

LAPORAN AKHIR
STUDI INDEPENDEN BERSERTIFIKAT
AI Mastery Program
Di Orbit Future Academy

Diajukan untuk memenuhi persyaratan kelulusan
Program MSIB MBKM

oleh :

Alif Yanuar Aditya Subagyo / 1202190187



PROGRAM STUDI S1 SISTEM INFORMASI
FAKULTAS REKAYASA INDUSTRI
UNIVERSITAS TELKOM
2022

Lembar Pengesahan S1 Sistem Informasi Universitas Telkom

***Rancang Bangun Sistem Deteksi Berita Hoax Menggunakan
Pendekatan Bidirectional Encoder Representations from Transformer
(BERT)***

Di Orbit Future Academy

oleh :

Alif Yanuar Aditya Subagyo / 1202190187

disetujui dan disahkan sebagai

Laporan Studi Independen Bersertifikat Kampus Merdeka

Bandung, 14 Juni 2022

Pembimbing Studi Independen Program Studi S1 Sistem Informasi Universitas
Telkom

A handwritten signature in blue ink, appearing to read 'Dita Pramesti', is written over a large, stylized blue triangle.

Dita Pramesti, S. Si., M. Si.

NIP: 20920012

Lembar Pengesahan

***Rancang Bangun Sistem Deteksi Berita Hoax Menggunakan
Pendekatan Bidirectional Encoder Representations from Transformer
(BERT)***

Di Orbit Future Academy

oleh :


Alif Yanuar Aditya Subagyo / 1202190187

disetujui dan disahkan sebagai

Laporan Studi Independen Bersertifikat Kampus Merdeka

Bandung, 14 Juni 2022

AI Coach



Kuncahyo Setyo Nugroho, S.Kom

NIP: 2201045

Abstraksi

AI Mastery Program yang dilaksanakan oleh Orbit Future Academy memiliki tujuan untuk meningkatkan kualitas hidup melalui inovasi, pendidikan, dan pelatihan keterampilan. Bidang pada perusahaan tersebut berkaitan dengan teknologi salah satunya *Artificial Intelligence*. Pelaksanaan Studi Independen ini dimulai dari bulan Februari hingga Juli 2022. Project akhir menjadi bagian penutup dari Studi Independen ini. Project akhir yang dibuat ini diambil dari topik klasifikasi teks yang mana masuk ke dalam domain *Natural Language Processing* (NLP). Project ini tentang mengidentifikasi teks berita hoax. Data dari laman website kominfo menyatakan bahwa terdapat 800.000 situs penyebar *hoax* dan *hate speech* di Indonesia. Pada eksperimen ini menggunakan dua algoritma yaitu *Long Short Term Memory* (LSTM) dan *Bidirectional Encoder Representations from Transformer* (BERT). Kedua algoritma tersebut dapat digunakan untuk mengklasifikasi teks. Dataset yang digunakan berjumlah 2.216 data dan 956 data. Pada kedua algoritma tersebut diambil nilai akurasi yang paling tinggi. Hasil tersebut diimplementasikan pada sistem informasi yang dirancang. Hasil akurasi yang didapatkan oleh LSTM sebesar 55% pada dataset berjumlah 956 data. Sedangkan BERT sebesar 89% dengan menggunakan *pre-trained model* Indolem pada *dataset* berjumlah 956 data. Project akhir ini berupa aplikasi berbasis website yang saat ini dapat diakses melalui *localhost*. Aplikasi yang diberi nama Sotaken ini dapat membedakan berita *hoax* dan valid. Aplikasi ini berbasis website. Aplikasi yang dibuat ini menjadi salah satu solusi sebagai wadah masyarakat yang ingin mengetahui atau memastikan berita yang didapatkannya itu tergolong berita fakta (*valid*)/*hoax*. Aplikasi yang diberi nama Sotaken ini diharapkan dapat membantu masyarakat dalam mendeteksi sebuah kebenaran berita yang tersebar di masyarakat.

Kata Kunci: Orbit Future Academy; Project Akhir; Hoax; Mengidentifikasi Teks Berita Hoax; Algoritma; Aplikasi berbasis Website

Kata Pengantar

Puji syukur kehadiran Allah SWT atas segala limpahan rahmat dan hidayah-Nya yang telah dilimpahkan kepada penulis. Dengan rahmat dan hidayah-Nya, penulis dapat menyelesaikan Laporan Akhir Studi Independen Bersertifikat AI Mastery Program di Orbit Future Academy dengan baik dan tepat waktu.

Kegiatan studi independen bersertifikat AI Mastery Program memiliki tujuan untuk memperkenalkan teknologi artificial intelligence (AI) dan membuat perangkat AI yang dapat bermanfaat dan memiliki dampak sosial. Program ini berjalan dari bulan Februari hingga Juli. Proses pengerjaan proyek akhir akan dilaksanakan dari bulan April hingga Juni. Kelancaran kegiatan studi independen ini tidak terlepas dari bantuan berbagai pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, penulis mengucapkan terima kasih kepada pihak-pihak yang telah membantu:

1. Panitia program Magang dan Studi Independent Bersertifikat Kampus Merdeka Batch 2.
2. Panitia AI Mastery Program Orbit Future Academy.
3. Bapak Ravi K. Menon selaku ketua program AI Mastery Program Orbit Future Academy.
4. Semua Coach yang mengajar AI Mastery Program Orbit Future Academy, terkhususnya Coach Cahyo sebagai Homeroom coach saya sekaligus Coach Domain NLP.
5. Semua anggota kelas Atlas yang telah memberi warna pada setiap pertemuan.
6. Semua anggota tim yang bersama telah bekerja keras dalam menyelesaikan proyek akhir.
7. Keluarga serta teman yang selalu memberi semangat dan dukungan selama kegiatan studi independen.

Laporan ini menjelaskan kegiatan studi independen berupa pembuatan proyek akhir pada AI Mastery Program di Orbit Future Academy. Semoga laporan akhir

ini dapat memberikan manfaat, baik berupa inspirasi maupun motivasi bagi para pembaca. Dalam proses pembuatan laporan tentu masih terdapat banyak kesalahan dan kekurangannya. Oleh karena itu, kritik dan saran bersifat membangun sangat kami harapkan demi perbaikan laporan ke depannya. Dengan adanya proyek akhir ini semoga dapat membantu sebagian masyarakat dalam menyaring sumber informasi yang ada.

Bandung, 14 Juni 2022

A handwritten signature in black ink, featuring a large, stylized 'A' followed by a series of loops and a final flourish.

Alif Yanuar Aditya Subagyo

Daftar Isi

Lembar Pengesahan S1 Sistem Informasi Universitas Telkom.....	i
Lembar Pengesahan.....	ii
Abstraksi	iii
Kata Pengantar	iv
Daftar Isi	vi
Daftar Tabel	viii
Daftar Gambar	ix
Bab I Pendahuluan	1
I.1 Latar belakang	1
I.2 Lingkup	2
I.3 Tujuan	4
Bab II Orbit Future Academy	5
II.1 Struktur Organisasi	5
II.2 Lingkup Pekerjaan	6
II.3 Deskripsi Pekerjaan.....	7
II.4 Jadwal Kerja.....	8
Bab III Sistem Deteksi Berita Hoax Menggunakan Pendekatan <i>Bidirectional Encoder Representations from Transformer</i> (BERT).....	10
III.1 Latar Belakang Proyek Akhir.....	10
III.2 Proses Pelaksanaan Proyek Akhir	11
III.3 Hasil Proyek Akhir.....	29
Bab IV Penutup	36

IV.1	Kesimpulan	36
IV.2	Saran.....	36
Bab V	Referensi.....	38
Bab VI	Lampiran A. TOR.....	40
Bab VII	Lampiran B. Log Activity.....	43
Bab VIII	Lampiran C. Dokumen Teknik.....	47

Daftar Tabel

Tabel 1. 1 Pembagian Peran PA.....	3
Tabel 2. 1 Agenda Kelas.....	8
Tabel 3. 1 Daftar Dataset.....	13
Tabel 3. 2 Keterangan Dataset 956 Data.....	15
Tabel 3. 3 Keterangan Dataset 2.216 Data.....	15
Tabel 3. 4 Contoh Case Folding.....	15
Tabel 3. 5 Contoh Cleansing.....	16
Tabel 3. 6 Pembagian Data 1 Untuk LSTM.....	17
Tabel 3. 7 Pembagian Data 2 Untuk LSTM.....	17
Tabel 3. 8 Parameter Training Word2Vec	20
Tabel 3. 9 Hyperparameter LSTM.....	21
Tabel 3. 10 Hasil Pembagian Data 1 Untuk BERT.....	23
Tabel 3. 11 Hasil Pembagian Data 2 Untuk BERT.....	23
Tabel 3. 12 Hyperparameter BERT	26
Tabel 3. 13 Ilustrasi Confusion Matrix	27
Tabel 3. 14 Performa Model LSTM.....	30
Tabel 3. 15 Performa Model BERT Berdasarkan pre-trained model dan Jumlah Dataset.....	30
Tabel 3. 16 Kelebihan dan Kelemahan Aplikasi Sotaken.....	34
Tabel 7. 1 Log Activity.....	43
Tabel 8. 1 Daftar Dataset Digunakan.....	48
Tabel 8. 2 Performa Model LSTM Berdasarkan Dataset.....	50
Tabel 8. 3 Performa Model BERT Berdasarkan Pre-trained model dan Dataset..	50
Tabel 8. 4 Deskripsi Pembagian Peran	54
Tabel 8. 5 User Interface dan Penjelasan	58
Tabel 8. 6 Kelebihan dan Kekurangan Aplikasi Website Sotaken	60

Daftar Gambar

Gambar 2. 1 Logo Orbit Future Academy	5
Gambar 2. 2 Struktur Organisasi OFA.....	6
Gambar 3. 1 Tahapan Pengerjaan Proyek Akhir.....	12
Gambar 3. 2 Wordcloud Hoax	14
Gambar 3. 3 Wordcloud Fakta	14
Gambar 3. 4 Arsitektur CBOW dan SG.....	18
Gambar 3. 5 Arsitektur LSTM.....	21
Gambar 3. 6 Arsitektur LSTM Untuk Klasifikasi.....	22
Gambar 3. 7 Arsitektur BERT	24
Gambar 3. 8 Tampilan Landing Page	31
Gambar 3. 9 Tampilan Kumpulan Top-Headline News	31
Gambar 3. 10 Tampilan About Us	32
Gambar 3. 11 Tampilan Demo Aplikasi	32
Gambar 3. 12 Tampilan History Latest Prediction	33
Gambar 3. 13 Tampilan Contoh Prediksi Berita 1 (Hoax).....	33
Gambar 3. 14 Tampilan Contoh Prediksi Berita 2 (Valid)	34
Gambar 8. 1 Grafik Dataset 956 Data.....	49
Gambar 8. 2 Grafik Dataset 2.216 Data.....	50
Gambar 8. 3 Langkah Ke – 1	51
Gambar 8. 4 Langkah Ke – 2	52
Gambar 8. 5 Langkah Ke – 3	52
Gambar 8. 6 Langkah Ke – 4	53
Gambar 8. 7 Langkah Ke – 5	53
Gambar 8. 8 Struktur Tim	54
Gambar 8. 9 Flowchart Dari Aplikasi Sotaken (Society Anti Fake News)	57

Bab I Pendahuluan

I.1 Latar belakang

Program MSIB (Magang Studi Independen Bersertifikat) merupakan sebuah program yang memberikan kesempatan kepada mahasiswa untuk mengasah dan mendapatkan keahlian, pengetahuan / wawasan, dan sikap di dalam dunia industri dengan cara bekerja dan belajar secara langsung dalam proyek atau suatu permasalahan yang nyata. Setelah mengikuti program ini, mahasiswa memiliki hak untuk mengkonversi ke mata kuliah yang ada di Perguruan Tinggi dengan maksimal 20 SKS. Banyak perusahaan yang ikut serta dalam program MSIB ini untuk menjembatani tujuan kemendikbud dalam program ini. Banyak perusahaan yang ikut serta berpartisipasi dalam program MSIB ini salah satunya yaitu PT. Orbit Ventura Indonesia (Orbit Future Academy).

Program MSIB Orbit Future Academy dilaksanakan secara daring (*online*) dari bulan Februari sampai Juli mendatang. Orbit Future Academy membuka dua posisi yaitu AI Mastery Program dan Foundation of AI and Life Skill for Gen-Z Melalui kedua program tersebut, mahasiswa mendapatkan kesempatan untuk mempelajari bidang *Artificial Intelligence* (AI) serta keterampilan hidup dan kewirausahaan secara lebih mendalam melalui pembelajaran yang aktif, dinamis, holistik, dan menyenangkan dengan tujuan di akhir program ini, mahasiswa dapat lebih unggul pada bidang masing – masing serta kontribusi secara aktif dan bernilai positif terhadap komunitas maupun masyarakat dan berpartisipasi aktif untuk membangun generasi muda menuju Indonesia Maju.

AI Mastery Program merupakan pelatihan *Artificial Intelligence* (AI) untuk pelajar yang memiliki tujuan tidak hanya memperkenalkan teknologi AI ke mahasiswa. Akan tetapi, juga memiliki kemungkinan supaya mahasiswa dapat mengangkat perangkat AI sehingga bisa membuat sesuatu yang menciptakan dampak sosial. Program ini berfokus pada komponen utama *Artificial Intelligence* (AI) diantaranya seperti *Data Science*, *Natural language Processing*, *Reinforcement Learning*, dan *Computer Vision*. Modul pembelajarannya meliputi *Artificial Intelligence Foundation*, *Python Foundation*, *Manajemen Data*, *Git*, &

Deployment, Machine Learning & Deep Learning, Data Science, Natural Language Processing, Reinforcement Learning, Computer Vision. Selain mempelajari materi – materi secara teori dan praktik, Orbit Future Academy memberikan *pre test, post test*, diskusi untuk mengetahui pengetahuan mahasiswa terkait dengan materi yang telah diberikan.

Project akhir menjadi penutup dari materi – materi yang telah disampaikan. Terkait dengan program studi independen ini, peneliti melakukan project akhir tentang klasifikasi teks berita *hoax* dengan menggunakan algoritma *deep learning* yaitu LSTM (*Long Short Term Memory*) dan BERT (*Bidirectional Encoder Representations from Transformers*). Pada dua algoritma tersebut akan dilakukan perbandingan hasil akurasi. Hasil akurasi menunjukkan lebih tinggi pada algoritma BERT dibanding dengan LSTM. BERT merupakan teknologi *open source* yang berbasis jaringan saraf untuk *pre-training Natural Language Processing (NLP)*. Dalam menggunakan BERT pada project akhir ini dimaksud untuk mendeteksi berita *hoax* yang terkandung dalam teks – teks berita. Berita *hoax* merupakan suatu informasi yang tidak terbukti kebenarannya. Akan tetapi, dibuat seakan – akan informasi tersebut benar adanya.

I.2 Lingkup

Studi Independen Artificial Intelligence (AI) Mastery Program yang diselenggarakan oleh PT. Orbit Ventura Indonesia (Orbit Future Academy). Program ini diselenggarakan dari bulan Februari hingga Juli mendatang. Sedangkan untuk pengerjaan project akhir dimulai dari bulan April hingga Juni dengan rincian sebagai berikut.

- Mencari permasalahan di sekitar : Minggu ke – 1,
(*problem scoping*) dan topik AI yang disukai dan dipilih 4 April – 8 April 2022
- Mencari referensi jurnal dan anggota tim : Minggu ke – 2,
11 April – 15 April 2022
- Fiksasi pembuatan Kelompok, membaca referensi dan Ide Project Akhir : Minggu ke – 3,
18 April – 22 April 2022

- Fiksasi Ide Project Akhir dan Pembagian Tugas Anggota Tim : Minggu ke – 4,
25 April – 29 April 2022
- Libur Nasional dan Cuti Bersama Hari Raya Idul Fitri : Minggu ke – 5,
2 Mei – 6 Mei 2022
- *Data Acquisition, Reprocessing, Modelling* : Minggu ke – 6,
9 Mei – 13 Mei 2022
- *Modelling*, Evaluasi, UI/UX : Minggu ke – 7,
16 Mei – 20 Mei 2022
- Hasil Sementara : Minggu ke – 8,
23 Mei – 27 Mei 2022
- Merapikan Coding dan *Deployment* : Minggu ke – 9,
30 Mei – 3 Juni 2022
- Menyelesaikan *Deployment* : Minggu ke – 10,
6 Juni – 10 Juni 2022
- Menyusun Laporan : Minggu ke – 11,
13 Juni – 17 Juni

Pada tiap progress tersebut tentunya terlibat secara individu maupun kelompok. Pengerjaan kelompok ini dilakukan dengan diskusi via chat maupun *Google Meet* agar mempermudah komunikasi didalam pengerjaan. Berikut Tabel 1.1 merupakan pembagian peran untuk pengerjaan proyek akhir.

