KLASIFIKASI SENTIMEN *TWEET* UNTUK MENDETEKSI KONTEN PADA PLATFORM TWITTER MENGGUNAKAN *NATURAL LANGUAGE PROCESSSING* (NLP)

LAPORAN KERJA PRAKTIK

Program Studi

S1 Teknik Komputer

Oleh:

Muhammad Azhar Ali

21410200020

FAKULTAS TEKNOLOGI DAN INFORMATIKA

UNIVERSITAS DINAMIKA

2024

KLASIFIKASI SENTIMEN *TWEET* UNTUK MENDETEKSI KONTEN PADA PLATFORM TWITTER MENGGUNAKAN *NATURAL LANGUAGE PROCESSING* (NLP)

Diajukan sebagai salah satu syarat untuk menyelesaikan

Mata Kuliah Kerja Praktik

Disusun Oleh:

Nama : MUHAMMAD AZHAR ALI

NIM : 21410200020

Program : S1 (Strata Satu)

Jurusan : Teknik Komputer

FAKULTAS TEKNOLOGI DAN INFORMATIKA

UNIVERSITAS DINAMIKA

2024

LEMBAR PENGESAHAN

LEMBAR PENGESAHAN

KLASIFIKASI SENTIMEN TWEET UNTUK MENDETEKSI KONTEN PADA PLATFORN TWITTER MENGGUNAKAN NATURAL LANGUAGE PROCESSING (NLP)

Laporan Kerja Praktik oleh

Muhammad Azhar Ali

NIM: 21410200020

Telah diperiksa, diuji, dan disetujui

Surabaya, 24 Juli 2024

Disetujui:

Pembimbing	Penyelia
cn=Pauladie Susanto, o=Universitas Dinamika, ou=PS S1 Teknik Komputer, email=pauladie@dinamika.ac.id, c=ID 2024.08.05 09:44:41 +07'00'	
Pauladie Susanto, S.Kom., M.T.	**Lutfi Dwimulya**
NIDN. 0729047501	

Mengetahui,

Ketua Program Studi S1 Teknik Komputer

cn=Pauladie Susanto, o=Universitas Dinamika, ou=PS S1 Teknik Komputer, email=pauladie@dinamika.ac.id, c=ID 2024.08.05 09:45:08 +07'00'

Pauladie Susanto, S.Kom., M.T.

NIDN. 0729047501

PERNYATAAN

PERSETUJUAN PUBLIKASI DAN KEASLIAN KARYA ILMIAH

Sebagai mahasiswa **Universitas Dinamika**, Saya :

Nama	: **Muhammad Azhar Ali**
NIM	: **21410200020**
Program Studi	: **S1 Teknik Komputer**
Fakultas	: **Fakultas Teknologi dan Informatika**
Jenis Karya	: **Laporan Kerja Praktik**
Judul Karya	: **KLASIFIKASI SENTIMEN *TWEET* UNTUK MENDETEKSI KONTEN PADA PLATFORM *TWITTER* MENGGUNAKAN *NATURAL LANGUAGE PROCESSING* (NLP)**

Menyatakan dengan sesungguhnya bahwa :

1. Demi pengembangan Ilmu Pengetahuan, Teknologi dan Seni, Saya menyetujui memberikan kepada **Universitas Dinamika** Hak Bebas Royalti Non-Eksklusif (***Non-Exclusive Royalty Free Right***) atas seluruh isi/sebagian karya ilmiah Saya tersebut diatas untuk disimpan, dialihmediakan, dan dikelola dalam bentuk pangkalan data (***database***) untuk selanjutnya didistribusikan atau dipublikasikan demi kepentingan akademis dengan tetap mencantumkan nama Saya sebagai penulis atau pencipta dan sebagai pemilik Hak Cipta.
2. Karya tersebut diatas adalah hasil karya asli Saya, bukan plagiat baik sebagian maupun keseluruhan. Kutipan, karya, atau pendapat orang lain yang ada dalam karya ilmiah ini semata-mata hanya sebagai rujukan yang dicantumkan dalam Daftar Pustaka Saya.
3. Apabila dikemudian hari ditemukan dan terbukti terdapat tindakan plagiasi pada karya ilmiah ini, maka Saya bersedia untuk menerima pencabutan terhadap gelar kesarjanaan yang telah diberikan kepada Saya.

Demikan surat pernyataan ini Saya buat dengan sebenar-benarnya.

Surabaya, 29 Juli 2024

Muhammad Azhar Ali
NIM : 21410200020

ABSTRAK

Twitter, media sosial dengan 18,45 juta pengguna pada tahun 2022, sering disalahgunakan untuk menyebarkan ujaran kebencian, yang berdampak negatif seperti diskriminasi dan konflik sosial. Penelitian ini bertujuan untuk mengembangkan model klasifikasi menggunakan teknik *Natural Language Processing* (NLP) yang dapat mendeteksi sentimen negatif dalam *tweet*. Dengan demikian, diharapkan dapat tercipta lingkungan *online* yang lebih sehat dan aman bagi pengguna Twitter. Model yang dihasilkan mampu mengidentifikasi dan mengklasifikasikan *tweet* negatif dengan tingkat akurasi hingga 94%, sehingga memberikan kontribusi signifikan dalam pencegahan penyebaran konten negatif di media sosial dan menekan tingkat diskriminasi.

KATA PENGANTAR

Segala puji bagi Allah SWT yang telah memberikan Kesehatan dan kemudahan sehingga penulis dapat menyelesaikan Kerja Praktik dengan baik. Laporan ini dibuat berdasarkan hasil proyek yang dilakukan selama lima bulan di PT Hacktivate Teknologi Indonesia. Laporan Kerja Praktik ini membahas tentang Klasifikasi sentimen *Tweet* untuk mendeteksi konten pada platform Twitter menggunakan *Natural Languange Processing* (NLP)

Penulis sangat berterima kasih kepada instruktur Kakak Aldo Lionel dan Ardin Febrianda selaku mentor yang telah memberikan materi tentang *Artificial Intellegence* dan *Cyber Security,* dan penulis juga sangat berterima kasih kepada Bapak Pauladie Susanto, S.Kom., M.T. selaku dosen pembimbing Kerja Praktik dan ketua program studi S1 Teknik Komputer yang telah memberikan izin kepada penulis untuk melaksanakan Kerja Praktik.

Sekian dari penulis apabila ada kata yang kurang berkenan kami memohon maaf yang sebesar besarnya.

Surabaya, 23 Juli 2024

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Saat ini banyak sekali media sosial yang dapat diakses dengan mudah oleh banyak orang dari berbagai umur, agama, ras dan suku. Salah satu media sosial yang cukup banyak digunakan adakah Twitter yang pada tahun 2022 mencapai 18,45 Juta pengguna menyebabkan kekhawatiran akan penyebaran ujaran kebencian. Ujaran kebencian yang meluas di media sosial dapat memberikan dampak negatif seperti deskriminasi, konflik sosial, dan genosida. Pengguna yang semakin banyak menyebabkan klasifikan sentimen pada *Tweet* yang diunggah oleh pengguna menjadi masalah yang perlu dipecahkan. Oleh sebab itu untuk mengklasifikasi *Tweet* dapat digunakan klasifikasi menggunakan *Natural Language Processing* atau NLP.

Natural Language Processing atau NLP adalah cabang kecerdasan buatan yang memungkinkan komputer memahami dan menganalisis bahasa manusia menggunakan Bahasa alami Dengan meningkatnya volume data teks dari media sosial, NLP menjadi krusial untuk memahami dan menganalisis teks secara otomatis.

