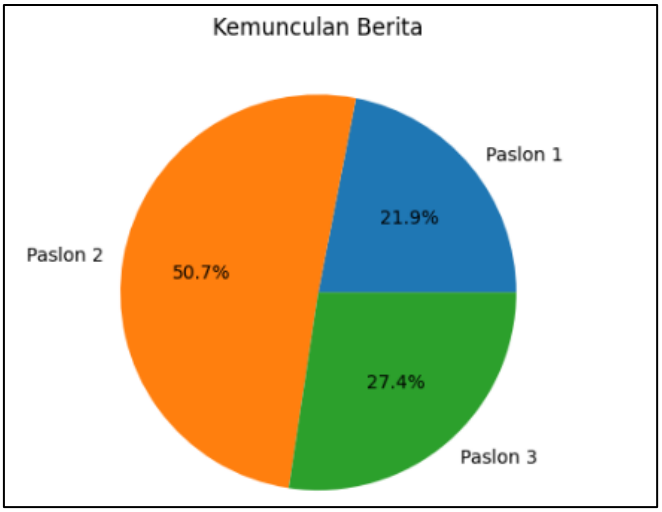
```python
def label_words_in_dataframe(df, column_name, filename):
    labels_df = pd.read_csv(filename)

    # Ubah DataFrame menjadi dictionary
    keyword_to_label = labels_df.set_index('keyword')['label'].to_dict()
    labeled_data = []

    for sentence in df[column_name]:
        # Pisahkan kalimat menjadi kata-kata
        words = sentence.split(' ')
        # Inisialisasi list kosong untuk menyimpan label kata dalam kalimat
        labels = []
        # Iterasi melalui kata-kata dalam kalimat
        for word in words:
            # Jika kata adalah kata kunci, berikan label yang sesuai
            if word in keyword_to_label:
                labels.append(keyword_to_label[word])

        # Gabungkan label menjadi string dan tambahkan ke data yang telah diberi label
        labeled_data.append(' '.join(labels))

    # Tambahkan data yang telah diberi label sebagai kolom baru ke DataFrame Anda
    df['category'] = labeled_data

    return df
```

```python
# MELAKUKAN LABELING TERHADAP BERITA

filename = 'label_map.csv'
df['merged_text_pre'] = df['merged_text_pre'].astype(str)
df = label_words_in_dataframe(df, 'merged_text_pre', filename)
df.head(10)
```

Gambar 4.6 Proses Lebeling

Kode fungsi label_words_in_dataframe bertujuan untuk menambahkan label pada kata-kata di kolom tertentu DataFrame berdasarkan pemetaan label kata kunci yang diberikan dalam file CSV. Fungsi ini pertama-tama membaca file CSV yang berisi kata kunci dan labelnya, kemudian mengubahnya menjadi kamus (keyword_to_label) dengan kata kunci sebagai kunci dan label sebagai nilai. Kemudian, untuk setiap kalimat di kolom yang ditentukan oleh nama_kolom di DataFrame, kalimat tersebut akan dipecah menjadi kata-kata individual. Fungsi tersebut kemudian memeriksa setiap kata untuk melihat apakah kata tersebut ada dalam kamus kata kunci, dan jika demikian, menambahkan tag yang sesuai ke daftar tag untuk frasa tersebut. Setelah memproses semua kata dalam kalimat, daftar label digabungkan dan ditambahkan ke daftar label_data. Terakhir, daftar label ini ditambahkan sebagai kolom

baru yang disebut "label" di DataFrame, dan DataFrame yang



Gambar 4.7 Hasil Lebeling kemunculan berita

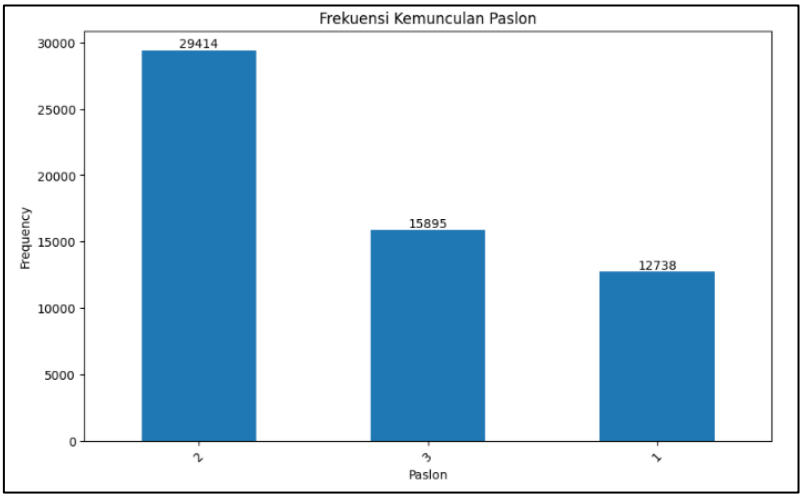
Tujuannya adalah menghitung persentase kemunculan setiap kata di kolom DataFrame tertentu, lalu menampilkannya sebagai diagram lingkaran. Pertama, jumlah kata dihitung dengan menjumlahkan semua nilai dalam word_counts.values(), yang dianggap kamus dengan kata sebagai kunci dan frekuensi kemunculan kata sebagai nilai. Kemudian dihitung tingkat kemunculan setiap kata dengan cara membagi jumlah kemunculan suatu kata dengan jumlah seluruh kata kemudian dikalikan dengan 100 untuk mendapatkan persentasenya. Persentase ini disimpan dalam daftar persentase. Selanjutnya, diagram lingkaran dibuat menggunakan plt.pie(), dengan persentase sebagai data, kata sebagai label, dan format persentase untuk setiap segmen diagram lingkaran (autopct='%1.1f%%') . Terakhir, diagram lingkaran diberi label “Kejadian berita” dan ditampilkan menggunakan plt.show(). Kode ini membantu memahami distribusi.

Pada tahap Proses ini untuk menganalisis dan memvisualisasikan frekuensi kemunculan kata-kata di kolom "label" pada bingkai data. Pertama, setiap kata dalam kolom dipisahkan menjadi kata-kata individual dan kemudian dihitung frekuensinya. Kata-kata tersebut

diurutkan dari yang paling sering hingga yang paling jarang, kecuali kata “Paslon” yang dikeluarkan dari daftar karena dianggap tidak ada artinya. Setelah proses penghitungan frekuensi selesai, kode ini membuat diagram batang yang menunjukkan 20 kata yang paling sering muncul. Bagan ini memiliki label "Lulus" dan "Frekuensi" masing-masing pada sumbu x dan y, dan diberi judul "Frekuensi Pelamar". Kata-kata pada sumbu x ditampilkan pada sudut 45 derajat untuk kemudahan membaca. Selain itu, setiap batang pada bagan disertai dengan label yang menunjukkan nilai frekuensi kata tersebut, ditempatkan di atas batang untuk memberikan informasi tambahan langsung. Grafik kemudian ditampilkan untuk visualisasi.

```
word_counts = (df['label'].str.split().explode()
                          .value_counts()
                          .sort_values(ascending=False)
                          .drop('Paslon'))


# Plotting the word frequency
plt.figure(figsize=(10, 6))
ax = word_counts[:20].plot(kind='bar')  # Get the Axes object for label placement
#word_counts[:20].plot(kind='bar')
plt.xlabel('Paslon')
plt.ylabel('Frequency')
plt.title('Frekuensi Kemunculan Paslon')
plt.xticks(rotation=45)
# Add value labels above each bar
for p in ax.patches:
    ax.annotate(f"{p.get_height():.0f}",  # Format the value as an integer
                (p.get_x() + p.get_width() / 2., p.get_height()),  # Place label above the bar
                ha='center', va='bottom')
plt.show()
```

Gambar 4.8proses kemunculan berita



Gambar 4.9 Hasil Lebeling Frekuensi Kemunculan berita paslon

Selanjutnya proses wordcloud adalah proses ini seringnya muncul beritaacara ,kata-kata ,kalimat yang sering muncul bertujuan untuk membuat dan menampilkan wordcloud yang menampilkan kata-kata yang paling sering muncul dalam teks yang dibersihkan (dalam hal ini teks disimpan di kolom "teks22" setelah proses pembersihan). Pertama, teks pada kolom “text22” digabungkan menjadi string menggunakan metode join(), yang kemudian digunakan untuk menghasulkan sebuah wordcloud dengan WordCloud. wordcloude ini menampilkan kata-kata yang paling sering muncul dalam teks dengan ukuran yang sebanding dengan frekuensi kemunculannya. Ukuran kata cloud diatur ke 800 x 400 piksel dengan latar belakang putih. wordcloud kemudian ditampilkan menggunakan imshow() dengan interpolasi bilinear untuk memastikan gambar halus, dan sumbu x dan y diatur agar tidak terlihat menggunakan plt.axis('off'). Bagan ini dimaksudkan untuk memberikan representasi visual dari kata-kata yang paling dominan atau sering muncul dalam teks yang dianalisis.

```
datax=df
datax['text22'] = datax['token_clean'].apply(lambda x: str(x).replace('[', '').replace(']', '
#datax.head()
text2 = ' '.join(datax['text22'].dropna().astype(str))

# Generate the word cloud
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(text2)

# Plot the word cloud
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Kata Yang Banyak Disebut')
plt.show()
```

Gambar 4.10 Proses Pembuatan Wordcloud

Gambar 4.11 Hasil kata Yang Sering muncul Pada Proses Wordcloud

Pada tahap ini proses selanjutnya nltk.tokenize di NLTK digunakan untuk membagi teks menjadi bagian-bagian yang lebih kecil, disebut token, berdasarkan aturan tertentu. Proses ini merupakan langkah awal dalam analisis teks, sehingga teks dapat diolah menjadi bentuk yang lebih mudah untuk diolah nantinya, seperti kata, frasa, atau kalimat. Dengan menggunakan nltk.tokenize, teks dapat dibagi menjadi token berdasarkan kata, frasa, atau bahkan karakter, tergantung kebutuhan analisis teks yang diinginkan.

```
[ ]  from nltk.tokenize import word_tokenize

     def tokenize_text(kalimat):
         if isinstance(kalimat, str):
             tokens = nltk.tokenize.word_tokenize(kalimat)
             return tokens
         else:
             return []

     df['token'] = df['news'].fillna('').apply(tokenize_text)
     df.head()
```

Gambar 4.12 Proses NLTK .Tokenize

	title	article_text	news	label	token
0	Hasto Ingin Debat Pilpres Pertarungan Gagasan ...	Sekjen PDIP sekaligus Sekretaris Tim Pemenang ...	hasto ingin debat pilpres pertarungan gagasan ...	Paslon 3	[hasto, ingin, debat, pilpres, pertarungan, ga...
1	Terima Kunjungan Gibran di Ponpes Al-Tsaqafah,...	Mantan Ketua Umum Pengurus Besar Nahdlatul Ula...	terima kunjungan gibran di ponpes al-tsaqafah ...	Paslon 2 Paslon 2 Paslon 2 Paslon 1	[terima, kunjungan, gibran, di, ponpes, al-tsa...
5	TKN Prabowo-Gibran Buka Pintu untuk Eks KSAD D...	Sekretaris Tim Kampanye Nasional (TKN) Prabowo...	tkn prabowo-gibran buka pintu untuk eks ksad d...	Paslon 2 Paslon 2 Paslon 2 Paslon 2	[tkn, prabowo-gibran, buka, pintu, untuk, eks,...
6	Gibran Dadakan Makan Gultik di Blok M Malam In...	Cawapres nomor urut 2 Gibran Rakabuming Raka m...	gibran dadakan makan gultik di blok m malam in...	Paslon 2 Paslon 2 Paslon 2 Paslon 2 Paslon 2 P...	[gibran, dadakan, makan, gultik, di, blok, m, ...
7	TPN Ganjar-Mahfud Anggap Dukungan Abuya Muhtad...	Sekretaris Tim Pemenangan Nasional (TPN) capre...	tpn ganjar-mahfud anggap dukungan abuya muhtad...	Paslon 3 Paslon 2 Paslon 2 Paslon 3 Paslon 3 P...	[tpn, ganjar-mahfud, anggap, dukungan, abuya, ...