Tabel 1. 1 Pembagian Peran PA

Anggota Kelompok	Peran
Alif Yanuar Aditya Subagyo	Deployment
Annisa Kunarji Sari	Modelling
Diah Siti Fatimah Azzahrah	Modelling
Muhammad Fachrizal Zulfi Hendra	UI/UX
Salsabila Zahrani Amril	Modelling

I.3 Tujuan

Tujuan atau hasil dari mengikuti MSIB dari awal hingga akhir adalah sebagai berikut.

1. Mahasiswa mendapatkan keterampilan untuk mengimplementasikan teori – teori tersebut ke dalam praktik.
2. Mahasiswa mendapatkan pengetahuan mengenai *Artificial Intelligence*, *Machine Learning*, *Deep Learning* dan ilmu lainnya.
3. Mahasiswa mendapatkan wadah untuk belajar dan mengembangkan diri melalui MSIB yang diikuti.
4. Mahasiswa dapat mengimplementasikan pengetahuan yang telah didapatkan pada project akhir dengan baik.

Bab II Orbit Future Academy

II.1 Struktur Organisasi



Gambar 2. 1 Logo Orbit Future Academy

Orbit Future Academy (OFA) didirikan pada tahun 2016 dengan tujuan untuk meningkatkan kualitas hidup melalui inovasi, edukasi, dan pelatihan keterampilan. Label atau *brand* Orbit merupakan kelanjutan dari warisan mendiang Prof. Dr. Ing. B. J. Habibie (presiden Republik Indonesia ke-3) dan istrinya, Dr. Hasri Ainun Habibie. Mereka berdua telah menjadi penggerak dalam mendukung perkembangan inovasi dan teknologi pendidikan di Indonesia. OFA mengkurasi dan melokalkan program/kursus internasional untuk *upskilling* atau *reskilling* pemuda dan tenaga kerja menuju pekerjaan masa depan. Hal ini sesuai dengan slogan OFA, yakni “*Skills-for-Future-Jobs*”.

Visi:

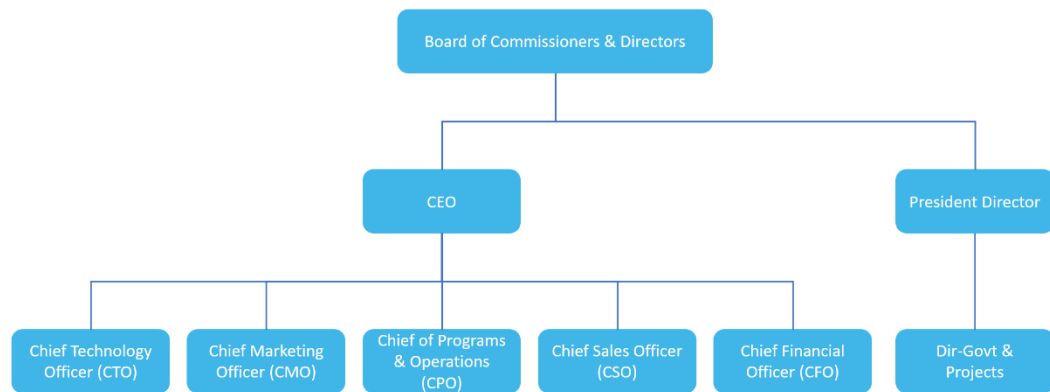
Memberikan pembelajaran berbasis keterampilan transformatif terbaik untuk para pencari kerja & pencipta lapangan kerja.

Misi:

1. Membangun jaringan Orbit Transformation Center (OTC) secara nasional untuk menyampaikan kurikulum keterampilan masa depan berbasis sertifikasi melalui Platform Konten Digital.

2. Secara proaktif bekerja dengan pemerintah & organisasi dengan mengubah tenaga kerja mereka agar sesuai dengan perubahan pekerjaan yang terjadi karena Industri 4.0.
3. Melatih pemuda dengan keterampilan kewirausahaan & mencocokkan mereka dengan peluang masa depan yang muncul di berbagai industri.
4. Menghubungkan jaringan inkubator dan akselerator yang dikurasi ke industri, investor, dan ekosistem start-up global.

Struktur organisasi OFA dapat dilihat pada Gambar 2.2.



Gambar 2. 2 Struktur Organisasi OFA

II.2 Lingkup Pekerjaan

Seorang fasilitator akan mendampingi kurang lebih 40 peserta MSIB (student) dalam satu kelas. Terdapat dua jenis fasilitator, yakni:

a. Homeroom Coach

Homeroom coach bertugas menyampaikan materi tentang dasar-dasar AI, memberikan penilaian pada student, dan mengarahkan *student* saat pengerjaan Proyek Akhir (PA).

b. Domain Coach

Domain coach bertugas menyampaikan materi tentang domain AI dan memberikan penilaian pada student.

Lingkup pekerjaan student adalah mengikuti kelas bersama homeroom atau domain coach, sesuai agenda kelas, hingga program selesai.

II.3 Deskripsi Pekerjaan

Berikut adalah deskripsi pekerjaan student sebelum pengerjaan PA:

- a. Mengikuti pre-test.
- b. Mengikuti kelas sesi pagi pada pukul 08.00 hingga 11.30 WIB.
- c. Mengikuti kelas sesi siang pada pukul 13.00 hingga 16.30 WIB.
- d. Mengulang materi yang telah disampaikan di kelas sesi pagi dan siang, setelah kelas sesi siang, selama 1 jam (*self-study*).
- e. Mengerjakan latihan individu atau kelompok yang diberikan oleh homeroom atau domain coach saat kelas berlangsung.
- f. Mengerjakan tugas yang diberikan homeroom atau domain coach hingga batas waktu tertentu.
- g. Mengerjakan *mini project* yang diberikan homeroom atau domain coach hingga batas waktu tertentu
- h. Mengikuti post-test.

Student memiliki peran *Deployment* selama pengerjaan PA, dengan deskripsi pekerjaan sebagai berikut:

- a. Merancang pembuatan website

Pada proses pembuatan website saya memiliki peran untuk mengambil alih pembuatan tampilan dan juga kerangka pada website. Pada pembuatan kerangka, saya menggunakan *framework* Flask sebagai *framework* website dengan menggunakan bahasa pemrograman python. Dan juga saya menggunakan *framework* Bootstrap sebagai pembantu kerangka CSS dalam pembuatan tampilan website. Menggunakan *library* SQLite sebagai sistem basis data untuk *history* teks pengujian sebelumnya.

b. Menyusun file – file berkaitan dengan deployment

Menyusun file yang berkaitan dengan deployment sebagai bukti proyek akhir yang sudah dibuat, sekaligus file penting seperti “requirement.txt”. dan juga file yg dibutuhkan dalam proses *publish to cloud*.

c. Mendeployment model AI pada website

Melakukan deployment model AI pada website yang berformat ekstensi “.h5” kepada website yang dibuat menggunakan *framework* Flask. Dengan menggunakan Flask sendiri, kita dapat mengekstrak kembali sebuah file yang dikonversi dari package berbahasa pemrograman Python. Ekstensi “.h5” sendiri adalah sebuah hasil model yang dibuat menggunakan *library* Tensorflow.

II.4 Jadwal Kerja

Program ini berlangsung setiap hari kerja (Senin sampai dengan Jumat) selama 8 jam per harinya, dengan rincian sebagai berikut:

Tabel 2. 1 Agenda Kelas

Pukul (WIB)	Durasi (jam)	Aktivitas
08.00 s.d. 11.30	3.5	Kelas Sesi Pagi
13.00 s.d. 16.30	3.5	Kelas Sesi Siang
16.30 s.d. 17.30	1	<i>Self-Study</i>

Program ini berlangsung dari bulan Februari 2022 sampai dengan bulan Juli 2022.

Bab III Sistem Deteksi Berita Hoax Menggunakan Pendekatan *Bidirectional Encoder Representations from Transformer (BERT)*

III.1 Latar Belakang Proyek Akhir

Seiring berkembangnya teknologi, kemudahan dalam mengakses berbagai hal melalui *internet* membawa dampak perubahan yang besar. Sebuah informasi dapat dengan mudah disebarluaskan hanya dengan hitungan detik. Tiap hari masyarakat mendapatkan informasi – informasi untuk mengetahui kabar terkini. Akan tetapi, adanya perkembangan teknologi ini ternyata juga mendapatkan dampak negatifnya yaitu muncul berita – berita palsu / *hoax* yang dilakukan oleh oknum – oknum tertentu. *Hoax* adalah suatu upaya untuk memanipulasi pembaca supaya terpengaruh pada opini yang dibawa. Salah satu contoh dari berita *hoax* yaitu meyakinkan sebuah kejadian yang sebenarnya tidak sesuai dengan fakta lapangan [1]. Banyak masyarakat yang menjadi purno dan panik ketika mendapatkan berita tersebut, sedangkan masyarakat belum mengetahui apakah berita tersebut termasuk fakta / *hoax*. Jumlah berita *hoax* yang tersebar saat ini sangat besar, tentu hal ini membuat masyarakat harus berhati - hati dalam mendapatkan informasi [2]. Data dari laman website kominfo mengatakan bahwa terdapat 800.000 situs penyebar *hoax* dan *hate speech* di Indonesia. Selain itu, berita *hoax* memberikan dampak negatif yang besar bagi masyarakat. *Hoax* merupakan efek samping dari era keterbukaan yang memiliki peluang untuk menciptakan perpecahan dan permusuhan karena dapat membuat masyarakat bingung akan sebuah kebenaran dari suatu informasi (kominfo.go.id, 2021).

Masyarakat Indonesia tercatat untuk pengguna internet mencapai 171 juta penduduk, dan 95% dari pengguna internet itu memanfaatkannya untuk beraktifitas di media sosial. Adanya aktivitas masyarakat yang cukup tinggi dalam bersosial media dan layanan aplikasi pesan mengakibatkan meningkatnya potensi terkena berita *hoax* [3]. Penelitian yang telah dilakukan Mastel menyatakan bahwa media paling banyak digunakan untuk penyebaran *hoax* yaitu situs web sebesar 34,90%, aplikasi *chatting* sebesar 62,80% dan melalui media sosial mencapai 92,40% [4].

Pada permasalahan tersebut tentunya membutuhkan upaya - upaya untuk menghentikan penyebaran berita *hoax* yang meresahkan masyarakat. Selain itu, adanya berita *hoax* ini juga merugikan masyarakat. Permasalahan ini apabila tidak dilakukan upaya - upaya untuk menghilangkan berita - berita *hoax* yang beredar akan memberikan dampak negatif seperti keributan, keresahan, perselisihan, ujaran kebencian, kecemasan, dan lain - lain.

Teknologi Artificial Intelligence pada ilmu *Natural Language Processing* (NLP) memberikan solusi untuk membangun sistem yang secara otomatis dapat mendeteksi suatu berita tergolong berita palsu/fakta [5]. Oleh karena itu, tentunya membutuhkan algoritma untuk menentukan berita yang didapatkan masuk ke berita fakta atau *hoax*. Klasifikasi berita *hoax* merupakan salah satu dari aplikasi kategorisasi teks [6]. Pada project akhir ini menggunakan algoritma LSTM dan BERT dimana hasil akurasi tertinggi akan digunakan pada pengerjaan project akhir ini. Nantinya, output dari project ini berupa suatu sistem yang dapat mendeteksi berita *hoax*.

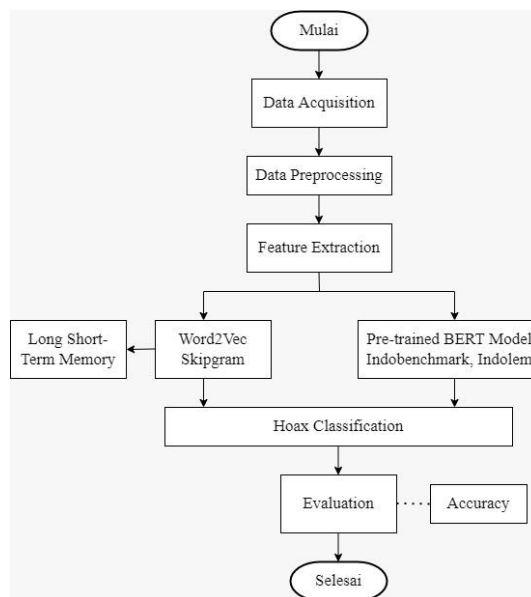
Dengan adanya klasifikasi berita *hoax* pemecahan masalah yang telah meresahkan masyarakat dapat terselesaikan. Masyarakat dapat membedakan berita yang *hoax* maupun tidak. Project yang kami beri nama Sotaken dapat membedakan berita *hoax* dan valid. Aplikasi ini berbasis website. Aplikasi yang dibuat ini menjadi salah satu solusi sebagai wadah masyarakat yang ingin mengetahui atau memastikan berita yang didapatkannya itu tergolong berita fakta(valid)/*hoax*.

Bagi user yang sudah terakses dengan aplikasi tersebut dan ingin menggunakan aplikasi tersebut. *User* hanya memasukkan teks berita yang ingin di *input* dan masukkan ke dalam website Sotaken. Setelah itu *output* yang dikeluarkan dari website adalah hasil klasifikasi dari berita yang telah diinputkan oleh *user*. Aplikasi yang diberi nama Sotaken ini diharapkan dapat membantu masyarakat dalam mendeteksi sebuah kebenaran berita yang tersebar di masyarakat.

III.2 Proses Pelaksanaan Proyek Akhir

Pengerjaan proyek akhir dimulai dengan pengumpulan data terlebih dahulu. Data yang akan digunakan pada pengerjaan proyek akhir ini adalah dataset publik.

Setelah mendapatkan data yang sesuai akan dilanjutkan ke tahap *preprocessing* data, pemodelan, evaluasi dan *deployment*. Metode awal yang akan digunakan untuk klasifikasi berita hoaks adalah *Long Short-Term Memory*, namun karena akurasi yang dihasilkan belum cukup baik, maka dilakukan pemodelan menggunakan pendekatan *Bidirectional Encoder Representations from Transformer* (BERT). Gambar 3.1 merupakan tahapan pengerjaan proyek akhir:



Gambar 3. 1 Tahapan Pengerjaan Proyek Akhir

Penjelasan lebih detail dari proses pengerjaan proyek akhir adalah sebagai berikut:

1. Persiapan Data

Persiapan data meliputi pengumpulan data, pemilihan fitur, pelabelan data, dan penggabungan data. Fitur yang akan digunakan adalah isi atau narasi berita dan labelnya. Sehingga pada proses pengumpulan data, data akan dicek terlebih dahulu kesesuaian dengan data yang dibutuhkan.

a. Metode pengumpulan data

Studi literatur dilakukan untuk mengumpulkan data dengan cara mencari literatur ilmiah, jurnal serta informasi yang tersedia di internet yang dapat dijadikan sebagai referensi dan landasan teori dalam pembuatan proyek akhir. Data yang akan digunakan merupakan *dataset* publik yang tersedia di internet, khususnya situs penyedia *dataset*

seperti Kaggle dan Github. Tabel 3.1 merupakan daftar *dataset* yang digunakan.

Tabel 3. 1 Daftar Dataset

Pemilik	Tautan Unduhan	Jumlah Data
Pierobeat	https://github.com/pierobeat/Hoax-News-Classification	250 berita <i>hoax</i> 250 berita valid
Jibran Fawaid	https://github.com/JibranFawaid/tumbackhoax-dataset	683 berita <i>hoax</i> 433 berita valid
Rahutomo, dkk	https://data.mendeley.com/datasets/p3hfgr5j3m/1	228 berita <i>hoax</i> 372 berita valid

b. Jumlah data

Jumlah dataset yang akan digunakan sebanyak 956 data dan 2,216 data. Dataset 956 data merupakan dataset hasil penggabungan dari dataset milik Rahutomo, dkk dan Pierobeat. Dataset tersebut memiliki jumlah kelas yang seimbang. Sedangkan dataset dengan jumlah 2,216 data, merupakan hasil penggabungan 956 data sebelumnya dengan dataset milik Jibran Fawaid. Pada dataset tersebut data yang memiliki label *hoax* sebanyak 1,161 data dan label *valid* sebanyak 1,055 data. Gambar 3.2 merupakan *word cloud* dari berita *hoax* dan Gambar 3.3 merupakan *word cloud* dari berita fakta.