Dalam proyek ini, teknik NLP diterapkan untuk mengembangkan model klasifikasi yang dapat mengidentifikasi dan mengklasifikasikan sentimen dalam tweet, khususnya untuk mendeteksi konten negatif. Hal ini bertujuan untuk menciptakan lingkungan *online* yang lebih sehat dan aman bagi pengguna Twitter.

1.2 Rumusan Masalah

Rumusan masalah pada proyek ini adalah sebagai berikut:

1. Bagaimana cara mengembangkan model klasifikasi sentimen yang efektif menggunakan teknik *Natural Language Processing* (NLP) untuk mendeteksi sentimen negatif dalam tweet di Twitter?
2. Seberapa akurat model NLP yang dikembangkan dalam mengidentifikasi dan mengklasifikasikan sentimen negatif dalam tweet?

3. Apa saja tantangan yang dihadapi dalam proses pengembangan dan penerapan model klasifikasi sentimen pada tweet di Twitter?

1.3 Batasan Masalah

Batasan masalah pada proyek ini adalah sebagai berikut

1. Dataset yang digunakan dalam pelatihan dan pengujian model dibatasi pada tweet yang ditulis dalam bahasa tertentu dan dikumpulkan dalam periode waktu tertentu.
2. Jumlah kosakata yang terbatas sehingga mempengaruhi hasil klasifikasi.

1.4 Tujuan

Berdasarkan uraian dari latar belakang dan rumusan masalah, maka dapat disimpulkan tujuan dari kerja praktik ini adalah untuk mengembangkan model klasifikasi sentiment yang efektif menggunakan *Teknik Natural Language Processing* (NLP) untuk mendeteksi sentiment Tweet.

1.5 Manfaat

Adapun manfaat dari kerja praktik ini adalah sebagai berikut:

1. Membantu mengurangi tingkat penyebaran ujaran kebencian di platform Twitter
2. Mendeteksi Tweet negatif yang banyak tersebar di platform Twitter.

BAB II
GAMBARAN UMUM PERUSAHAAN

2.1 Latar Belakang Perusahaan

Perusahaan yang memberikan tugas untuk kerja praktik ini adalah Hacktiv8 yang merupakan sebuah Perusahaan *bootcamp* yang bergerak dalam pengembangan *Fullstack*, ilmu data, Pemasaran kinerja, dan Pengembangan Golang. Hacktiv8 memiliki beberapa kantor cabang diantaranya di Jakarta, Tangerang, dan Surabaya. Logo dari Hacktiv8 bisa dilihat dalam gambar 2.1.

Gambar 2 1 Logo Hacktiv8

2.2 Identitas Perusahaan

Nama Instansi : PT Hacktivate Teknologi Indonesia

Alamat : Alamat: Jl. Sultan Iskandar Muda No.7, Kebayoran Lama, Jakarta Selatan, DKI Jakarta 12240

No. Telepon : (021) 8067 5787

Website : https://www.hacktiv8.com/

Email : halo@hacktiv8.com

2.3 Visi Perusahaan

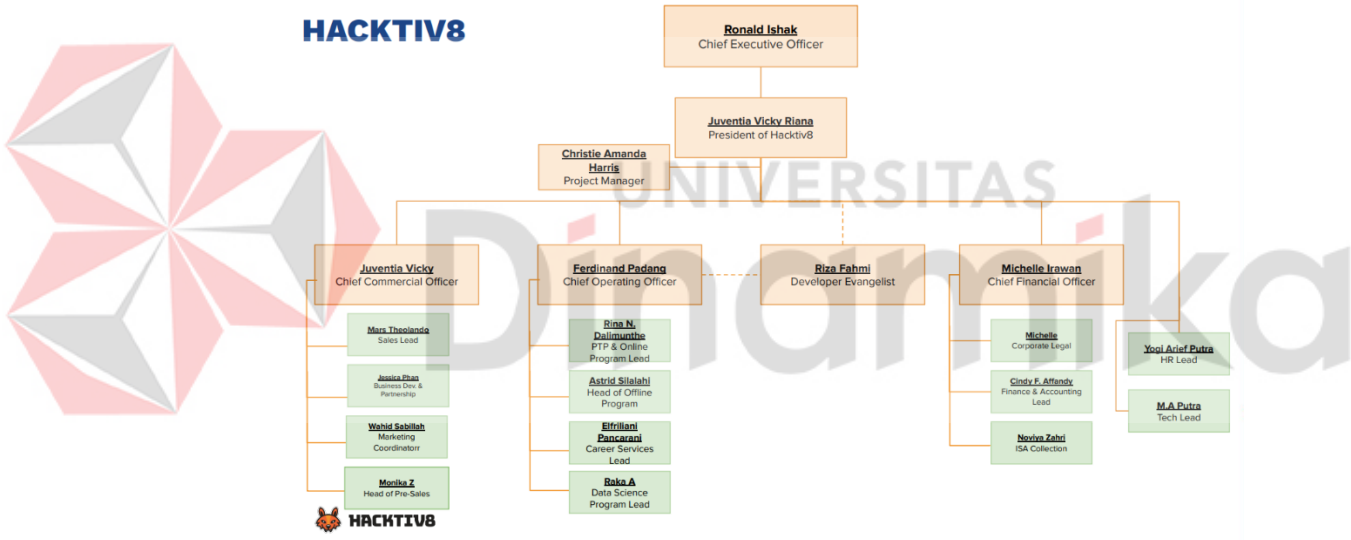
Hacktiv8 memiliki visi untuk meningkatkan taraf hidup dan karir orang banyak dalam bidang teknologi melalui kursus pemrograman intensif selama 12 minggu.

2.4 Misi Perusahaan

Misi Perusahaan Hacktiv8 adalah membawa perubahan pada industri digital di Indonesia melalui pendidikan teknologi untuk mendorong transformasi digital.

2.5 Struktur Organisasi Perusahaan

Struktur organisasi dari Hacktiv8 dapat dilihat dalam gambar 2.2. pada posisi teratas terdapat *Chief Executive Officer* (CEO) yang diisi oleh Ronald Ishak. Dibawah CEO terdapat President of Hacktiv8 yaitu Juventia Vicky Riana. Kemudian di beawahnya terdapat beberapa bagian diantaranya Project Manager, Chief Commercial Officer, Chief Operating Officer, Developer Evangelist, dan Chief Financial Officer.



Gambar 2 2 Struktur Perusahaan Hacktiv8

BAB III
LANDASAN TEORI

3.1 Twitter

Twitter atau yang sekarang disebut X adalah sebuah media sosial dan layanan jejaring sosial yang dijalankan oleh Perusahaan Twitter. Inc dan sekarang diambil alih oleh Perusahaan X Corp. di Twitter pengguna dapat mengunggah beberapa jenis konten diantaranya teks, gambar, dan video. Pengguna juga dapat memberikan *like* atau suka, memposting ulang, memberikan komentar, hingga mengirim pesan ke sesama pengguna. Twitter dapat diakses di berbagai jenis perangkat dan cara, diantaranya dapat diakses melalui peramban web dan melalui aplikasi di perangkat seluler semisal Android dan IOS.

Twitter yang merupakan salah satu dari banyak media sosial yang tersebar di dunia maya membuatnya menjadi salah satu tempat mengekspresikan diri yang banyak disalahgunakan sehingga menjadi pemantik konflik. Mudahnya pengguna untuk membuat akun memunculkan banyak akun-akun palsu yang kemudian menuliskan banyak sentimen negatif dan ujaran kebencian yang akhirnya merugikan banyak pihak karena banyak diantara pelaku yang "berlindung" dibalik kebebasan berekspresi, padahal penghinaan dalam wujud pencemaran nama baik adalah *character assassination* atau pembunuhan karakter. Karena itu penting untuk kita dapat menyaring konten-konten yang mengandung sentimen negatif sehingga dapat menciptakan lingkungan *online* yang sehat dan aman bagi semua pengguna Twitter.