Gambar 4.13 Hasil Proses Pemberian Token

Gambar ini menggambarkan proses hasil pemberian token yang diterapkan pada data kandidat pemilu, dengan fokus pada tahap persiapan data teks. Proses dimulai dengan kolom "berita" yang berisi teks artikel terkait kandidat pemilu. Untuk memastikan tidak ada nilai yang hilang, semua nilai NaN di kolom "berita" diisi dengan string kosong menggunakan metode fillna('). Setelah itu, dilakukan tokenisasi teks menggunakan fungsi tokenize_text, yang bertujuan untuk memecah teks artikel menjadi unit-unit kata atau token. Hasil dari tokenisasi ini disimpan dalam kolom baru bernama "token". Selanjutnya, proses pembersihan lebih lanjut dilakukan untuk menghapus stopwords dari token-token ini, menghasilkan kolom "token_clean" yang berisi token-token yang telah dibersihkan dari kata-kata tidak bermakna. Data yang telah diproses ini kemudian digunakan untuk analisis K-Means, yang mengelompokkan data berdasarkan kemiripan token-token yang ada, memungkinkan pengelompokan artikel-artikel berita berdasarkan pola teks yang muncul terkait dengan kandidat pemilu. Proses ini penting untuk memahami isu-isu utama dan sentimen yang muncul dalam liputan berita mengenai kandidat pemilu.

Setelah mendapatkan hasil token proses selanjutnya adalah proses stopword untuk memproses stopword menghapus kalimat yang tidak bermakna yang ada pada kolom token.

```
#proses stopword menghapus kalimat yang tidak bermakna yang ada pada kolom token
def stopword_text(tokens):
  #proses pembersihan kata
    stopwords = pd.read_csv('stopword.txt', header=None)[0].tolist()
    cleaned_tokens = []
    for token in tokens:
        if token not in stopwords:
            cleaned_tokens.append(token)
    return cleaned_tokens

df['token_clean'] = df['token'].apply(stopword_text)
```

Gambar 4.14 Proses stopword

Tujuan dari proses ini adalah untuk membersihkan kolom “Token” dari kata-kata yang tidak masuk akal atau yang sering disebut dengan stopword. Pertama, fungsi stopword_text(tokens) membaca daftar stopwords dari file "stopword.txt" dan menyimpannya dalam format daftar. File "stopword.txt" ini digunakan oleh pd.read_csv dengan parameter header=None yang menunjukkan bahwa file tersebut tidak memiliki header. Daftar akhiran ini kemudian diubah menjadi daftar menggunakan [0].tolist(), dengan [0] merujuk pada kolom pertama dari DataFrame yang dihasilkan. Fungsi ini kemudian memproses setiap kata dalam tag masukan, yaitu daftar kata. Setiap karakter diperiksa untuk melihat apakah karakter tersebut ada dalam daftar kata-kata berhenti; jika tidak, maka akan ditambahkan ke daftar "clean_tokens". Daftar clean_tokens berisi token yang telah dibersihkan dari stopwords. Fungsi ini akhirnya mengembalikan 'clean_tokens'. Kolom "token_clean" baru kemudian ditambahkan ke df DataFrame menggunakan fungsi stopword_text untuk setiap entri di kolom "token" menggunakan metode apply. Artinya setiap baris pada kolom token diproses untuk menghilangkan stop word dan hasilnya disimpan pada kolom token_clean.

Pada tahap lebeling selanjutnya adalah menampilkan proses kerja dan hasil dari analisis sentimen

```
# Membuat class untuk membaca kamus lexicon positif dan negatif
lexicon_positive = dict()
with open("/content/Kamus_Positif.csv", "r") as csvfile:
  reader = csv.reader(csvfile, delimiter=",")
  for row in reader:
    lexicon_positive[row[0]] = int(row[1])


lexicon_negative = dict()
with open("/content/Kamus_Negatif.csv", "r") as csvfile:
  reader = csv.reader(csvfile, delimiter=",")
  for row in reader:
    lexicon_negative[row[0]] = int(row[1])


# Membuat fungsi untuk mempolarisasi sentimen berdasarkan kamus lexicon yang ada
def sentiment_analysis(text):
  score = 0
  for word in text:
    if (word in lexicon_positive):
      score = score + lexicon_positive[word]
  for word in text:
    if (word in lexicon_negative):
      score = score + lexicon_negative[word]
  polarity = ''
  if (score > 0 ):
    polarity = "Positif"
  elif (score < 0 ):
    polarity = "Negatif"
  else :
    polarity = "Netral"
  return score, polarity
```

Gambar 4.15 Proses Mangasilkan Analisis Sentimen

Proses yang diberikan membuat sistem untuk melakukan analisis sentimen berdasarkan kamus kosakata positif ,negatif dan netral. Pertama, dua kamus (`lexikon_ Positif` dan `lexikon_ Negatif`) dibuat dengan membaca data dari file CSV yang berisi kata-kata positif dan negatif beserta skornya masing-masing. Fungsi “analisis_sentimen” kemudian digunakan untuk menganalisis teks tertentu, yang dikatakan telah dikodekan ke dalam kata-kata individual. Fungsi ini menghitung skor sentimen dengan menjumlahkan nilai kata yang ditemukan dalam kamus positif dan negatif. Jika sebuah kata ditemukan dalam kamus positif, skornya ditambahkan ke variabel "skor", dan hal yang sama dilakukan untuk kata-kata dalam kamus negatif. Setelah menghitung skor total, fungsi tersebut menentukan polaritas sentimen: “Positif” jika skornya positif, “Negatif” jika skornya negatif, dan “Netral” jika skornya 0. Fungsi tersebut kemudian mengembalikan poin yang sesuai dan polarisasi.

```
result = df['token'].apply(sentiment_analysis)
result = list(zip(*result))
df["polarity_score"] = result[0]
df["token"] = result[1]
print(df["token"].value_counts())
df.shape
df.head(5)
```

```
token
Negatif    5593
Positif     611
Netral       27
Name: count, dtype: int64
```

Gambar 4.16 Hasil Lebeling Analisis Sentimen

dataframe melakukan proses yang di perbarui untuk menambahkan kolom baru yang bernama polarity_score yang berisi skor sentimen dari hasil analisis. Kolom token juga diperbarui sesuai dengan polaritas sentimen (positif, negatif, atau netral). Output nilai value_counts() di kolom token ditampilkan

Hasil dari analisis sentimen yang di lakukan pada tahap data selection / scraping mendapatkan hasil data yang diperoleh sebanyak 9.192 .untuk data yang tidak memiliki lebel atau tidak sesuai kriteria akan diberhentikan bekerja dan di berikan dengan menggunakan (df = df.dropna()) untuk menghentikan proses kerja nya dan mendapatkan hasil sentiment yang di butuhkan dah hasil yang muncul sebanyak:

Negative sebanyak (5593), positif sebanyak (611), netral sebanyak (27)

Dan hasil ini nantinya akan di proses lagi untuk mendapatkan hasil pengkelompokan cluster setiap data dari proses metode K-Means.

4.4 K-Means

Pen golahan data selanjutnya adalah dengan menggunakan k-means untuk melihat cluster persebaran data yang ada pada dokumen text.

```
[ ] from sklearn.feature_extraction.text import TfidfVectorizer
    from sklearn.cluster import KMeans

[ ] documents = df['merged_text_pre'].values.astype("U")

    vectorizer = TfidfVectorizer(stop_words='english')
    features = vectorizer.fit_transform(documents)

    k = 3
    model = KMeans(n_clusters=k, init='k-means++', max_iter=100, n_init=1)
    model.fit(features)

    df['cluster'] = model.labels_

    df.head()
```

Gambar 4.17 Proses K-means Cluster

Hal pertama yang dilakukan adalah mengimport fitur ekstraksi kata TFIDF yang dilakuakn untuk mengubah kalimat atau kata yang ada pada dokumen memnjadi sebuah nilau angka yang bisa dibaca oleh mesin. Penggunaan algoritma K-means dengan mengimpirtnya dari library sclearn. Pengubahan nilau pada dokumen text menjadi nilai atau value dengan unicode. Serta mengelolahnya dengan tdidf vektorize.

Nilai k=3 adalah kelompok kluster yang diingin kan untuk membagi kelaster menjadi tiga bagian. Dengan iterasi pengelolahan sebanyak 100 kali untuk menghasilkan nilai cestoin yang stabil, nilai yang didapatkan dimasukan kedalam kolom bernana cluster.

	merged_text_pre	cluster
0	hasto ingin debat pilpres pertarungan gagasan ...	2
1	terima kunjungan gibran di ponpes al-tsaqafah ...	2
2	contoh format dan isi surat pernyataan anggota...	2
3	ribu personel gabungan dikerahkan amankan nat...	2
4	amarah ukraina usai rusia bakal gelar pilpres ...	2

Gambar 4.18 Proses Pengkelompokan Cluster

Untuk mengtahui seperti nilai atau kalimat dan kata apa yang terdapat pada masing-masing cluster digunakan kode berikut untuk melihatnya.

```
clusters = df.groupby('cluster')

for cluster in clusters.groups:
    f = open('cluster'+str(cluster)+ '.csv', 'w') # create csv file
    data = clusters.get_group(cluster)[['merged_text_pre']] # get title and overview columns
    f.write(data.to_csv(index_label='id')) # set index to id
    f.close()

print("Cluster centroids: \n")
order_centroids = model.cluster_centers_.argsort()[:, ::-1]
terms = vectorizer.get_feature_names_out()
```

```
Cluster centroids:
```

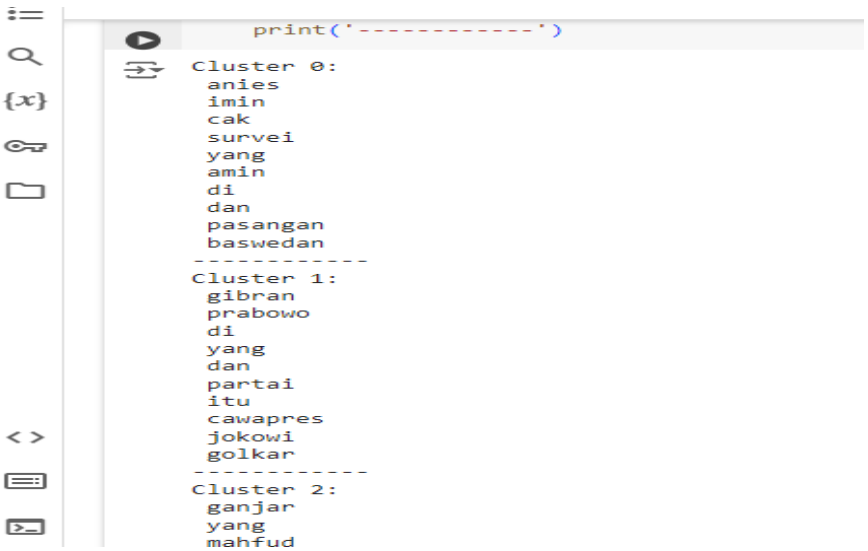
```
[ ] for i in range(k):
        print("Cluster %d:" % i)
        for j in order_centroids[i, :10]:
            print (' %s' % terms[j])
        print('------------')
```

Gambar 4.19 Proses kluster Pada Fungsi Grupby

Fungsi grupby untuk mengelompokan nilai atau kata pada masing-masing kluster untuk ditampilkan. Kata yang ditampilkan adalah kaya yang letaknya paling dekat dengan centroid sehingga tiap-tiap kluster dapat dilihat informasi yang berguna.

melakukan pengelompokan teks dengan algoritma K-Means. Proses kode mengelompokkan data di DataFrame berdasarkan kolom cluster. Setiap cluster disimpan dalam file CSV terpisah, dengan nama file sesuai dengan nomor cluster. Kode tersebut kemudian mencetak centroid di setiap cluster. Centroid ini diperoleh dari model K-Means yang telah dilatih sebelumnya dimana urutan centroid diambil berdasarkan argumen terkecil dari pusat cluster. Kata kunci (istilah) diekstraksi menggunakan vektorizer yang mengubah teks menjadi fitur numerik. Kode-kode tersebut menunjukkan kata kunci terpenting di setiap cluster dan memberikan gambaran umum tentang karakteristik utama teks di setiap cluster.