3.3 merupakan jumlah data tiap kelas pada dataset yang berjumlah 2.216 data.

Tabel 3. 2 Keterangan Dataset 956 Data

Kelas	Jumlah
<i>Hoax</i>	478
Valid	478

Tabel 3. 3 Keterangan Dataset 2.216 Data

Kelas	Jumlah
<i>Hoax</i>	1,161
Valid	1,055

2. Preprocessing

Preprocessing data atau pembersihan data merupakan tahap pembersihan data teks dari karakter-karakter tertentu atau hal yang dapat mengganggu proses pemodelan. Preprocessing data meliputi:

a. Case folding

Tahap pertama dalam *preprocessing* data adalah *case folding*. *Case folding* adalah proses mengubah semua karakter menjadi huruf kecil.

Tabel 3.4 merupakan contoh *case folding*.

Tabel 3. 4 Contoh Case Folding

Masukan	Keluaran
Ikan lele disebut-sebut mengandung 3.000 sel kanker. Kabar tersebut beredar di media sosial. Ikan air tawar itu juga dianggap sebagai ikan paling jorok karena memakan segala macam jenis kotoran.	ikan lele disebut-sebut mengandung 3.000 sel kanker. kabar tersebut beredar di media sosial. ikan air tawar itu juga dianggap sebagai ikan paling jorok karena memakan segala macam jenis kotoran.

b. *Cleansing*

Proses *cleansing* merupakan tahap untuk menghilangkan karakter atau atribut yang tidak berpengaruh dalam proses klasifikasi [7]. Pada tahap ini akan dilakukan pembersihan terhadap data yang memiliki atribut tanda baca, angka (0-9), URL, tautan, simbol-simbol dan karakter-karakter lainnya. Tabel 3.5 merupakan contoh *cleansing*.

Tabel 3. 5 Contoh Cleansing

Masukan	Keluaran
Ikan lele disebut-sebut mengandung 3.000 sel kanker. Kabar tersebut beredar di media sosial. Ikan air tawar itu juga dianggap sebagai ikan paling jorok karena memakan segala macam jenis kotoran.	ikan lele disebutsebut mengandung sel kanker kabar tersebut beredar di media sosial ikan air tawar itu juga dianggap sebagai ikan paling jorok karena memakan segala macam jenis kotoran

3. *Train Test Split Data*

Pada pengujian model akan dilakukan 2 jenis dataset dengan jumlah yang berbeda. *Dataset* pertama berjumlah 956 data, dan dataset kedua berjumlah 2216 data. *Dataset* yang telah melewati proses *preprocessing* akan dibagi menjadi 2 yaitu data latih dan data uji dengan perbandingan 80:20. Data uji akan dibagi kembali untuk dijadikan data validasi pada proses *k-fold validation* sebanyak 10 kali. Penjelasan mengenai detail jumlah akan dijelaskan pada bagian pemodelan.

4. Pemodelan Menggunakan LSTM

Pemodelan tahap pertama adalah menggunakan metode *Long- Short Term Memory* (LSTM) dan *word embedding word2vec*. Berikut merupakan penjelasan mengenai *word embedding word2vec* dan metode LSTM.

a. Pembagian data latih dan data uji

Detail mengenai jumlah pembagian data uji, data latih dan data validasi dapat dilihat pada Tabel 3.6 dan Tabel 3.7.

1) Pembagian data pertama

Data pertama memiliki jumlah data sebanyak 956 data. Data akan dibagi menjadi data uji, data latih dan data validasi. Perbandingan yang digunakan saat pembagian data adalah 80:20. Tabel 3.6 merupakan jumlah dari masing-masing hasil dari pembagian data.

Tabel 3. 6 Pembagian Data 1 Untuk LSTM

Data	Jumlah Data
Data Latih	573
Data Uji	192
Data Validasi	191

2) Pembagian data kedua

Data kedua memiliki jumlah data sebanyak 2.216 data. Pembagian dilakukan sama seperti pada dataset pertama. Tabel 3.7 merupakan jumlah dari masing-masing hasil pembagian data.

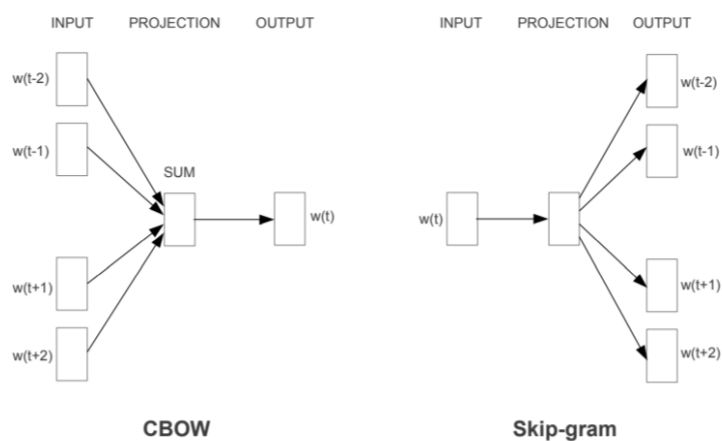
Tabel 3. 7 Pembagian Data 2 Untuk LSTM

Data	Jumlah Data
Data Latih	1.329
Data Uji	443
Data Validasi	444

b. Word Embedding

Word embedding merupakan proses memetakan kata-kata dalam bentuk vektor. Kata-kata yang memiliki hubungan semantik dengan kata lain akan dipetakan dalam nilai vektor yang saling berdampingan [8]. *Skip-gram* (SG) dan *Continuous Bag of Words* (CBOW) merupakan bagian dari jenis algoritma *word embedding* [9]. CBOW melakukan prediksi pada kata target dengan cara

menggunakan kata yang berada di sebelah kiri dan kanan kata target pada batas *window* dengan nilai tertentu. Sedangkan pada SG, bekerja dengan cara menggunakan sebuah kata untuk memperkirakan kata-kata yang dapat berada di bagian kanan dan kiri kata tersebut, dalam batas *window* dengan nilai tertentu. *Window* merupakan kernel untuk mendapatkan masukan dan kata target, *window* akan digeser dari awal hingga akhir dalam rangkaian kata. Pada proyek ini akan digunakan *Skip-gram* sebagai *word embedding*. SG dianggap memiliki kinerja yang lebih baik pada kumpulan data besar [9]. Gambar 3.4 merupakan arsitektur CBOW dan SG.



(Mikolov, 2013)

Gambar 3. 4 Arsitektur CBOW dan SG

Sebelum masuk ke tahap *word embedding* akan dilakukan *preprocessing* kembali, yaitu *one hot encoding*, *tokenizing data*, *text to sequences*, *padding* dan *truncate*. Setelah beberapa langkah tersebut akan dilanjutkan ke tahap *word embedding*. Berikut merupakan detail dari setiap langkah-langkahnya:

- 1) *One hot encoding*

Data kategorik merupakan jenis data yang terdiri dari variabel atau data dari hasil pengelompokan berdasarkan kategori yang telah ditentukan. Data kategorik berjenis tipe data string, sehingga tidak dapat diukur menggunakan angka atau bilangan. Contoh data kategorik adalah ‘sangat baik’, ‘baik’, ‘tidak baik’. Tipe data kategorik tidak dapat diolah oleh program komputer, sehingga perlu dilakukan perubahan dari tipe data kategorik menjadi tipe data numerik, proses tersebut dikenal sebagai *encoding*. *One hot encoding* merupakan metode *encoding* yang merepresentasikan data bertipe kategori sebagai vektor biner yang bernilai integer 0 dan 1, elemen yang memiliki nilai kategori akan diberi nilai 1, dan elemen lain akan diberi nilai 0. Pada dataset yang digunakan akan dilakukan *one hot encoding* pada kolom label.

2) *Tokenizing*

Tokenizing atau tokenisasi merupakan proses untuk memecah setiap kata pada kalimat menjadi kata demi kata. Proses *tokenizing* dilakukan dengan menggunakan library dari *Keras*. Urutan kata demi kata yang telah dipecah akan dibagi menjadi daftar token dan memiliki indeks.

3) *Text to Sequences*

Setelah melakukan tokenisasi, langkah selanjutnya adalah mengubah setiap kalimat dalam data ke dalam bentuk *sequence* atau urutan. Kalimat akan diubah menjadi daftar urutan dari indeks token.

4) *Padding*

Setelah melalui tahap *text to sequences*, langkah selanjutnya adalah membuat setiap *sequence* memiliki panjang yang sama. Hal tersebut dapat dilakukan dengan menambahkan padding pada setiap *sequence*. Sebelum menambahkan padding, hal yang perlu diketahui adalah menetapkan panjang maksimal

yang akan digunakan dari seluruh vektor kalimat yang ada. Padding akan menambahkan nilai 0 secara sufiks maupun prefiks sesuai dengan panjang maksimum yang telah didefinisikan sebelumnya. Apabila *sequence* melebihi panjang maksimum, maka *sequence* tersebut akan dipotong agar *sequence* memiliki panjang yang sesuai dengan panjang maksimal. Panjang maksimal yang digunakan pada tahap ini adalah 512.

5) *Model Word Embedding*

Setelah penambahan *indeks* dan *padding* maka akan dilanjutkan ke dalam tahap *word embedding*. Pemodelan dalam melakukan *word embedding* dibantu dengan menggunakan library gensim. Berikut Tabel 3.8 merupakan parameter yang digunakan dalam melakukan *training word2vec*.

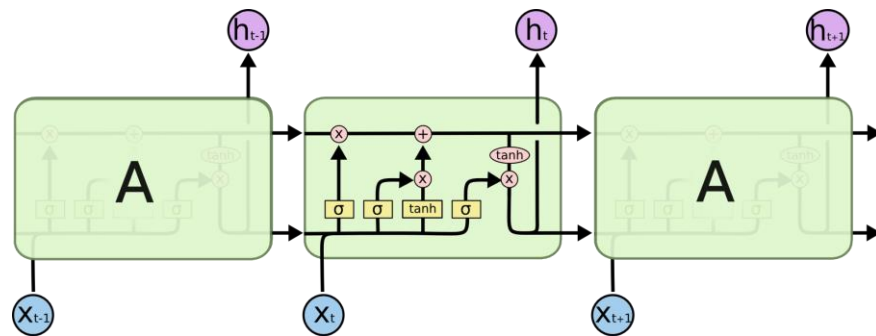
Tabel 3. 8 Parameter Training Word2Vec

Min Count	Window	Epoch	Algoritma pelatihan (sg)	Size
1	5	10	<i>Skip-gram</i>	5

c. *Long Short Term-Memory*

Long Short-Term Memory (LSTM) merupakan salah satu jenis arsitektur dari *Recurrent Neural Network* (RNN), yang pertama kali diperkenalkan oleh Hochreiter dan Schmidhuber. RNN memiliki kelemahan yaitu, tidak dapat mengingat informasi saat proses pelatihan ketika informasi yang disimpan terlalu banyak [10]. Kekurangan tersebut dapat diatasi dengan LSTM, karena LSTM mampu mengingat informasi yang dibutuhkan dan menghapus informasi yang tidak dibutuhkan lagi. Perbedaan dari RNN dan LSTM adalah pada kemampuan dalam mengelola informasi. LSTM memiliki empat gerbang yaitu *forget gate*, *input gate*, *input*

modulation gate dan *output gate*. LSTM juga mempunyai *internal cell state* yang memiliki kemampuan untuk menyimpan informasi hasil penyaringan dari unit sebelumnya. Gerbang pertama pada LSTM adalah *forget gate*. *Forget gate* memiliki fungsi untuk melupakan atau menghapus informasi yang tidak lagi relevan. Gerbang selanjutnya adalah *input gate* yang memiliki fungsi untuk menambahkan informasi hasil seleksi dari *forget gate*. Pada *input gate* terdapat istilah *input modulation gate* yang memiliki fungsi untuk memodulasi informasi agar dapat mengurangi kecepatan konvergensi data *zero-mean*. Terakhir adalah *output gate* yang memiliki fungsi menghasilkan data yang aktual. Gambar 3.5 merupakan arsitektur LSTM.



Gambar 3. 5 Arsitektur LSTM

Pada pemodelan yang akan dilakukan akan menggunakan metode LSTM dengan *hyperparameter* dapat dilihat pada Tabel 3.9 sebagai berikut.

Tabel 3. 9 Hyperparameter LSTM

Hyperparameter	Value
Loss function	Categorical-crossentropy
Optimizer	Adam
Number of epochs	10
Batch Size	8

Berikut pada Gambar 3.6 merupakan arsitektur LSTM yang digunakan untuk melakukan klasifikasi.

Model: "sequential_2"

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, None, 5)	57585
lstm_2 (LSTM)	(None, 100)	42400
dense_2 (Dense)	(None, 2)	202
Total params: 100,187		
Trainable params: 100,187		
Non-trainable params: 0		

Gambar 3. 6 Arsitektur LSTM Untuk Klasifikasi

5. Pemodelan Menggunakan BERT

Pemodelan kedua adalah menggunakan pendekatan *Bidirectional Encoder Representations from Transformer* (BERT). Pada pemodelan ini akan menggunakan *pretrained model* dari BERT yaitu Indobenchmark dan Indolem. Berikut merupakan langkah-langkah pemodelan dan penjelasan mengenai metode BERT.

a. Pembagian data latih dan data uji

Detail mengenai jumlah pembagian data uji, data latih dan data validasi dapat dilihat pada Tabel 3.10 dan Tabel 3.11.

1) Pembagian data pertama

Sama seperti data yang digunakan saat pemodelan menggunakan LSTM. Data pertama memiliki jumlah data sebanyak 956 data. Perbandingan pembagian antara data latih dan data uji adalah 80:20 dan perbandingan pembagian antara data uji dan data validasi adalah 50:50. Tabel 3.10 merupakan jumlah dari masing-masing hasil dari pembagian data.

Tabel 3. 10 Hasil Pembagian Data 1 Untuk BERT

Data	Jumlah Data
Data Latih	764
Data Uji	96
Data Validasi	96

2) Pembagian data kedua

Data kedua memiliki jumlah data sebanyak 2.216 data. Pembagian dilakukan sama seperti pada dataset pertama. Tabel 3.11 merupakan jumlah dari masing-masing hasil pembagian data.

Tabel 3. 11 Hasil Pembagian Data 2 Untuk BERT

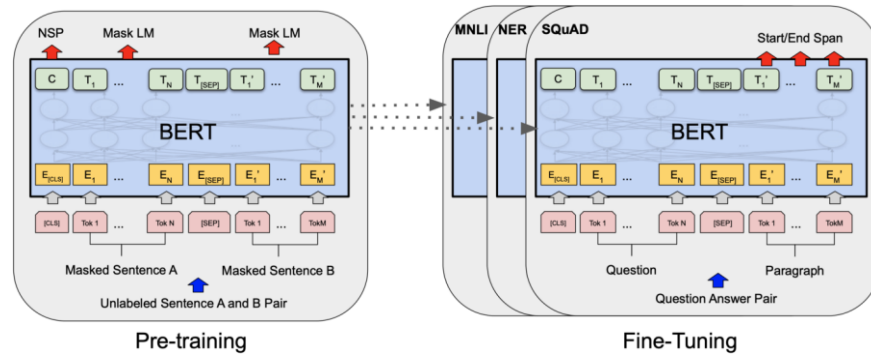
Data	Jumlah Data
Data Latih	1772
Data Uji	222
Data Validasi	222

b. BERT

Bidirectional Encoder Representations from Transformers (BERT) merupakan model *deep learning* untuk pemrosesan bahasa alami atau *Natural Language Processing* yang dikembangkan oleh [11]. dan para peneliti di Google AI Language pada tahun 2018. BERT dibangun menggunakan pendekatan Transformer. Transformer adalah sebuah model yang dapat menghindari pengulangan dan menerapkan mekanisme *attention* untuk memahami ketergantungan secara menyeluruh antara *input* dan *output* [12].