3.2 Artificial Intelligence

Artificial Intelligence atau dalam Bahasa Indonesia disebut kecerdasan buatan adalah cabang dari ilmu komputer yang berfokus pada pengembangan sistem yang dapat melakukan pekerjaan yang sebelumnya hanya dapat dilakukan oleh kecerdasan manusia. Ini mencakup kemampuan seperti pembelajaran, pemahaman, penalaran, dan adaptasi. kecerdasan kuatan dapat ditemukan di banyak aplikasi, dari asisten virtual, hingga kendaraan otonom, di mana ia akan membantu dalam otomatisasi proses dan pengambilan keputusan. Teknik utama dalam AI termasuk machine learning, yang memungkinkan sistem belajar dari data, dan deep learning,

yang menggunakan jaringan saraf tiruan untuk memproses informasi secara lebih kompleks dan mendalam.

3.3 Natural Language Processing

Natural Language Processing atau biasa disingkat NLP adalah sebuah cabang dari kecerdasan buatan yang berfokus pada interaksi antara komputer dan Bahasa manusia. NLP memungkinkan komputer untuk memahami, menafsirkan, dan menghasilkan Bahasa yang digunakan manusia secara natural, baik dalam bentuk teks maupun ucapan. Teknik NLP mencakup banyak proses seperti analisis teks, dan pemahaman makna pada teks tersebut.

Salah satu penerapan NLP adalah penggunaan untuk mendeteksi sentimen, yang bertujuan untuk mengidentifikasi dan mengklasifikasi emosi dan opini yang dituliskan dalam sebuah teks. Penerapan untuk deteksi sentiment inilah yang mendasari penggunaan NLP untuk digunakan untuk mendeteksi sentimen pada Twitter. Dalam konteks pendeteksian sentimen, NLP dapat digunakan untuk mendeteksi kata-kata atau frasa yang mengindikasikan penghinaan, kritik, emosi, dan sejenisnya.

3.4 Convolutional Neural Network

Convolutional Neural Network atau disingkan CNN adalah salah satu jaringan syaraf buatan yang dirancang khusus untuk mengenali pola dan fitur dalam data. NLP awalnya dikembangkan untuk *Computer Vision* atau analisis data visual, namun, mereka juga terbukti efektif dalam NLP. Dalam konteks NLP, CNN digunakan untuk menangkap pola lokal dalam teks, seperti kata-kata dan frasa yang mengandung informasi penting. CNN sangat baik dalam memproses data dengan struktur spasial, yang dalam kasus teks bisa berupa urutan kata. Salah satu kelebihan CNN dalam NLP adalah kemampuannya untuk mengurangi jumlah parameter model melalui penggunaan pooling, yang membantu dalam menangkap informasi kontekstual tanpa terlalu bergantung pada panjang teks input. Hal ini membuat CNN sangat efisien dalam menangani teks yang panjang dan bervariasi.

3.5 Long Short-Term Memory

Long Short-Term Memory (LSTM) adalah jenis Recurrent Neural Network (RNN) yang dirancang untuk mengatasi masalah ketergantungan jangka panjang dalam data urutan, seperti teks. LSTM memiliki mekanisme khusus yang disebut "gates" yang memungkinkan mereka untuk mempertahankan atau melupakan informasi sesuai kebutuhan. Ini membuat LSTM sangat efektif dalam menangani teks yang memiliki hubungan kontekstual jangka panjang, seperti dalam kalimat atau paragraf yang kompleks. Dalam NLP, LSTM digunakan secara luas untuk tugas-tugas seperti penerjemahan mesin, pemodelan bahasa, dan analisis sentimen, di mana memahami urutan kata dalam konteks keseluruhan sangat penting.

3.6 Python

Python merupakan bahasa pemrograman tinggi yang bisa melakukan eksekusi sejumlah instruksi multi guna secara langsung dengan metode *Object Oriented Programming* dan juga menggunakan semantik dinamis untuk memberikan tingkat keterbacaan *syntax*. Sebagai bahasa pemrograman tinggi, python dapat dipelajari dengan mudah karena telah dilengkapi dengan manajemen memori otomatis.

Python sering digunakan dalam pengembangan situs web dan perangkat lunak, analisis data, visualisasi data, serta otomatisasi tugas. Karena kemudahannya dalam pembelajaran, bahasa pemrograman ini banyak dipilih oleh orang-orang non-programmer seperti ilmuwan dan akuntan untuk menyelesaikan tugas sehari-hari mereka, seperti mengelola keuangan.

3.7 Google Colab

Google Colab adalah platform berbasis web yang biasa digunakan untuk menulis, menjalankan, dan berbagi kode Python. Platform ini dirancang untuk digunakan oleh analis, pengembang, peneliti, penyidik yang bekerja di bidang *data science* dan *machine learning* dengan menyediakan lingkungan komputasi yang fleksibel dan mudah diakses tanpa biaya. Google Colab juga memiliki kemampuan

untuk menjalankan *Jupyter Notebook* secara langsung dari peramban web tanpa konfigurasi apapun.

BAB IV
DESKRIPSI PEKERJAAN

4.1 Kerja Praktik

Kerja praktik atau biasa disebut Magang, adalah suatu kegiatan yang dirancang untuk memberikan pengalaman langsung kepada mahasiswa dalam lingkungan kerja professional. Tujuan dari kegiatan Kerja Praktik adalah untuk menghubungkan pembelajaran secara teoritis di bangku perkuliahan dengan kerja nyata di luar perkuliahan. Dalam kerja praktik mahasiswa bisa mendapatkan pemahaman yang lebih jauh tentang bidang yang yang digeluti juga mengasah keterampilan yang mungkin diperlukan di dunia kerja.

Kerja praktik pada umumnya merupakan penugasan mahasiswa pada proyek atau tugas tertentu yang relevan dengan bidang studi mahasiswa, dimana mereka memiliki kesempatan untuk menerapkan pengetahun akademis yang sudah di dapat di bangku perkuliahan di situasi nyata. Kegiatan ini juga bertujuan untuk meningkatkan kesiapan kerja mahasiswa setelah lulus.

Selama kerja praktik, mahasiswa biasanya bekerja di bawah mentor atau supervisor berpengalaman yang ada pada Perusahaan atau instansi tempat mereka melakukan kerja praktik. Selain itu kerja praktik juga sering kali melibatkan evaluasi berkala untuk mengukur kemampuan dan pencapaian mahasiswa, seta memberikan umpan balik yang konstruktif untuk pengembangan lebih lanjut.

4.2 Metode Pembelajaran selama Kerja Praktik

Dalam pelaksanaan kerja praktik, beberapa metode pembelajaran diterapkan untuk memastikan pemahaman yang medalam terhadap materi dan penerapan yang efektif dari pengetahuan yang diperoleh. Metode-metode ini meliputi sebagai berikut:

1. **Akses Materi**

   Dalam pembelajaran yang dilakukan selama kerja praktik berlangsung, website kode.id menjadi sumber utama untuk mengakses materi. Dalam web tersebut tersedia semua materi yang akan dipelajari dan diurutkan sesuai urutan pembelajaran.

