TESIS

MODEL MULTILINGUAL NAMED ENTITY RECOGNITION UNTUK EKSTRAKSI LOKASI DAN WAKTU KEBAKARAN HUTAN

Disusun oleh:

Nama : **Hafidz Sanjaya**
NIM : **22.55.2287**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2024

TESIS

MODEL MULTILINGUAL NAMED ENTITY RECOGNITION UNTUK EKSTRAKSI LOKASI DAN WAKTU KEBAKARAN HUTAN

MULTILINGUAL NAMED ENTITY RECOGNITION MODEL FOR LOCATION AND TIME EXTRACTION OF FOREST FIRES

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama : **Hafidz Sanjaya**
NIM : **22.55.2287**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2024

HALAMAN PENGESAHAN

MODEL MULTILINGUAL NAMED ENTITY RECOGNITION UNTUK EKSTRAKSI LOKASI DAN WAKTU KEBAKARAN HUTAN

MULTILINGUAL NAMED ENTITY RECOGNITION MODEL FOR LOCATION AND TIME EXTRACTION OF FOREST FIRES

Dipersiapkan dan Disusun oleh

Hafidz Sanjaya

22.55.2287

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Senin, 03 Juni 2024

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 03 Juni 2024
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

MODEL MULTILINGUAL NAMED ENTITY RECOGNITION UNTUK EKSTRAKSI LOKASI DAN WAKTU KEBAKARAN HUTAN

MULTILINGUAL NAMED ENTITY RECOGNITION MODEL FOR LOCATION AND TIME EXTRACTION OF FOREST FIRES

Dipersiapkan dan Disusun oleh

Hafidz Sanjaya

22.55.2287

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Senin, 03 Juni 2024

Pembimbing Utama	**Anggota Tim Penguji**
Prof. Dr. Kusrini, M.Kom.	**Tonny Hidayat, S.Kom., M.Kom. Ph.D.**
NIK. 190302106	**NIK. 190302182**
	Alva Hendi Muhammad, S.T., M.Eng., Ph.D.
Pembimbing Pendamping	**NIK. 190302493**
Dr. Kumara Ari Yuana, S.T., M.T.	**Prof. Dr. Kusrini, M.Kom.**
NIK. 190302575	**NIK. 190302106**

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 03 Juni 2024
Direktur Program Pascasarjana

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Hafidz Sanjaya
NIM : 22.55.2287
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Model Multilingual Named Entity Recognition Untuk Ekstraksi Lokasi dan Waktu Kebakaran Hutan

Dosen Pembimbing Utama : Prof. Dr. Kusrini, M.Kom.
Dosen Pembimbing Pendamping : Dr. Kumara Ari Yuana, S.T., M.T.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 03 Juni 2024
Yang Menyatakan,

Hafidz Sanjaya

HALAMAN PERSEMBAHAN

Segala puji bagi Allah SWT karena berkat Rahmat dan Hidayah-Nya sehingga saya diberikan kekuatan dan kemampuan untuk bisa menyelesaikan tesis ini dengan baik. Tesis ini saya persembahkan untuk :

1. Kedua orang tua terkasih Bapak Aang Suryana (alm.) dan Ibu Amung Muntasiroh, terima kasih banyak telah memberikan doa, nasehat dan dukungan yang tiada hentinya.
2. Kedua mertua Bapak Abdul Rojak dan Ibu Nunung Nurlaela, terima kasih telah memberikan dukungan kepada saya.
3. Istri tercinta Hasna Faujiyah, S.Kom., terima kasih telah memberikan bantuan, motivasi dan dukungan serta sabar menemani dalam setiap langkah.
4. Anak-anakku tersayang Hisyam Mujahid dan Hakeem Ikhsanul Akbar.
5. Untuk kakak Anan Bahrul Khoir, S.Ud., M.A. dan kedua adik Wulan Sari dan Fitria Komalasari, terima kasih sudah mendukung dan selalu mengirim doa untuk kelancaran penulisan tesis ini.
6. Serta semua pihak yang telah membantu tersusunnya tesis ini yang tidak dapat saya sebutkan satu persatu.

HALAMAN MOTTO

"Sesungguhnya Allah beserta orang-orang yang sabar" (QS. Al-Baqarah/2:153)

"Barang siapa yang tidak mampu menahan lelahnya belajar, maka ia harus mampu menahan perihnya kebodohan" (Imam Syafi'i)

"Menuntut ilmu adalah takwa. Menyampaikan ilmu adalah ibadah. Mengulang-ulang ilmu adalah zikir. Mencari ilmu adalah jihad" (Imam Ghazali)

KATA PENGANTAR

Dengan mengucapkan Alhamdulillah segala puji dan syukur penulis panjatkan atas kehadirat Allah SWT, karena berkat rahmat dan hidayah-Nya penyusunan tesis yang berjudul "Model Multilingual Named Entity Recognition untuk Ekstraksi Lokasi dan Waktu Kebakaran Hutan" ini dapat diselesaikan guna memenuhi salah satu syarat dalam menyelesaikan pendidikan magister pada Program Studi S2 PJJ Informatika Universitas Amikom Yogyakarta. Banyak hambatan yang dihadapi dalam penyusunannya, namun berkat pertolongan dan kehendak-Nya penulis dapat menyelesaikan tesis ini. Untuk itu, pada kesempatan ini penulis ingin menyampaikan terima kasih yang sebesar-besarnya kepada :

1. Bapak Prof. Dr. M. Suyanto, M.M., selaku Rektor Universitas Amikom Yogyakarta yang telah memberikan motivasi dan arahan kepada penulis.
2. Ibu Prof. Dr. Kusrini, M.Kom., selaku Direktur Pascasarjana Universitas Amikom Yogyakarta sekaligus dosen pembimbing utama yang telah memberikan bimbingan, arahan serta kepercayaan dan dukungan untuk memperoleh beasiswa Silvanus Universitas Amikom Yogyakarta.
3. Bapak Dr. Kumara Ari Yuana, S.T., M.T., selaku dosen pembimbing pendamping yang telah memberikan bimbingan, arahan dan dukungan kepada penulis.
4. Bapak Ibu Staf Admisi Pascasarjana Universitas Amikom Yogyakarta yang telah melayani dan memfasilitasi kegiatan akademik.

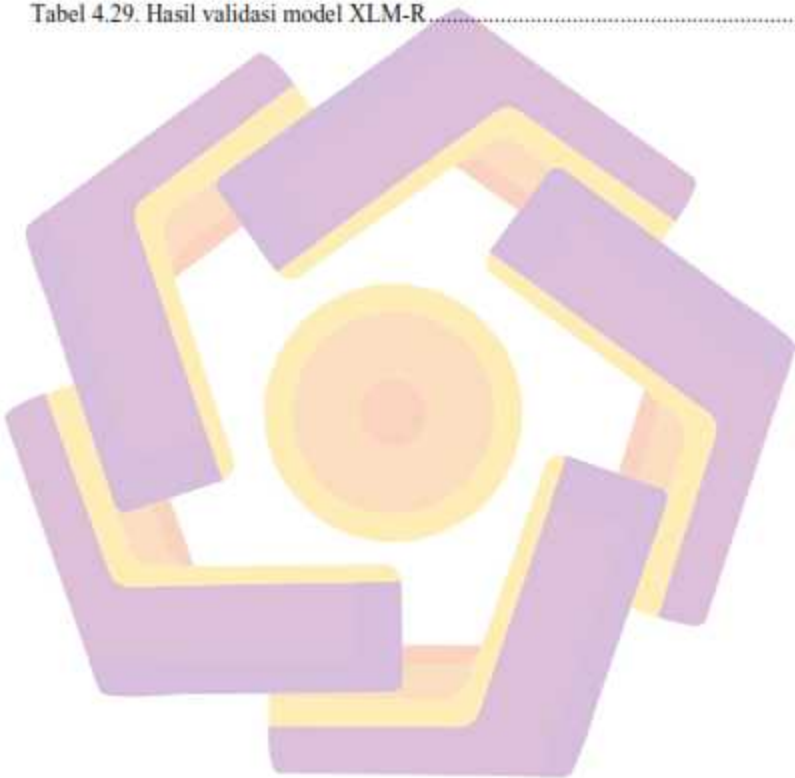
5. Tim Silvanus Universitas Amikom Yogyakarta yang telah memberikan masukan dan dukungan kepada penulis.
6. Rekan-rekan Mahasiswa S2 PJJ Informatika Universitas Amikom Yogyakarta yang saya cintai dan saya banggakan serta segenap pihak-pihak terkait yang tidak bisa penulis sebutkan satu persatu namanya. Semoga amal dan kebaikan kalian dibalas oleh Allah SWT.

Yogyakarta, 25 September 2024

Penulis

DAFTAR ISI

DAFTAR TABEL



DAFTAR GAMBAR

INTISARI

Kelangkaan dan terbatasnya akses kepada *dataset* multibahasa dapat menghambat pelatihan model pengenalan entitas bernama, terutama untuk penerapan penginderaan media sosial dalam manajemen bencana, seperti peringatan dini terjadinya kebakaran hutan mengingat dampaknya yang dapat mengancam keanekaragaman hayati termasuk manusia. Selain itu, membuat dan mendapatkan *dataset* multibahasa berkualitas tinggi untuk mengekstrak informasi lokasi dan waktu kebakaran hutan menggunakan pengenalan entitas bernama juga memerlukan sumber daya dan upaya yang cukup besar. Sehingga, mengatasi langkanya ketersediaan *dataset* pengenalan entitas bernama multibahasa dalam peringatan dini terjadinya kebakaran hutan merupakan salah satu pendekatan yang mungkin dilakukan untuk mengurangi sumber daya dan upaya agar menjadi lebih efisien.

Oleh karena itu, penelitian ini akan menggunakan suatu publik *dataset* dalam Bahasa Indonesia untuk disempurnakan pada beberapa *pre-trained* model multibahasa berbasis BERT seperti Multilingual BERT Cased , Multilingual BERT Uncased dan XLM-RoBERTa untuk membandingkan kinerjanya dalam mengekstraksi atau mengenali entitas lokasi dan waktu kebakaran hutan secara multibahasa dari teks media sosial seperti Twitter (Sekarang "X").

Hasil pelatihan menunjukkan XLM-RoBERTa memperoleh performa *fine-tuning* terbaik dengan *accuracy* 98,59%, *precision* 91,89%, *recall* 92,73% dan *f1-score* 92,31% serta memperoleh performa terbaik pada pengujian dengan akurasi 98,53% dalam melakukan klasifikasi token. Hasil validasi *tweet* secara manual juga menunjukkan bahwa XLM-RoBERTa memperoleh akurasi tertinggi pada semua bahasa yang divalidasi yaitu Bahasa Indonesia 92,32%, Bahasa Inggris 73,97%, Bahasa Spanyol 77,45%, Italia 78,39% dan Slovakia 96,50%.