Arsitektur BERT memiliki *multi-layer bidirectional* Transformer berbasis *encoder* [11]. BERT mendefinisikan *L* sebagai

jumlah dari *layer*, H sebagai *hidden size*, dan A sebagai jumlah dari *self-attention heads*. Gambar 3.7 merupakan arsitektur BERT



(Devlin et al, 2019)

Gambar 3. 7 Arsitektur BERT

c. Pre-trained Model BERT

BERT memiliki kelebihan untuk melakukan *Transfer Learning*. *Transfer learning* merupakan sebuah metode dimana sebuah model yang telah dilatih untuk suatu tugas, digunakan kembali sebagai titik awal untuk model pada tugas yang baru. BERT telah menyediakan *pretrained* model yang dapat dimanfaatkan dalam menyelesaikan tugas di bidang *Natural Language Processing* tanpa harus membuat model dari awal [11]. Pada pembuatan proyek akhir ini *pretrained* model yang akan digunakan adalah IndoBERT dan Indolem. Pada penelitian Willie et al menjelaskan bahwa IndoBERT dilatih menggunakan 250 juta kalimat [13]. IndoBERT memiliki 4 jenis model yaitu indobert-base, indobert-large, indobert-lite-base, indobert-lite-large. Model yang akan digunakan adalah indobert-base. Sedangkan pada penelitian Koto et al Indolem dilatih menggunakan Wikipedia sebanyak 74 kata, artikel berita 55 juta kata, dan web *corpus* sebanyak 90 juta kata [14].

d. Klasifikasi

Data yang telah dibersihkan pada tahap *preprocessing* akan dilanjutkan ke tahap klasifikasi. Dalam melakukan klasifikasi,

BERT akan melakukan 2 langkah pemrosesan yaitu *pre-training* dan *fine tuning*.

1) Pre-training

Terdapat 2 proses yang dilalui pada tahap *pre-training* yaitu *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP). Pada tahap *Masked Language Modeling*, BERT akan mengambil sebuah kalimat sebagai masukan dan menutupi 15% kata dari kalimat tersebut secara acak menggunakan token [MASK], langkah selanjutnya model akan memprediksi kata yang ditutupi. Sedangkan *Next Sentence Prediction*, model BERT akan mempelajari hubungan antar kalimat. BERT melakukan prediksi apakah kalimat B merupakan kalimat aktual yang menghasilkan kalimat A. Model akan memiliki kalimat A dan kalimat B pada proses *pre-training*, sebanyak 50% kalimat B merupakan kalimat berikutnya setelah kalimat A dan akan dilabeli dengan isNext, sedangkan 50% lagi merupakan kalimat acak dan dilabeli dengan NotNext [15].

Sebelum memasuki *fine tuning*, masukan akan melewati beberapa proses yaitu:

a) Tokenizer

Pada proses ini kalimat akan dipecah menjadi kata per kata untuk menghasilkan masukan yang sesuai. BERT menggunakan *WordPiece tokenizer* untuk menghasilkan kamus kata atau kumpulan *vocabulary*.

b) Token khusus

BERT memiliki token khusus yang digunakan agar masukan dapat dipahami oleh model BERT, token khusus yang digunakan untuk tugas klasifikasi diantaranya adalah token [CLS], [SEP], dan [PAD]. Token [CLS] ditambahkan di awal kalimat, token [SEP] digunakan diakhir kalimat untuk

memisahkan segmen teks, dan [PAD] ditambahkan pada kalimat yang kurang dari panjang maksimum.

c) Token embedding

Selanjutnya, pada proses ini kalimat akan dicocokkan dengan id pada *dictionary* yang dihasilkan saat proses *tokenizer*. Id tersebut akan disimpan sebagai id token.

d) Segment embedding

Segment embedding digunakan untuk mengetahui urutan kalimat. Kalimat pertama akan diberi angka 0, sedangkan *padding* akan diberikan angka 1.

e) Position embedding

Position embedding digunakan untuk mengetahui posisi kata dalam sebuah kalimat.

2) Fine tuning

Pada langkah ini, model BERT akan dilakukan *fine tuning* dengan menggunakan *hyperparameter* pada Tabel 3.12.

Tabel 3. 12 Hyperparameter BERT

Hyperparameter	Value
Loss function	Categorical-crossentropy
Learning rate	5e-5
Optimizer	Adam
Number of epochs	10
Batch Size	8

6. Evaluasi

Evaluasi merupakan cara untuk mengetahui kualitas dari proses yang telah dilakukan. Metode evaluasi yang digunakan pada project akhir ini yaitu *confusion matrix*. *Confusion matrix* merupakan alat yang digunakan untuk melakukan sebuah analisis terhadap seberapa baik klasifikasi yang telah dihasilkan dan dapat mengenali tuple dari kelas yang

berbeda. Berikut merupakan Tabel 3.13 yang mengilustrasikan *Confusion Matrix* [16].

Tabel 3. 13 Ilustrasi *Confusion Matrix*

		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FN
	Negatif	FP	TN

Pada *confusion matrix* memiliki beberapa evaluasi yang sering digunakan antara lain *Accuracy*, *Precision*, *Recall*, dan *F1 Score*. Pada project ini akan menggunakan dua evaluasi yaitu *Accuracy* dan *F1 Score*.

7. Deployment

Pada proses deployment ini kita melakukan implementasi model AI pada aplikasi berbasis website. Dikarenakan model AI yang kita buat menggunakan bahasa pemrograman python, maka kita menggunakan *framework* untuk aplikasi website yang berbasis bahasa pemrograman python. Disini kita menggunakan *framework* Flask yang dimana *framework* tersebut sangat efisien dan juga membantu para developer ataupun para pengembang website yang ingin mengembangkan websitenya dengan sistem pemrograman python.

a. Flask

Flask adalah sebuah *framework* python untuk menghandle atau mendevelop sebuah pemrograman bahasa python, namun ia sebuah *microframework*. Dimana *framework* ini membentuk sebuah kerangka kerja dalam suatu *website* seperti tampilan hingga *routing* sebuah *website*. Pada *website* yang kita kembangkan Flask berfungsi sebagai pembuat kerangka dan juga sebagai pembaca dari suatu model yang sudah kita buat. Dengan menggunakan flask, model berformat “.h5”

dapat terbaca, yang dimana format tersebut adalah sebuah ekstensi dari hasil pembuatan model package TensorFlow.

b. SQLite

SQLite merupakan sebuah *software* atau sistem manajemen relational database yang bersifat *open-source*, *serverless*, dan *portabel* yang digunakan untuk mempermudah akses dan mengelola penyimpanan data. Pengguna dapat dengan mudah menggunakan SQLite tanpa perlu melakukan instalasi *environment* atau konfigurasi apapun karena SQLite merupakan stand-alone *software* yang dibuat untuk disematkan ke dalam sebuah aplikasi. SQLite juga tidak menggunakan model arsitektur *client-server* seperti MySQL dan PostgreSQL karena seluruh programnya berisi *library* Bahasa pemrograman C yang menyematkan database ke dalam sebuah aplikasi. Pada *website* yang kita kembangkan, SQLite ini berfungsi sebagai database *History* percobaan teks yang dilakukan pada *website*. Dengan begitu, kita dapat melihat beberapa *History* percobaan dari seorang user.

c. Bootstrap

Bootstrap adalah framework HTML, CSS dan JavaScript yang berguna untuk menyederhanakan pengembangan halaman web. Pada umumnya, Bootstrap itu sendiri digunakan untuk mengimplementasikan berbagai pilihan warna, ukuran, *font*, dan *layout* yang ada dalam *framework* tersebut kedalam sebuah *website*. Sebagai sebuah *framework*, Bootstrap menyediakan *template* untuk mendefinisikan *style* dasar seluruh HTML. Hal ini mempermudah dalam pembuatan website tanpa harus mendefinisikan *style attribute* untuk setiap elemen HTML secara berulang. Bootstrap sangat berpengaruh dalam pembuatan tampilan *website* yang kita kembangkan. Dengan menggunakan Bootstrap, *slicing* tampilan jauh lebih efisien dan tidak memakan waktu banyak dalam pembuatannya.

Dalam tahap deployment, kita bertujuan untuk memasang pada cloud Heroku. Namun adanya kendala dalam *size storage cloud* yang terbatas dan juga salah satu file kami yang ukurannya menyentuh 422mb, membuat kami tidak dapat melakukan deployment pada *cloud* yang tersedia. Dengan begitu, tahap deployment dapat dikatakan hanya dapat berjalan pada *localhost*.

8. Hambatan

Beberapa hambatan yang dialami selama pembuatan proyek akhir diantaranya adalah:

- a. Terdapat beberapa kendala saat pencarian dataset diantaranya dataset kurang sesuai dengan yang dibutuhkan, dimana pada kasus ini dataset tidak merepresentasikan isi berita.
- b. Pemodelan menggunakan LSTM belum mendapatkan akurasi yang baik.
- c. Saat melakukan pelatihan model menggunakan pendekatan *deep learning* mengalami kendala pada proses komputasi karena sumber daya yang dibutuhkan cukup besar.
- d. Deployment website belum dapat dilakukan pada cloud dikarenakan ukuran model yang dibuat cukup besar hingga menyentuh 422 mb.

III.3 Hasil Proyek Akhir

Pada project ini model AI yang telah dibuat akan diimplementasikan dalam bentuk aplikasi berbasis website dengan nama Sotaken. Sotaken merupakan aplikasi yang dapat membantu masyarakat dalam mengidentifikasi berita *hoax*. Berikut merupakan hasil proyek akhir yang meliputi model AI yang telah dikembangkan dan hasil implementasi website.

1. Evaluasi dan Analisis

Eksperimen dilakukan dengan membandingkan kinerja nilai *F1 Score* menggunakan metode LSTM, IndoBERT dan Indolem. Pada model LSTM

dengan *word embedding word2vec* akan dilakukan dua kali percobaan menggunakan 2 dataset dengan jumlah berbeda. Sedangkan, pada model menggunakan *pre-trained* model BERT akan membandingkan 2 dataset dan membandingkan hasil akurasi dari IndoBERT dan Indolem.

Performa dari kedua model AI yang telah dirancang dapat dilihat pada Tabel 3.14 dan Tabel 3.15 sebagai berikut.

Tabel 3. 14 Performa Model LSTM

Algoritma	Data	Word Embedding	Akurasi (%)
LSTM	956	<i>Skip-gram</i>	55
LSTM	2,216	<i>Skip-gram</i>	49

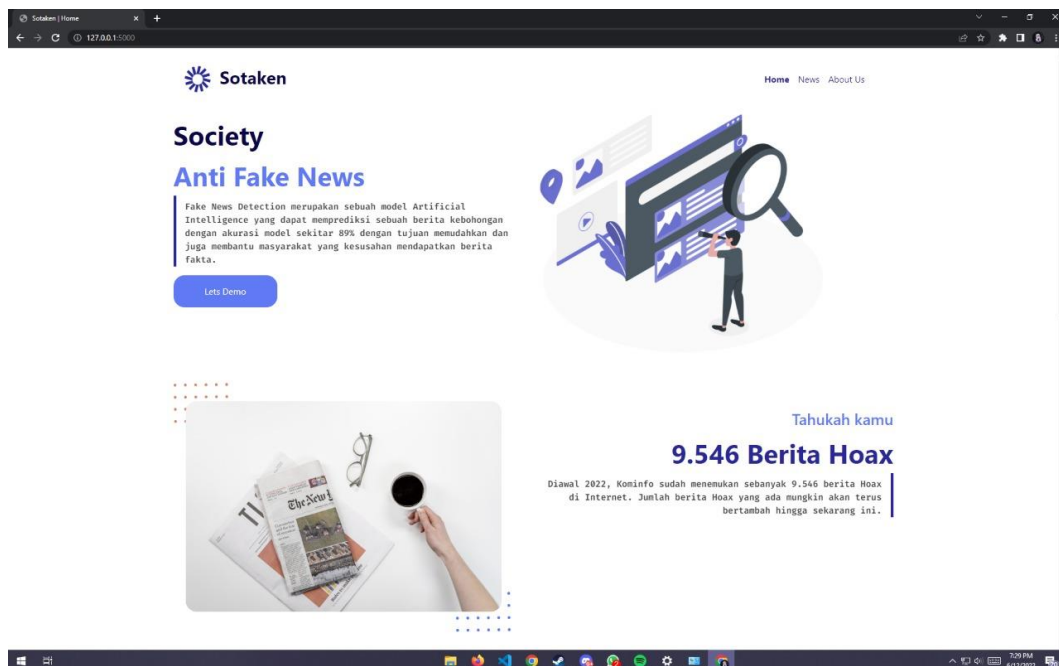
Tabel 3. 15 Performa Model BERT Berdasarkan pre-trained model dan Jumlah Dataset

Algoritma	Data	Pretrained Model	Akurasi (%)
BERT	956	Indolem	89
BERT	2,216	Indolem	86
BERT	956	IndoBERT	84
BERT	2,216	IndoBERT	88

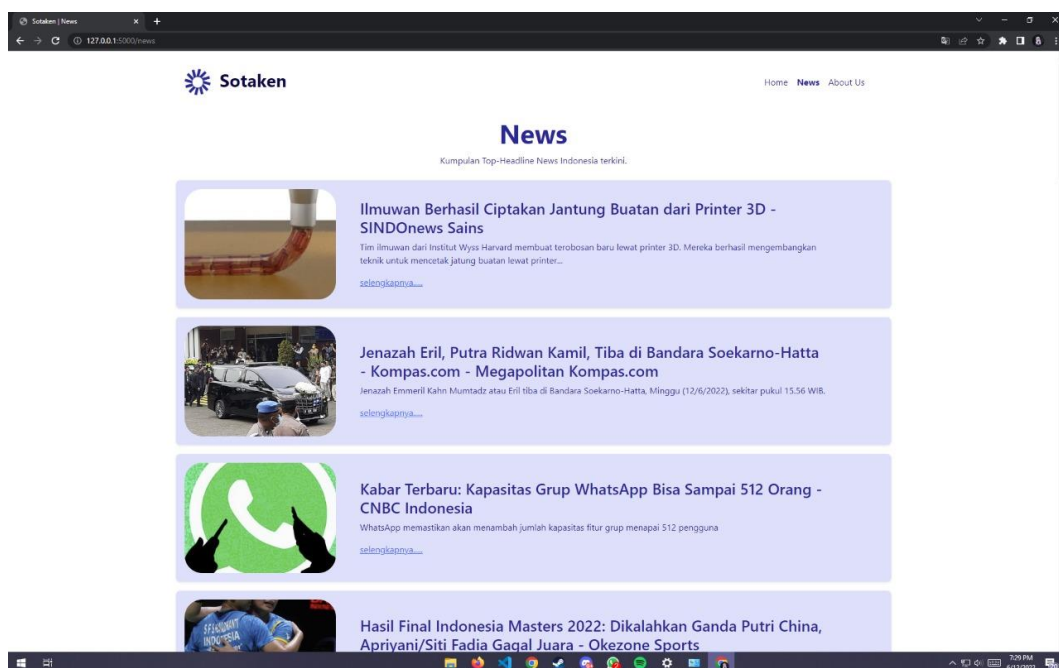
Berdasarkan hasil eksperimen yang telah dilakukan, hasil kinerja dari metode LSTM dan *word embedding word2vec* memiliki akurasi tertinggi yaitu 55% dengan menggunakan dataset yang berjumlah 956 data. Sedangkan pada metode menggunakan *pretrained* model BERT, akurasi tertinggi diperoleh menggunakan *pretrained* model Indolem dengan 956 data. Sehingga pada kelanjutan pengerjaan project akhir ini menggunakan model BERT.