2. **Pembelajaran Online**

Pembelajaran online pada kerja praktik dilakukan menggunakan platform Google Meet, sehingga memungkinkan interaksi langsung antara mahasiswa dan instruktur. Di sesi ini mahasiswa dapat menanyakan berbagai hal terkait materi dan non-materi. Pembelajaran online memastikan semua mahasiswa dapat mengakses materi dan memahaminya.

3. **Kelas Besar**

Dalam pembelajaran online terdapat beberapa jenis kelas, salah satunya adalah kelas besar. Kelas ini terdiri dari dua kelas yang masing masing memiliki instruktur sendiri. Kelas ini memungkinkan semua peserta untuk belajar Bersama. Sesi ini dibuat untuk memberikan pemahaman umum, berbagi pengetahuan, dan melakukan diskusi kelompok yang melibatkan semua peserta dalam kelas.

4. **Kelas Mentoring**

Selain kelas besar, peserta juga dibagi menjadi beberapa kelas kecil yang disebut kelas mentoring. Kelas ini dibuat untuk memberikan pemahaman yang lebih mendalam terhadap materi dari mentor yang berbeda. Kelas ini dikelola oleh seorang mentor yang bertugas memberikan dukungan, umpan balik, dan arahan khusus yang sesuai ke masing-masing peserta sehingga proses pembelajaran bisa dilakukan dengan lebih intensif.

5. **Pembelajaran Mandiri**

Dalam pelaksaan kerja praktik ini, pembelajaran mandiri juga merupakan salah satu metode yang digunakan selain kelas besar dan kelas mentoring. Pembelajaran ini dilakukan oleh peserta secara individu untuk mengeksplorasi dan memahami materi secara mendalam. Mahasiswa pada sesi ini diharapkan dapat membaca berbagai literatur terkait, melakukan penelitian, dan mengasah keterampilan secara mandiri.

6. **Pengumpulan Tugas**

Dalam pengumpulan tugas, Google Classroom merupakan platform utama yang digunakan. Peserta dapat mengunggah tugas dan melihat tenggat waktu yang diberikan sehingga bisa mengerjakan tugas yang diberikan melalui platform ini sehingga peserta bisa menyelesaikan tugas

dengan tepat waktu. Platform ini juga memudahkan pengelolaan tugas dan pemberian nilai oleh mentor dan instruktur.

4.3 Deskripsi Proyek

a. Tujuan Proyek

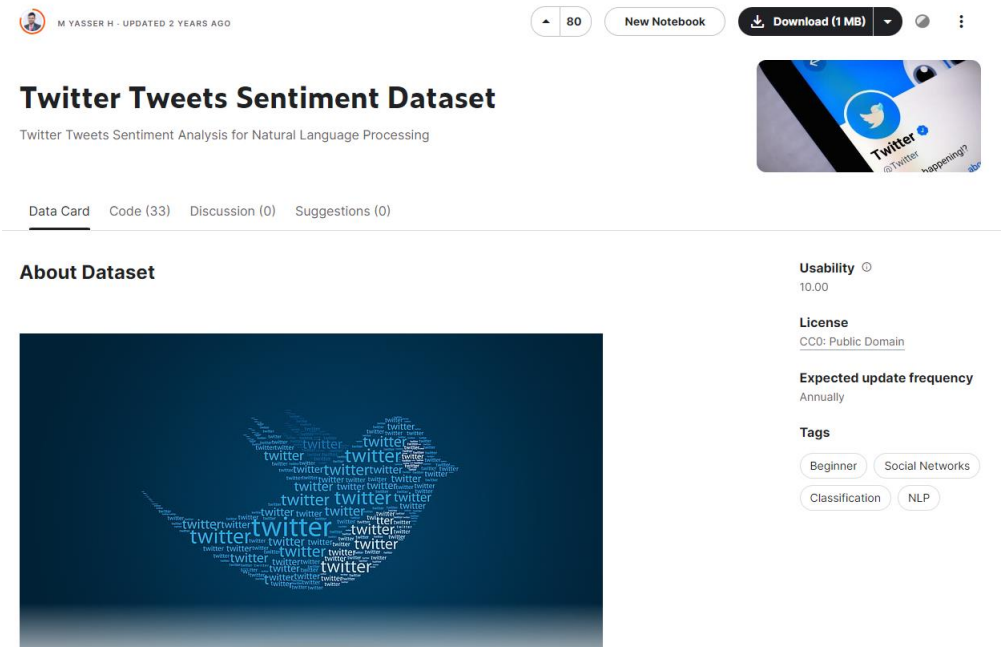
Proyek ini bertujuan untuk melakukan klasifikasi sentimen pada tweet menggunakan *natural language processing* atau NLP. Proyek ini berfokus pada penerapan teknik NLP untuk model klasifikasi yang mampu mengidentifikasi mengembangkandan mengklasifikasikan tweet dengan sentimen negatif. Upaya ini diharapkan dapat berkontribusi dalam menciptakan lingkungan online yang lebih sehat dan aman bagi pengguna Twitter.

b. Ruang Lingkup Proyek

Ruang lingkup dari proyek ini terbatas pada klasifikasi data sentimen Twitter yang diperoleh dari situs Kaggle. Klasifikasi ini hanya mencakup data umum yang tidak ada jenis konten tertentu yang dipakai sebagai data sentimen.

4.4 Data yang digunakan

Data yang digunakan dalam proyek kali ini diperoleh dari situs Kaggle, yang merupakan platfrom yang biasa digunakan untuk berbagi dataset publik. Dataset ini disediakan dalam bentuk file CSV dan berisi kumpulan Tweet yang telah dikategorikan berdasarkan sentimen mereka. Penggunaan dataset dari Kaggle ini mempermudah akses data dikarenakan data yang telah tersusun secara terstruktur. Hal ini memungkinkan peneliti berfokus pada klasifikasi sentimen tanpa perlu mengumpulkan data secara manual dari platform Twitter.



Gambar 4 1 Platform Kaggle sebagai sumber data

Dataset yang digunakan dalam analisis ini berisi 27421 Tweet yang telah diberi label sesuai dengan sentimen yang relevan. Data ini terdiri dari 12199 sentimen positif, 9940 sentimen netral, dan 5282 sentimen negatif. Data yang diambil hanya sebatas teks sentimen, sehingga data lainnya seperti waktu, tanggal, dan informasi pengguna tidak ada dalam data ini. Dengan begitu klasifikasi ini difokuskan pada data sentimen.

4.5 Data Loading and Cleaning

Tujuan: Memuat data kedalam Program dan membersihkan teks dengan mengubahnya menjadi huruf kecil, menghapus karakter khusus dan angka, serta menghilangkan stopwords. Langkah-langkah ini mempersiapkan teks untuk analisis dengan menghilangkan noise dan fokus pada informasi yang relevan.