Kata kunci: Pengenalan Entitas Bernama, Kebakaran Hutan, Ekstraksi Informasi, Penginderaan Media Sosial, BERT

ABSTRACT

The scarcity and limited access to multilingual datasets can hinder the training of named entity recognition models, especially for the application of social media sensing in disaster management, such as early warning of forest fires considering the impact that can threaten biodiversity including humans. Additionally, creating and obtaining high-quality multilingual datasets to extract forest fire location and time information about forest fires using named entity recognition also requires considerable resources and effort. Thus, overcoming the scarce availability of multilingual named entity recognition datasets in early warning of forest firest is one of the possible approaches to reduce resources and efforts to become more efficient.

Therefore, this research will use a public dataset in Indonesian, to be fine-tuning on several pre-trained BERT-based multilingual models such as Multilingual BERT Cased, Multilingual BERT Uncased and XLM-RoBERTa to compare their performance in extracting fire locations and times of forest fire from social media texts such as Twitter (Now "X").

The training results show that XLM-RoBERTa obtained the best fine-tuning performance with accuracy of 98.59%, precision of 91.89%, recall of 92.73% and f1-score of 92.31%, as well as the best performance on testing with accuracy of 98.53% in conducting token classification. The results of manual tweet validation also showed that XLM-RoBERTa obtained the highest accuracy in all validated languages such as Indonesian of 92,32%, English of 73,97%, Spanish of 77,45%, Italian of 78,39% and Slovak of 96,50%.

Keyword: Named Entity Recognition, Forest Fires, Information Extraction, Social Media Sensing, BERT

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Kebakaran hutan menjadi salah satu isu lokal dan global yang semakin meningkat dalam beberapa tahun terakhir dan dapat memberikan dampak serius terhadap ekosistem hutan dengan menghasilkan emisi gas rumah kaca sehingga dapat menyebabkan perubahan iklim, mengancam keanekaragaman hayati dan polusi udara (MacCarthy *dkk*., 2023). Selain itu, kebakaran hutan juga mengancam kesehatan manusia, ditunjukkan dengan tingginya jumlah korban kebakaran hutan secara global sebanyak 310 orang dari bulan Januari 2023 hingga September 2023 dibandingkan tahun-tahun sebelumnya (Salas, 2023). Upaya manajemen bencana seperti kebakaran hutan memiliki urgensi yang tinggi mengingat isu ini semakin meningkat dan berdampak pada skala lokal maupun global. Di tingkat global, beberapa negara saling bekerja sama dan tergabung dalam proyek penanggulangan bencana kebakaran hutan yang bernama SILVANUS.

Proyek SILVANUS sendiri merupakan konsorsium internasional yang didalamnya terdapat berbagai macam tenaga ahli dari empat benua yang bertujuan untuk menanggulangi ancaman kebakaran hutan dan meningkatkan ketangguhan hutan terhadap perubahan iklim (Silvanus Amikom, 2022). Proyek SILVANUS didanai oleh program *Green Deal EU Horizon 2020* dan dikoordinasikan oleh Università Telematica Pegaso, Itali (https://cordis.europa.eu/project/id/101037247). Proyek internasional ini

melibatkan 49 mitra dari 18 negara seperti Amerika Serikat, Spanyol, Italia, Slovakia dan lainnya, termasuk Indonesia yang diwakili oleh Universitas AMIKOM Yogyakarta. Proyek ini melibatkan berbagai teknologi mutakhir yang terintegrasi big data untuk menganalisis data dan informasi dalam manajemen bencana kebakaran hutan.

Dalam managemen bencana kebakaran hutan, data dan informasi terjadinya kebakaran hutan secara global bisa diperoleh dari berbagai sumber termasuk penginderaan jauh (*remote sensing*). Teknologi tersebut telah banyak digunakan dalam manajemen bencana khususnya pada tahap kesiapsiagaan, peringatan dini, tanggap darurat, pemantauan dan pemulihan pasca bencana (Kaku, 2019). Dimana satelit mendeteksi kebakaran hutan yang terjadi di bumi sebagai titik panas (*hotspot*). Titik panas sendiri merupakan salah satu indikator kebakaran hutan yang mendeteksi suatu lokasi memiliki suhu relatif tinggi dibandingkan suhu disekitarnya (Pasal 1 Angka 9 Permenhut No. P 12/ PMenhut-II/ 2009). Secara global, pemantauan titik panas salah satunya bisa menggunakan satelit MODIS dan VIIRS yang memiliki jarak *buffer* hingga 5 kilometer (Indradjad *dkk.*, 2019).

Pendeteksian titik panas dari satelit telah digunakan sebagai manajemen bencana kebakaran hutan. Namun, tingkat akurasi titik panas terhadap kejadian kebakaran dari satelit MODIS dan VIIRS cukup rendah, kurang dari 50% berdasarkan perbandingan jumlah kejadian yang diperoleh satelit dengan yang terjadi di lapangan (Indradjad *dkk.*, 2019). Terdapat beberapa kasus dimana penggunaan data satelit juga terkendala oleh keterbatasan yang diberlakukan seperti resolusi spektral, geografis dan waktu sensor observasi bumi, hal ini terutama

terlihat dalam aplikasi manajemen bencana yang membutuhkan respon *real-time* atau hampir *real-time* (Li *dkk.*, 2017). Informasi mengenai lokasi dan waktu terjadinya kebakaran hutan juga bisa diperoleh salah satunya dari media sosial (Suganda Girsang & Noveta, 2022).

Saat ini, data dari platform media sosial telah dimanfaatkan untuk mengatasi kekurangan penginderaan jauh untuk tanggap darurat di berbagai tahapan manajemen bencana yang meliputi mitigasi, kesiapan, respon dan pemulihan pasca bencana (Phengsuwan *dkk.*, 2021). Selain itu, pada era *big data* sekarang ini muncul paradigma baru dimana manusia digunakan sebagai sensor yang adaptif dan hemat biaya untuk menyampaikan pemikiran mereka tentang dunia nyata melalui penggunaan media sosial (Zhang *dkk.*, 2019). Oleh karena itu, data dari sosial media dapat diolah sehingga menjadi informasi yang dapat digunakan oleh otoritas dan masyarakat untuk merespons dan mengambil tindakan yang diperlukan dalam konteks manajemen bencana.

Twitter (sekarang "X") berpotensi menjadi platform media sosial untuk melakukan analisis *big data* dalam konteks manajemen bencana karena memiliki kemampuan dapat memberikan informasi dan peringatan secara *real time* mengenai peristiwa yang sedang berlangsung (Shah *dkk.*, 2021). Twitter menghasilkan informasi berupa teks selama terjadinya bencana yang menunjukkan tempat dan waktu kejadian secara tepat. *Named entity recognition* biasanya digunakan dalam ekstraksi informasi untuk mengidentifikasi entitas pada data berbasis teks (Jati *dkk.*, 2020). Tugas pemrosesan bahasa alami, seperti *named entity recognition* dibatasi kapasitasnya untuk diterapkan pada serangkaian bahasa yang terbatas. Umumnya,

metode ini berhasil untuk bahasa yang banyak digunakan dalam situasi darurat secara global, seperti bahasa Inggris. Namun demikian, hal itu kurang mampu menangkap kondisi lokal secara tepat dan kesulitan dalam mengikuti arus informasi media sosial yang cepat dan multibahasa. Kelangkaan dan terbatasnya akses kepada *dataset* multibahasa di sektor manajemen bencana dapat menghambat pelatihan model *machine learning*. Selain itu, membuat *dataset* berkualitas tinggi juga memerlukan sumber daya dan upaya yang besar.

Berbagai penelitian *named entity recognition* telah dilakukan dalam sektor manajemen bencana. Sebuah studi yang memanfaatkan *Stanford Named Entity Recognition* (*Stanford NER*) untuk mengekstraksi lokasi berbasis konten untuk menyelidiki penerapan media sosial seperti Twitter (Sekarang "X") dalam memfasilitasi evakuasi selama musim kebakaran hutan di Amerika Barat (Li *dkk.*, 2021). Pemanfaatan data media sosial memungkinkan pembuatan peta evakuasi secara *real-time* yang menyediakan perencanaan evakuasi yang lebih cepat dan dapat diandalkan di daerah yang terkena dampak dibandingkan dengan peta evakuasi tradisional.

Stanford NER juga telah digunakan pada konten (data teks) di Twitter untuk memprediksi lokasi bencana alam di Indonesia (Suganda Girsang & Noveta, 2022). Enam kelas lokasi seperti PROP, KAB, KEC, KEL, STREET, POI dilatih berdasarkan tingkat regional di Indonesia dan memperoleh akurasi sebesar 85,65%. Selain itu, menghasilkan akurasi 87,5% dalam pemetaan bencana alam menggunakan ArcGIS berdasarkan evaluasi geolokasi. Penelitian ini juga

menghasilkan data spasial yang dapat digunakan untuk menemukan lokasi bencana alam pada peta.

Penelitian lain membandingkan berbagai model berbasis *Recurrent Neural Network (RNN)* seperti LSTM, Bidirectional LSTM, dan GRU dengan penyematan kata GloVe sebagai model *named entity recognition* untuk mengekstraksi nama orang, organisasi, dan tempat dari *tweet* pengguna dalam bahasa Inggris (Eligüzel *dkk.*, 2022). Ketiga model tersebut dikombinasikan dengan berbagai fungsi aktivasi dan *optimizer* yang berbeda untuk menentukan kombinasi terbaik. Hasil penelitian menunjukkan bahwa model Bidirectional LSTM dengan aktivasi *softmax* dan *optimizer Nadam* mencapai tingkat kinerja tertinggi dengan *f1-score* sebesar 0.92.

Korpus bencana alam yang dianotasi dalam bahasa Inggris telah dibuat dengan entitas yang terdiri dari Lokasi Geografis (Loc), Bencana Alam (Haz) dan Metode Penelitian (Met). Korpus ini digunakan untuk membandingkan kinerja dari 12 model seperti BERT-CRF, ALBERT-CRF, XLNet-CRF, BERT-BiLSTM, BERT-BiLSTM-CRF, ALBERT-BiLSTM-CRF, XLNet-BiLSTM -CRF, BERT-BiGRU-CRF, ALBERT-BiGRU-CRF, XLNet-BiGRU-CRF, BiGRU-CRF, dan BiLSTM-CRF (Sun *dkk.*, 2022). Model XLNet-BiLSTM-CRF memiliki kinerja terbaik dalam pra-pelatihan, lapisan *encoding* dan lapisan *decoding*. Model tersebut memperoleh *precision*, *recall*, dan *f1-score* masing-masing sebesar 92,80%, 91,74%, dan 92,27%.