2. Implementasi Sistem

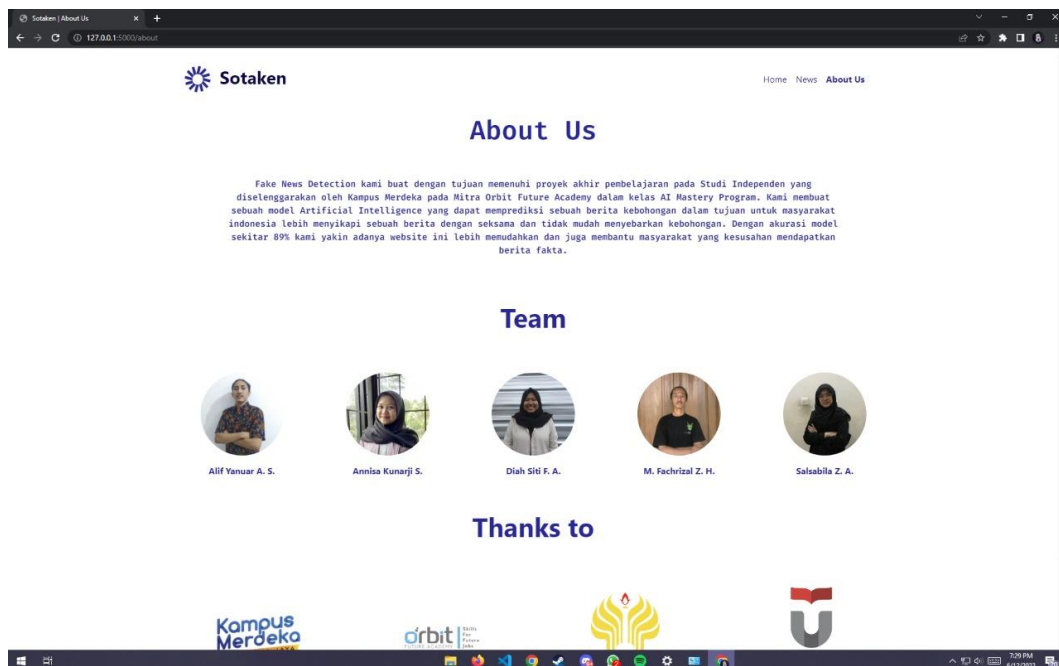
Setiap aplikasi tentu akan memiliki tampilan yang dimana akan mempermudah *user* dalam menggunakan aplikasi tersebut. *User Interface* pada aplikasi Sotaken dapat dilihat sebagai berikut.



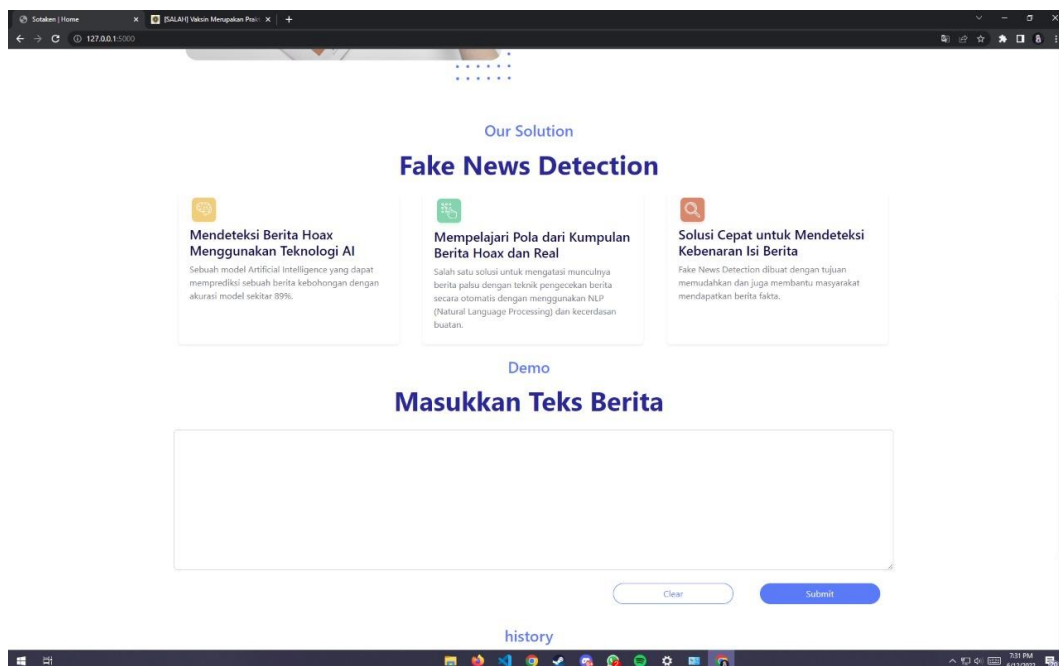
Gambar 3. 8 Tampilan Landing Page



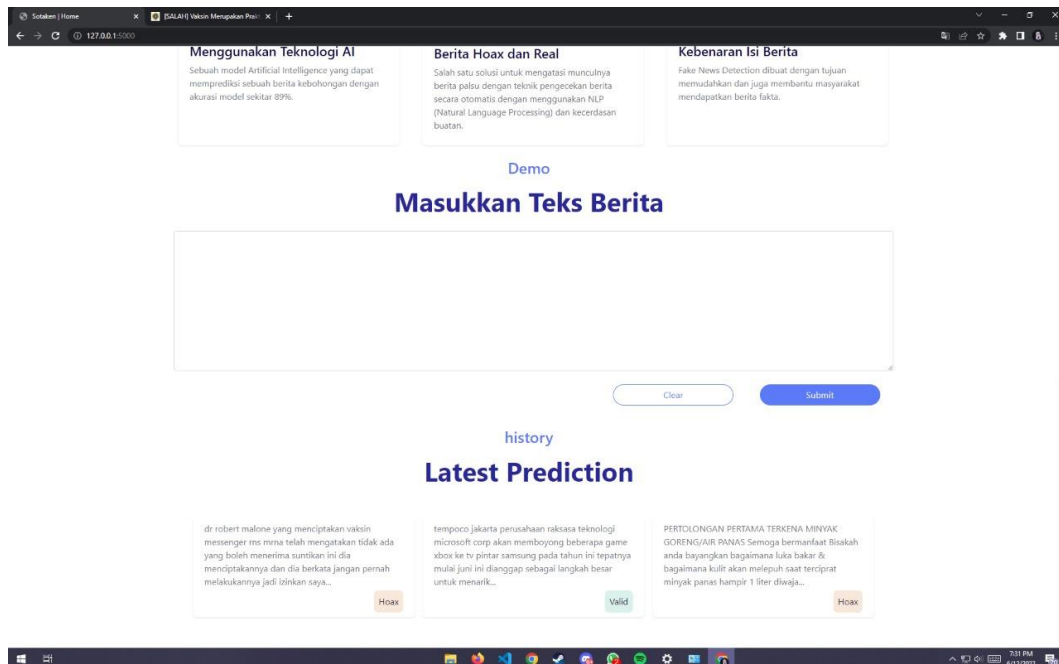
Gambar 3. 9 Tampilan Kumpulan Top-Headline News



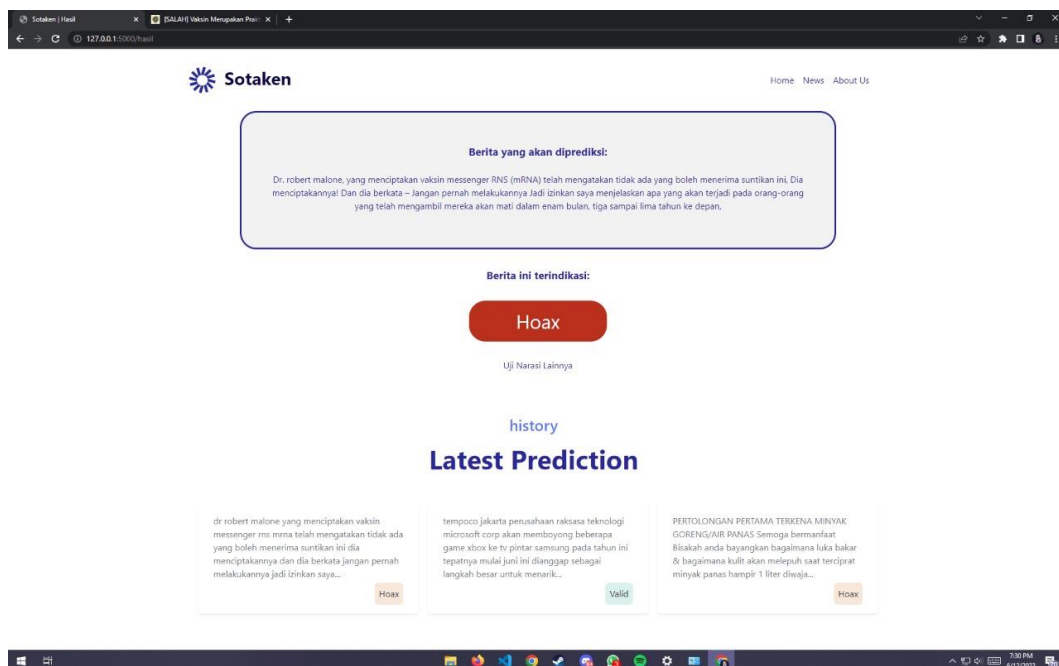
Gambar 3. 10 Tampilan About Us



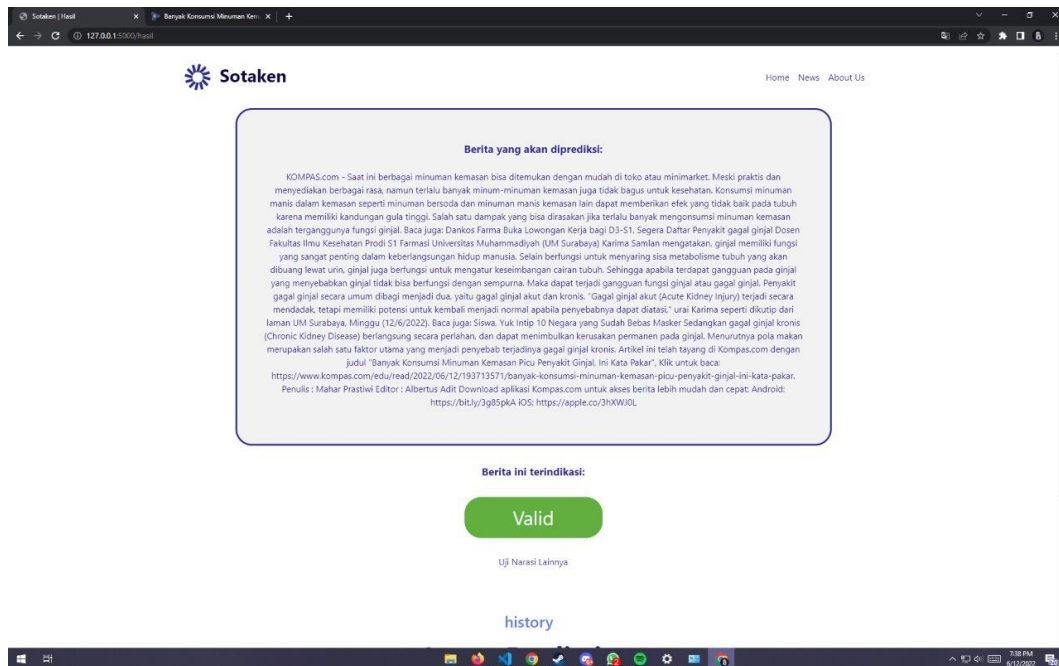
Gambar 3. 11 Tampilan Demo Aplikasi



Gambar 3. 12 Tampilan History Latest Prediction



Gambar 3. 13 Tampilan Contoh Prediksi Berita 1 (Hoax)



Gambar 3. 14 Tampilan Contoh Prediksi Berita 2 (Valid)

Setiap aplikasi tentu masih terdapat kelebihan dan kekurangan didalamnya. Kelebihan dan kelemahan pada aplikasi Sotaken dapat dilihat pada Tabel 3.16.

Tabel 3. 16 Kelebihan dan Kelemahan Aplikasi Sotaken

Kelebihan	Kekurangan
Mendeteksi berita <i>hoax</i> / fakta dengan input kalimat berita	Saat ini, hanya dapat diakses di <i>local</i>
User Interface pada aplikasi ini mudah digunakan dan dimengerti	Input kalimat berita pada aplikasi hanya 512 saja ini tentunya membuat tidak maksimal untuk mendeteksi berita.
Berbasis website yang mudah untuk diakses	Belum dapat dipasang pada cloud dan masih dalam tahap <i>localhost</i>
Model dapat melakukan tokenisasi yang lebih spesifik pada setiap katanya	Membutuhkan encoding dari bantuan GPU untuk melakukan tokenisasi yang spesifik jika dilakukan pada <i>localhost</i>

Saat ini project akhir yang output hasilnya berupa Aplikasi *Sotaken* berbasis website hanya bisa diakses di *local*. Pengembangan aplikasi di masa depan diharapkan dapat di *deployment* ke publik supaya *user* dengan lebih mudah untuk menggunakannya.

Bab IV Penutup

IV.1 Kesimpulan

Berdasarkan kegiatan studi independen yang diikuti oleh penulis di Orbit Future Academy, dapat disimpulkan sebagai berikut.

- Penulis mempelajari mengenai Artificial Intelligence secara mendalam dan juga mempelajari berbagai algoritma *Machine Learning* maupun *Deep Learning*.
- Penulis mendapatkan pengalaman, pengetahuan, serta *challenge* pada pengerjaan project akhir ini di dalam tim.
- Project akhir ini mengenai klasifikasi text berita *hoax*. Algoritma yang digunakan untuk proses pengerjaan yaitu LSTM dan BERT.
- Model dengan akurasi terbaik diperoleh menggunakan *pre-trained* model BERT yaitu Indolem dengan 956 data. Akurasi yang dihasilkan adalah 89%.
- Model yang telah dibuat diimplementasikan dalam bentuk website menggunakan framework Flask, database SQLite, dan Bootstrap.
- Project akhir ini menghasilkan *output* sebuah aplikasi berbasis website yang diberi nama Sotaken (Society Anti Fake News).

IV.2 Saran

Dari hasil selama penulis mengikuti kegiatan studi independen kampus merdeka, penulis memberikan saran supaya studi independen dapat dilaksanakan dengan lancar dan baik untuk kedepannya, penulis memiliki harapan diantaranya sebagai berikut.

- Program studi independen memberikan pengetahuan yang sangat bagus dan bermanfaat tentunya untuk dimasa yang akan datang.
- Hubungan sesama tim selalu terjaga keharmonisannya supaya dapat terciptanya suasana kerjasama yang baik di dalam pengerjaan project akhir.
- Project Akhir penulis yang berjudul “**Sistem Deteksi Berita Hoax Menggunakan Pendekatan *Bidirectional Encoder Representations from***

Transformer (BERT)” tentunya masih jauh dari kata sempurna. Tetapi, kami berusaha memberikan yang terbaik. Tidak hanya itu, laporan yang kami buat juga belum sempurna. Maka dari itu, kritik dan saran dari para pembaca yang bersifat membangun supaya kami dapat lebih baik lagi kedepannya dalam pengerjaan project akhir maupun laporan.

Bab V Referensi

- [1] F. N. Rozi and D. H. Sulistyawati, “Klasifikasi Berita Hoax Pilpres Menggunakan Metode Modified K-Nearest Neighbor Dan Pembobotan Menggunakan TF-IDF,” *Konvergensi*, vol. 15, no. 1, pp. 1–10, 2019, doi: 10.30996/konv.v15i1.2828.
- [2] R. K. Putri and M. Athoillah, “Identifikasi Berita Hoax Terkait Virus Corona Menggunakan Long Short-Term Memory,” *Semin. Nas. Has. Ris. dan Pengabd.*, pp. 506–513, 2022, [Online]. Available: <https://snhrp.unipasby.ac.id/prosiding/index.php/snhrp/article/view/354/298>.
- [3] R. Yunanto, A. P. Purfini, and A. Prabuwisesa, “Survei Literatur: Deteksi Berita Palsu Menggunakan Pendekatan Deep Learning,” *J. Manaj. Inform.*, vol. 11, no. 2, pp. 118–130, 2021, doi: 10.34010/jamika.v11i2.5362.
- [4] D. M. Almas Zakirah, “Pengaruh Hoax di Media Sosial Terhadap Preferensi Sosial Politik Remaja di Surabaya,” *J. Mediakita*, vol. 4, no. 1, 2020, doi: 10.30762/mediakita.v4i1.2446.
- [5] E. I. Setiawan and I. Lestari, “Stance Classification Pada Berita Berbahasa Indonesia Berbasis Bidirectional LSTM,” *J. Intell. Syst. Comput.*, vol. 3, no. 1, pp. 41–48, 2021, doi: 10.52985/insyst.v3i1.148.
- [6] R. Wati, “Penerapan Algoritma Naive Bayes Dan Particle Swarm Optimization Untuk Klasifikasi Berita Hoax Pada Media Sosial,” *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 5, no. 2, pp. 9–14, 2020, doi: 10.33480/jitk.v5i2.1034.
- [7] E. Y. Hidayat, R. W. Hardiansyah, and A. Affandy, “Analisis Sentimen Twitter untuk Menilai Opini Terhadap Perusahaan Publik Menggunakan Algoritma Deep Neural Network,” *J. Nas. Teknol. dan Sist. Inf.*, vol. 7, no. 2, pp. 108–118, 2021, doi: 10.25077/teknosi.v7i2.2021.108-118.
- [8] C. K. N. Paputungan and A. Jacobus, “Sentiment Analysis of Social Media Users Using Long-Short Term Memory Method Analisis Sentimen

- Pengguna Sosial Media Menggunakan Metode Long Short Term Memory,” *J. Tek. Elektro dan Komput.*, vol. 10, no. 2, pp. 99–106, 2021.
- [9] M. Wankhade, A. C. S. Rao, and C. Kulkarni, *A survey on sentiment analysis methods, applications, and challenges*. Springer Netherlands, 2022.
 - [10] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
 - [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” 2019.
 - [12] A. Vaswani *et al.*, “Attention is all you need,” *31st Conf. Neural Inf. Process. Syst. (NIPS 2017)*, Long Beach, CA, USA, pp. 1–15, 2017.
 - [13] B. Wilie *et al.*, “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” 2020, [Online]. Available: <http://arxiv.org/abs/2009.05387>.
 - [14] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” *Proc. of the 28th Int. Conf. Comput. Linguist.*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.
 - [15] H. K. Putra, M. Arif Bijaksana, and A. Romadhony, “Deteksi Penggunaan Kalimat Abusive Pada Teks Bahasa Indonesia Menggunakan Metode IndoBERT,” *e-Proceeding Eng.*, vol. Vol.8, No., no. 2, pp. 3028–3038, 2021.
 - [16] N. S. N. Salam, A. A. Supianto, and A. R. Perdanakusuma, “Analisis Sentimen Opini Mahasiswa Terhadap Saran Kuesioner Penilaian Kinerja Dosen dengan Menggunakan TF-IDF dan K-Nearest Neighbor,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 6, pp. 6148–6156, 2019, [Online]. Available: <http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/5649>.