Proses: Menampilkan Data Head, Data Info, Distribusi Label, dan Deskripsi Data.

```
================================================================
============================== Data Head ==============================
       textID                                               text  \
0  cb774db0d1                I`d have responded, if I were going
1  549e992a42      Sooo SAD I will miss you here in San Diego!!!
2  088c60f138                          my boss is bullying me...
3  9642c003ef                     what interview! leave me alone
4  358bd9e861   Sons of ****, why couldn`t they put them on t...

                        selected_text sentiment
0  I`d have responded, if I were going   neutral
1                            Sooo SAD  negative
2                          bullying me  negative
3                       leave me alone  negative
4                        Sons of ****,  negative
================================================================
```

Gambar 4 2 Data Head

```
================================================================
============================== Data Info ==============================
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27481 entries, 0 to 27480
Data columns (total 4 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   textID         27481 non-null  object
 1   text           27480 non-null  object
 2   selected_text  27480 non-null  object
 3   sentiment      27481 non-null  object
dtypes: object(4)
memory usage: 858.9+ KB
None
================================================================
```

Gambar 4 3 Data Info

```
================================================================
======================== Distribusi Label ========================
text
 I`d have responded, if I were going                                                                                  1
Watchin Hollyoaks...poor Justin!                                                                                      1
WAH. i`m gonna miss bowie people  ESP YOU ALYANNA BONDOC <3 and Cesar. D:< I dunno if I can survive without my SOSA!  1
 aww  that`s horrible! xD                                                                                             1
My new camera... http://tinyurl.com/l8pde3 ... RIP my hot pink Polaroid i733                                          1
                                                                                                                     ..
 oooh lush. i cant sunbathe i burn way to easily even with sun cream   im great thanks lovely sunny day no?           1
hmz... second most popular page on this governmental site is the 404 page... fail                                     1
 I like it too  I hadn`t seen the clip before, though; pretty cool!                                                   1
I never order chips any more due to how unhealthy they are, but getting a burrito from Chipotle or Qdoba doesn`t feel right without em  1
   All this flirting going on - The ATG smiles. Yay.  ((hugs))                                                       1
Name: count, Length: 27480, dtype: int64
================================================================
```

Gambar 4 4 Distribusi Label

```
================================================================
============================== Data Description ==============================
            textID                                  text selected_text  \
count        27481                                 27480         27480
unique       27481                                 27480         22463
top     cb774db0d1   I`d have responded, if I were going          good
freq             1                                     1           199
```

Gambar 4 5 Deskripsi Data

4.6 Exploratory Data Analysis

Tujuan: memberikan pemahaman awal yang mendalam tentang dataset yang akan dianalisis. Ini adalah langkah awal yang penting dalam proses analisis data, di mana analis memeriksa struktur dasar data, termasuk jenis variabel, jumlah observasi, dan kehadiran data yang hilang atau outliers.

Proses: Dalam proses EDA ini, kami memulai dengan menghitung rata-rata jumlah kata dan ukuran kosakata untuk memahami karakteristik dasar teks dalam dataset. Selanjutnya, distribusi frekuensi kata dianalisis untuk mengidentifikasi kata-kata yang paling umum dan pola penggunaan kata. Visualisasi wordcloud digunakan untuk memberikan representasi visual dari kata-kata yang paling sering muncul. Kami juga melakukan analisis sentimen untuk menentukan perasaan atau emosi yang diekspresikan dalam teks, di mana kami menerapkan fungsi pada kolom 'polarity' untuk membuat kolom baru 'sentiment_category' yang mengkategorikan sentimen sebagai positif, negatif, atau netral. Berdasarkan hasil ini, kami membuat DataFrame baru yang hanya berisi kolom yang diinginkan, yang kemudian ditampilkan untuk review lebih lanjut. Akhirnya, kami menggambarkan distribusi kategori sentimen dalam pie chart untuk memberikan gambaran visual mengenai proporsi masing-masing kategori sentimen dalam dataset.

```
Average word count after cleaning: 7.237299879654279
```

Gambar 4 6 Rata-rata jumlah kata

```
Vocabulary size: 27770
```

Gambar 4 7 Ukuran kosakata

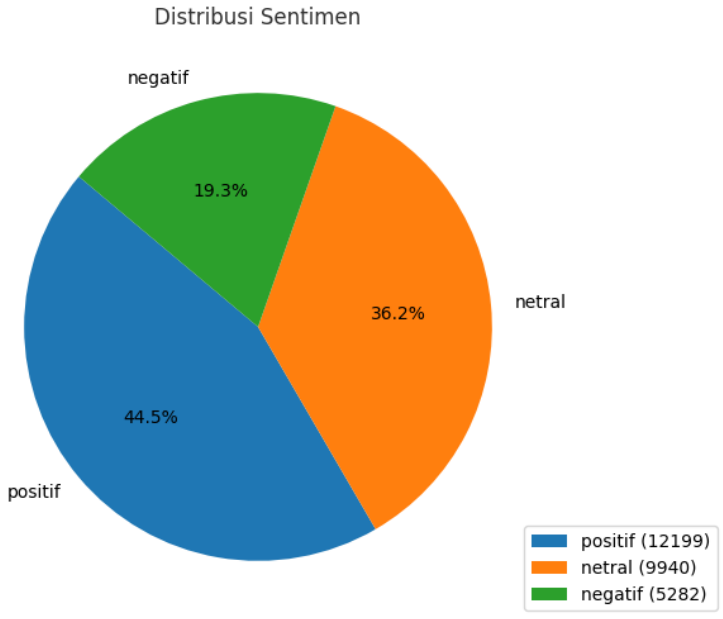
```
Most common words: [('im', 3055), ('day', 2086), ('good', 1549),
```

Gambar 4 8 Distribusi Frekuensi Data

Gambar 4 9 Word Cloud

	cleaned_text	polarity	sentiment_category
0	id responded going	0.000000	netral
1	sooo sad miss san diego	-0.500000	negatif
2	boss bullying	0.000000	netral
3	interview leave alone	0.000000	netral
4	sons couldnt put releases already bought	0.000000	netral
5	httpwwwdothebouncycomsmf shameless plugging be...	1.000000	positif
6	feedings baby fun smiles coos	0.300000	positif
7	soooo high	0.160000	positif
9	journey wow u became cooler hehe possible	0.050000	positif
10	much love hopeful reckon chances minimal p im ...	0.200000	positif
11	really really like song love story taylor swift	0.350000	positif
12	sharpie running dangerously low ink	0.000000	netral
13	want go music tonight lost voice	0.000000	netral
14	test test lg env	0.000000	netral
15	uh oh sunburned	0.000000	netral
16	sok trying plot alternatives speak sigh	0.000000	netral
17	ive sick past days thus hair looks wierd didnt...	-0.482143	negatif
18	back home gonna miss every one	0.000000	netral
19	hes	0.000000	netral
20	oh marly im sorry hope find soon	-0.500000	negatif

Gambar 4 10 Dataframe baru



Gambar 4 11 Pie Chart Distribusi Sentimen

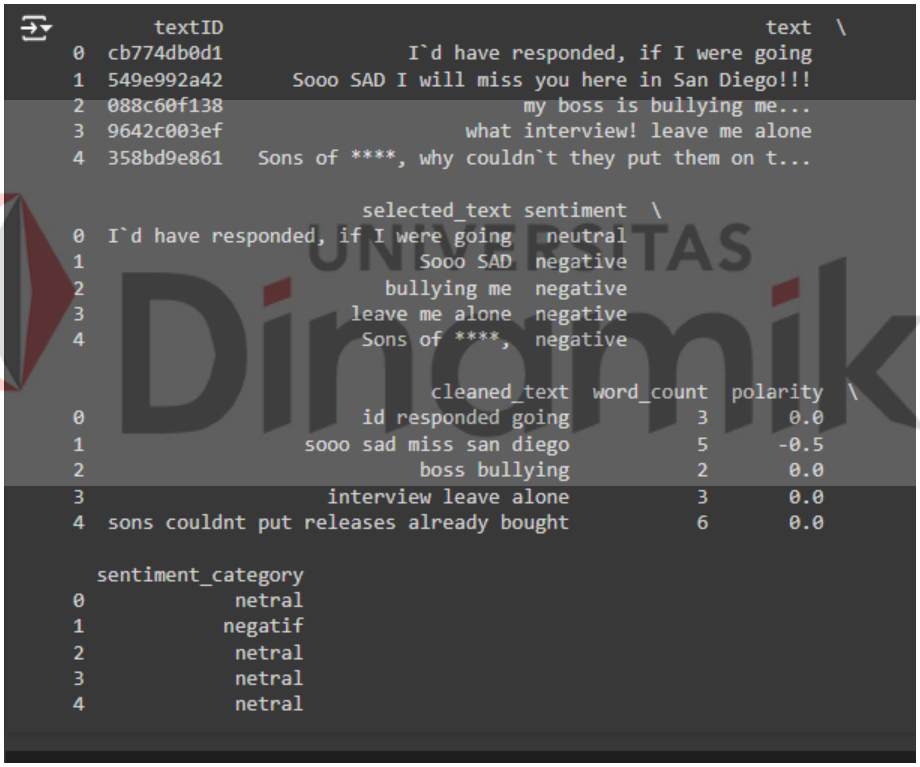
4.7 Feature Engineering

Dalam Feature Engineering ada beberapa proses yang dilakukan:

a. One-Hot Encoding pada Kolom 'Sentiment_Category'

Tujuan: Mengonversi data kategori menjadi format numerik biner yang dapat digunakan oleh algoritma machine learning.