Hasilnya menunjukkan kata kunci dalam dua kelompok: cluster 0 berisi kata kunci seperti "anies", "imin", "cak", "survei" dan "baswedan", sedangkan cluster 1 berisi kata kunci seperti "gibran", "prabowo", " di" dan "tentu saja". Ini menunjukkan topik atau topik utama yang dibahas dalam setiap kelompok teks.



Gambar 4.20 Hasil Culster Pengkelompokan k-means

.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil penelitian dan proses yang sudah di jelaskan dapat di simpulkan bahwa perancangan dan implementasi dari sistem analisis sentimen terhadap pemilu 2024 di twitter mengunakan algoritma k-means dapat di tarik kesimpulan. sentimen data yang berkaitan dengan isu pemilu dari mulai 01-10-2023 sampai dengan 30-12-2023 dengan data yang diperoleh sebanyak 9.192 data. Hasil tersebut mendapatkan hasil sebesar 21,9% untuk kemunculan berta paslon 1, 50,7% untuk kemuncukan berita paslon 2 dan 27,4% kemunculan berita untuk paslon 3. Kemunculan berita dari akun twitter detik.com ,dengan hasil analisis ini mendapatkan hasil posisi yang tertinggi yaitu 50,7%. Jika dihitung secar manual maka ada terdapat 4.596 kali kemunculan berita yang membahas paslon 2. Sedangkan untuk kemunculan berita paslon 1 sebanyak 2.013 kali kemunculan berita yang berkaitan dengan paslon 1. Dan untuk paslon 3 sebanyak 2.518 kali kemunculan berita yang berkatan dengan paslon 3. Dan mendapatkan hasil analisis sentimen sebanyak, tanggapan Negatif 5593, dan tanggapan Positif 611 dan juga tanggapan Netral 27.

5.2 Saran

Proses algoritma k-means ini sudah menunjukkan sebuah hasil yang cukup baik dalam proses nya untuk pengelompokan data ke dalam cluster. Penelitian ini perlu dilakukan sebuah evaluasi lainnya yang nantinya berguna nutuk mengetahui apakah masih menghasilkan hasil cukup baik atau tidak.

Adapun saran dari penulisan selanjutnya untuk penelitian ini yang berkaitan dengan analisis sentiment dengan motode K-means adalah sebagai berikut:

1) Analisis ini dikembangkan untuk membuat labelisasi secara otomatis lagi.

2) bisa Meningkatkan sarana untuk belajar meningkatkan koleksi data latih, koleksi korpus stopwords, dan koleksi korpus stemming sehingga bisa dapat meningkatkan akurasi dari analisis sentimen.

3) Penelitian ini juga dapat lebih diimplementasikan kesebuah aplikasi sehingga proses dilakukan secara otomatis.

DAFTAR PUSTAKA

Aditama, M. I., Irfan Pratama, R., Hafizzana, K., Wiwaha, U., & Rakhmawati, N. A. (n.d.). *Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin COVID-19*.

Amrullah, A. A., Tantoni, A., Hamdani, N., Taufik, R., Bau, R. L., Ahsan, M. R., & Utami, E. (n.d.). *PROSIDING SEMINAR NASIONAL MULTI DISIPLIN ILMU & CALL FOR PAPERS UNISBANK (SENDI_U) KE-2 Tahun 2016 Kajian Multi Disiplin Ilmu dalam Pengembangan IPTEKS untuk Mewujudkan Pembangunan Nasional Semesta Berencana (PNSB) sebagai Upaya Meningkatkan Daya Saing Global REVIEW ATAS ANALISIS SENTIMEN PADA TWITTER SEBAGAI REPRESENTASI OPINI PUBLIK TERHADAP BAKAL CALON PEMIMPIN*.

Annur, H. (2018). KLASIFIKASI MASYARAKAT MISKIN MENGGUNAKAN METODE NAÏVE BAYES. In *Agustus* (Vol. 10, Issue 2).

Azis, H., Tangguh Admojo, F., & Susanti, E. (n.d.). Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah Performance Comparison Analysis of Classification Methods on the Multiclass Dataset of Bows. In *Agustus* (Vol. 19, Issue 3).

Duei Putri, D., Nama, G. F., & Sulistiono, W. E. (2022). Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier. *Jurnal Informatika Dan Teknik Elektro Terapan*, *10*(1).

Faroek, D., Yusuf, M., & Syatauw, G. (2023). Sentiment Analysis of the Popularity of Parties Supporting the 2024 Presidential Candidates on Twitter Using the Naive Bayes Classifier Algorithm. *Antivirus : Jurnal Ilmiah Teknik Informatika*, *17*(2), 216–227.

Firdaus, A., & Firdaus, W. I. (2021). Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi : (Sebuah Ulasan). *Jurnal JUPITER*, *13*(1), 66.

Fridom Mailo, F., Lazuardi, L., Manajemen dan kebijakan Kesehatan Fakultas Kedokteran, D., Masyarakat dan Keperawatan Universitas Gadjah Mada, K., Sistem Informasi Manajemen Kesehatan Fakultas Kedokteran, D., Masyarakat dan Keperawatan, K., & Gadjah Mada, U. (2019). Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia. In *Jurnal Sistem Informasi Kesehatan Masyarakat Journal of Information Systems for Public Health* (Vol. 4, Issue 1).

Fuadi, W., Razi, A., & Fariadi, D. (2022). Automasi Penentuan Tren Topik Skripsi Menggunakan Algoritma K-Means Clustering. *Serambi Engineering*, *VII*(2)

Han, J., Kamber, M., Pei, J., & Kaufmann, M. (n.d.). *[DATA MINING: CONCEPTS AND TECHNIQUES 3RD EDITION] 2 Data Mining: Concepts and Techniques Third Edition*.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1), 1–34.

Ilmawan, L. B., & Mude, M. A. (2020a). Perbandingan Metode Klasifikasi Support Vector Machine dan Naïve Bayes untuk Analisis Sentimen pada Ulasan Tekstual di Google Play Store. *ILKOM Jurnal Ilmiah*, *12*(2), 154–161.

Laurensz, B., & Sediyono, E. (2021). Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19 (Analysis of Public Sentiment on Vaccination in Efforts to Overcome the Covid-19 Pandemic). *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *10*(2), 118–123

Luthfiansyah Dan, R., Wasito, B., Program, S., & Sistem, I. (n.d.). *ANALISIS SENTIMEN TERHADAP PARA KANDIDAT PRESIDEN 2024 BERDASARKAN NETIZEN PENGGUNA TWITTER DENGAN METODE DATA MINING DAN TEXT MINING*.

Miquel Yosafat, & Jatmika, J. (2024). Implementasi Text Clustering Terkait Pilpres 2024 Menggunakan Metode K-Means. *JURNAL SAINS DAN KOMPUTER*, *8*(01), 6–12.

Muhammad Imam Ghozali, Wibowo Harry Sugiharto, A. F. I. (2022). Analisis Sentimen Pinjaman Online Di Media Sosial Twitter Menggunakan Metode Naive Bayes. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *33*(1), 1–12.

Nisa, A., Darwiyanto, E., & Asror, I. (n.d.). *Analisis Sentimen Menggunakan Naive Bayes Classifier dengan Chi-Square Feature Selection Terhadap Penyedia Layanan Telekomunikasi*. N. Jannah and T. Yulianto, "Mengelompokkan Siswa Berprestasi Akademik Dengan," vol. 2, no. 2, 2016

Oon Wira Yuda, Darmawan Tuti, Lim Sheih Yee, & Susanti. (2022a). Penerapan Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest. *SATIN - Sains Dan Teknologi Informasi*, *8*(2), 122–131.

Program, E. I., Sistem, S., Kampus, I. A., & Bogor, K. (2019). *Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes*. *VII*(1).

Puad, S., & Susilo Yuda Irawan, A. (2023). ANALISIS SENTIMEN MASYARAKAT PADA TWITTER TERHADAP PEMILIHAN UMUM 2024 MENGGUNAKAN ALGORITMA NAÏVE BAYES. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 7, Issue 3).

Putra Sugiarto, A. (2020a). *Seminar Nasional Hasil Penelitian dan Pengabdian 2020 IBI DARMAJAYA Bandar Lampung*.

Rahman, A. T. (n.d.). *Coal Trade Data Clusterung Using K-Means (Case Study PT. Global Bangkit Utama)*.

Rerung, R. R. (2018). Penerapan Data Mining dengan Memanfaatkan Metode Association Rule untuk Promosi Produk. *Jurnal Teknologi Rekayasa*, *3*(1), 89.

Rohmawati, N. W., Defiyanti, S., Jajuli, M., & Informatika Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang Jl Ronggowaluyo Telukjambe Timur Karawang, T. H. (2015a). IMPLEMENTASI ALGORITMA K-MEANS DALAM PENGKLASTERAN MAHASISWA PELAMAR BEASISWA. In *Jurnal Ilmiah Teknologi Informasi Terapan: Vol. I* (Issue 2).

Susanto, A., Asy, H., & Sri Hardjanto, U. (2016). ANALISIS PELAKSANAAN TUGAS DAN WEWENANG PANITIA PENGAWAS PEMILU DALAM PEMILIHAN UMUM PRESIDEN DAN WAKIL PRESIDEN YANG DEMOKRATIS DI KOTA SEMARANG TAHUN 2014. In *DIPONEGORO LAW REVIEW* (Vol. 5, Issue 2).

S. R. Elisabet, A. M Khairul, R. Rahmaddeni, and N. U. Aniq, "Perbandingan Algoritma Svm Dan Nbc Dalam Analisa Sentimen Pilkada Pada Twitter," CSRID J., vol. 13, no. 3, pp. 169–179,

2021.

Tineges, R., Triayudi, A., & Sholihati, I. D. (2020a). Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM). *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *4*(3), 650.

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *4*(2), 437.

Wanto, A., Siregar, M. N. H., Windarto, A. P., Hartama, D., Ginantra, N. L. W. S. R., Napitupulu, D., Negara, E. S., Lubis, M. R., Dewi, S. V., & Limbong, C. P. T. (2020). Data Mining: Algoritma dan Implementasi. Medan: Yayasan bKita Menulis.

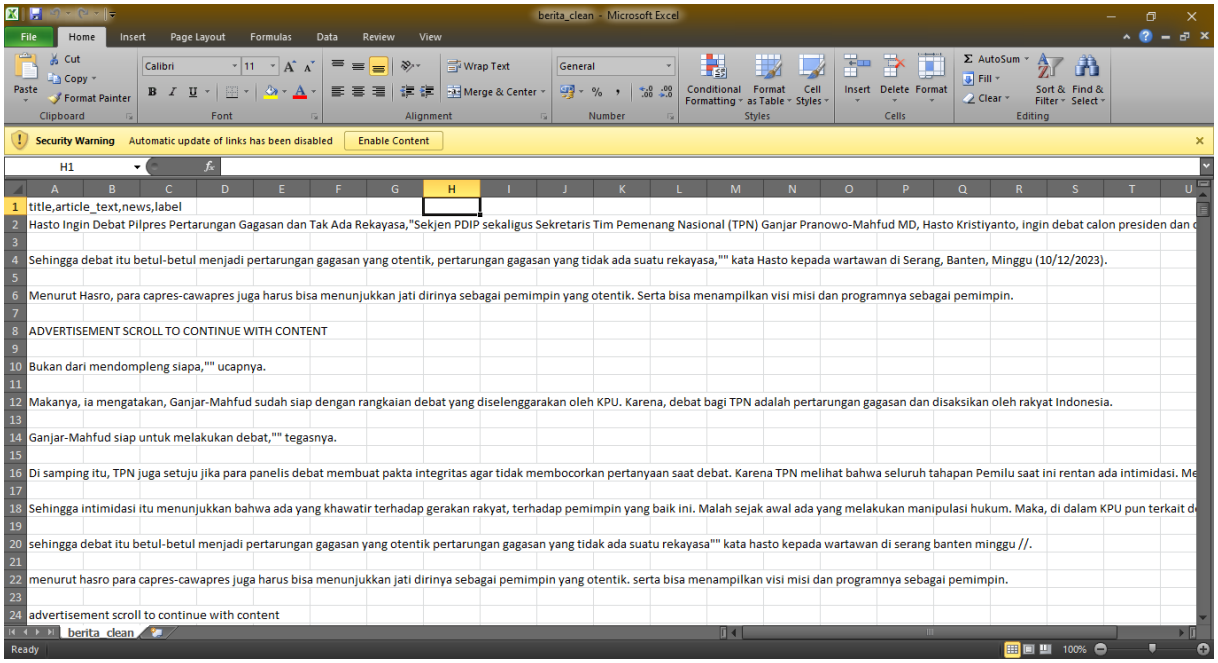
Wandani, A. (2021). Sentimen Analisis Pengguna Twitter pada Event Flash Sale Menggunakan Algoritma K-NN , Random Forest , dan Naive Bayes. 5(September), 651–665.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (n.d.). *Algorithms: The basic methods Most of these slides (used with permission) are based on the book: Data Mining: Practical Machine Learning Tools and Techniques*.