TERM OF REFERENCE (TOR)
STUDI INDEPENDEN BERSERTIFIKAT
AI MASTERY PROGRAM
DI ORBIT FUTURE ACADEMY

A. Rincian Program

AI Mastery Program adalah program pelatihan *Artificial Intelligence* (AI) daring yang bertujuan untuk memperkenalkan teknologi dan perangkat AI kepada pelajar, sehingga diharapkan mereka dapat mengembangkan produk AI yang memiliki dampak sosial. Program ini berfokus pada komponen utama AI, seperti Data Science (DS), Natural Language Processing (NLP), Computer Vision (CV), dan Reinforcement Learning (RL).

B. Tujuan Program

Tujuan yang diharapkan setelah peserta mengikuti program ini:

1. Mampu memahami apa itu AI, penerapan dan pemanfaatannya.
2. Mampu memahami terkait tiga domain utama AI (DS, NLP, dan CV).
3. Mampu mengelaborasi kemampuan terkait AI dengan bidang lain.
4. Mampu memahami pentingnya data dalam AI.
5. Mampu membuat project AI yang berdampak sosial.
6. Mampu menulis kode dengan bahasa pemrograman Python.
7. Mampu memahami operasi dan logika sederhana pada Python.
8. Mampu membuat *project* Python.
9. Mampu melakukan kolaborasi secara interaktif dengan Git/Github.
10. Mampu membuat *repository* di akun Git/Github.
11. Mampu membuat portfolio dengan Git/Github.
12. Mampu menganalisis algoritma Machine Learning (ML) yang paling sesuai.
13. Mampu membuat model ML.

14. Mampu memahami dan menerapkan algoritma ML untuk membantu kehidupan.
15. Mampu membuat model Deep Learning (DL).
16. Mampu membuat kode program untuk pengujian model data science.
17. Mampu melakukan pengujian model dan analisis.
18. Mampu membuat ramalan dan prediksi berdasarkan data.
19. Mampu mengolah data yang besar untuk membuat keputusan.
20. Mampu men-clustering untuk memetakan pola.
21. Mampu membuat dokumentasi hasil pengujian model DS.
22. Mampu Membuat model DS dengan ML & DL.
23. Memahami NLP.
24. Mampu membuat model pengenalan suara.
25. Mampu membuat chatbot.
26. Mampu membuat project terkait dengan RL.
27. Mampu mengkombinasikan dan membuat project terkait AI, IoT, dan sensor.
28. Mampu mengaplikasikan konsep RL dan diterapkan bersama domain AI lain.
29. Mampu memahami dan membuat project terkait CV.
30. Mampu menggunakan teknologi terkait Computer Vision.
31. Mampu mengembangkan project CV untuk kepentingan sosial.
32. Mampu membuat model ML dan DL untuk berbagai kasus.
33. Mampu men-deploy model menggunakan Heroku dan atau menggunakan layan Machine learning as a service (MLaaS).

C. Jadwal dan Tempat Pelaksanaan

Jadwal pelaksanaan tertera dalam tabel berikut:

Pukul (WIB)	Durasi (jam)	Aktivitas
08.00 s.d. 11.30	3.5	Kelas Sesi Pagi
13.00 s.d. 16.30	3.5	Kelas Sesi Siang
16.30 s.d. 17.30	1	<i>Self-Study</i>

Kelas akan diselenggarakan secara daring melalui aplikasi *video conference*.

D. Peserta

Peserta program ini adalah mahasiswa yang berasal dari Perguruan Tinggi di bawah Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia.

E. Uraian Tugas Peserta

Selama mengikuti program ini, peserta diharuskan:

1. Mengikuti program dari awal hingga selesai.
2. Mematuhi aturan program.
3. Mematuhi aturan kelas yang dibuat bersama *homeroom* atau *domain coach*.
4. Mengikuti kelas dengan presensi minimal 85%.
5. Membuat laporan harian dan mingguan di *website* Kampus Merdeka.
6. Menyelesaikan Proyek Akhir (PA) beserta laporannya.

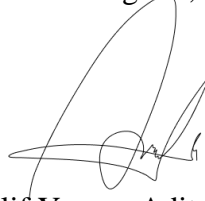
Homeroom Coach,



Kuncahyo Setyo Nugroho, S.Kom
NIP: 2201045

Bandung, 14 Juni 2022

Peserta Program,



Alif Yanuar Aditya Subagyo
NIM: 1202190187

Bab VII Lampiran B. Log Activity

Tabel 7.1 ini merupakan log activity yang dimulai dari pengerjaan Proyek Akhir (PA) sampai selesai.

Tabel 7. 1 Log Activity

Minggu/Tgl	Kegiatan	Hasil
Minggu ke – 1, 21 Feb – 25 Feb 22	Pembelajaran Secara Teori dan Praktik Menggunakan Bahasa Pemrograman Python	<ul style="list-style-type: none"> Memahami tentang logika dan konsep teknologi <i>artificial intelligence</i> Mengetahui dasar – dasar pemrograman bahasa Python Machine Learning dan Deep Learning Memahami implementasi AI pada kegidupan nyata Mengetahui fungsionalitas AI terhadap efisiensi kerja.
Minggu ke – 2, 28 Feb – 4 Mar 22		
Minggu ke – 3, 7 Mar – 11 Mar 22		
Minggu ke – 4, 14 Mar – 18 Mar 22		
Minggu ke – 5, 21 Mar – 25 Mar 22		Mengetahui lebih dalam mengenai domain – domain yang ada di artificial intelligence diantara lain Data Science, Computer Vision, Natural Language Processing dan Reinforcement Learning. Selain itu, juga memahami Manajemen Data, Git, dan Deployment yang mana bermanfaat untuk pengerjaan project.
Minggu ke – 6, 28 Mar – 1 Apr 22		
Minggu ke – 7, 4 Apr – 8 Apr 22	Mencari Permasalahan di Sekitar (<i>Problem Scoping</i>) dan Topik AI	Topik AI yang disukai yaitu NLP. Untuk permasalahannya

	yang Disukai Untuk Dipilih sebagai topik model AI yang akan digunakan.	berfokus pada lingkungan sekitar
Minggu ke – 8, 11 Apr – 15 Apr 22	Mencari Referensi Jurnal sebagai sumber pembelajaran untuk proyek akhir dan juga mempelajari beberapa materi LSTM dan juga pembelajaran unit testing sebagai bekal proyek akhir.	Memikirkan ide dan referensi jurnal untuk project akhir. Memahami beberapa <i>mini project</i> sebagai sumber referensi baru.
Minggu ke – 9, 18 Apr – 22 Apr 22	Fiksasi Pembuatan Kelompok, Membaca Referensi dan Ide Project Akhir	Pembuatan kelompok terdiri dari 5 mahasiswa yang berasal dari kelas berbeda diantaranya Atlas, Persevere dan Better. Lalu diskusi mengenai ide project. Hasilnya terdapat dua pilihan yaitu Data Science dan Natural Language Processing
Minggu ke – 10, 25 Apr – 29 Apr 22	Fiksasi Ide Project Akhir dan Pembagian Tugas Anggota Tim	Fiksasi ide project akhir diambil dari domain Natural Language Processing (NLP), dengan tema project akhir tentang berita hoax. Pembagian tugas anggota tim ada 3 diantaranya Modelling, UI/UX, dan Deployment.

Minggu ke – 11, 2 Mei – 6 Mei 22	Libur Nasional dan Cuti Bersama Hari Raya Idul Fitri	Tetap mencari dataset publik, metode, jurnal referensi terkait dengan project akhir.
Minggu ke – 12, 9 Mei – 13 Mei 22	<i>Data Acquisition, Reprocessing, Modelling</i>	<i>Dataset</i> yang digunakan pada <i>modelling</i> merupakan <i>dataset public</i> yang ada di github dan data Mendeley. Modelling menggunakan algoritma LSTM + Word2Vec.
Minggu ke – 13, 16 Mei – 20 Mei 22	Membantu proses <i>modelling</i> dan juga membuat wireframe sebagai rancangan awal website.	Modelling menggunakan algoritma LSTM + Word2Vec. Akan tetapi, hasil yang didapatkan belum cukup bagus. Hasil fiksasi UI untuk dilakukan <i>High Fidelity</i> oleh UI/UX Designer.
Minggu ke – 14, 23 Mei – 27 Mei 22	<i>Slicing</i> tampilan website berdasarkan UI yang sudah dibuat sebelumnya. Dan juga melakukan brainstorming kepada tim mengenai perubahan UI yang dilakukan.	Selesai membuat <i>slicing</i> terhadap website yang sedang dikembangkan.
Minggu ke – 15, 30 Mei – 3 Jun 22	Merapikan Coding dan Deployment, sekaligus menyambungkan website kepada model yang sudah dibuat.	Dapat menyambungkan model AI kepada website yang sedang dikembangkan.

Minggu ke – 16, 6 Jun – 10 Jun 22	Menyelesaikan Deployment, Mengerjakan Laporan Akhir	Progress dari project akhir ini sudah masuk ke tahap merapikan hasil deployment. Mulai mengerjakan laporan akhir.
Minggu ke – 17, 13 Jun – 17 Jun 22	Menyusun dan Menyelesaikan Laporan	Menyelesaikan laporan akhir dari bab 1 hingga bab 8. Selain itu, mengcompile kembali link – link googlecolab algoritma LSTM dan BERT untuk dimasukkan ke dalam laporan. Setelah itu, meminta ttd homeroom coach dan dosen pembimbing.

Bab VIII Lampiran C. Dokumen Teknik

1. AI Project Cycle

a. Problem Scoping

Seiring berkembangnya teknologi, kemudahan dalam mengakses berbagai hal melalui *internet* membawa dampak perubahan yang besar. Sebuah informasi dapat dengan mudah disebarluaskan hanya dengan hitungan detik. Tiap hari masyarakat mendapatkan informasi - informasi untuk mengetahui kabar terkini. Akan tetapi, adanya perkembangan teknologi ini ternyata juga mendapatkan dampak negatifnya yaitu muncul berita - berita palsu / *hoax* yang dilakukan oleh oknum - oknum tertentu. *Hoax* adalah suatu upaya untuk memanipulasi pembaca supaya berpengaruh pada opini yang dibawa. Banyak masyarakat yang menjadi purno dan panik ketika mendapatkan berita tersebut, sedangkan masyarakat belum mengetahui apakah berita tersebut termasuk fakta / *hoax*. Data dari laman website kominfo mengatakan bahwa terdapat 800.000 situs penyebar *hoax* dan *hate speech* di Indonesia. Selain itu, berita *hoax* memberikan dampak negatif yang besar bagi masyarakat. *Hoax* merupakan efek samping dari era keterbukaan yang memiliki peluang untuk menciptakan perpecahan dan permusuhan karena dapat membuat masyarakat bingung akan sebuah kebenaran dari suatu informasi (kominfo.go.id, 2021).

Pada permasalahan tersebut tentunya membutuhkan upaya - upaya untuk menghentikan penyebaran berita *hoax* yang meresahkan masyarakat. Selain itu, adanya berita *hoax* ini juga merugikan masyarakat. Permasalahan ini apabila tidak dilakukan upaya - upaya untuk menghilangkan berita - berita *hoax* yang beredar akan memberikan dampak negatif seperti keributan, keresahan, perselisihan, ujaran kebencian, kecemasan, dan lain - lain.

Dengan adanya klasifikasi berita *hoax* pemecahan masalah yang telah meresahkan masyarakat dapat terselesaikan. Masyarakat dapat membedakan berita yang *hoax* maupun tidak. Proyek yang kami beri nama Sotaken dapat membedakan berita *hoax* dan valid. Aplikasi ini berbasis website. Aplikasi yang dibuat ini menjadi salah satu solusi sebagai wadah masyarakat yang ingin

mengetahui atau memastikan berita yang didapatkannya itu tergolong berita fakta(valid)/hoax. Aplikasi yang diberi nama Sotaken ini diharapkan dapat membantu masyarakat dalam mendeteksi sebuah kebenaran berita yang tersebar di masyarakat.

b. Data Acquisition

Data yang akan digunakan merupakan *dataset* publik yang tersedia di internet, khususnya situs penyedia *dataset* seperti Kaggle dan Github. Tabel 8.1 merupakan daftar *dataset* yang digunakan.

Tabel 8. 1 Daftar Dataset Digunakan

Pemilik	Tautan Unduhan	Jumlah Data
Pierobeat	https://github.com/pierobeat/Hoax-News-Classification	250 berita <i>hoax</i> 250 berita valid
Jibrán Fawaid	https://github.com/JibránFawaid/turnbackhoax-dataset	683 berita <i>hoax</i> 433 berita valid
Ruhtomo, dkk	https://data.mendeley.com/datasets/p3hfgr5j3m/1	228 berita <i>hoax</i> 372 berita valid

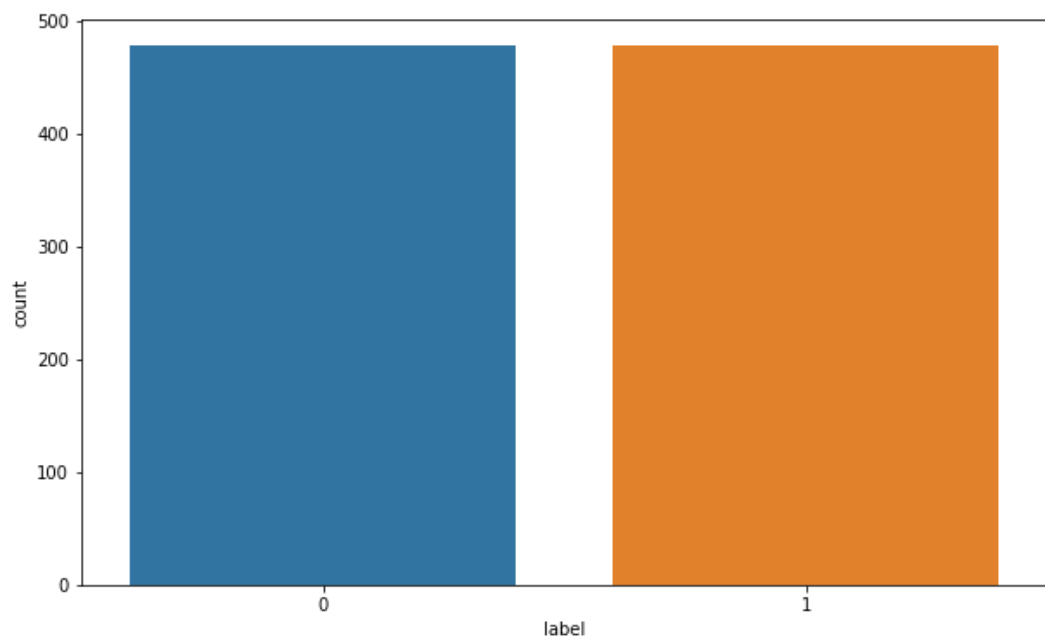
Jumlah dataset yang akan digunakan sebanyak 956 data dan 2,216 data. Dataset 956 data merupakan dataset hasil penggabungan dari dataset milik Ruhtomo dkk dan Pierobeat. Dataset tersebut memiliki jumlah kelas yang seimbang. Sedangkan dataset dengan jumlah 2,216 data, merupakan hasil penggabungan 956 data sebelumnya dengan dataset milik Jibrán Fawaid. Pada dataset tersebut data yang memiliki label *hoax* sebanyak 1,161 data dan label *valid* sebanyak 1,055 data.