Proses: Dalam one-hot encoding, setiap kategori dalam kolom 'sentiment_category' (misalnya, positif, negatif, netral) diubah menjadi kolom biner terpisah. Jika suatu teks memiliki kategori tertentu, kolom biner yang sesuai diberi nilai 1, sementara kolom lainnya diberi nilai 0. Ini menghasilkan representasi numerik dari kategori yang dapat diproses oleh model.



	textID	text	selected_text	sentiment	cleaned_text	word_count	polarity	sentiment_category
0	cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral	id responded going	3	0.0	netral
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative	sooo sad miss san diego	5	-0.5	negatif
2	088c60f138	my boss is bullying me...	bullying me	negative	boss bullying	2	0.0	netral
3	9642c003ef	what interview! leave me alone	leave me alone	negative	interview leave alone	3	0.0	netral
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	Sons of ****,	negative	sons couldnt put releases already bought	6	0.0	netral

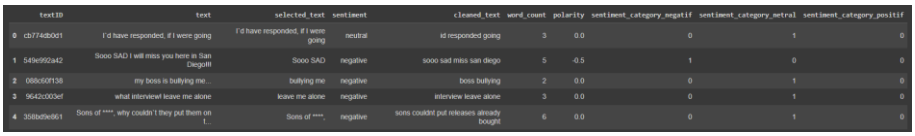
Gambar 4 12 Proses one-hot encoding

b. Mengubah Tipe Data One-Hot Encoding menjadi Integer

Tujuan: Meningkatkan efisiensi penyimpanan dan komputasi.

Proses: Setelah proses one-hot encoding, hasilnya sering dalam format tipe data boolean atau float. Mengubah tipe data ini menjadi integer dapat

mengurangi penggunaan memori dan meningkatkan kecepatan komputasi selama pelatihan model berlangsung.



	textID	text	selected_text	sentiment	cleaned_text	word_count	polarity	sentiment_category_negatif	sentiment_category_netral	sentiment_category_positif
0	cb774db0d1	I'd have responded, if I were going	I'd have responded, if I were going	neutral	id responded going	3	0.0	0	1	0
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative	sooo sad miss san diego	5	-0.5	1	0	0
2	088c60f138	my boss is bullying me...	bullying me	negative	boss bullying	2	0.0	0	1	0
3	9642c003ef	what interviewl leave me alone	leave me alone	negative	interview leave alone	3	0.0	0	1	0
4	358bd9e861	Sons of ****, why couldn't they put them on t...	Sons of ****,	negative	sons couldnt put releases already bought	6	0.0	0	1	0

Gambar 4 13 Hasil pengubahan tipe data menjadi integer

c. Menentukan Panjang Teks yang Paling Sering Muncul (Modus)

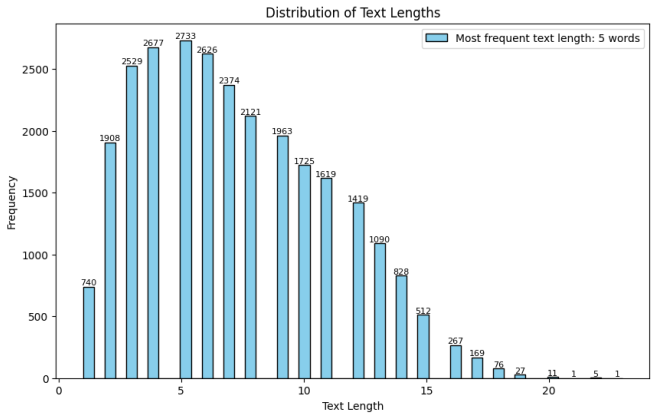
Tujuan: Memahami distribusi panjang teks dalam dataset untuk membantu dalam langkah-langkah preprocessing lebih lanjut.

Proses: Modus panjang teks adalah panjang teks yang paling sering muncul dalam dataset. Menentukan nilai ini membantu dalam memahami karakteristik dataset, seperti apakah teks-teks tersebut cenderung pendek atau panjang, yang dapat memengaruhi bagaimana kita mengelola teks untuk analisis lebih lanjut.

d. Analisis Distribusi Panjang Teks

Tujuan: Mengidentifikasi variasi panjang teks dan menentukan parameter yang optimal untuk padding.

Proses: Analisis distribusi panjang teks dilakukan dengan menghitung distribusi panjang teks dalam dataset. Ini membantu dalam memahami seberapa banyak teks yang panjang atau pendek, yang berguna untuk menentukan nilai maxlen yang tepat.

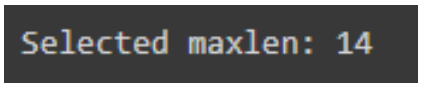


Gambar 4 14 Grafik Distribusi Panjang teks

e. Menentukan Maxlen

Tujuan: Menetapkan panjang maksimum teks setelah padding untuk memastikan konsistensi dalam input model.

Proses: Berdasarkan analisis distribusi panjang teks, maxlen dipilih sebagai batas panjang teks yang akan digunakan dalam padding. Nilai ini sering kali dipilih untuk mencakup sebagian besar teks dalam dataset tanpa membuat banyak padding tambahan, sehingga mengoptimalkan penggunaan memori dan komputasi.



Gambar 4 15 Menentukan Maxlen

f. Tokenisasi Teks

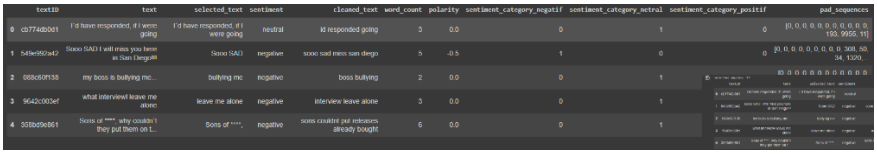
Tujuan: Mengonversi teks menjadi format numerik dengan setiap kata direpresentasikan sebagai angka unik.

Proses: Dalam tokenisasi, setiap kata atau token dalam teks diberi nomor unik. Proses ini menghasilkan representasi numerik dari teks yang dapat diolah oleh model *machine learning*. Tokenisasi juga memungkinkan penghapusan kata-kata yang jarang muncul atau tidak relevan, yang dapat mengurangi kompleksitas data.

g. Padding Sequences

Tujuan: Menyamakan panjang semua teks agar sesuai dengan maxlen yang telah ditentukan.