Zaki Hariansyah, M. (2022). Implementasi Metode Multinomial Naive Bayes pada Analisis Sentimen Terhadap Layanan Aplikasi Livin by Mandiri Implementation of Naive Bayes Multinomial Method on Sentiment Analysis of Livin by Mandiri Application Services. In *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI) Jakarta-Indonesia*.

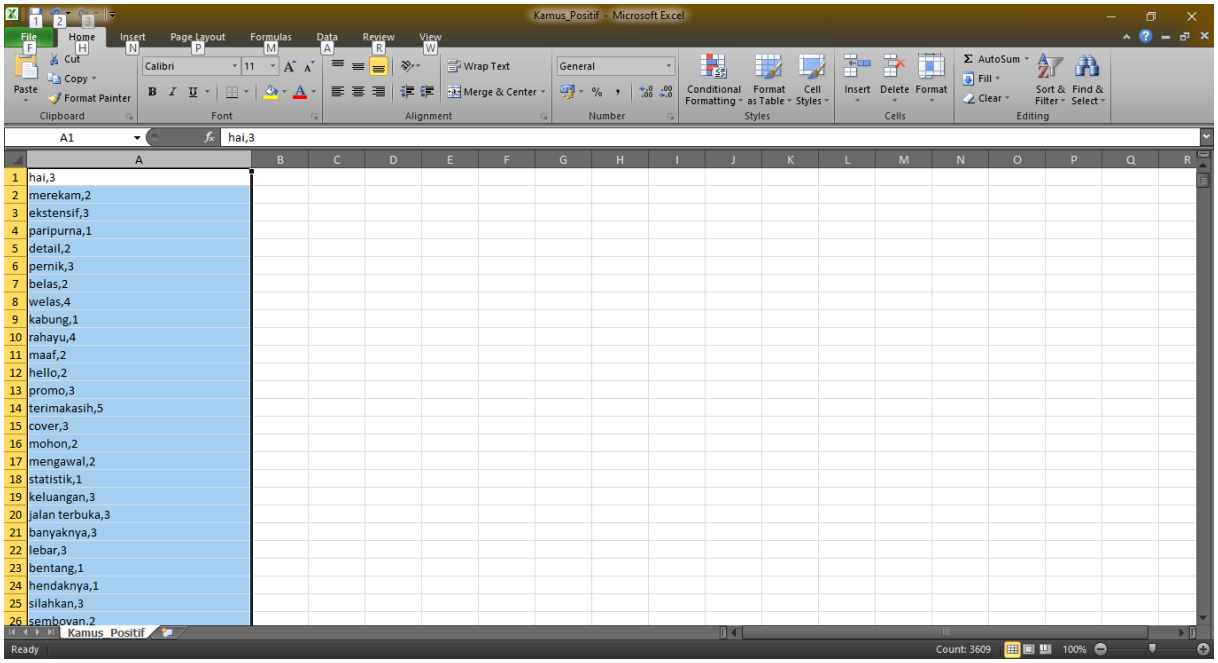
LAMPIRAN

1.DATA CLEAN PEMILU 2024



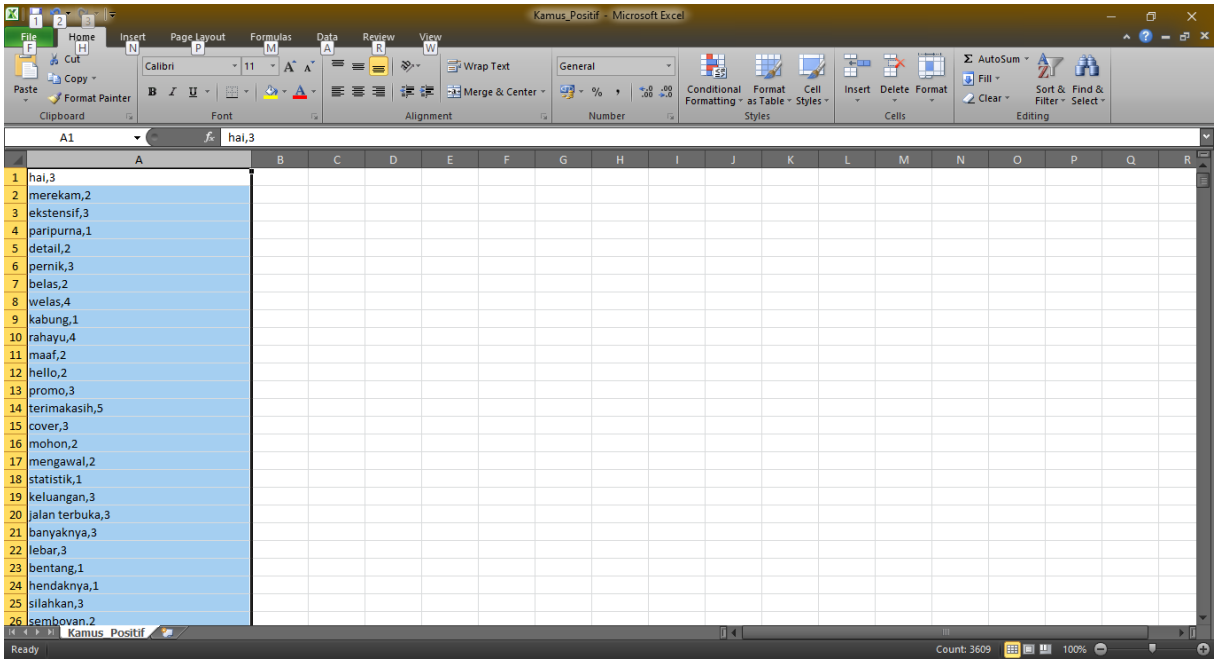
	A
1	title,article_text,news,label
2	Hasto Ingin Debat Pilpres Pertarungan Gagasan dan Tak Ada Rekayasa,"Sekjen PDIP sekaligus Sekretaris Tim Pemenang Nasional (TPN) Ganjar Pranowo-Mahfud MD, Hasto Kristiyanto, ingin debat calon presiden dan c
3	
4	Sehingga debat itu betul-betul menjadi pertarungan gagasan yang otentik, pertarungan gagasan yang tidak ada suatu rekayasa,"" kata Hasto kepada wartawan di Serang, Banten, Minggu (10/12/2023).
5	
6	Menurut Hasro, para capres-cawapres juga harus bisa menunjukkan jati dirinya sebagai pemimpin yang otentik. Serta bisa menampilkan visi misi dan programnya sebagai pemimpin.
7	
8	ADVERTISEMENT SCROLL TO CONTINUE WITH CONTENT
9	
10	Bukan dari mendompleng siapa,"" ucapnya.
11	
12	Makanya, ia mengatakan, Ganjar-Mahfud sudah siap dengan rangkaian debat yang diselenggarakan oleh KPU. Karena, debat bagi TPN adalah pertarungan gagasan dan disaksikan oleh rakyat Indonesia.
13	
14	Ganjar-Mahfud siap untuk melakukan debat,"" tegasnya.
15	
16	Di samping itu, TPN juga setuju jika para panelis debat membuat pakta integritas agar tidak membocorkan pertanyaan saat debat. Karena TPN melihat bahwa seluruh tahapan Pemilu saat ini rentan ada intimidasi. Me
17	
18	Sehingga intimidasi itu menunjukkan bahwa ada yang khawatir terhadap gerakan rakyat, terhadap pemimpin yang baik ini. Malah sejak awal ada yang melakukan manipulasi hukum. Maka, di dalam KPU pun terkait d
19	
20	sehingga debat itu betul-betul menjadi pertarungan gagasan yang otentik pertarungan gagasan yang tidak ada suatu rekayasa"" kata hasto kepada wartawan di serang banten minggu //.
21	
22	menurut hasro para capres-cawapres juga harus bisa menunjukkan jati dirinya sebagai pemimpin yang otentik. serta bisa menampilkan visi misi dan programnya sebagai pemimpin.
23	
24	advertisement scroll to continue with content

2.DATA KAMUS POSITIF



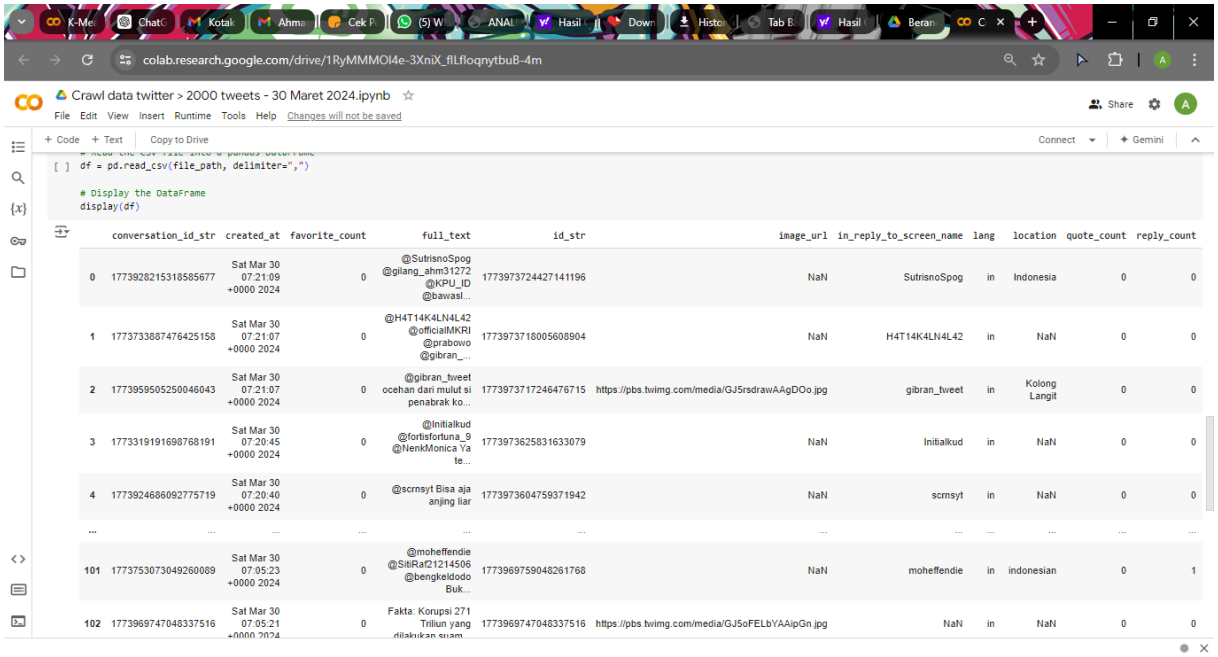
	A
1	hai,3
2	merekam,2
3	ekstensif,3
4	paripurna,1
5	detail,2
6	pernik,3
7	belas,2
8	welas,4
9	kabung,1
10	rahayu,4
11	maaf,2
12	hello,2
13	promo,3
14	terimakasih,5
15	cover,3
16	mohon,2
17	mengawal,2
18	statistik,1
19	keluangan,3
20	jalan terbuka,3
21	banyaknya,3
22	lebar,3
23	bentang,1
24	hendaknya,1
25	silahkan,3
26	sembovan,2