Pre-trained model multibahasa berbasis BERT juga dilatih untuk Pengenalan Sebutan Lokasi (*Location Mention Recognition*) di Twitter menggunakan berbagai konfigurasi, *dataset*, bahasa dan kedekatan geografis. Hal

ini dilakukan dengan mengasumsikan bahwa tidak ada data bahasa target yang tersedia (pengaturan *zero-shot*) atau hanya ada sejumlah data bahasa target yang terbatas (pengaturan *few-shot*) dalam *dataset* pelatihan (Suwaileh *dkk.*, 2022). Hasil penelitian menunjukkan bahwa penggunaan *pre-trained* model multibahasa tanpa data dari bahasa target memberikan hasil yang memuaskan. Namun, menggabungkan sejumlah kecil data dalam bahasa target berkisar antara 263 hingga 356 dapat meningkatkan kinerja secara signifikan. Selain itu, memberi label pada sekitar 500 *tweet* kejadian bencana menghasilkan kinerja model yang dapat diterima dengan *f1-score* lebih dari 0.7.

Keseluruhan penelitian menunjukkan bahwa manajemen bencana menggunakan pendekatan *named entity recognition* efektif dalam mengekstrak informasi dari teks media sosial seperti Twitter, khususnya *pre-trained* model multibahasa berbasis BERT yang mampu mengenali nama tempat yang tidak ada di dalam *dataset* (*out-of-vocabulary*). Selain itu, model tersebut juga mampu mengenali nama tempat secara multibahasa (*multilingual*) meskipun model dilatih menggunakan satu bahasa saja. Hal itu menunjukkan *pre-trained* model multibahasa efektif dan mampu untuk mengekstrak informasi lokasi dan waktu dari Twitter, namun diperlukan adanya perbandingan antara *pre-trained* model multibahasa untuk mencari model mana yang memiliki akurasi terbaik dalam mengekstrak lokasi dan waktu terjadinya kebakaran hutan.

Oleh karena itu, penelitian ini akan menganalisis penggunaan suatu publik *dataset* untuk tujuan umum dalam Bahasa Indonesia yang dilatih pada tiga *pre-trained* model multibahasa berbasis BERT yang tersedia di *huggingface* yaitu

Multilingual BERT Cased (mBERT Cased), Multilingual BERT Uncased (mBERT Uncased) dan XLM-RoBERTa (XLM-R) serta mengevaluasi model mana yang memiliki performa terbaik dalam mengenali entitas lokasi dan waktu. Selain itu, model multibahasa yang dibangun juga divalidasi secara manual dalam mengekstrak lokasi dan waktu dari *tweet* kebakaran hutan pada Bahasa Indonesia, Inggris, Spanyol, Italia dan Slovakia untuk mengetahui keakuratan model dalam mengekstrak lokasi dan waktu kebakaran hutan dari Twitter (X). Kelima bahasa tersebut dipilih untuk mendemonstrasikan deteksi kebakaran hutan secara global dari data media sosial pada proyek SILVANUS.

1.2. Rumusan Masalah

Berdasarkan uraian latar belakang masalah diatas, maka rumusan masalah pada penelitian ini adalah:

a. Berapa hasil performa *fine-tuning* dari *pre-trained* model multibahasa berbasis BERT dalam mengenali entitas lokasi dan waktu?
b. Berapa nilai akurasi pengujian dari model multilingual *named entity recognition* dalam mengenali entitas lokasi dan waktu?
c. Berapa nilai akurasi validasi dari model multilingual *named entity recognition* dalam mengekstrak lokasi dan waktu pada *tweet* kebakaran secara multibahasa?

1.3. Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

a. *Dataset* publik yang digunakan untuk *fine-tuning* model adalah Nergrit Corpus dalam Bahasa Indonesia (PT Gria Inovasi Teknologi, 2019).
b. *Pre-trained* model multibahasa yang digunakan adalah model berbasis BERT yaitu Multilingual BERT Cased (mBERT Cased), Multilingual BERT Uncased (mBERT Uncased) dan XLM-RoBERTa (XLM-R) yang tersedia di *huggingface*.
c. Jumlah bahasa yang digunakan untuk validasi *tweet* sebanyak 5 bahasa, yaitu bahasa Indonesia, Inggris, Spanyol, Italia dan Slovakia sesuai dengan pilot proyek Silvanus Amikom dan ketersediaan validator model.
d. Performa *fine-tuning* yang dievaluasi adalah *precision, recall, accuracy* dan *f1-score*.

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

a. Mengetahui performa *fine-tuning* dari *pre-trained* model multibahasa berbasis BERT dalam mengenali entitas lokasi dan waktu.
b. Mengetahui nilai akurasi pengujian dari model multilingual *named entity recognition* dalam mengenali entitas lokasi dan waktu.
c. Mengetahui nilai akurasi validasi model multilingual *named entity recognition* dalam mengekstrak lokasi dan waktu kebakaran hutan pada teks *tweet* secara multibahasa.

1.5. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat diantaranya:

a. Bagi peneliti, mengetahui performa *fine-tuning* beberapa *pre-trained* model multibahasa berbasis BERT dalam mengenali entitas lokasi dan waktu.
b. Memberikan pengetahuan baru mengenai *pre-trained* model multibahasa yang dilatih menggunakan *dataset* publik dalam Bahasa Indonesia untuk mengenali entitas lokasi dan waktu kebakaran hutan secara multibahasa seperti Bahasa Indonesia, Inggris, Spanyol, Italia dan Slovakia.
c. Model multilingual *named entity recognition* yang dibangun dapat diterapkan pada sistem tanggap darurat bencana kebakaran hutan berbasis *social media sensing* untuk mengekstrak lokasi dan waktu kebakaran hutan dari teks media sosial secara multibahasa.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

(Li *dkk.*, 2021) melakukan penelitian berbasis data tentang penggunaan media sosial dalam membantu evakuasi selama musim kebakaran hutan di Amerika Barat. Penelitian ini dilakukan berdasarkan analisis terhadap 53,990 tweet yang relevan untuk mengungkap pola penyebaran informasi melalui *network analysis* dan Standford NER yang dikembangkan oleh (Finkel *dkk.*, 2005). Kelebihan dari penelitian ini menunjukkan efektivitas penggunaan media sosial dalam membantu evakuasi dan juga memberikan pedoman untuk studi di masa depan tentang ekstraksi informasi prioritas tinggi dari media sosial untuk kebencanaan. Namun, penelitian ini juga memiliki beberapa kelemahan seperti *dataset* yang didasarkan pada tweet di Amerika Barat, yang bisa saja bias terhadap demografi pengguna Twitter tertentu sehingga terbatas pada kebakaran hutan di Amerika Barat dan tidak dapat diterapkan ke bahasa lain.

(Suganda Girsang & Noveta, 2022) mengusung konsep prediksi lokasi bencana alam di Indonesia menggunakan *Standford Named Entity Recognition* (Standford NER) pada data Twitter. Metode yang digunakan adalah *support vector machine* (SVM) untuk klasifikasi delapan jenis bencana alam yang data-datanya diperoleh dari Twitter, yang kemudian diekstrak lokasinya menggunakan Standford NER dengan enam kelas lokasi berdasarkan level regional di Indonesia yaitu Provinsi (PROP), Kabupaten (KAB), Kecamatan (KEC), Kelurahan (KEL), Jalan

(STREET) dan POI (*Place-Of-Interest*). Hasil pelatihan model untuk ekstraksi lokasi memperoleh akurasi sebesar 85.65%. Selain itu, hasil evaluasi geolokasi menggunakan peta ArcGIS menunjukkan efektivitas metode ini dengan tingkat akurasi sebesar 87,5%. Namun, penelitian ini belum mampu untuk mengenali lokasi yang tidak ada di dataset (termasuk pulau-pulau kecil) sehingga dimasukkan dalam kategori POI dan belum bisa mendapatkan konteks lokasi bencana alam apabila terdapat nama pengguna yang sama dengan nama lokasi. Selain itu, model yang dikembangkan tidak disebutkan mampu untuk mengekstrak lokasi pada bahasa yang berbeda atau multibahasa.

(Eligüzel *dkk.*, 2022) mengkaji penggunaan *named entity recognition* pada *tweet* selama bencana gempa dengan menggunakan berbagai model berbasis *Recurrent Neural Network* (RNN) seperti GRU, LSTM, dan bidirectional LSTM yang dilengkapi dengan berbagai fungsi aktivasi (*Tanh, Elu, Softmax, Relu, Sigmoid, Softplus, Linear, Softsign*) dan algoritma optimisasi (*Adam, SGD, Nadam, Adamax, RMSprop, Adagrad, Adadelta*). Selain itu, penyematan kata GloVe digunakan sebagai representasi inputan pada saat pelatihan. Data yang digunakan berasal dari *tweet* terkait gempa bumi Nepal tahun 2015. Hasil penelitian ini menunjukkan efektivitas model dalam mengekstrak lokasi yang ditunjukkan dengan *precision* sebesar 0.94, *recall* sebesar 0.94, *f1-score* sebesar 0.92 dan *hamming loss* sebesar 0.059 pada model bidirectional LSTM (biLSTM) dengan fungsi aktivasi *softmax* dan fungsi optimasi *Nadam*. Namun, model ini juga memiliki beberapa kelemahan karena memerlukan jumlah data yang besar untuk "belajar" secara efektif, jika tidak maka performa algoritma bisa menurun. Selain

itu, model yang digunakan tidak bisa mengenali lokasi dari bahasa yang berbeda dari *dataset* atau secara multibahasa.

(Suwaileh *dkk.*, 2022) berfokus pada pengenalan penyebutan tempat atau *Location Mention Recognition* (LMR) pada *tweet* bencana. Penelitian ini menggunakan model *bert-large-cased* dan *bert-base-multilingual-cased* untuk melatih model menggunakan berbagai pengaturan, *dataset*, bahasa, dan kedekatan geografis. Hasil penelitian menunjukkan bahwa penggunaan sedikitnya 263 sampai 356 *tweet* pelatihan dari bahasa target (pengaturan *few-shot*) meningkatkan kinerja dalam pengaturan multibahasa (*multilingual*) dibandingkan pengaturan *zero-shot* yang sama sekali tidak menggunakan data target sebagai pelatihan. Selain itu, memberi label pada sekitar 500 *tweet* bencana target menghasilkan kinerja LMR yang dapat diterima dengan *f1-score* lebih dari 0.7 dalam pengaturan monolingual. Kelebihan dari penelitian ini adalah model multibahasa yang dihasilkan dapat digunakan dalam berbagai pengaturan dan bahasa, yang membuatnya berfungsi dalam berbagai situasi bencana. Namun, penelitian ini juga memiliki beberapa kelemahan seperti model yang digunakan mungkin tidak sepenuhnya akurat dalam mengidentifikasi tempat dalam *tweet* bencana, terutama jika tweet tersebut berbahasa lain yang secara topologis huruf berbeda atau jika krisis tersebut terjadi di daerah yang jauh dari daerah yang digunakan untuk pelatihan. Selain itu, penelitian ini hanya menggunakan satu model multibahasa saja sehingga diperlukan adanya perbandingan dengan model multibahasa lainnya untuk mencari model multibahasa mana yang memiliki kinerja lebih baik.