Proses: Padding sequences melibatkan penambahan nilai nol (atau nilai lainnya) pada akhir atau awal teks yang lebih pendek dari maxlen. Hal ini memastikan bahwa semua teks dalam dataset memiliki panjang yang sama, yang diperlukan untuk memproses data dalam batch oleh *model neural network*, karena model memerlukan input dengan dimensi yang konsisten.



	textID	text	selected_text	sentiment	cleaned_text	word_count	polarity	sentiment_category_negatif	sentiment_category_netral	sentiment_category_positif	pad_sequences
0	cb774db0d1	I'd have responded, if I were going	I'd have responded, if I were going	neutral	id responded going	3	0.0	0	1	0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 193, 9955, 11]
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative	sooo sad miss san diego	5	-0.5	1	0	0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 308, 59, 34, 1329]
2	088c60f138	my boss is bullying me...	bullying me	negative	boss bullying	2	0.0	0	1	0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
3	9642c003ef	what interview! leave me alone	leave me alone	negative	interview leave alone	3	0.0	0	1	0	
4	358bd9e861	Sons of ****, why couldn't they put them on t...	Sons of ****,	negative	sons couldnt put releases already bought	6	0.0	0	1	0	

Gambar 4 16 Hasil dari proses Padding Sequence

4.8 Model Architecture Definition

Dalam Model Architecture Definition ada beberapa proses yang dilakukan antara lain.

a. Mengubah Kolom 'pad_sequences' Menjadi Numpy Array

Tujuan: Mengonversi data yang telah diproses menjadi format yang efisien untuk digunakan dalam model machine learning.

Proses: Pada langkah ini, kolom 'pad_sequences' yang berisi data teks dalam bentuk urutan angka dipadukan menjadi numpy array. Numpy array adalah struktur data yang efisien dan cepat untuk komputasi, yang sangat berguna dalam deep learning. Dengan mengonversi ke numpy array, kita mempersiapkan data untuk diinput ke dalam model machine learning dengan lebih mudah dan efisien.

b. Membagi Data Menjadi Data Pelatihan dan Data Validasi

Tujuan: Membagi dataset menjadi dua bagian. Ini penting untuk memastikan model dapat melakukan generalisasi terhadap data baru.

Proses: membagi dataset x dan y menjadi data pelatihan dan validasi yaitu 20% data digunakan untuk validasi dan 80% sisanya untuk pelatihan. Juga Menjamin hasil pembagian yang konsisten dan dapat direproduksi.

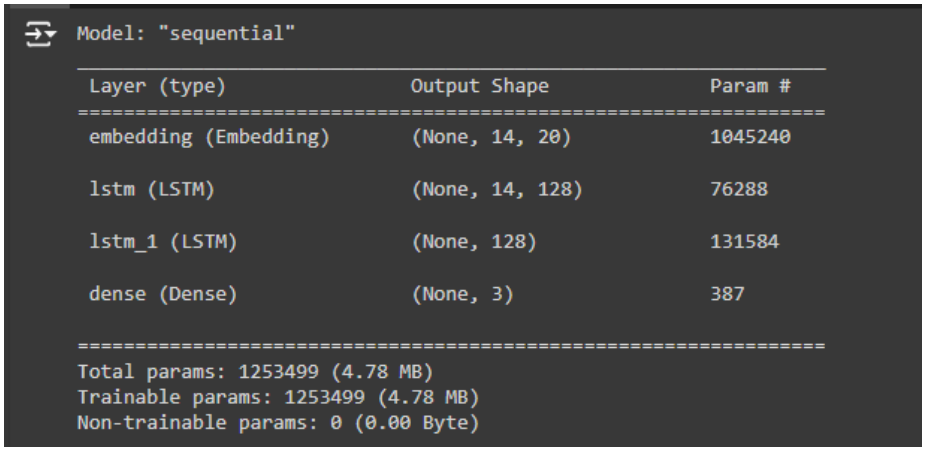
	textID	text	selected_text	sentiment
0	cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative
2	088c60f138	my boss is bullying me...	bullying me	negative
3	9642c003ef	what interview! leave me alone	leave me alone	negative
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	Sons of ****,	negative
5	28b57f3990	http://www.dotheboucy.com/smf - some shameles...	http://www.dotheboucy.com/smf - some shameles...	neutral
6	6e0c6d75b1	2am feedings for the baby are fun when he is a...	fun	positive
7	50e14c0bb8	Soooo high	Soooo high	neutral
9	fc2cbefa9d	Journey!? Wow... u just became cooler. hehe....	Wow... u just became cooler.	positive
10	2339a9b08b	as much as i love to be hopeful, i reckon the...	as much as i love to be hopeful, i reckon the ...	neutral
11	16fab9f95b	I really really like the song Love Story by Ta...	like	positive
12	74a76f6e0a	My Sharpie is running DANGERously low on ink	DANGERously	negative
13	04dd1d2e34	i want to go to music tonight but i lost my vo...	lost	negative
14	bbe3cbf620	test test from the LG enV2		
15	8a939bfb59	Uh oh, I am sunburned		
16	3440297f8b	S`ok, trying to plot alternatives as we speak...		
17	919fa93391	i`ve been sick for the past few days and thus...		
18	af3fed7fc3	is back home now gonna miss every one		
19	40e7becabf	Hes just not that into you		
20	04d17ef61e	oh Marly, I`m so sorry!! I hope you find her...		

Gambar 4 17 Tampilan Hasil pembagian data

c. Membangun Model Arsitektur LSTM

Tujuan: Membangun arsitektur neural network berbasis LSTM untuk memproses urutan teks dan melakukan klasifikasi sentimen. Model ini dirancang untuk menangkap hubungan temporal dalam data teks dan memberikan output prediksi dalam bentuk kategori sentimen.

Proses: Menginisialisasi model sebagai urutan lapisan yang bertingkat. Kemudian melakukan *Layer Embedding*. Lapisan LSTM pertama mengembalikan urutan lengkap output, bukan hanya output akhir. Lapisan LSTM kedua hanya meneruskan output terakhir ke lapisan berikutnya. Kemudian Lapisan output dengan 3 unit, sesuai dengan jumlah kategori sentimen, menggunakan fungsi aktivasi 'softmax' untuk memberikan probabilitas kelas.

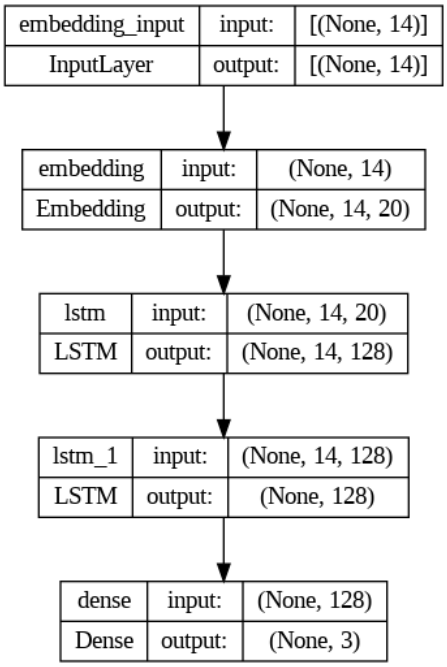


Gambar 4 18 Ringkasan Model

d. Mengompilasi Model

Tujuan: Mengompilasi model untuk menentukan fungsi loss, optimizer, dan metrik yang akan digunakan selama pelatihan. Ini adalah langkah penting sebelum melatih model, karena menentukan bagaimana model akan belajar dan dievaluasi.