DATA KAMUS NEGATIF



	A
1	hai,3
2	merekam,2
3	ekstensif,3
4	paripurna,1
5	detail,2
6	pernik,3
7	belas,2
8	welas,4
9	kabung,1
10	rahayu,4
11	maaf,2
12	hello,2
13	promo,3
14	terimakasih,5
15	cover,3
16	mohon,2
17	mengawal,2
18	statistik,1
19	keluangan,3
20	jalan terbuka,3
21	banyaknya,3
22	lebar,3
23	bentang,1
24	hendaknya,1
25	silahkan,3
26	sembovan,2

CRAWLING DATA

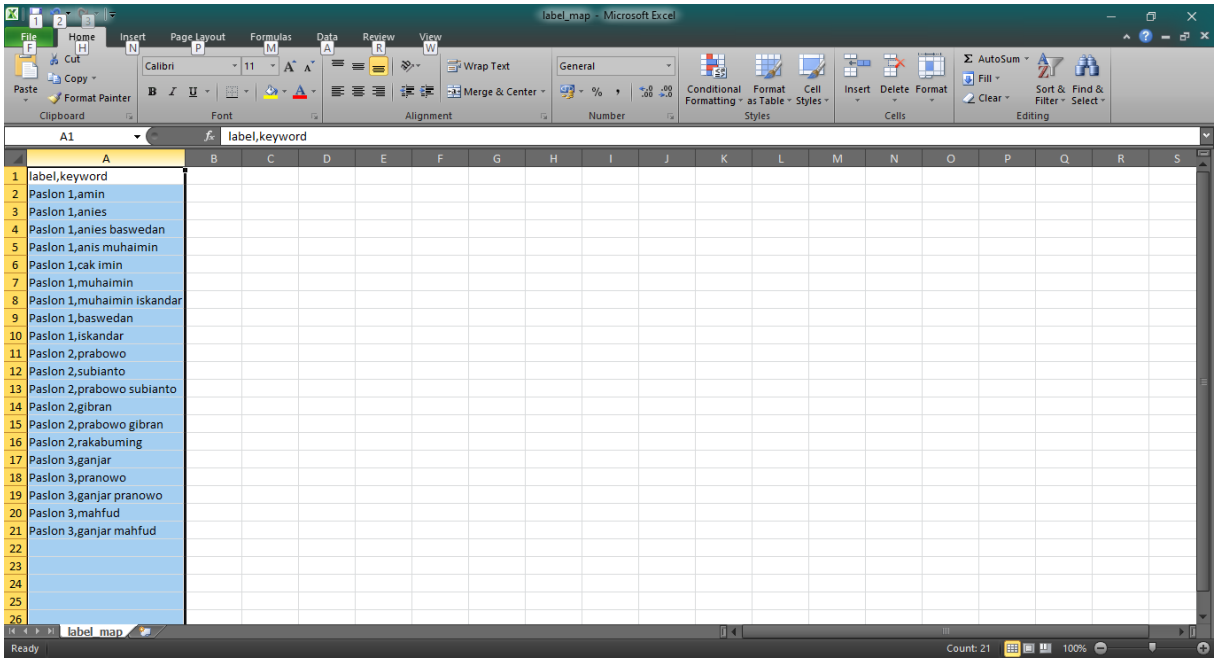


Crawl data twitter > 2000 tweets - 30 Maret 2024.ipynb

```
df = pd.read_csv(file_path, delimiter=",")

# Display the DataFrame
display(df)
```

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count
0	1773928215318585677	Sat Mar 30 07:21:09 +0000 2024	0	@SutrisnoSpog @gilang_ahm31272 @KPU_ID @bawasl...	1773973724427141196	NaN	SutrisnoSpog	in	Indonesia	0	0
1	1773733887476425158	Sat Mar 30 07:21:07 +0000 2024	0	@H4T14K4LN4L42 @officialMKRI @prabowo @gibran_...	1773973718005608904	NaN	H4T14K4LN4L42	in	NaN	0	0
2	1773959505250046043	Sat Mar 30 07:21:07 +0000 2024	0	@gibran_tweet ocehan dari mulut si penabrak ko...	1773973717246476715	https://pbs.twimg.com/media/GJ5rsdrawAAgDOo.jpg	gibran_tweet	in	Kolong Langit	0	0
3	1773319191698768191	Sat Mar 30 07:20:45 +0000 2024	0	@Initialkud @fortisfortuna_9 @NenkMonica Ya te...	1773973625831633079	NaN	Initialkud	in	NaN	0	0
4	1773924686092775719	Sat Mar 30 07:20:40 +0000 2024	0	@scrnsyt Bisa aja anjing liar	1773973604759371942	NaN	scrnsyt	in	NaN	0	0
...	...	...	...	...	...	...	...	...	...	...	...
101	1773753073049260089	Sat Mar 30 07:05:23 +0000 2024	0	@moheffendie @SitiRaf21214506 @bengkeldodo Buk...	1773969759048261768	NaN	moheffendie	in	indonesian	0	1
102	1773969747048337516	Sat Mar 30 07:05:21 +0000 2024	0	Fakta: Korupsi 271 Triliun yang dilakukan suam...	1773969747048337516	https://pbs.twimg.com/media/GJ5oFELbYAAipGn.jpg	NaN	in	NaN	0	0

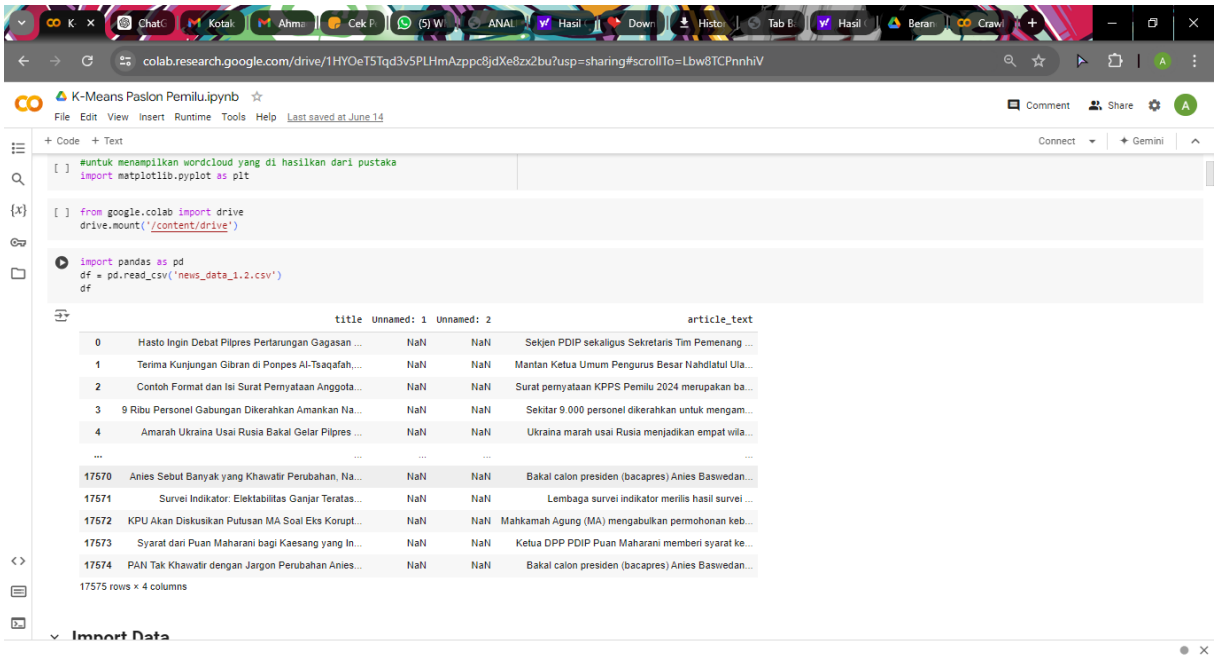
LEBELING DATA



	A
1	label,keyword
2	Paslon 1,amin
3	Paslon 1,anies
4	Paslon 1,anies baswedan
5	Paslon 1,anis muhaimin
6	Paslon 1,cak imin
7	Paslon 1,muhaimin
8	Paslon 1,muhaimin iskandar
9	Paslon 1,baswedan
10	Paslon 1,iskandar
11	Paslon 2,prabowo
12	Paslon 2,subianto
13	Paslon 2,prabowo subianto
14	Paslon 2,gibran
15	Paslon 2,prabowo gibran
16	Paslon 2,rakabuming
17	Paslon 3,ganjar
18	Paslon 3,pranowo
19	Paslon 3,ganjar pranowo
20	Paslon 3,mahfud
21	Paslon 3,ganjar mahfud
22	
23	
24	
25	
26	

TAMPILAN PADA SISTEM

1. INPUT DATA



```
#untuk menampilkan wordcloud yang di hasilkan dari pustaka
import matplotlib.pyplot as plt
```

```
from google.colab import drive
drive.mount('/content/drive')
```

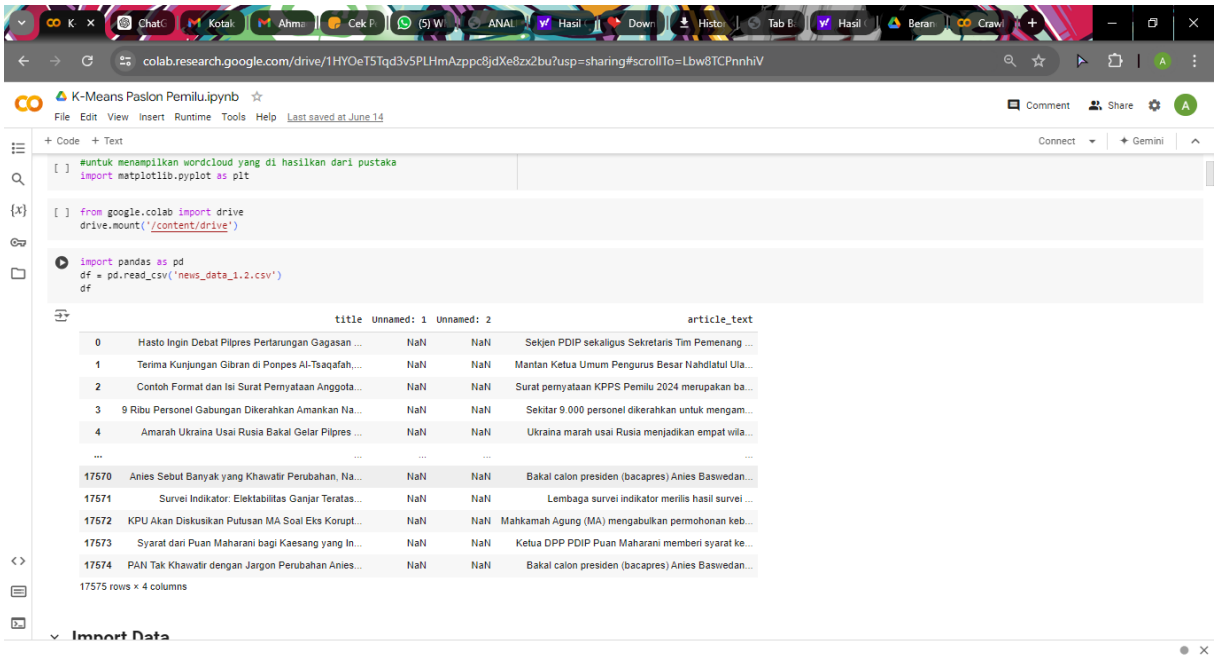
```
import pandas as pd
df = pd.read_csv('news_data_1.2.csv')
df
```

	title	Unnamed: 1	Unnamed: 2	article_text
0	Hasto Ingin Debat Pilpres Pertarungan Gagasan ...	NaN	NaN	Sekjen PDIP sekaligus Sekretaris Tim Pemenang ...
1	Terima Kunjungan Gibran di Ponpes Al-Tsaqafah,...	NaN	NaN	Mantan Ketua Umum Pengurus Besar Nahdlatul Ula...
2	Contoh Format dan Isi Surat Pernyataan Anggota...	NaN	NaN	Surat pernyataan KPPS Pemilu 2024 merupakan ba...
3	9 Ribu Personel Gabungan Dikerahkan Amankan Na...	NaN	NaN	Sekitar 9.000 personel dikerahkan untuk mengam...
4	Amarah Ukraina Usai Rusia Bakal Gelar Pilpres ...	NaN	NaN	Ukraina marah usai Rusia menjadikan empat wila...
...	...	...	...	...
17570	Anies Sebut Banyak yang Khawatir Perubahan, Na...	NaN	NaN	Bakal calon presiden (bacapres) Anies Baswedan...
17571	Survei Indikator: Elektabilitas Ganjar Teratas...	NaN	NaN	Lembaga survei indikator merilis hasil survei ...
17572	KPU Akan Diskusikan Putusan MA Soal Eks Korupt...	NaN	NaN	Mahkamah Agung (MA) mengabulkan permohonan keb...
17573	Syarat dari Puan Maharani bagi Kaesang yang In...	NaN	NaN	Ketua DPP PDIP Puan Maharani memberi syarat ke...
17574	PAN Tak Khawatir dengan Jargon Perubahan Anies...	NaN	NaN	Bakal calon presiden (bacapres) Anies Baswedan...