(Sun *dkk.*, 2022) mengusulkan 12 model *named entity recognition* untuk mengekstrak informasi lokasi dan bencana alam seperti BERT-CRF, ALBERT-CRF, XLNet-CRF, BERT-BiLSTM, BERT-BiLSTM-CRF, ALBERT-BiLSTM-CRF, XLNet-BiLSTM-CRF, BERT-BiGRU-CRF, ALBERT-BiGRU-CRF, XLNet-BiGRU-CRF, BiGRU-CRF, dan BiLSTM-CRF untuk membandingkan kinerja, kelebihan dan kekurangannya. Hasil penelitian ini menunjukkan model XLNet-BiLSTM-CRF berhasil mengekstrak tiga jenis entitas bencana alam dengan efisien dan akurat ditunjukkan dengan *precision* sebesar 92.80%, *recall* sebesar 91.74%, dan *f1-score* sebesar 92.27% dengan waktu pelatihan selama 681 detik. Model ini memiliki kinerja sangat baik dibandingkan dengan model lainnya, akan tetapi model ini memiliki beberapa kelemahan seperti membutuhkan korpus yang telah dilabeli secara khusus untuk bencana alam dan tidak dapat digunakan untuk menangani bahasa yang berbeda atau multibahasa.

(Berragan *dkk.*, 2023) merancang lima model seperti BERT, DistilBERT, RoBERTa, CRF+biLSTM dan CRF+biLSTM-*basic* yang dibuat khusus sebagai model *named entity recognition* dan dibandingkan dengan tiga *pre-built model* yaitu SpaCy (*small*), SpaCy (*large*) dan Stanza untuk ekstraksi nama tempat di Inggris Raya berdasarkan data Wikipedia. Model yang paling baik dalam penelitian ini adalah BERT dengan nilai *f1-score* sebesar 0.939, dibandingkan *pre-built* model Stanza dengan nilai *f1-score* sebesar 0.730. Model BERT menunjukkan kelebihannya dalam menangani masalah *out-of-vocabulary* (OOV) dengan menangkap nama tempat yang tidak ada di *dataset*. Namun, penelitian ini juga memiliki beberapa keterbatasan karena model ini diuji menggunakan artikel

wikipedia yang mungkin tidak mencerminkan variasi dan kompleksitas teks bahasa alami yang tidak baku dan tidak terstruktur, serta penelitian ini tidak membahas bagaimana model ini dapat digunakan atau divalidasi untuk menangani bahasa yang berbeda atau multibahasa.

2.2. Keaslian Penelitian

Tabel 2.1. Matriks literatur review dan posisi penelitian
Model Multilingual Named Entity Recognition Untuk Ekstraksi Lokasi dan Waktu Kebakaran Hutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Data-driven investigations of using social media to aid evacuations amid Western United States wildfire season	Lingyao Li, Zihui Ma, Tao Cao, *Elsevier*, 2021	Memanfaatkan *Stanford Named Entity Recognition* (Stanford NER) untuk mengekstraksi lokasi berbasis konten untuk menyelidiki penerapan media sosial seperti Twitter (Sekarang "X") dalam memfasilitasi evakuasi selama musim kebakaran hutan di Amerika Barat	Pemanfaatan data media sosial memungkinkan pembuatan peta evakuasi secara *real-time* yang menyediakan perencanaan evakuasi yang lebih cepat dan dapat diandalkan di daerah yang terkena dampak dibandingkan dengan peta evakuasi tradisional	Tidak disebutkan performa yang diperoleh oleh Standford NER dalam mengekstrak lokasi. Penelitian ini juga memiliki beberapa kelemahan seperti *dataset* yang didasarkan pada tweet di Amerika Barat, yang bisa saja bias terhadap demografi pengguna Twitter tertentu sehingga terbatas pada kebakaran hutan di Amerika Barat dan tidak dapat diterapkan ke bahasa lain	Penelitian yang dilakukan penulis tidak menggunakan *pre-built* model seperti Standford NER, akan tetapi menggunakan *dataset* publik berbahasa Indonesia dan tiga *pre-trained* model multibahasa berbasis BERT seperti mBERT Cased, mBERT Uncased dan XLM-RoBERTa untuk mengekstrak lokasi dan waktu kebakaran hutan dari *tweet* secara multibahasa
2	Location Prediction using Named Entity Recognition for Indonesia Natural	Abba Suganda Girsang, et. al, *Elsevier*, 2022	Menggunakan Stanford NER yang digunakan pada konten (data teks) di Twitter untuk	Enam kelas lokasi seperti PROP, KAB, KEC, KEL, STREET, POI dilatih berdasarkan tingkat regional di	Penelitian ini belum mampu untuk mengenali lokasi yang tidak ada di dataset (termasuk pulau-pulau kecil) sehingga	Penelitian yang dilakukan penulis tidak menggunakan *pre-built* model seperti Standford NER, akan tetapi menggunakan *dataset* publik

Tabel 2.1. Matriks literatur review dan posisi penelitian
Model Multilingual Named Entity Recognition Untuk Ekstraksi Lokasi dan Waktu Kebakaran Hutan (lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	Disasters in Data Twitter		memprediksi lokasi bencana alam di Indonesia	Indonesia dan memperoleh akurasi sebesar 85,65%. Selain itu, menghasilkan akurasi 87,5% dalam pemetaan bencana alam menggunakan ArcGIS berdasarkan evaluasi geolokasi	dimasukkan dalam kategori POI dan belum bisa mendapatkan konteks lokasi bencana alam apabila terdapat nama pengguna yang sama dengan nama lokasi. Selain itu, model yang dikembangkan tidak disebutkan mampu untuk mengekstrak lokasi pada bahasa yang berbeda atau multibahasa	berbahasa Indonesia dan tiga *pre-trained* model multibahasa berbasis BERT seperti mBERT Cased, mBERT Uncased dan XLM-RoBERTa untuk mengekstrak lokasi dan waktu kebakaran hutan dari *tweet* secara multibahasa
3	Application Of Named Entity Recognition On Tweets During Earthquake Disaster: A Deep Learning-Based Approach	Nazmiye Eliguzel et. al., *Springer*, 2022	Membandingkan berbagai model berbasis *Recurrent Neural Network (RNN)* seperti LSTM, Bidirectional LSTM, dan GRU dengan penyematan kata GloVe sebagai model *named entity recognition* untuk mengekstraksi nama orang, organisasi, dan tempat dari *tweet*	Hasil penelitian menunjukkan bahwa model Bidirectional LSTM dengan aktivasi *softmax* dan *optimizer Nadam* mencapai tingkat kinerja tertinggi dengan *f1-score* sebesar 0.92	Model yang dibangun memiliki beberapa kelemahan karena memerlukan jumlah data yang besar untuk "belajar" secara efektif, jika tidak maka performa algoritma bisa menurun. Selain itu, model yang digunakan tidak bisa mendapatkan konteks lokasi dari bahasa yang berbeda atau multibahasa.	Penelitian yang dilakukan penulis tidak menggunakan model berbasis RNN, akan tetapi menggunakan *dataset* publik berbahasa Indonesia dan tiga *pre-trained* model multibahasa berbasis BERT seperti mBERT Cased, mBERT Uncased dan XLM-RoBERTa untuk mengekstrak lokasi dan waktu kebakaran hutan dari *tweet* secara multibahasa

Tabel 2.1. Matriks literatur review dan posisi penelitian
Model Multilingual Named Entity Recognition Untuk Ekstraksi Lokasi dan Waktu Kebakaran Hutan (lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
			pengguna dalam bahasa Inggris			
4	When a disaster happens, we are ready: Location mention recognition from crisis tweets	Reem Suwaileh et. al., *Elsevier*, 2022	Melatih *pre-trained* model multibahasa berbasis BERT untuk Pengenalan Sebutan Lokasi (*Location Mention Recognition*) di Twitter menggunakan berbagai konfigurasi, *dataset*, bahasa dan kedekatan geografis	Hasil penelitian menunjukkan bahwa penggunaan *pre-trained* model multibahasa tanpa data dari bahasa target memberikan hasil yang memuaskan. Namun, menggabungkan sejumlah kecil data dalam bahasa target berkisar antara 263 hingga 356 dapat meningkatkan kinerja secara signifikan. Selain itu, memberi label pada sekitar 500 *tweet* kejadian bencana menghasilkan kinerja model yang dapat diterima dengan *f1-score* lebih dari 0.7	Penelitian ini memiliki beberapa kelemahan seperti model yang digunakan mungkin tidak sepenuhnya akurat dalam mengidentifikasi tempat dalam *tweet* bencana, terutama jika *tweet* tersebut berbahasa lain yang secara topologis huruf berbeda atau jika krisis tersebut terjadi di daerah yang jauh dari daerah yang digunakan untuk pelatihan. Selain itu, penelitian ini hanya menggunakan satu model multibahasa saja sehingga diperlukan adanya perbandingan dengan model multibahasa lainnya untuk mencari model multibahasa mana yang	Penelitian yang dilakukan penulis tidak hanya menggunakan Multilingual BERT saja, akan tetapi menggunakan *dataset* publik berbahasa Indonesia dan tiga *pre-trained* model multibahasa berbasis BERT seperti mBERT Cased, mBERT Uncased dan XLM-RoBERTa untuk mengekstrak lokasi dan waktu kebakaran hutan dari *tweet* secara multibahasa yang divalidasi pada Bahasa Indonesia, Inggris, Spanyol, Italia dan Slovakia