Proses: Dalam langkah ini, digunakan optimizer 'adam', yang mengoptimalkan kecepatan dan efektivitas pelatihan dengan menggabungkan keunggulan dari *optimizers* 'AdaGrad' dan 'RMSprop'. Fungsi loss yang dipilih adalah 'categorical_crossentropy', yang cocok untuk masalah klasifikasi multikelas di mana label target dalam format *one-hot encoded*. Fungsi ini mengukur perbedaan antara distribusi probabilitas prediksi model dan label sebenarnya, dan tujuannya adalah meminimalkan nilai loss ini selama proses pelatihan. Selain itu, metrik 'accuracy' digunakan untuk mengevaluasi kinerja model, memberikan informasi tentang persentase prediksi yang benar dari model, sehingga memberikan gambaran langsung mengenai seberapa baik model mengenali kelas yang benar dari data input.



Gambar 4 19 Diagram Arsitektur Model

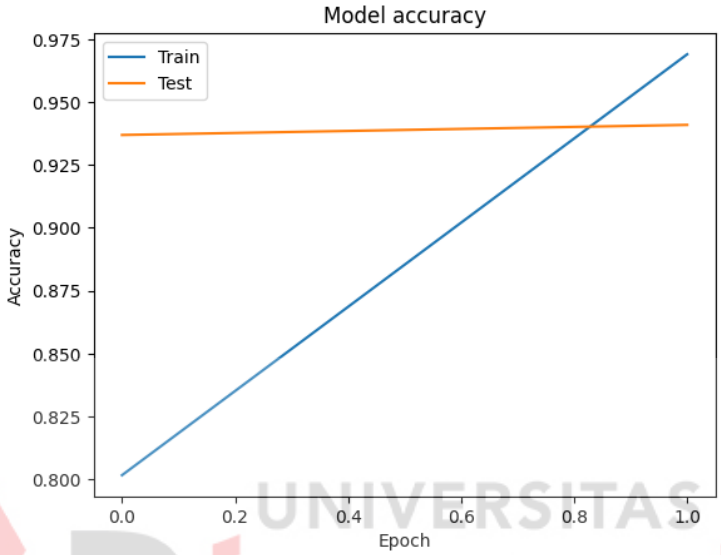
4.9 Model Evaluation

Tujuan: Evaluasi model dilakukan untuk mengukur kinerja model setelah pelatihan, dengan fokus pada akurasi dan kemampuan klasifikasi. Proses ini mencakup analisis visual dan metrik kuantitatif untuk memahami sejauh mana model telah belajar dari data pelatihan dan seberapa baik ia dapat memprediksi data yang belum pernah dilihat.

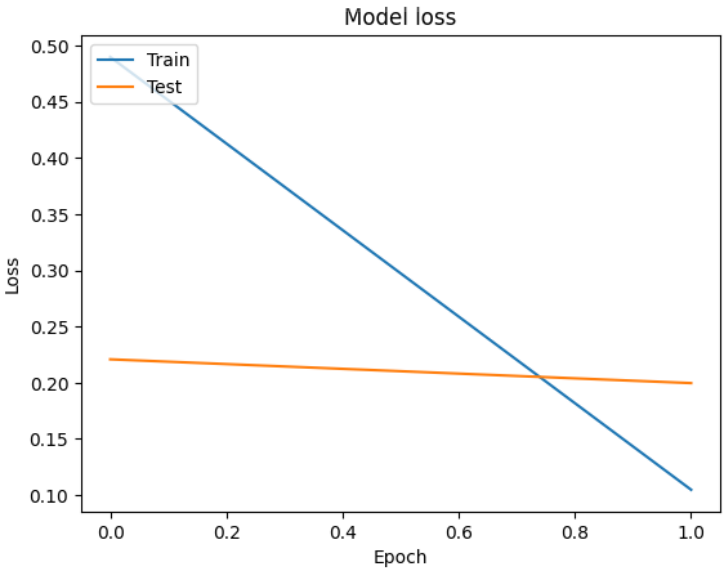
Proses:

1. Visualisasi Akurasi dan Loss:
    - Grafik akurasi dan loss untuk data pelatihan dan validasi diplot untuk melihat tren selama *epoch*, membantu mengidentifikasi *overfitting* atau *underfitting*.
2. Evaluasi Klasifikasi:

- Model memprediksi label pada data validasi. Prediksi ini dikonversi dari probabilitas ke label kelas sebenarnya.
- Metrik klasifikasi seperti *precision*, *recall*, dan *f1-score* dihitung untuk masing-masing kelas menggunakan "classification_report".
- Akurasi keseluruhan model dihitung menggunakan "accuracy_score", memberikan persentase prediksi yang benar.



Gambar 4 20 Model Accuracy



Gambar 4 21 Model Loss

```
Classification Report:
              precision    recall  f1-score   support

    negative       0.93      0.86      0.89      1083
     neutral       0.97      0.95      0.96      1986
    positive       0.93      0.97      0.95      2416

    accuracy                           0.94      5485
   macro avg       0.94      0.93      0.93      5485
weighted avg       0.94      0.94      0.94      5485


Accuracy Score: 0.940929808568824
```

Gambar 4 22 Classification Report

BAB V

PENUTUP

5.1 Kesimpulan

Dalam Penelitian kali ini, kami menerapkan proses pengolahan data dan machine learning untuk mengklasifikasi sentimen teks. Langkah-langkah utama meliputi *exploratory data analysis* (EDA), *feature engineering*, dan Pembangunan serta pelatihan model LSTM. Selanjutnya proses evaluasi menunjukkan bagaimana performa model dalam mengklasifikasi sentimen pada data validaasi.

Kesimpulannya, proyek ini berhasil mengembangkan dan mengevaluasi model yang mampu mengklasifikasikan sentimen teks dengan akurasi hingga 94%. Model ini dapat lebih ditingkatkan dengan teknik tuning *hyperparameter*, penggunaan dataset yang lebih besar, atau arsitektur model yang lebih kompleks. Evaluasi hasil menunjukkan bahwa model mampu mengenali pola dalam data dan memberikan prediksi yang akurat, meskipun masih terdapat ruang untuk perbaikan lebih lanjut.

5.2 Saran

1. Peningkatan Kualitas Data dengan mengumpulkan data yang lebih banyak dan beragam untuk melatih model, hal ini dapat membantu dalam menangani variasi dalam data teks
2. *Tuning Hyperparamenter* dengan melakukan eksperimen dengan berbagai *hyperparameter*, seperti ukuran *embedding*, jumlah unit dalam layer LSTM, dan jenis *optimizer*.
3. Ekspansi Model Arsitektur dengan Pertimbangkan untuk menggunakan model yang lebih kompleks seperti Bidirectional LSTM, GRU, atau bahkan model Transformer yang telah menunjukkan performa superior dalam banyak tugas NLP.
4. Setelah model mencapai kinerja yang baik, pertimbangkan untuk menerapkannya dalam sistem produksi yang nyata.

DAFTAR PUSTAKA

Edwin, P., Eko, S., & Budhi, W. (2016). KEBIJAKAN HUKUM PIDANA DALAM UPAYA PENEGAKAN TINDAK PIDANA PENCEMARAN NAMA BAIK MELALUI TWITTER. *Diponegoro Law Journal*.

Fitri, L. (2022). Analisis Sentimen Twitter Terhadap Kebijakan Ppkm Di Tengah Pandemi Covid-19 Menggunakan Mode LSTM. *Journal of Information System, Applied, Management, Accounting and Research*.

Primandani, A., Iphang, P., & Muhammad, H. A. (2023). Klasifikasi Sentimen Publik Terhadap Jenis Vaksin Covid-19 yang Tersertifikasi WHO Berbasis NLP dan KNN. *Media Informtika Budidarma*.