Dataset yang memiliki isi label yang berbeda akan diubah dan disamakan isinya agar memudahkan proses penggabungan data. Label yang akan

digunakan yaitu 1 (satu) untuk berita *hoax* dan 0 (nol) untuk berita valid atau fakta.

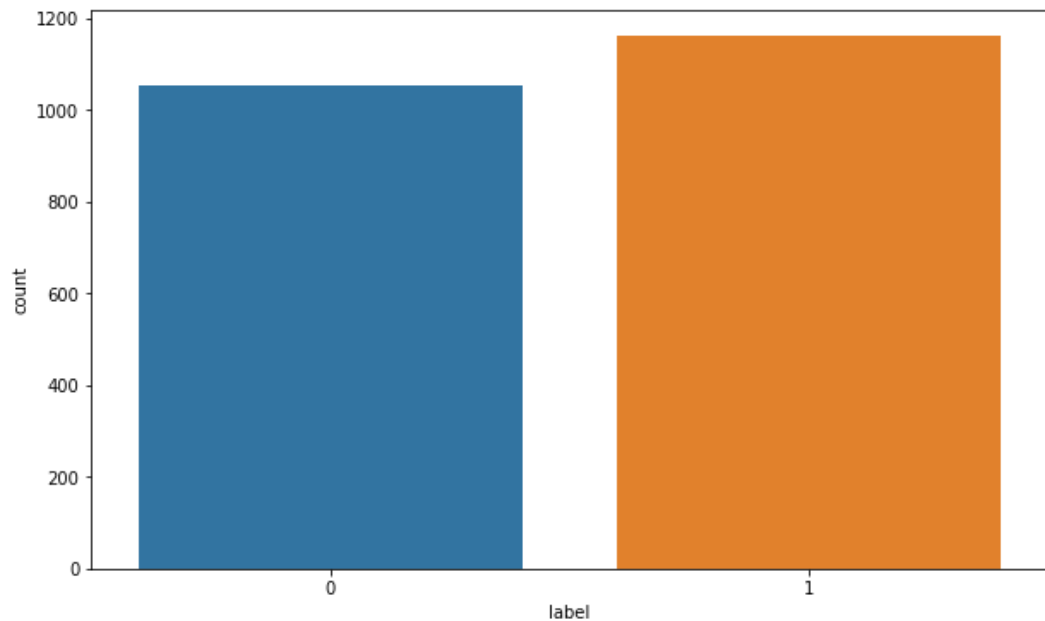
c. Data Exploration

Terdapat 2 dataset yang akan digunakan untuk tahap pemodelan, dataset pertama berjumlah 956 data dan dataset kedua berjumlah 2,216 data. Gambar 8.1 merupakan grafik dari dataset yang berjumlah 956 data dan Gambar 8.2 merupakan grafik dari dataset yang berjumlah 2,216 data.



Gambar 8. 1 Grafik Dataset 956 Data

Pada grafik tersebut terlihat bahwa data dengan label *hoax* berjumlah 478 dan data dengan label fakta berjumlah 478. Terlihat bahwa data *hoax* dan fakta seimbang.



Gambar 8. 2 Grafik Dataset 2.216 Data

Dataset kedua terlihat pada Gambar 8.2 bahwa dataset yang memiliki 2,216 data memiliki kelas yang tidak seimbang.

d. Modelling

Algoritma yang digunakan pada project ini diantaranya LSTM dan *pretrained model* BERT yaitu IndoBERT dan Indolem. Pada kedua algoritma tersebut dilakukan perbandingan untuk memilih hasil akurasi yang terbaik dan tertinggi. Kedua algoritma tersebut dapat digunakan untuk mengklasifikasi teks.

e. Evaluation

Performa dari kedua model AI yang telah dirancang dapat dilihat pada Tabel 8.2 dan Tabel 8.3 sebagai berikut.

Tabel 8. 2 Performa Model LSTM Berdasarkan Dataset

Algoritma	Data	Word Embedding	Akurasi (%)
LSTM	956	<i>Skip-gram</i>	55
LSTM	2,216	<i>Skip-gram</i>	49

Tabel 8. 3 Performa Model BERT Berdasarkan Pre-trained model dan Dataset

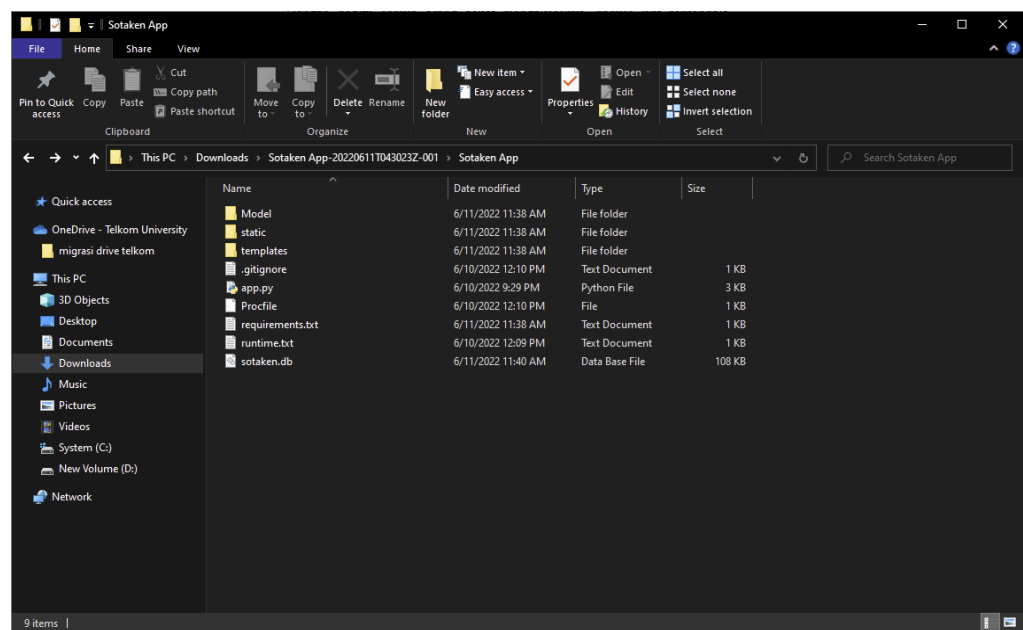
Algoritma	Data	Pretrained Model	Akurasi (%)
BERT	956	Indolem	89
BERT	2,216	Indolem	86
BERT	956	IndoBERT	84
BERT	2,216	IndoBERT	88

f. Deployment

Tujuan dari deployment yaitu untuk menyebarkan aplikasi yang telah dikerjakan. Pada dasarnya kita bertujuan untuk melakukan *deployment* pada *cloud*. Dengan begitu semua orang dapat mengaksesnya, namun kita terkendala dalam melakukan hal *push to cloud*. Diakibatkan salah satu file yaitu model yang sudah dibuat mempunyai ukuran yang cukup besar. Hal tersebut membuat kita tidak bisa melakukan *deployment* pada sebuah *cloud*. Kita memutuskan untuk sementara ini dijalankan pada *localhost*. *Resource* mengenai Aplikasi ini dapat diakses pada <https://bit.ly/SotakenApp>.

Langkah - langkah untuk dapat menggunakan Aplikasi tersebut yaitu:

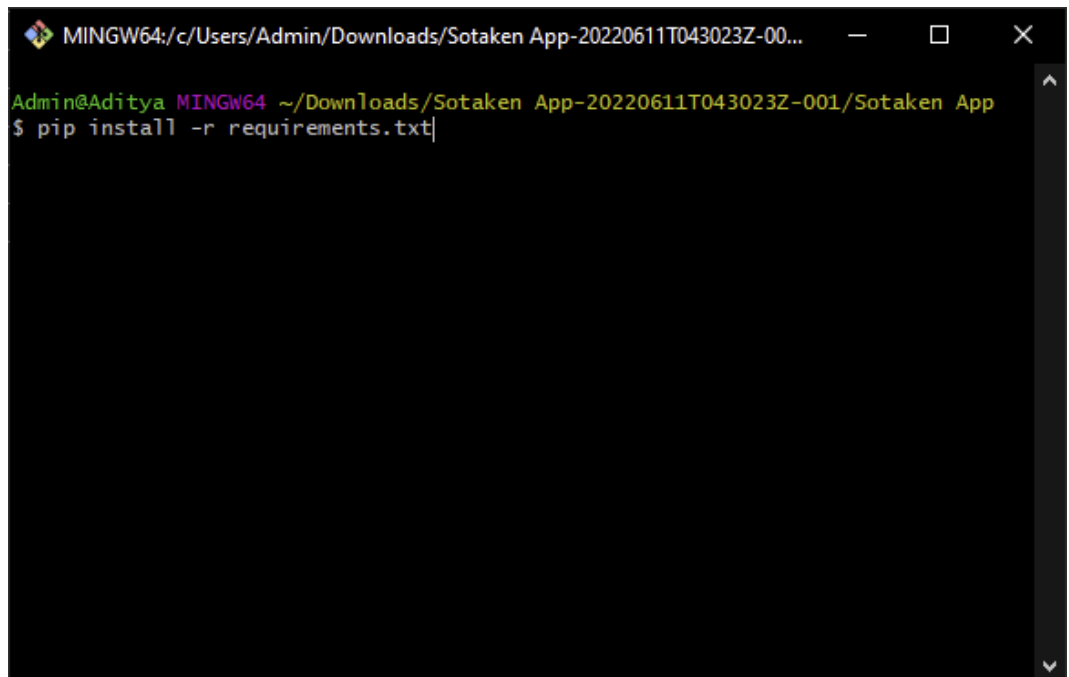
- Download lalu buka directory yang telah di extract.



Gambar 8. 3 Langkah Ke - 1

- Buka dan jalankan perintah untuk melakukan instalasi *requirement* seperti berikut:

```
pip install -r requirements.txt
```

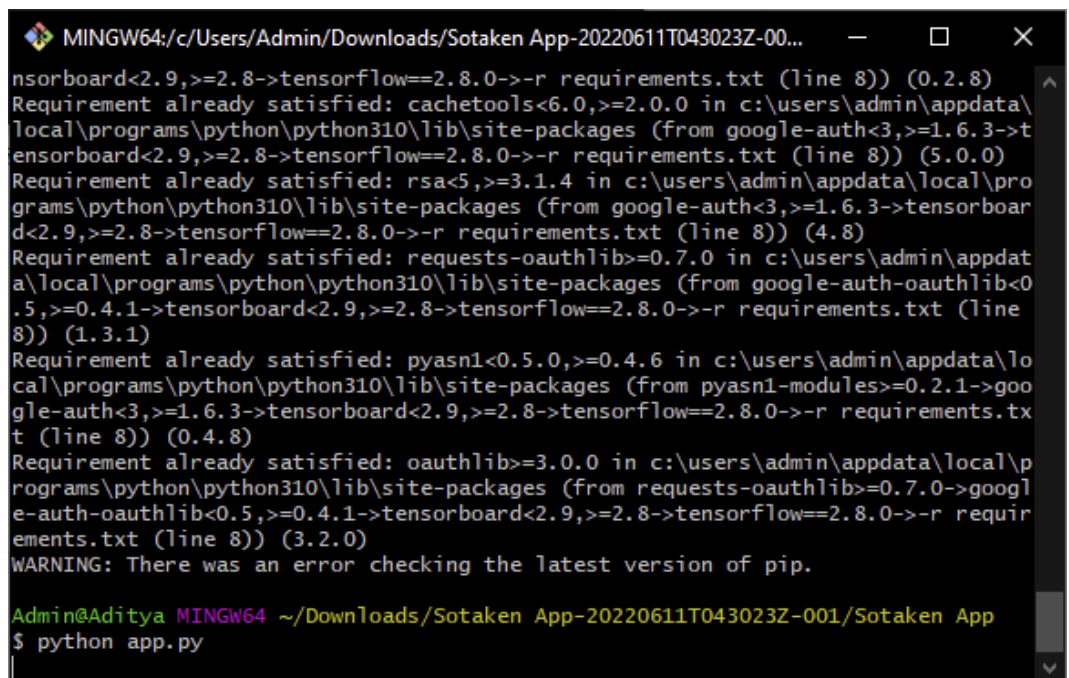


```
Admin@Aditya MINGW64 ~/Downloads/Sotaken App-20220611T043023Z-001/Sotaken App
$ pip install -r requirements.txt
```

Gambar 8. 4 Langkah Ke - 2

- Setelah melakukan *instalasi*, maka buka terminal pada directory tersebut lalu jalankan perintah:

```
python app.py
```



```
tensorflow==2.8.0->r requirements.txt (line 8)) (0.2.8)
Requirement already satisfied: cachetools<6.0,>=2.0.0 in c:\users\admin\appdata\local\programs\python\python310\lib\site-packages (from google-auth<3,>=1.6.3->tensorflow<2.9,>=2.8->tensorflow==2.8.0->r requirements.txt (line 8)) (5.0.0)
Requirement already satisfied: rsa<5,>=3.1.4 in c:\users\admin\appdata\local\programs\python\python310\lib\site-packages (from google-auth<3,>=1.6.3->tensorflow<2.9,>=2.8->tensorflow==2.8.0->r requirements.txt (line 8)) (4.8)
Requirement already satisfied: requests-oauthlib<0.7.0 in c:\users\admin\appdata\local\programs\python\python310\lib\site-packages (from google-auth-oauthlib<0.5,>=0.4.1->tensorflow<2.9,>=2.8->tensorflow==2.8.0->r requirements.txt (line 8)) (1.3.1)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in c:\users\admin\appdata\local\programs\python\python310\lib\site-packages (from pyasn1-modules<0.2.1->google-auth<3,>=1.6.3->tensorflow<2.9,>=2.8->tensorflow==2.8.0->r requirements.txt (line 8)) (0.4.8)
Requirement already satisfied: oauthlib<3.0.0 in c:\users\admin\appdata\local\programs\python\python310\lib\site-packages (from requests-oauthlib<0.7.0->google-auth-oauthlib<0.5,>=0.4.1->tensorflow<2.9,>=2.8->tensorflow==2.8.0->r requirements.txt (line 8)) (3.2.0)
WARNING: There was an error checking the latest version of pip.

Admin@Aditya MINGW64 ~/Downloads/Sotaken App-20220611T043023Z-001/Sotaken App
$ python app.py
```

Gambar 8. 5 Langkah Ke - 3

- Setelah terlihat bahwa sudah jalan pada localhost, maka buka ip sebagai berikut pada web browser kita:

`http://127.0.0.1:5000/`

```

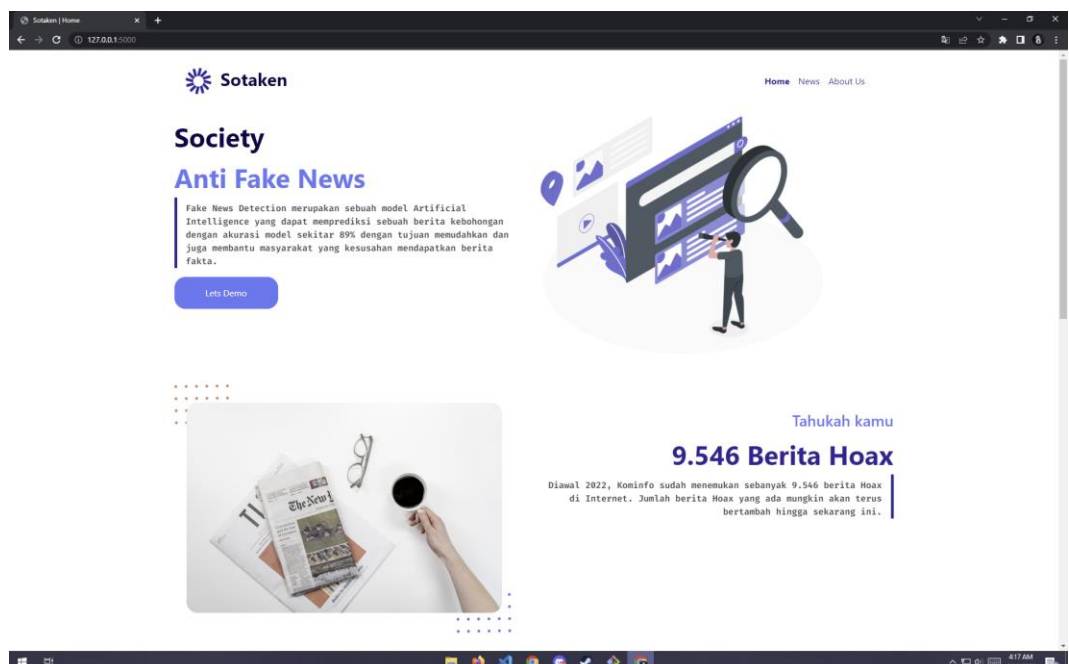
2022-06-12 03:58:41.379669: W tensorflow/core/common_runtime/gpu/gpu_device.cc:1850] Cannot dlopen some GPU libraries. Please make sure the missing libraries mentioned above are installed properly if you would like to use GPU. Follow the guide at https://www.tensorflow.org/install/gpu for how to download and setup the required libraries for your platform.
Skipping registering GPU devices...
2022-06-12 03:58:41.380037: I tensorflow/core/platform/cpu_feature_guard.cc:151] This TensorFlow binary is optimized with oneAPI Deep Neural Network Library (oneDNN) to use the following CPU instructions in performance-critical operations:
AVX AVX2
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
All PyTorch model weights were used when initializing TFBertForSequenceClassification.

Some weights or buffers of the TF 2.0 model TFBertForSequenceClassification were not initialized from the PyTorch model and are newly initialized: ['classifier.weight', 'classifier.bias']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
* Debugger is active!
* Debugger PIN: 445-643-043
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)

```

Gambar 8. 6 Langkah Ke - 4

- Maka dengan seperti itu kita sudah dapat mengakses Aplikasi Sotaken

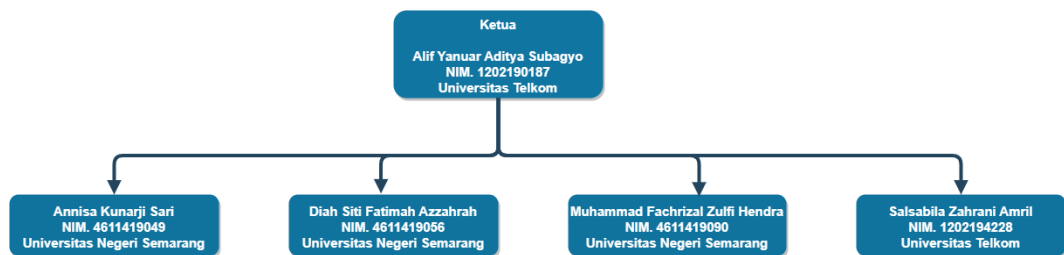


Gambar 8. 7 Langkah Ke - 5

Untuk video tutorial deployment localhost lebih lengkapnya dapat dilihat pada link berikut: <https://bit.ly/SotakenDeployment>

2. Profil Tim dan Deskripsi Pembagian Tugas

Struktur tim dapat dilihat pada Gambar 8.8



Gambar 8. 8 Struktur Tim

Deskripsi Pembagian Peran dan Tupoksi

Didalam tim tentunya terdapat pembagian peran dan tupoksi (tugas pokok dan fungsi) untuk memperlancar dan menyelesaikan project akhir dengan tepat waktu. Deskripsi tersebut dapat dilihat pada Tabel 8.4.

Tabel 8. 4 Deskripsi Pembagian Peran

Anggota Kelompok	Peran	Tupoksi (Tugas Pokok dan Fungsi)
Alif Yanuar Aditya Subagyo	Deployment	<ul style="list-style-type: none"> - Merancang pembuatan website berkaitan dengan html dan css dengan menggunakan framework Flask - Menyusun file - file berkaitan dengan deployment untuk dimasukkan ke dalam google drive - Mendeployment model AI pada website
Annisa Kunarji Sari	Modelling	<ul style="list-style-type: none"> - Mencari dataset publik yang jumlahnya > 500 - Mencoba mengganti algoritma LSTM dengan

		BERT - Melakukan modelling pada BERT
Diah Siti Fatimah Azzahrah	Modelling	- Melakukan modelling pada algoritma LSTM (Long Short Term Memory) - Membandingkan hasil modelling algoritma LSTM dan BERT. - Melakukan penggantian dataset untuk algoritma LSTM
Muhammad Fachrizal Zulfi Hendra	UI/UX	- Merancang pembuatan User Interface (UI) - Mencari color palette yang cocok untuk tampilan UI - Membuat flowchart alur website
Salsabila Zahrani Amril	Modelling	- Membantu mencari nilai akurasi yang bagus dengan mengganti parameter yang ada - Melakukan modelling pada LSTM - Melakukan modelling pada BERT

3. Deskripsi Aplikasi

a. Nama dan Fungsi Aplikasi

Nama aplikasi dari project yang kami buat yaitu **Sotaken (Society Anti Fake News)**. Fungsi atau kegunaan dari aplikasi yang kami buat untuk mengetahui apakah berita yang kita dapatkan itu termasuk berita hoax atau fakta. Masyarakat mendapatkan dampak negatif dari penyebaran berita hoax di lingkungan. Selain itu, penyebaran berita hoax itu terjadi secara masif dan cepat. Maka dari itu, pembuatan aplikasi ini merupakan upaya dalam memerangi berita hoax yang beredar. Tentunya dalam pembuatan aplikasi ini memiliki target user yaitu masyarakat.

Cara kerja aplikasi ini sebagai berikut.

1. Ketika *user* membuka website ini maka tampilan awal akan menuju di landing page.
2. Lalu klik button “Let’s demo”.
3. Tulis teks berita yang ingin *user* tau apakah masuk ke berita *hoax* / fakta. Kemudian klik button “Submit”.
4. Setelah itu, akan muncul tampilan hasil dari *input* teks berita. *Output* yang dikeluarkan dari website adalah hasil klasifikasi dari berita yang telah diinputkan oleh *user* (pengguna).
5. Apabila ingin mengulangi menginput teks berita, maka klik “Uji narasi lainnya” .
6. Selain itu, pada aplikasi website tersebut terdapat *history* dari pengecekan berita sebelumnya.

b. Jenis Aplikasi dan *Specific Requirement*

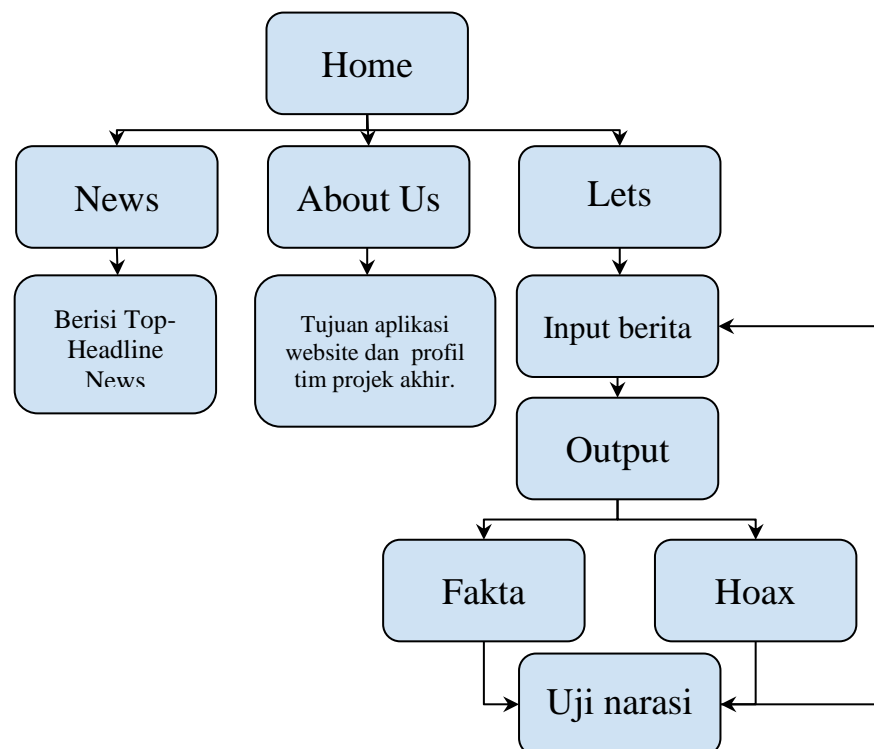
Aplikasi Sotaken (Society Anti Fake News) ini dirancang berbasis website. Pada dasarnya aplikasi yang berbasis website ini tidak begitu membutuhkan spesifikasi khusus, karena dia akan disimpan pada *cloud*. Namun, untuk saat ini kita hanya bisa hingga pada tahap *localhost*. Untuk menjalankan aplikasi ini kita membutuhkan *package* Transformers dalam proses tokenisasi nya. *Package* Transformers itu sendiri membutuhkan *package* Tensorflow dan PyTorch. *Package* Tensorflow dan PyTorch itu sendiri membutuhkan komputasi tingkat tinggi untuk mengelola matrik yang digunakan untuk pemodelan BERT. Komputasi tersebut melibatkan beberapa *requirements*:

- Python 3.10
 - *Library* Flask 2.1.1
 - *Library* nltk 3.7
 - *Library* ScikitLearn 1.1.1
 - *Library* Seaborn 0.11.2
 - *Library* Tensorflow 2.8.0
 - *Library* Transformers 4.19.2

- pip 21.0 atau terbaru
- Ubuntu 16.04 atau terbaru (64-bit)
- macOS 10.12.6 (Sierra) atau terbaru (64-bit)
- Windows 7 atau terbaru (64-bit) (hanya Python 3.10)
 - Microsoft Visual C++ Redistributable for Visual Studio 2015, 2017 and 2019
- Komputasi GPU seperti CUDA®-enabled card (untuk Ubuntu dan Windows)

c. User Interface

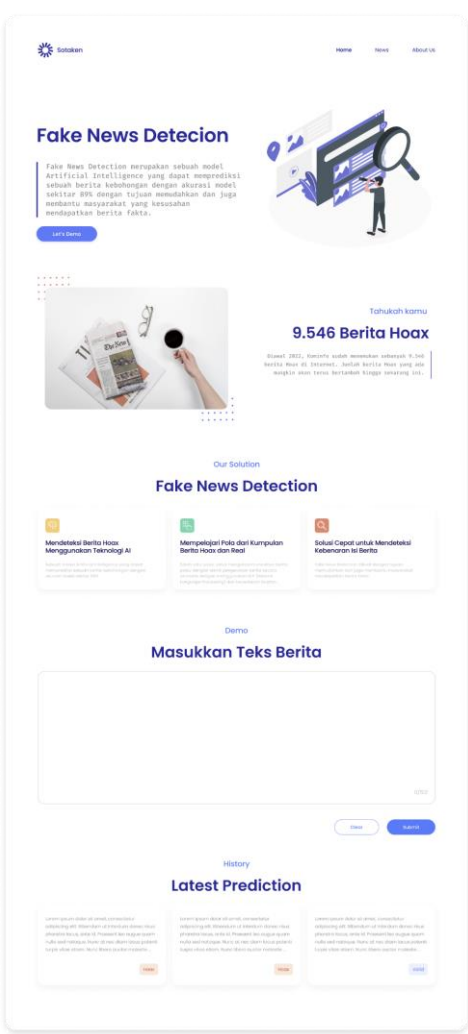
Flowchart merupakan diagram yang menampilkan langkah-langkah untuk melakukan sebuah proses dari suatu program. Setiap langkah digambarkan dalam bentuk diagram dan dihubungkan dengan garis atau arah panah. Flowchart dari aplikasi Sotaken dapat dilihat di Gambar 8.9 Flowchart dari Aplikasi Sotaken (Society Anti Fake News)

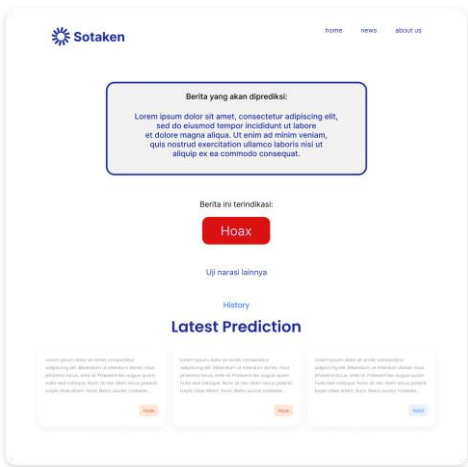
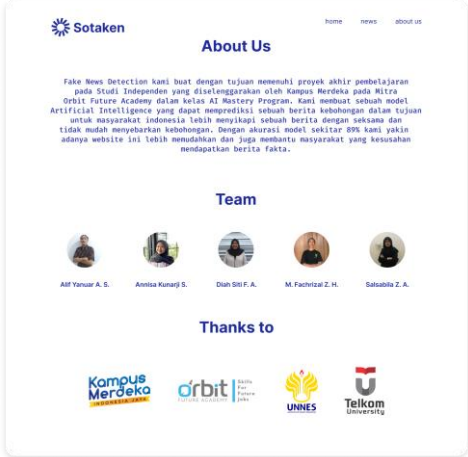



Gambar 8. 9 Flowchart Dari Aplikasi Sotaken (Society Anti Fake News)

User Interface (UI) merupakan tampilan visual dari sebuah produk yang mempertemukan sistem dengan pengguna (*user*). Berikut dapat dilihat pada Tabel 8.5 terkait User Interface dari aplikasi Sotaken (*Society Anti Fake News*) beserta penjelasan dari fitur - fitur yang ada di aplikasi Sotaken.

Tabel 8. 5 User Interface dan Penjelasan

Laman	Design UI
<p>Landing Page</p> <p>Pada tampilan awal ini, kita dapat langsung melakukan demo untuk mencoba <i>Artificial Intelligence</i> yang sudah dibuat dengan menggunakan pemodelan BERT. Ketika <i>user</i> klik “Lets Demo” maka halaman akan langsung otomatis scroll kepada tempat yang telah disediakan. Beberapa hasil percobaan sebelumnya juga kita tampilkan pada halaman ini.</p>	

<p>Output</p> <p>Setelah menginput kalimat berita yang ingin tahu apakah tergolong berita <i>hoax</i>/fakta (valid), maka tampilan <i>output</i> akan seperti pada gambar di samping.</p>	
<p>About Us</p> <p>Fitur ini menampilkan tujuan dari pembuatan aplikasi website, dan profil dari tim project akhir.</p>	
<p>News</p> <p>Tampilan ini berisi berita - berita yang Top-Headline News Indonesia</p>	

d. Keterangan Lainnya

Dalam pembuatan aplikasi Sotaken tentu masih jauh dari kata sempurna. Tabel 8.6 merupakan kelebihan dan kekurangan dari aplikasi website Sotaken.

Tabel 8. 6 Kelebihan dan Kekurangan Aplikasi Website Sotaken

Kelebihan	Kekurangan
Mendeteksi berita <i>hoax</i> / fakta dengan input kalimat berita	Saat ini, hanya dapat diakses di <i>local</i>
User Interface pada aplikasi ini mudah digunakan dan dimengerti	Input kalimat berita pada aplikasi hanya 512 saja ini tentunya membuat tidak maksimal untuk mendeteksi berita.
Berbasis website yang mudah untuk diakses	Belum dapat dipasang pada cloud dan masih dalam tahap <i>localhost</i>
Model dapat melakukan tokenisasi yang lebih spesifik pada setiap katanya	Membutuhkan encoding dari bantuan GPU untuk melakukan tokenisasi yang spesifik jika dilakukan pada <i>localhost</i>

Saat ini project akhir yang output hasilnya berupa Aplikasi Sotaken berbasis website hanya bisa diakses di *local*. Pengembangan aplikasi di masa depan diharapkan dapat di *deployment* ke publik supaya *user* dengan lebih mudah untuk menggunakannya.