Tabel 2.1. Matriks literatur review dan posisi penelitian
Model Multilingual Named Entity Recognition Untuk Ekstraksi Lokasi dan Waktu Kebakaran Hutan (lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
					memiliki kinerja lebih baik	
5	Deep learning-based methods for natural hazard named entity recognition	Junlin Sun et. al. *scientific reports. 2022*	Mengusulkan 12 model *named entity recognition* seperti BERT-CRF, ALBERT-CRF, XLNet-CRF, BERT-BiLSTM, BERT-BiLSTM-CRF, ALBERT-BiLSTM-CRF, XLNet-BiLSTM-CRF, BERT-BiGRU-CRF, ALBERT-BiGRU-CRF, XLNet-BiGRU-CRF, BiGRU-CRF, dan BiLSTM-CRF untuk membandingkan kinerja, kelebihan dan kekurangannya dalam mengekstrak informasi lokasi dan bencana alam	Hasil penelitian ini menunjukkan model XLNet-BiLSTM-CRF berhasil mengekstrak tiga jenis entitas bencana alam dengan efisien dan akurat ditunjukkan dengan *precision* sebesar 92.80%, *recall* sebesar 91.74%, dan *f1-score* sebesar 92.27% dengan waktu pelatihan selama 681 detik	Model pada penelitian memiliki beberapa kelemahan seperti membutuhkan korpus yang telah dilabeli secara khusus untuk bencana alam dan tidak dapat disesuaikan untuk menangani bahasa yang berbeda atau multibahasa	Penelitian yang dilakukan penulis tidak menggunakan model monolingual, akan tetapi menggunakan *dataset* publik berbahasa Indonesia dan tiga *pre-trained* model multibahasa berbasis BERT seperti mBERT Cased, mBERT Uncased dan XLM-RoBERTa untuk mengekstrak lokasi dan waktu kebakaran hutan dari *tweet* secara multibahasa
6	Transformer Based Named Entity Recognition For Place Name	Cillian Berragan, et. al. *Taylor &*	Merancang lima model seperti BERT, DistilBERT, RoBERTa,	Model yang paling baik dalam penelitian ini adalah BERT dengan nilai *f1-score* sebesar	Penelitian ini memiliki beberapa keterbatasan karena model ini diuji menggunakan artikel	Penelitian yang dilakukan penulis tidak menggunakan model monolingual, akan tetapi menggunakan *dataset*

Tabel 2.1. Matriks literatur review dan posisi penelitian
Model Multilingual Named Entity Recognition Untuk Ekstraksi Lokasi dan Waktu Kebakaran Hutan (lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	Extraction From Unstructured Text	*Francis*. 2023	CRF+biLSTM dan CRF+biLSTM-*basic* yang dibuat khusus sebagai model *named entity recognition* dan dibandingkan dengan tiga *pre-built* model yaitu SpaCy (*small*), SpaCy (*large*) dan Stanza untuk ekstraksi nama tempat di Inggris Raya berdasarkan data Wikipedia	0.939, dibandingkan *pre-built* model Stanza dengan nilai *f1-score* sebesar 0.730. Model BERT menunjukkan kelebihannya dalam menangani masalah *out-of-vocabulary* (OOV) dengan menangkap nama tempat yang tidak ada di *dataset*	Wikipedia yang mungkin tidak mencerminkan variasi dan kompleksitas teks bahasa alami yang tidak baku dan tidak terstruktur, serta penelitian ini tidak membahas bagaimana model ini dapat disesuaikan atau ditingkatkan untuk menangani bahasa yang berbeda atau multibahasa	publik berbahasa Indonesia dan tiga *pre-trained* model multibahasa berbasis BERT seperti mBERT Cased, mBERT Uncased dan XLM-RoBERTa untuk mengekstrak lokasi dan waktu kebakaran hutan dari *tweet* secara multibahasa

2.3. Landasan Teori

2.3.1 Kebakaran Hutan dan Lahan

Kebakaran hutan dan lahan juga dikenal sebagai karhutla adalah salah satu masalah lingkungan yang sering terjadi dan dianggap penting sehingga menarik perhatian baik di tingkat lokal maupun global (Agustiar *dkk.*, 2020). Kebakaran hutan didefinisikan sebagai suatu kejadian dimana api melahap bahan bakar bervegetasi yang terjadi di dalam hutan yang menjalar secara bebas dan tidak terkendali, sedangkan kebakaran lahan terjadi dikawasan non-hutan (Wahyudi, 2021). Penyebab kebakaran hutan dan lahan dapat dikelompokkan menjadi tiga yaitu faktor kondisi bahan bakar dan kadar air yang relatif rendah, faktor iklim berupa suhu, kelembaban, angin dan curah hujan. Suhu yang tinggi dapat menyebabkan bahan bakar mengering sehingga mudah terbakar. Daerah dengan kelembaban tinggi dapat mengurangi kemungkinan terjadinya karhutla, angin mempengaruhi kecepatan menjalarnya api dan curah hujan mempengaruhi jumlah kadar air dalam bahan bakar dan faktor sosial dan budaya masyakarat yang menggunakan api sebagai kegiatan dalam pembukaan lahan, pembakaran liar, perambahan hutan dan kurangnya kesadaran masyarakat akan bahaya api (Arisanty *dkk.*, 2020).

2.3.2 Media Sosial

Media sosial banyak digunakan dan sangat popular dan hampir semua orang menggunakan media sosial. Selain itu, media sosial juga dapat digunakan untuk mengirimkan informasi bencana termasuk kebakaran hutan. Berdasarkan

(Aini & Basuki, 2020) media sosial *online* merupakan sebuah media yang beroperasi dengan bantuan teknologi berbasis platform (web, mobile, desktop) yang membuat perubahan dalam hal komunikasi yang dahulu hanya dapat satu arah dan berubah menjadi dua arah atau dapat disebut sebagai dialog interaktif. Media sosial juga merupakan tempat, layanan dan alat bantu yang memungkinkan setiap orang terhubung sehingga dapat berekspresi dan berbagi dengan individu lainnya menggunakan bantuan internet. Selain itu, media sosial juga bisa digunakan sebagai sensor dengan saling mengirimkan informasi keadaan fisik di dunia nyata sehingga bisa dianalisis untuk mendukung pembuatan keputusan dalam berbagai sektor termasuk bencana alam (Shi *dkk.*, 2022). Saat ini, pada era *big data* muncul paradigma baru dimana manusia digunakan sebagai sensor yang adaptif dan hemat biaya untuk menyampaikan pemikiran mereka tentang dunia nyata melalui penggunaan media sosial yang disebut sebagai penginderaan media sosial (*social media sensing*) (Zhang *dkk.*, 2019).

2.3.3 Natural Language Processing

Natural Language Processing (NLP) adalah rangkaian dari teknik komputasi untuk menganalisa dan merepresentasikan teks yang terjadi secara alami pada satu atau lebih tingkat analisis linguistik dengan tujuan mencapai pemrosesan bahasa yang menyerupai manusia untuk berbagai tugas atau aplikasi (Nugraha *dkk.*, 2020). NLP merupakan pengembangan dari *artificial intelligence* atau kecerdasan buatan yang diharapkan dapat mempelajari bahasa yang digunakan oleh manusia. Saat ini, implementasi NLP sudah banyak

digunakan oleh manusia seperti Google Translate, Google Assistant, Siri, Alexa dan sebagainya. NLP memiliki banyak potensi untuk dikembangkan lebih lanjut, salah satunya adalah *named entity recognition* untuk mengenali entitas data dari informasi media sosial khususnya dalam mengekstrak lokasi dan waktu kebakaran hutan yang akan dibahas lebih dalam pada penelitian ini.

2.3.4 Named Entity Recognition

Named Entity Recognition (NER) merupakan salah satu jenis teknik dari *Natural Language Processing* (NLP) yang berguna untuk mengenali entitas dari sebuah teks dan mengklasifikasikannya ke dalam label yang telah ditentukan sebelumnya (Sharma *dkk.*, 2022). NER banyak diterapkan pada sistem *question answering*, *information extraction*, *co-reference resolution*, *topic modeling* dan lain- lain (Yang *dkk.*, 2022).

Gambar 2.1. Contoh hasil *named entity recognition*

Secara umum, terdapat dua pendekatan yang biasa diterapkan pada NER, yaitu *non-machine learning* dan *machine learning*. Pendekatan *non-machine learning* terdiri dari beberapa metode seperti *rule-base, lexicon, statistical based* dan *ontology*, pendekatan ini lebih menekankan pada pendefinisian aturan ataupun formula matematis dalam melakukan pengenalan entitas. Sedangkan pendekatan *machine learning* terdiri dari *traditional*

machine learning dan *deep learning* yang menggunakan sejumlah data untuk mempelajari pola dari setiap data, sehingga dapat memprediksi label atau entitas dari data sebelumnya (Budi & Suryono, 2023).

2.3.5 Transfer Learning

Transfer learning adalah metode menggunakan jaringan saraf yang sudah dilatih sebelumnya lalu mengurangi jumlah parameter dengan cara mengambil beberapa bagian dari model yang sudah dilatih untuk digunakan dalam mengenali model baru (Raaijmakers, 2022). Didasari oleh fakta bahwa manusia dapat menerapkan pengetahuan yang dipelajari sebelumnya untuk memecahkan masalah baru dengan lebih cepat dan dengan solusi yang lebih baik. Jaringan saraf sangat bergantung pada jumlah data untuk mencapai kinerja yang tinggi. Berikut adalah alasan mengapa *transfer learning* digunakan:

1. Masalah data, *deep learning* membutuhkan banyak data untuk bisa mendapatkan hasil yang baik. Selain itu, membutuhkan banyak waktu untuk membuat ataupun mendapatkan data berlabel khusus jika dilakukan oleh manusia.
2. Masalah komputasi, *deep learning* membutuhkan perangkat keras canggih untuk melatih jaringan saraf yang sangat banyak sehingga akan sangat mahal dan membutuhkan waktu berhari-hari serta perlu dilakukan proses berulang untuk mendapatkan hasil yang optimal.

2.3.6 Transformers

Transformers merupakan arsitektur yang didasarkan untuk menarik perhatian (*attention*) dan menarik ketergantungan global antara *input-output* serta menggantikan lapisan yang paling umum digunakan dalam arsitektur *encoder* & *decoder* (Vaswani *dkk.*, 2017). Model Transformers menunjukkan seni baru dalam kualitas terjemahan bahasa, sementara itu Transformers juga dapat dilatih secara signifikan lebih cepat daripada arsitektur berbasis *Recurrent Neural Network* atau *Convolutional Neural Network*. Transformers adalah dua pilar utama NLP modern, salah satu model transformers *pre-trained* yang paling populer yaitu *Bidirectional Encoder Representations from Transformers* (BERT). Transformers dapat mengkonversi pemahaman yang diperoleh dengan mekanisme yang bernama *self-attention mechanism*. *Self-attention mechanism* merupakan cara Transformers untuk mengubah atau mengkonversikan kata-kata terkait sehingga diproses melalui mekanisme Transformers (Vaswani *dkk.*, 2017). Terdapat dua mekanisme pada Transformer, yaitu:

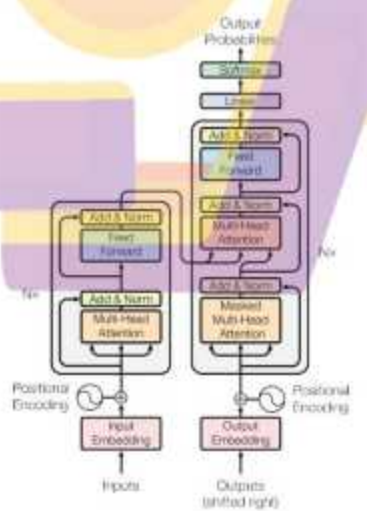
1) Encoder

Encoder berguna untuk membaca keseluruhan teks input secara bersama. *Encoder* mempunyai *stack* (tumpukan) dari N = 6 *layers* yang identik. Setiap layer memiliki dua sub-layer yaitu *self-attention layer* dan *feed-forward neural network*. *Self-attention layer* dalam *encoder* berguna untuk membantu *node* agar tidak fokus dengan kata yang sedang dilihat tetapi juga untuk

mendapatkan konteks semantik dari kata tersebut. Setiap posisi di *encoder* dapat menangani semua posisi di layer sebelumnya.