17575 rows × 4 columns

Import Data

PROSES PREPROCESSING DATA



K-Means Paslon Pemilu.ipynb

File Edit View Insert Runtime Tools Help Last saved at June 14

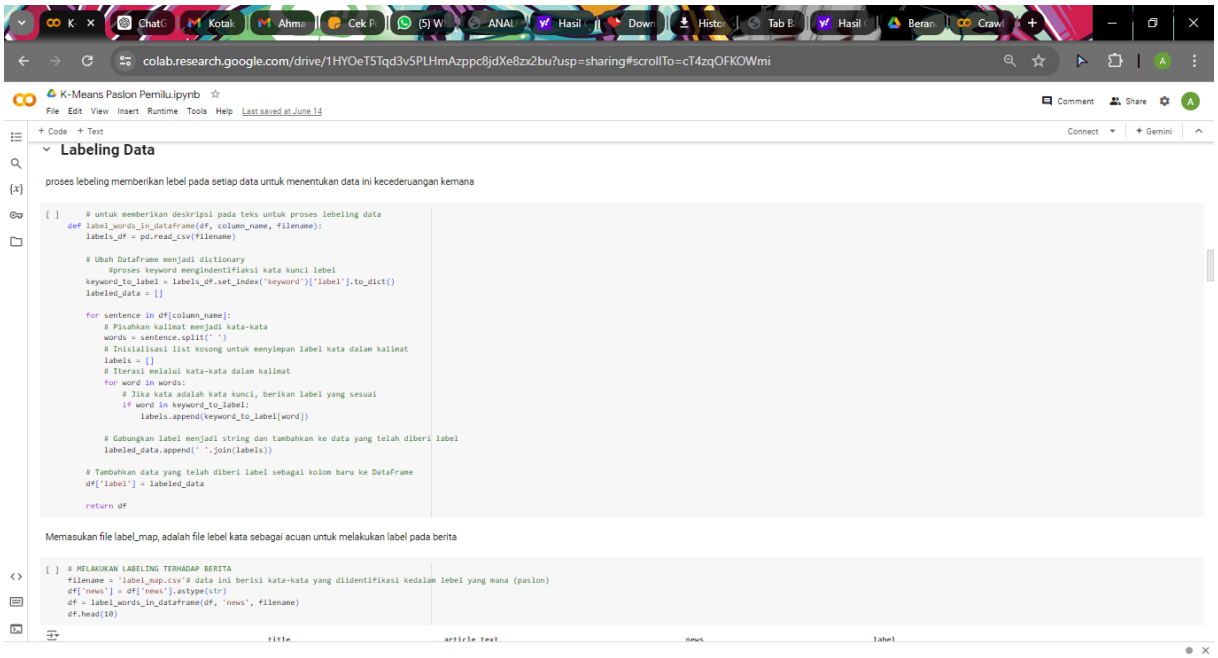
```
[ ] #untuk menampilkan wordcloud yang di hasilkan dari pustaka
    import matplotlib.pyplot as plt

[ ] from google.colab import drive
    drive.mount('/content/drive')

    import pandas as pd
    df = pd.read_csv('news_data_1.2.csv')
    df
```

	title	Unnamed: 1	Unnamed: 2	article_text
0	Hasto Ingin Debat Pilpres Pertarungan Gagasan ...	NaN	NaN	Sekjen PDIP sekaligus Sekretaris Tim Pemenang ...
1	Terima Kunjungan Gibran di Ponpes Al-Tsaqafah,...	NaN	NaN	Mantan Ketua Umum Pengurus Besar Nahdlatul Ula...
2	Contoh Format dan Isi Surat Pernyataan Anggota...	NaN	NaN	Surat pernyataan KPPS Pemilu 2024 merupakan ba...
3	9 Ribu Personel Gabungan Dikerahkan Amankan Na...	NaN	NaN	Sekitar 9.000 personel dikerahkan untuk mengam...
4	Amarah Ukraina Usai Rusia Bakal Gelar Pilpres ...	NaN	NaN	Ukraina marah usai Rusia menjadikan empat wila...
...	...	...	...	...
17570	Anies Sebut Banyak yang Khawatir Perubahan, Na...	NaN	NaN	Bakal calon presiden (bacapres) Anies Baswedan...
17571	Survei Indikator: Elektabilitas Ganjar Teratas...	NaN	NaN	Lembaga survei indikator merilis hasil survei ...
17572	KPU Akan Diskusikan Putusan MA Soal Eks Korupt...	NaN	NaN	Mahkamah Agung (MA) mengabulkan permohonan keb...
17573	Syarat dari Puan Maharani bagi Kaesang yang In...	NaN	NaN	Ketua DPP PDIP Puan Maharani memberi syarat ke...
17574	PAN Tak Khawatir dengan Jargon Perubahan Anies...	NaN	NaN	Bakal calon presiden (bacapres) Anies Baswedan...

17575 rows × 4 columns

Import Data

PROSES LEBELING DATA



K-Means Paslon Pemilu.ipynb

File Edit View Insert Runtime Tools Help Last saved at June 14

Labeling Data

proses lebeling memberikan lebel pada setiap data untuk menentukan data ini kecederuangan kemana

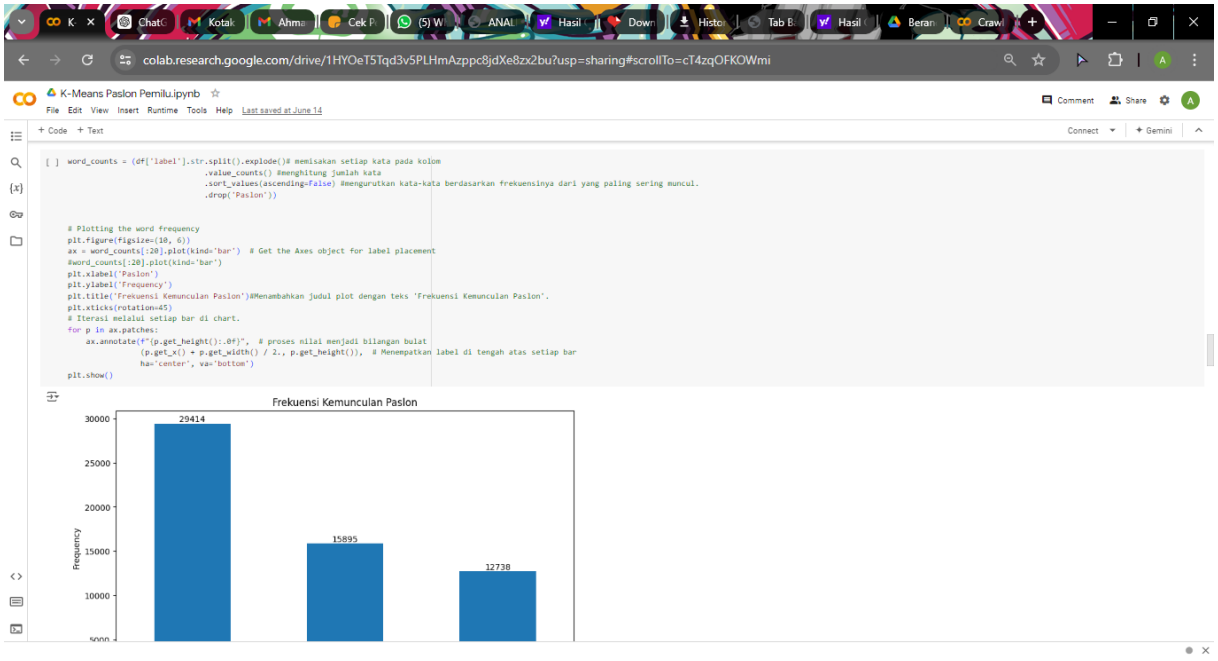
```
[ ]  # untuk memberikan deskripsi pada teks untuk proses lebeling data
     def label_words_in_dataframe(df, column_name, filename):
         labels_df = pd.read_csv(filename)

         # Ubah DataFrame menjadi dictionary
             #proses keyword mengindentifiaksi kata kunci lebel
         keyword_to_label = labels_df.set_index('keyword')['label'].to_dict()
         labeled_data = []

         for sentence in df[column_name]:
             # Pisahkan kalimat menjadi kata-kata
             words = sentence.split(' ')
             # Inisialisasi list kosong untuk menyimpan label kata dalam kalimat
             labels = []
             # Iterasi melalui kata-kata dalam kalimat
             for word in words:
                 # Jika kata adalah kata kunci, berikan label yang sesuai
                 if word in keyword_to_label:
                     labels.append(keyword_to_label[word])

             # Gabungkan label menjadi string dan tambahkan ke data yang telah diberi label
             labeled_data.append(' '.join(labels))

         # Tambahkan data yang telah diberi label sebagai kolom baru ke DataFrame
         df['label'] = labeled_data

         return df
```

Memasukan file label_map, adalah file lebel kata sebagai acuan untuk melakukan label pada berita