2) *Decoder*

Decoder berguna untuk menghasilkan urutan *output* yang berupa prediksi. *Decoder* juga mempunyai *stack* (tumpukan) sebanyak N = 6 *layers* yang identik. Setiap layer terdiri dari dua *sub-layer* seperti yang ada pada *encoder,* tetapi didalam *decoder* ada penambahan *attention layer* diantara dua *layer* tersebut. Dua *layer* tersebut ditambah dengan *attention layer* berguna untuk membantu *node* saat mendapatkan *key content* yang membutuhkan *attention* (Vaswani *dkk.*, 2017) dengan melakukan *multi-head attention* pada *output* dari *encoder. Self attention layer* di *decoder* membuat setiap posisi di-*decoder* dapat menangani semua posisi sebelumnya dan posisi saat itu (Vaswani *dkk.*, 2017).



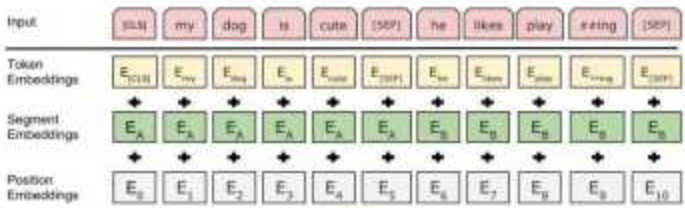
Gambar 2.2. Arsitektur Transformers

Langkah langkah berikut menunjukkan proses yang terjadi pada *encoder* dan *decoder* (Vaswani *dkk.*, 2017):

1. Setiap input kata yang memasuki *encoder* diubah menjadi sebuah *list vector* menggunakan *embeddings,* karena *self-attention layer* tidak membedakan urutan kata-kata pada sebuah kalimat. *Positional encoding* ditambahkan untuk menunjukkan posisi dari setiap kata. Setiap vektor dari kata *input* memiliki ukuran sebesar 512. Proses ini hanya berlaku di *encoder* yang berada paling bawah, sehingga *encoder* lainya dapat menerima *output* dari *encoder* pertama (Vaswani *dkk.*, 2017).
2. *Input* vektor melewati dua layer yang ada setiap *encoder* yaitu *self-attention layer* dan *feed-forward neural network*. *Self-attention layer* dibuat tiga vektor yang terbagi dalam masing masing vektor *input* yaitu *query*, *key*, dan *value vector*. Tiga vektor ini dibuat dengan mengalikan *embeddings*. Dimensi dari tiap vektor masing masing yaitu 64. *Self- attention layer* dari tiap kata dihitung dengan mengalikan *query vector* dan *key vector* (Vaswani *dkk.*, 2017). Setelah itu, nilai *self-attention* dibagi 8, karena nilai 8 adalah akar kuadrat dari dimensi tiap vektor memiliki nilai ukuran 64. Nilai *self-attention* juga dihitung dengan *softmax* sehingga setiap *value vector* akan dikali dengan nilai *softmax*. Hasilnya *value vector* dijumlahkan dan menjadi output dari *self-attention layer*. Output dari *self-attention layer* kemudian masuk ke *feed-forward* untuk setiap posisi (Vaswani *dkk.*, 2017).

2.3.7 BERT

Model *Word Embedding* pada NLP memiliki satu permasalahan, yaitu model tidak dapat menangkap makna bersifat polisemi, yaitu kata dalam bentuk yang sama tetapi memiliki arti yang berbeda, misalnya saja kata "bisa" yang berarti "dapat" tetapi juga bisa berkmakna "racun yang dihasilkan oleh binatang ular" (Young *dkk.*, 2018). Hal ini dikarenakan *Word Embedding* masih memperlakukan kata secara individual, walaupun sudah bisa melihat hubungan antara masing-masing kata lainnya. Tantangan selanjutnya pada NLP yaitu bagaimana model tidak hanya bisa memaknai suatu kata, tetapi juga bisa mengenali kata dalam konteksnya, misalnya saja ingin memprediksi kata selanjutnya dari suatu kalimat. Isu ini dapat diatasi dengan model seperti *Recurrent Neural Network* (RNN) dan *Long Short- Term Memory* (LSTM) (Young *dkk.*, 2018). Kemudian, sampai di mana kita diperkenalkan dengan *Transformers*, yaitu model atensi (*Attention Model*) yang melampaui model RNN atau LSTM karena model atensi ini memungkinkan lapisan *decoder* untuk mengenali langsung *hidden state* dari *encoder*-nya sendiri, sehingga setiap *decoder* yang ada bisa langsung mengerjakan bagiannya secara paralel dibandingkan arsitektur *neural network* yang bersifat sekuensial seperti RNN dan LSTM yang dimana setiap *decoder* harus menunggu *hidden state* dari *decoder* sebelumnya (Vaswani *dkk.*, 2017). Model inilah yang kemudian diadopsi oleh Google BERT (Devlin *dkk.*, 2018).



Gambar 2.3. Representasi Input pada Arsitektur BERT

Representasi *input* BERT ditampilkan pada Gambar 2.3. Berikut merupakan langkah-langkah representasi input dalam BERT (Khalid *dkk.*, 2023):

1. Tokenisasi : Membagi teks menjadi token-token yang terdiri dari kata-kata. BERT menggunakan tokenisasi *WordPiece*, yang berarti beberapa token dapat dibagi lagi menjadi sub-token.
2. *Token Embeddings* : BERT menambahkan dua token khusus ke awal dan akhir setiap kalimat, yaitu [CLS] dan [SEP]. Token [CLS] digunakan untuk merepresentasikan kalimat secara keseluruhan yang berada di awal kalimat, sedangkan token [SEP] di akhir kalimat digunakan untuk memisahkan kalimat dalam *input* yang berbeda dari urutan *input*.
3. Konversi Token menjadi ID : Setiap token dalam *input* kemudian dikonversi menjadi ID token yang sesuai menggunakan kamus token yang telah ditetapkan. Selanjutnya, setiap ID token dikonversi menjadi vektor dengan mengambil nilai *embedding* dari matriks *embedding* kata yang telah dilatih sebelumnya. Matriks *embedding* menggambarkan setiap kata dalam ruang vektor yang terdiri dari banyak dimensi.

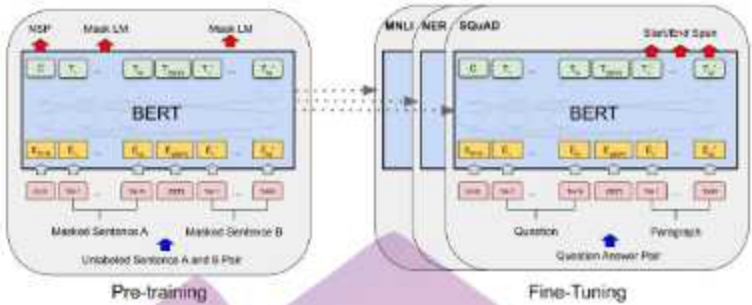
4. *Segment Embeddings* : Jika *input* terdiri dari dua kalimat, setiap token dalam *input* harus ditandai sebagai milik kalimat pertama atau kedua. Ini dilakukan dengan memberikan segmen ID ke setiap token, tergantung pada kalimat mana yang mengandung token tersebut.
5. *Position Embedding* : BERT menggunakan *Position Embedding* untuk menambahkan informasi posisi absolut ke dalam representasi token. Ini dilakukan dengan menambahkan vektor posisional yang telah ditentukan sebelumnya ke setiap vektor token.

BERT dapat mempelajari hubungan kontekstual antara kata-kata dalam sebuah kalimat yang telah dilatih pada data teks yang besar. Secara khusus, BERT dilatih pada dua tugas yaitu *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP). Dalam MLM, beberapa kata dalam sebuah kalimat secara acak diganti dengan token [MASK], dan model harus memprediksi kata asli. Tugas ini membantu model memahami makna kata dalam konteks. Sedangkan, dalam NSP, model diberikan dua kalimat dan harus memprediksi apakah kalimat kedua mungkin mengikuti kalimat pertama. Tugas ini membantu model memahami hubungan antar kalimat (Devlin *dkk.*, 2018).

Lapisan *Encoder* dari *Transformers* dilatih oleh BERT secara dua arah (*bidirectional*), dalam arti tidak hanya membaca sebuah susunan teks berdasarkan arah (kiri-ke-kanan atau kanan-ke-kiri). Hasil *output* dari model BERT ini adalah distribusi probabilitas hasil perhitungan fungsi *softmax*, yaitu pemodelan bahasa yang bersifat *unsupervised*, tanpa diberi label tetapi

memahami secara kontekstual dari kumpulan teks yang ada. BERT hanya akan memanfaatkan bagian *Encoder* dari *Transformers*.

BERT menggunakan dua paradigma pelatihan yaitu *pre-training* dan *fine tuning* yang ditunjukkan pada Gambar 2.4. *Pre-training* termasuk *unsupervised learning* karena model dilatih pada *unlabelled dataset* untuk mengekstrak pola. Model ini dilatih pada BooksCorpus (800M kata) dan English Wikipedia (2.5B kata) oleh Google. Proses *pre-training* terdiri dari dua tugas, yaitu *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP). Sedangkan, selama proses *fine tuning*, model dilatih kembali pada tugas *downstream* dengan data berlabel. *Fine-tuning* melibatkan penyesuaian parameter dari model BERT yang sudah dilatih (*pre-trained*) pada tugas tertentu dengan menggunakan data berlabel untuk mengoptimalkan kinerja model pada tugas tersebut. *Fine-tuning* dilakukan dengan menambahkan lapisan khusus untuk tugas di atas model BERT yang sudah dilatih sebelumnya (*pre-trained*) dan kemudian melatih seluruh model dari awal hingga akhir (*end-to-end*) pada data tugas khusus tersebut. Jumlah parameter di lapisan khusus jauh lebih kecil dari model BERT yang sudah dilatih sebelumnya (*pre-trained*). Selama *fine-tuning*, model dilatih dengan tingkat pembelajaran yang lebih kecil dibandingkan saat *pre-training*. Hal ini karena model yang sudah dilatih sebelumnya telah belajar fitur umum bahasa dan lapisan khusus tugas hanya perlu mempelajari fitur dari tugas *downstream* (Sun *dkk.*, 2020).



Gambar 2.4. Arsitektur BERT

BERT sangat mahal dalam hal komputasi pada tahap *pre-training*, tetapi akan sangat optimal ketika sudah sampai pada tahap *fine tuning* untuk tugas khusus atau *downstream task* (Salminen *dkk.*, 2020). Oleh karena itu Google sudah melakukan *pre-training* dengan sumber daya komputasi yang dimilikinya, sehingga para pengembang dapat memanfaatkannya cukup dengan melakukan *fine tuning*.

2.3.8 Multilingual BERT (mBERT)

Multilingual BERT atau mBERT merupakan salah satu variasi model BERT yang telah dilatih sebelumnya menggunakan korpus besar dengan bahasa yang berbeda termasuk bahasa Indonesia dan bahasa Inggris. Hal ini berarti bahwa model telah belajar untuk merepresentasikan arti kata dan kalimat dalam berbagai bahasa dan dapat disesuaikan pada tugas yang melibatkan teks dalam salah satu bahasa tersebut. Multilingual BERT sangat bagus dalam transfer model lintas bahasa *zero-shot*, misalnya penggunaan *dataset* berlabel pada tugas *named entity recognition* dalam satu bahasa juga digunakan untuk menyempurnakan model dalam bahasa lain (Pires *dkk.*, 2019).

Disebutkan bahwa model tersebut dapat menemukan pasangan terjemahannya yang secara topologis mirip, namun representasi tersebut menunjukkan kekurangan sistematis yang mempengaruhi pasangan bahasa tertentu (Pires *dkk.*, 2019). Multilingual BERT memiliki arsitektur yang sama dengan BERT-*base* yaitu 12-*layer*, 768-*hidden*, 12-*heads*, dan 110M parameter. Multilingual BERT memiliki dua model yaitu *cased* dan *uncased*, dimana model *cased* dilatih pada korpus wikipedia 104 bahasa yang berbeda dan membedakan antara huruf besar dan kecil dalam teks sehingga dapat memberikan lebih banyak informasi tentang arti kata-kata dalam beberapa bahasa. Sedangkan, model *uncased* dilatih pada korpus dengan 102 bahasa yang berbeda dan sebelum diproses oleh model, teks harus diubah menjadi huruf kecil (Pires *dkk.*, 2019).

Multilingual BERT dapat digunakan untuk berbagai tugas pemrosesan bahasa alami, seperti klasifikasi teks, *named entity recognition*, *question-answering* dan masih banyak lagi. Dapat disempurnakan (*fine-tuned*) pada tugas tertentu menggunakan *dataset* berlabel dan mencapai kinerja terbaik pada banyak *benchmarks*. Salah satu manfaat menggunakan Multilingual BERT adalah dapat menangani banyak bahasa dalam satu model. Hal ini dapat menghemat banyak waktu dan sumber daya, karena model yang sama dapat digunakan dalam berbagai bahasa (Devlin *dkk.*, 2018).

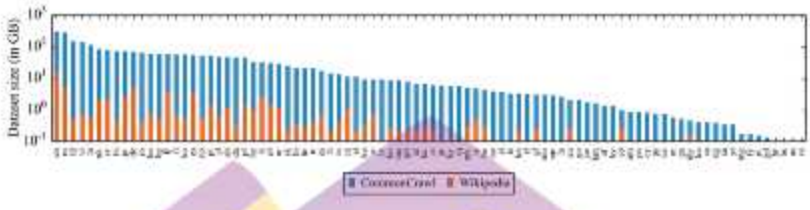
Tabel 2.2. Perbandingan BERT dan Multilingual BERT

Pre-Trained Model	Spesifikasi
BERT-*base Uncased*	12-*layer*, 768-*hidden*, 12-*heads*, 110M *parameters*. *Trained on lower-cased English text.*
BERT-*base Cased*	12-layer, 768-hidden, 12-heads, 110M parameters. *Trained on cased English text.*
BERT-*large Uncased*	24-layer, 1024-hidden, 16-heads, 340M parameters. *Trained on lower-cased English text.*
BERT-*large Cased*	24-layer, 1024-hidden, 16-heads, 340M parameters. *Trained on cased English text.*
Multilingual BERT *Uncased*	12-layer, 768-hidden, 12-heads, 110M parameters. *Trained on lower-cased text in the top 102 languages with the largest Wikipedias*
Multilingual BERT *Cased*	12-layer, 768-hidden, 12-heads, 110M parameters. *Trained on cased text in the top 104 languages with the largest Wikipedias*

2.3.9 XLM-RoBERTa (XLM-R)

XLM-RoBERTa atau *Cross Lingual Model – RoBERTa* adalah pengembangan dari XLM dan mBERT yang merupakan penelitian *Natural Language Processing* (NLP) multibahasa. XLM-RoBERTa menggunakan *transformer-based multilingual mask language model* (MLM) yang sudah dilakukan *pre-trained* pada teks dengan 100 bahasa dan mampu menunjukkan performa yang sangat baik pada *cross-lingual classification, sequence labeling,* dan *question answering* (Conneau *dkk.*, 2020). XLM-RoBERTa merupakan salah satu algoritma *machine learning* di dalam *library transformers*, sebuah *library* yang menyediakan ribuan *pre-tained model* untuk melakukan pekerjaan seperti klasifikasi teks, *named entity recognition, information extraction, question answering*, dan lain-lain. Pengembangan XLM-RoBERTa memiliki tujuan untuk meningkatkan kemampuan mesin atau komputer dalam

melakukan *multilingual Natural Language Processing* (NLP) terutama pada *low-resource languages*.



Gambar 2.5. Jumlah data GiB (skala log) untuk bahasa yang muncul di korpus Wiki-100 yang digunakan untuk mBERT dan XLM-100 dibandingkan CC-100 yang digunakan untuk XLM-R (Conneau *dkk.*, 2020)

Gambar 2.6. Pratinjau ukuran data 27 bahasa dari 100 bahasa pada korpus CC-100

Tidak seperti XLM, pengembangan XLM-RoBERTa justru menghindari metode yang digunakan oleh XLM yaitu *Translation Language Modeling* (TLM). XLM-RoBERTa menggunakan metode *Mask Language Modeling* (MLM) karena pengembangannya lebih fokus pada *unsupervised learning*. MLM merupakan bagian dari *Transformer-based model* yang memiliki tujuan untuk memprediksi token yang hilang dari sebuah input dan merekonstruksi ulang urutan input yang sesungguhnya. Namun, MLM tidak memiliki akses terhadap keseluruhan input tersebut, melainkan hanya memiliki

akses terhadap *masked token*. Konsep MLM ini juga digunakan pada BERT, mBERT, RoBERTa dan XLM (MLM pada XLM dikombinasikan dengan *language modeling* yang lain).

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Adapun jenis, sifat dan pendekatan yang digunakan pada penelitian ini adalah :

a) Jenis Penelitian

Jenis penelitian yang digunakan pada penelitian ini adalah penelitian eksperimental. Penelitian eksperimental adalah metode sistematis guna membangun hubungan yang mengandung fenomena sebab akibat untuk mencari pengaruh perlakukan tertentu terhadap yang lain dalam kondisi terkendali. Dalam penelitian ini eksperimen dilakukan dengan menggunakan suatu objek berupa *dataset* publik dalam Bahasa Indonesia yang disempurnakan (*fine-tuning*) pada tiga *pre-trained* model multibahasa berbasis BERT sebagai model multilingual *named entity recognition* untuk mengekstraksi lokasi dan waktu. Ketiga model tersebut kemudian dievaluasi dan divalidasi untuk menilai keakuratan dalam mengekstrak lokasi dan waktu.

b) Sifat Penelitian

Penelitian yang dilakukan bersifat deskriptif, karena menggambarkan suatu objek yang diteliti dan menjabarkan hasil eksperimen yang dilakukan. Dimana, penelitian ini menganalisis penggunaan suatu *dataset* publik dalam bahasa Indonesia yang

disempurnakan (*fine-tuning*) pada tiga *pre-trained* model multibahasa berbasis BERT sehingga diketahui efektivitas penggunaan *dataset* publik dan melihat model mana yang memiliki hasil evaluasi dan validasi terbaik dalam mengekstrak lokasi dan waktu.

c) Pendekatan Penelitian

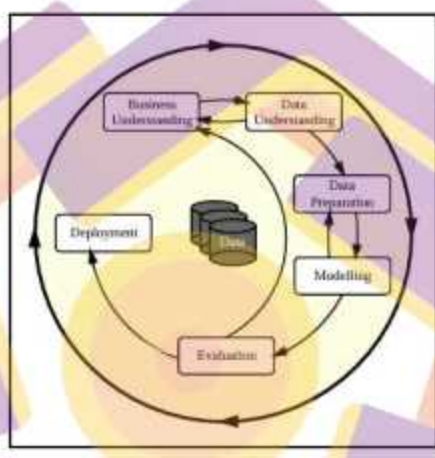
Penelitian ini menggunakan pendekatan kuantitatif karena penelitian menghasilkan nilai yang objektif dalam skala numerik. Penelitian yang dilakukan menghasilkan hasil evaluasi *fine-tuning* model berdasarkan perhitungan matematis *confusion matrix* seperti *accuracy, precision, recall* dan *f1-score*. Selain itu, penelitian juga menghasilkan nilai *accuracy* berdasarkan validasi kinerja model menggunakan beberapa sampel *tweet* dalam 5 bahasa.

3.2. Metode Pengumpulan Data

Dataset utama pada penelitian ini dibatasi berdasarkan *dataset* format *arrow* yang diambil dari *huggingface*. Data dikumpulkan dalam bentuk *sequence tokens* disertai label (*tags*). Dimana, *tokens* merupakan pecahan kata-kata dari suatu kalimat, sedangkan *tags* merupakan id label entitas untuk setiap pecahan kata. Setiap data pada *dataset* disimpan dalam bentuk larik (*array*). Selain itu, beberapa *tweet* dengan kata kunci "kebakaran hutan" dalam 5 bahasa diperoleh dari Twitter (X) menggunakan teknik *scrapping* dengan memanfaatkan *tweet-harvest* dalam bahasa pemrograman *python*.

3.3. Metode Analisis Data

Metode analisis data pada penelitian ini menggunakan model standar CRISP-DM (*Cross-Standard Industry for Data Mining*) untuk mendapatkan hasil evaluasi dan validasi model *named entity recognition*. Metode ini terdiri dari 6 (enam) tahapan (Wirth & Hipp, 2000), diantaranya:



Gambar 3.1. Tahapan CRISP-DM

1. *Business Understanding*

Pada tahap pertama dilakukan penentuan tujuan penelitian dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan berdasarkan rumusan masalah, menerjemahkan tujuan dan batasan menjadi formula dari permasalahan *data mining* serta menyiapkan strategi awal untuk mencapai tujuan. Tahapan ini juga dapat dilakukan melalui diskusi dengan pakar dan studi literatur.

2. *Data Understanding*

Tahap kedua melakukan pengumpulan data yang diperlukan, lalu melakukan identifikasi terhadap data untuk mencari pengetahuan awal (*prior knowledge*) dan menganalisis data tersebut sehingga dapat mencapai tujuan.