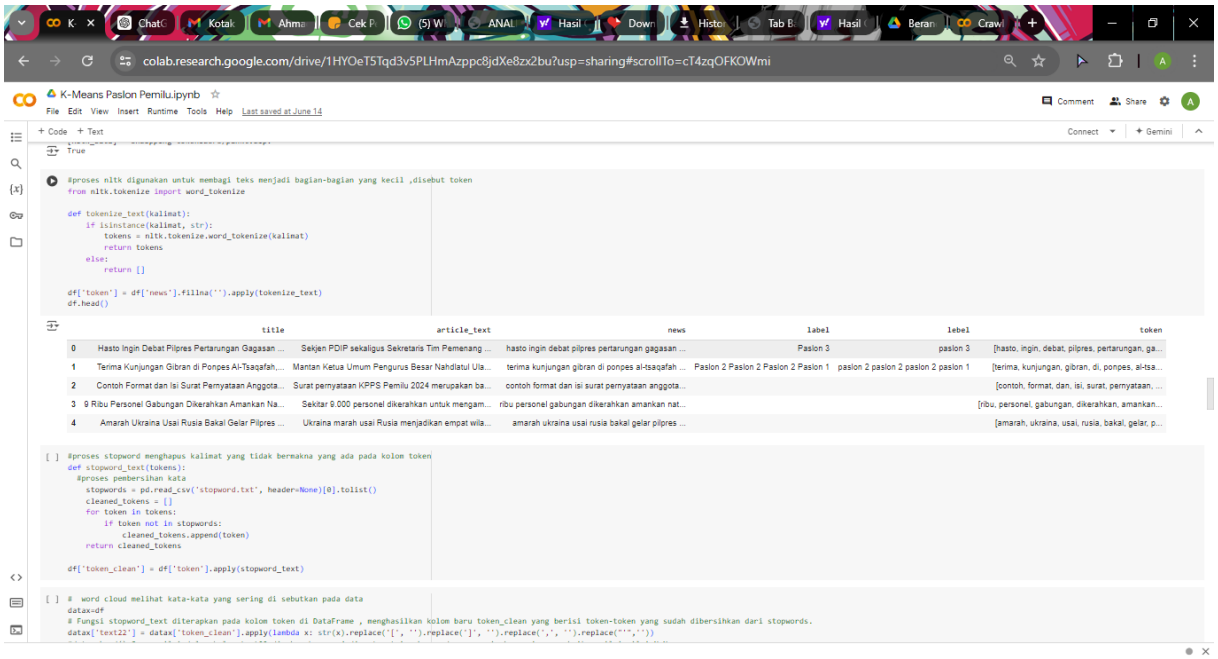
```
[ ] # MELAKUKAN LABELING TERHADAP BERITA
    filename = 'label_map.csv'# data ini berisi kata-kata yang diidentifikasi kedalam lebel yang mana (paslon)
    df['news'] = df['news'].astype(str)
    df = label_words_in_dataframe(df, 'news', filename)
    df.head(10)
```

	title	article_text	news	label

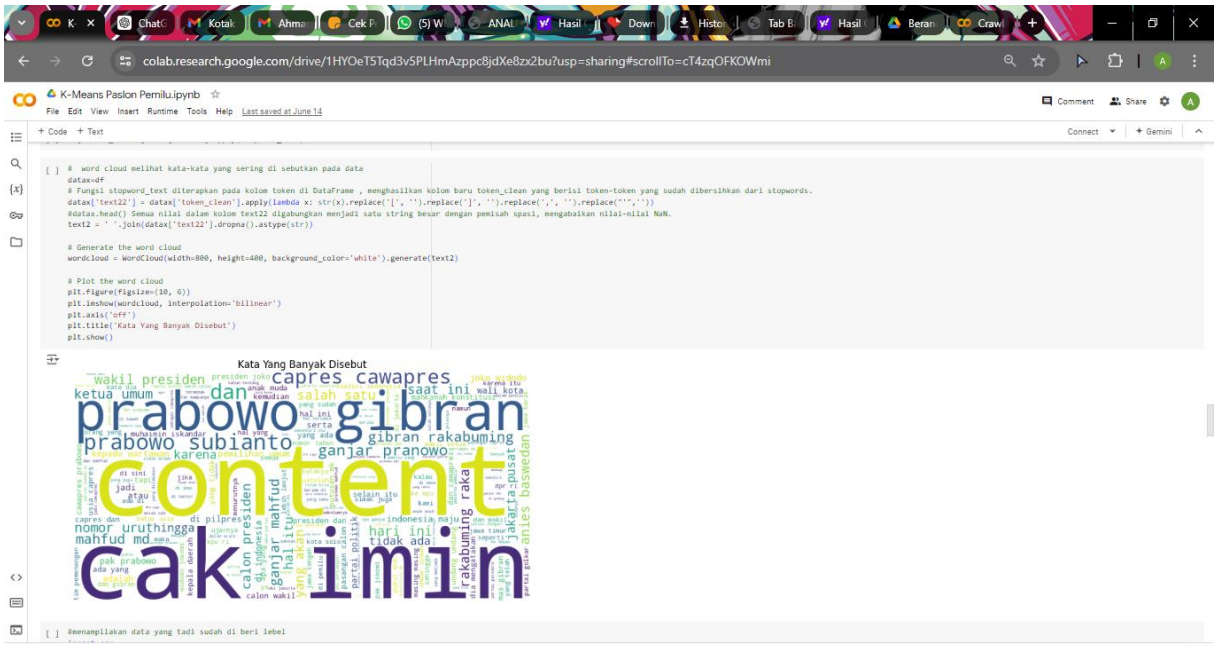
PROSES KEMUNCULAN FREKUANSI KATA



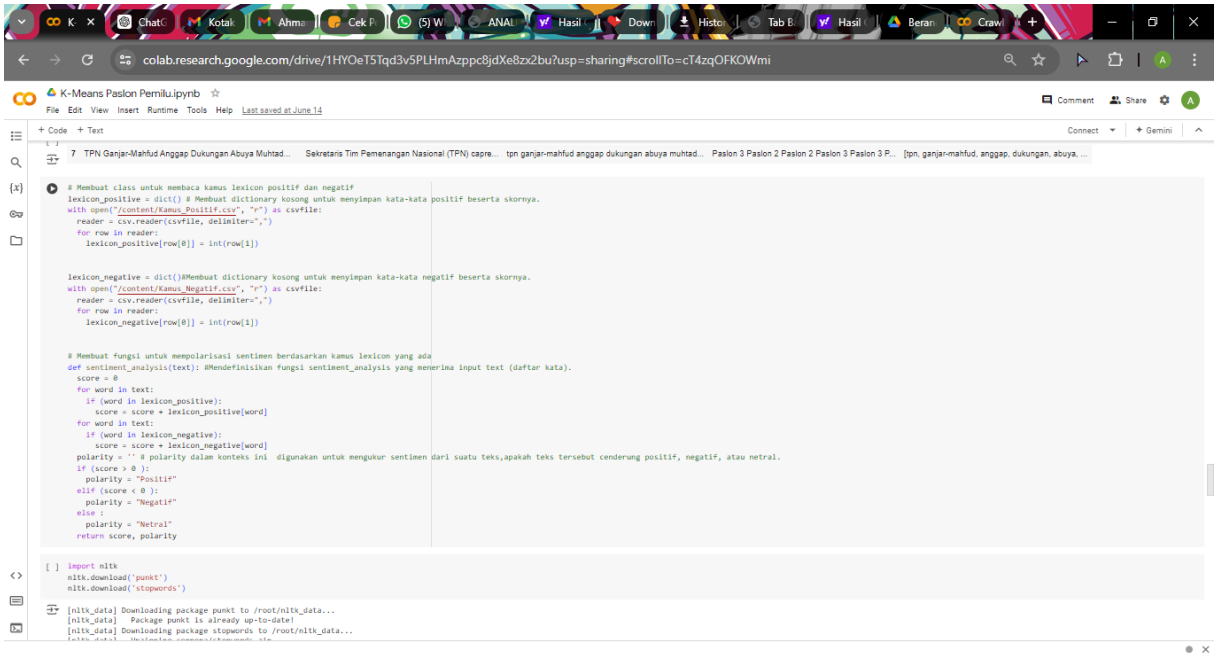
PROSES TOKENIZE



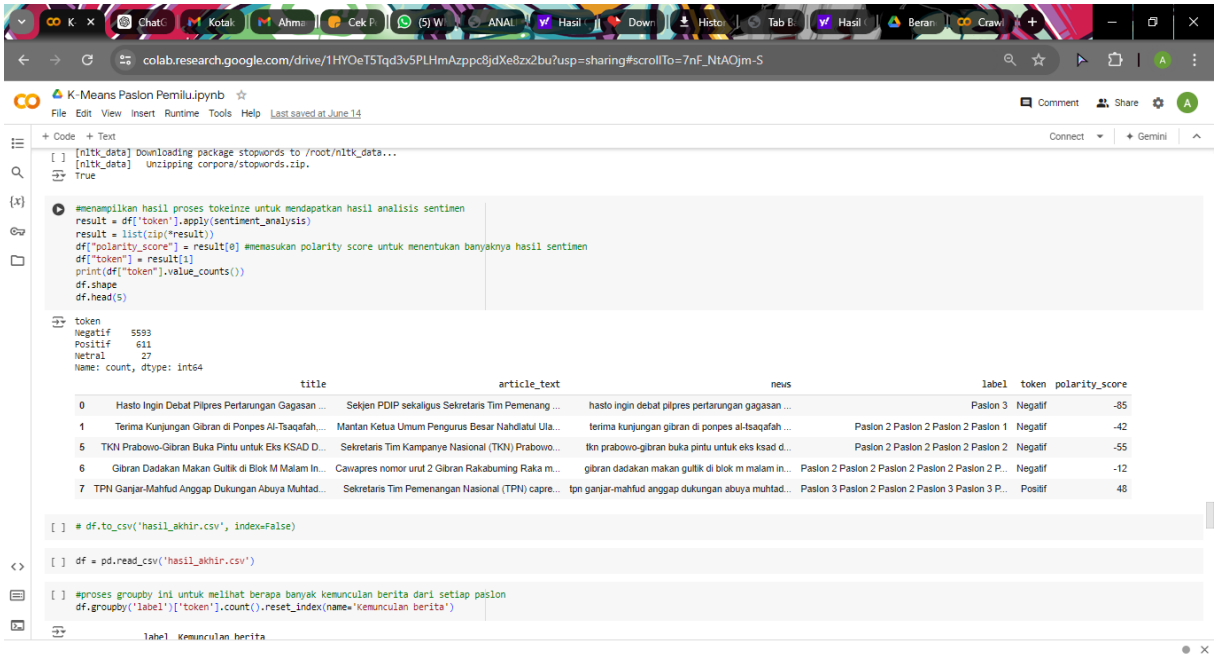
MENAMPILKAN WORDCLOUD



PROSES ANALISIS SENTIMEN



HASIL ANALISIS SENTIMEN



K-Means Paslon Pemilu.ipynb

File Edit View Insert Runtime Tools Help Last saved at June 14

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
#menampilkan hasil proses tokeinze untuk mendapatkan hasil analisis sentimen
result = df['token'].apply(sentiment_analysis)
result = list(zip(*result))
df["polarity_score"] = result[0] #memasukan polarity score untuk menentukan banyaknya hasil sentimen
df["token"] = result[1]
print(df["token"].value_counts())
df.shape
df.head(5)
```

```
token
Negatif    5593
Positif     611
Netral       27
Name: count, dtype: int64
```

	title	article_text	news	label	token	polarity_score
0	Hasto Ingin Debat Pilpres Pertarungan Gagasan ...	Sekjen PDIP sekaligus Sekretaris Tim Pemenang ...	hasto ingin debat pilpres pertarungan gagasan ...	Paslon 3	Negatif	-85
1	Terima Kunjungan Gibran di Ponpes Al-Tsaqafah,...	Mantan Ketua Umum Pengurus Besar Nahdlatul Ula...	terima kunjungan gibran di ponpes al-tsaqafah ...	Paslon 2 Paslon 2 Paslon 2 Paslon 1	Negatif	-42
5	TKN Prabowo-Gibran Buka Pintu untuk Eks KSAD D...	Sekretaris Tim Kampanye Nasional (TKN) Prabowo...	tkn prabowo-gibran buka pintu untuk eks ksad d...	Paslon 2 Paslon 2 Paslon 2 Paslon 2	Negatif	-55
6	Gibran Dadakan Makan Gultik di Blok M Malam In...	Cawapres nomor urut 2 Gibran Rakabuming Raka m...	gibran dadakan makan gultik di blok m malam in...	Paslon 2 Paslon 2 Paslon 2 Paslon 2 Paslon 2 P...	Negatif	-12
7	TPN Ganjar-Mahfud Anggap Dukungan Abuya Muhtad...	Sekretaris Tim Pemenangan Nasional (TPN) capre...	tpn ganjar-mahfud anggap dukungan abuya muhtad...	Paslon 3 Paslon 2 Paslon 2 Paslon 3 Paslon 3 P...	Positif	48

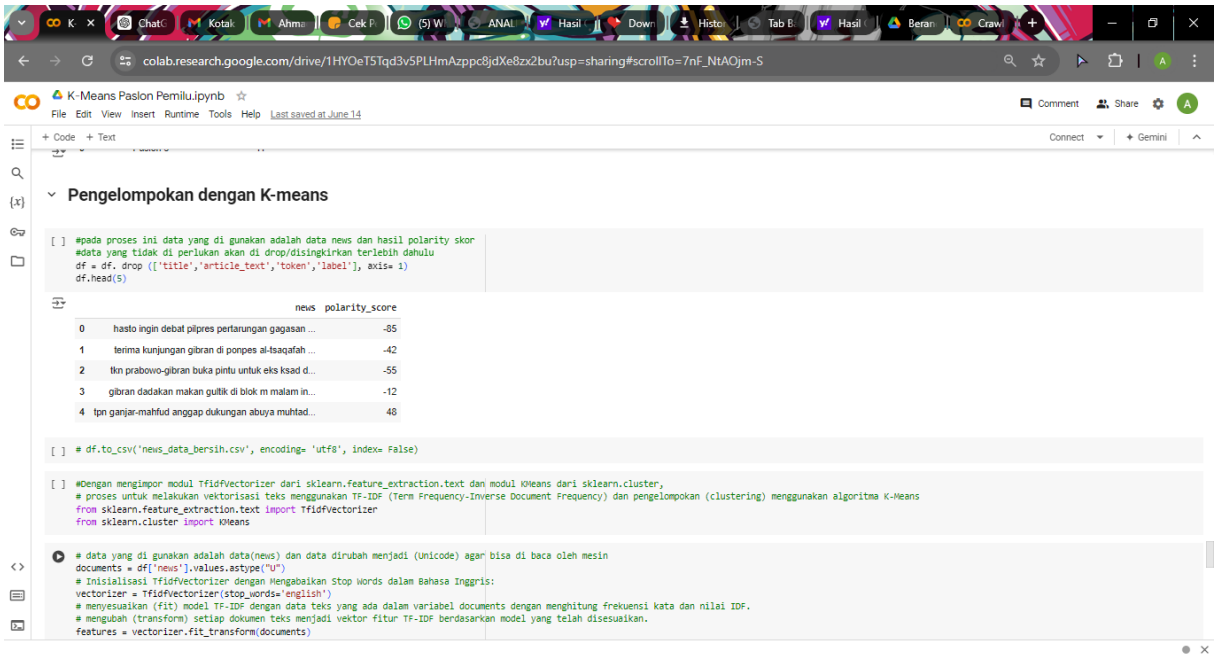
```
# df.to_csv('hasil_akhir.csv', index=False)
```

```
df = pd.read_csv('hasil_akhir.csv')
```

```
#proses groupby ini untuk melihat berapa banyak kemunculan berita dari setiap paslon
df.groupby('label')['token'].count().reset_index(name='Kemunculan berita')
```

```
label  Kemunculan berita
```

PROSES PENGKELOMPKAN K-MEANS



K-Means Paslon Pemilu.ipynb

File Edit View Insert Runtime Tools Help Last saved at June 14

Pengelompokan dengan K-means

```
#pada proses ini data yang di gunakan adalah data news dan hasil polarity skor
#data yang tidak di perlukan akan di drop/disingkirkan terlebih dahulu
df = df. drop (['title','article_text','token','label'], axis= 1)
df.head(5)
```

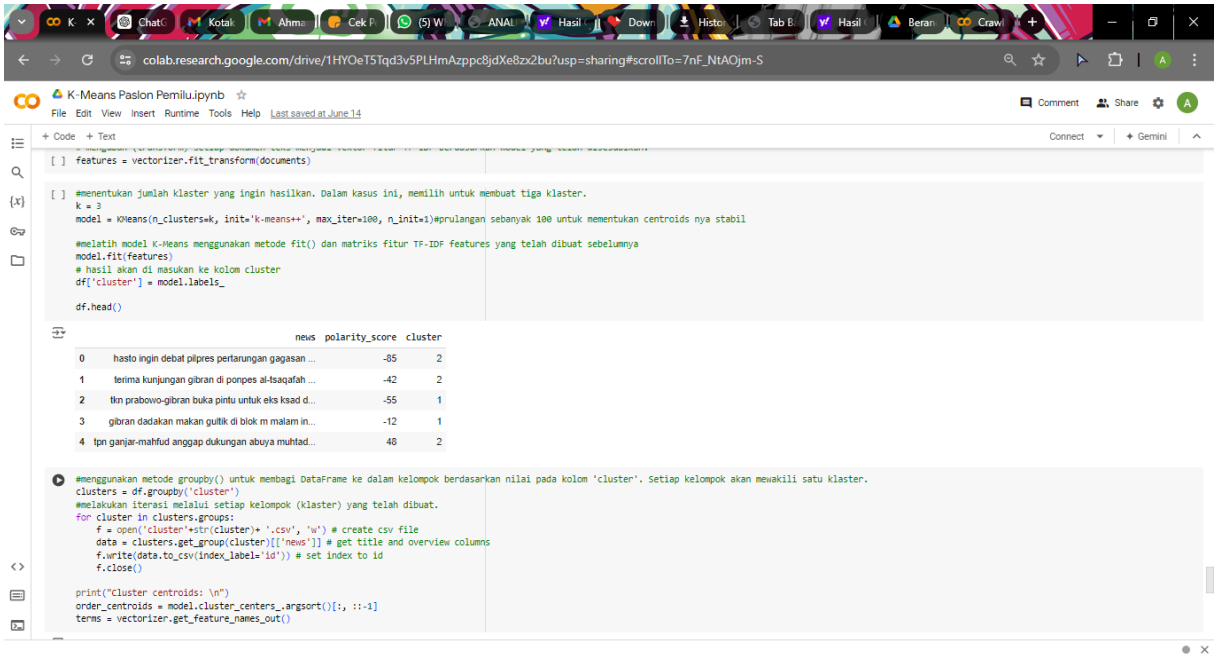
	news	polarity_score
0	hasto ingin debat pilpres pertarungan gagasan ...	-85
1	terima kunjungan gibran di ponpes al-tsaqafah ...	-42
2	tkn prabowo-gibran buka pintu untuk eks ksad d...	-55
3	gibran dadakan makan gultik di blok m malam in...	-12
4	tpn ganjar-mahfud anggap dukungan abuya muhtad...	48

```
# df.to_csv('news_data_bersih.csv', encoding= 'utf8', index= False)
```

```
#Dengan mengimpor modul TfidfVectorizer dari sklearn.feature_extraction.text dan modul KMeans dari sklearn.cluster,
# proses untuk melakukan vektorisasi teks menggunakan TF-IDF (Term Frequency-Inverse Document Frequency) dan pengelompokan (clustering) menggunakan algoritma K-Means
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
```

```
# data yang di gunakan adalah data(news) dan data dirubah menjadi (Unicode) agar bisa di baca oleh mesin
documents = df['news'].values.astype("U")
# Inisialisasi TfidfVectorizer dengan Mengabaikan Stop Words dalam Bahasa Inggris:
vectorizer = TfidfVectorizer(stop_words='english')
# menyesuaikan (fit) model TF-IDF dengan data teks yang ada dalam variabel documents dengan menghitung frekuensi kata dan nilai IDF.
# mengubah (transform) setiap dokumen teks menjadi vektor fitur TF-IDF berdasarkan model yang telah disesuaikan.
features = vectorizer.fit_transform(documents)
```

HASIL PENGKELOMPOKAN K-MEANS



K-Means Paslon Pemilu.ipynb

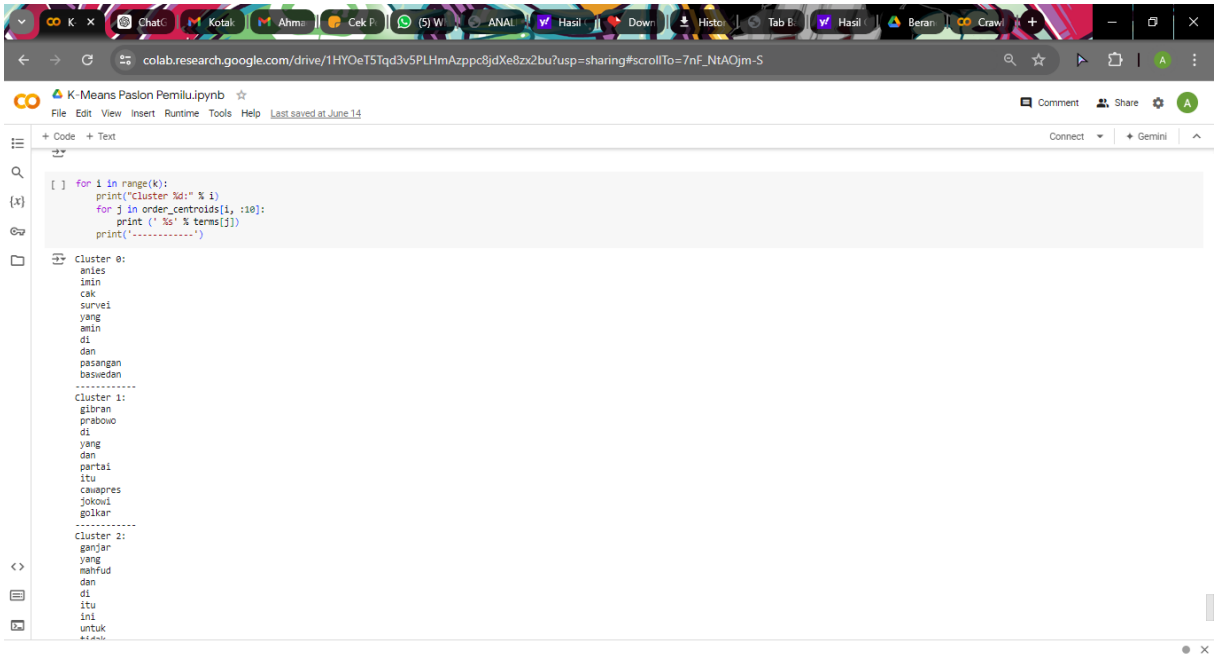
File Edit View Insert Runtime Tools Help Last saved at June 14

```
features = vectorizer.fit_transform(documents)
```

```
#menentukan jumlah klaster yang ingin hasilkan. Dalam kasus ini, memilih untuk membuat tiga klaster.
k = 3
model = KMeans(n_clusters=k, init='k-means++', max_iter=100, n_init=1)#prulangan sebanyak 100 untuk mementukan centroids nya stabil

#melatih model K-Means menggunakan metode fit() dan matriks fitur TF-IDF features yang telah dibuat sebelumnya
model.fit(features)
# hasil akan di masukan ke kolom cluster
df['cluster'] = model.labels_

df.head()
```

	news	polarity_score	cluster
0	hasto ingin debat pilpres pertarungan gagasan ...	-85	2
1	terima kunjungan gibran di ponpes al-tsaqafah ...	-42	2
2	tkn prabowo-gibran buka pintu untuk eks ksad d...	-55	1
3	gibran dadakan makan gultik di blok m malam in...	-12	1
4	tpn ganjar-mahfud anggap dukungan abuya muhtad...	48	2

```
#menggunakan metode groupby() untuk membagi DataFrame ke dalam kelompok berdasarkan nilai pada kolom 'cluster'. Setiap kelompok akan mewakili satu klaster.
clusters = df.groupby('cluster')
#melakukan iterasi melalui setiap kelompok (klaster) yang telah dibuat.
for cluster in clusters.groups:
    f = open('cluster'+str(cluster)+ '.csv', 'w') # create csv file
    data = clusters.get_group(cluster)[['news']] # get title and overview columns
    f.write(data.to_csv(index_label='id')) # set index to id
    f.close()

print("Cluster centroids: \n")
order_centroids = model.cluster_centers_.argsort()[:, ::-1]
terms = vectorizer.get_feature_names_out()
```



K-Means Paslon Pemilu.ipynb

File Edit View Insert Runtime Tools Help Last saved at June 14

```
for i in range(k):
    print("Cluster %d:" % i)
    for j in order_centroids[i, :10]:
        print (' %s' % terms[j])
    print('------------')
```

```
Cluster 0:
 anies
 imin
 cak
 survei
 yang
 amin
 di
 dan
 pasangan
 baswedan
------------
Cluster 1:
 gibran
 prabowo
 di
 yang
 dan
 partai
 itu
 cawapres
 jokowi
 golkar
------------
Cluster 2:
 ganjar
 yang
 mahfud
 dan
 di
 itu
 ini
 untuk
```