3. *Data Preparation*

Pada tahap ketiga dilakukan pengolahan data sehingga dapat menghasilkan model sesuai kebutuhan. Tahap ini sangat menentukan hasil pemodelan, karena model yang baik dan akurat berawal dari data yang dipersiapkan dengan baik.

4. *Modelling*

Pada tahap ini menentukan dan mengaplikasikan teknik pemodelan yang sesuai dan menentukan parameter model untuk mengoptimalkan hasil. Jika diperlukan, proses dapat kembali ke tahapan *data preparation* sehingga spesifikasi kebutuhan model sesuai dengan tujuan.

5. *Evaluation*

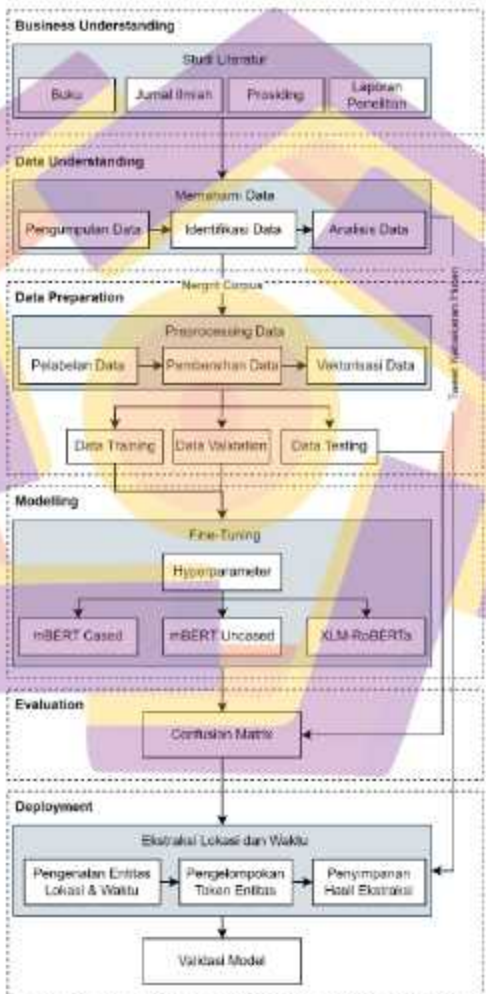
Tahap ini mengevaluasi model-model yang dibangun untuk mendapatkan kualitas dan performa sebelum digunakan atau disebarkan (*deployment*). Jika diperlukan, proses ini bisa kembali ke tahap awal untuk menyesuaikan ulang spesifikasi kebutuhan sehingga sesuai dengan tujuan.

6. *Deployment*

Sebelum model benar-benar digunakan dan disebarkan, pada tahap ini model terlebih dahulu divalidasi berdasarkan skenario nyata (riil) menggunakan beberapa validator untuk mengetahui tingkat akurasi model dalam mengenali entitas lokasi dan waktu. Terakhir, model dengan akurasi terbaik dapat diterapkan sepenuhnya.

3.4. Alur Penelitian

Dalam penelitian ini, diperlukan adanya tahapan-tahapan yang diurutkan secara sistematis agar pelaksanaan penelitian dapat berjalan dengan baik. Rangkaian alur pada penelitian ini dapat dilihat pada Gambar 3.2.



Gambar 3.2. Alur Penelitian

Adapun penjelasan mengenai alur penelitian diatas, yaitu:

1. *Business Understanding*

Tahap ini mengidentifikasi perkembangan sistem tanggap darurat bencana yang menggunakan *social media sensing* berdasarkan *literature review* untuk merumuskan masalah dan menentukan tujuan penelitian. Strategi yang digunakan untuk menyelesaikan tujuan penelitian adalah pemanfaatan *dataset* publik dalam Bahasa Indonesia dengan menggunakan pendekatan *transfer learning* dalam sistem manajemen bencana kebakaran hutan secara multibahasa untuk mengekstrak lokasi dan waktu kebakaran hutan pada tingkat regional dan global dengan penelitian ini sebagai kajian untuk mencari model multilingual *named entity recognition* terbaik. Pada tahap ini juga ditentukan strategi-strategi untuk mendapatkan data yang dibutuhkan.

2. *Data Understanding*

Tahap ini melakukan pengumpulan data yang dibutuhkan, yaitu *dataset nergrit corpus* dari *huggingface* dan beberapa data *tweet* dari Twitter (X). Pertama, *dataset nergrit corpus* diidentifikasi dan dianalisis data dan label entitasnya untuk memperoleh informasi sehingga memenuhi spesifikasi kebutuhan dalam mengekstrak lokasi dan waktu. Kedua, beberapa *tweet* dengan kata kunci "kebakaran hutan" dalam 5 bahasa dikumpulkan menggunakan metode *scrapping* dari Twitter (X) untuk diidentifikasi dan dianalisis sehingga dapat digunakan untuk validasi model.

3. *Data Preparation*

Pengolahan data pada tahap ini dilakukan berdasarkan hasil analisis dari tahap sebelumnya untuk menyiapkan data ke pemodelan. Pengolahan data pada *dataset nergrit corpus* dilakukan dengan pelabelan data, pembersihan data dan vektorisasi data. Sedangkan, beberapa *tweet* hasil *scrapping* disiapkan dengan hanya memilih data pada atribut yang berisi teks media sosial.

4. *Modelling*

Pada tahap ini, pemodelan dilakukan dengan cara *fine-tuning* pada beberapa *pre-trained* model multibahasa berbasis BERT dengan menggunakan *data train* dan *data validation* yang sudah disipakan di tahap sebelumnya. Tahapan ini menghasilkan 3 (tiga) model multilingual *named entity recognition* yang mampu mengenali entitas lokasi dan waktu dari data teks.

5. *Evaluation*

Model multilingual *named entity recognition* yang dihasilkan di tahap sebelumnya kemudian diuji dengan menggunakan *data testing* untuk mengevaluasi performa masing-masing model dalam melakukan klasifikasi token entitas lokasi dan waktu. Alat ukur yang digunakan untuk mendapatkan performa model adalah *confusion matrix* yang menghasilkan nilai akurasi.

6. *Deployment*

Pada tahap ini model multilingual *named entity recognition* digunakan pada skenario nyata untuk mengekstrak lokasi dan waktu dari *tweet* kebakaran hutan dengan cara memanggil masing-masing model. Setelah model dipanggil, kemudian model melakukan pengenalan entitas lokasi dan waktu dari *tweet* kebakaran hutan yang menghasilkan token (kata) entitas. Entitas lokasi atau waktu yang dikenali

biasanya terdiri dari beberapa token, sehingga perlu untuk dilakukan pengelompokan token pada masing-masing entitas. Hasil dari pengelompokan token entitas kemudian disimpan sebagai hasil ekstraksi lokasi dan waktu dari *tweet* kebakaran hutan. Terakhir, hasil ekstraksi lokasi dan waktu divalidasi secara manual oleh beberapa validator untuk mengetahui tingkat akurasi model multilingual *named entity recognition* dalam mengekstrak lokasi dan waktu dari *tweet* kebakaran hutan dalam lima bahasa.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Data Understanding

4.1.1. Pengumpulan Data

Dataset utama yang digunakan untuk *fine-tuning* model pada penelitian ini adalah *nergrit corpus* yang diambil menggunakan *library datasets* dari *huggingface* (https://huggingface.co/datasets/id_nergrit_corpus). Tampilan *dataset nergrit corpus* di *huggingface* dapat dilihat pada Gambar 4.1.

Gambar 4.1. Pratinjau *dataset nergrit corpus* di *huggingface*

Huggingface sendiri merupakan sumber pustaka terbuka yang mengembangkan berbagai perkakas termasuk model, *library* dan *dataset* untuk aplikasi *deep learning*. Proses pengambilan *dataset nergrit corpus* dilakukan dengan menggunakan perintah sebagai berikut:

```
from datasets import load_dataset
nergrit = load_dataset("id_nergrit_corpus", 'ner')
```

Pada tahapan ini beberapa *tweet* dengan kata kunci "kebakaran hutan" dalam 5 bahasa diperoleh menggunakan metode *scrapping* dari Twitter (X) dimana jumlah sampel data yang diperoleh dibatasi kurang lebih sebanyak 500 data saja untuk masing-masing bahasa. Proses *scrapping tweet* menggunakan *library tweet-harvest* dan dieksekusi di *Google Colabs* sehingga menghasilkan berkas csv (*comma separated values*).

Gambar 4.2. Pratinjau hasil *scrapping tweet* dalam Bahasa Indonesia

Rekapitulasi hasil *scrapping tweet* pada 5 bahasa disertai kata kuncinya dapat dilihat pada Tabel 4.1.

Tabel 4.1. Rekapitulasi hasil *scrapping tweet*

No	Bahasa Tweet	Kata Kunci	Jumlah Data
1	Bahasa Indonesia	Kebakaran Hutan	482
2	Bahasa Inggris	Forest Fire	511
3	Bahasa Spanyol	Incendio Forestal	510
4	Bahasa Italia	Incendi Forestali	509
5	Bahasa Slovakia	Lesné Požiare	314

Tabel 4.1 memperlihatkan perbedaan jumlah data yang diperoleh disebabkan karena perbedaan tren kebakaran hutan dan relevansi kata kunci yang digunakan pada masing-masing bahasa untuk memperoleh *tweet*. Pada penelitian ini kata kunci kebakaran hutan yang digunakan untuk memperoleh *tweet* dibatasi berdasarkan hasil terjemahan kata "kebakaran hutan" dalam Bahasa Indonesia ke bahasa tujuan dengan menggunakan *google translate*.

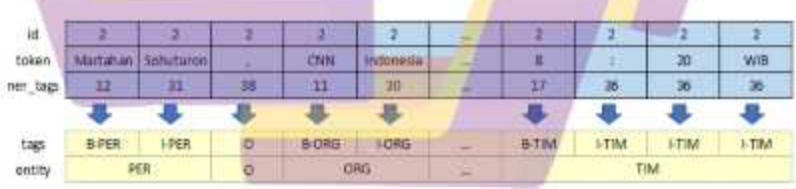
Proses *scrapping tweet* dilakukan dengan menggunakan perintah sebagai berikut:

```
# Install Node.js (because tweet-harvest built using Node.js)
!pip install pandas

!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings

!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-
repo.gpg.key | sudo gpg --dearmor -o
/etc/apt/keyrings/nodesource.gpg

!NODE_MAJOR=20 && echo "deb [signed-
by=/etc/apt/keyrings/nodesource.gpg]
https://deb.nodesource.com/node_$NODE_MAJOR.x nodistr main" |
sudo tee /etc/apt/sources.list.d/nodesource.list

!sudo apt-get update
!sudo apt-get install nodejs -y

!node -v

twitter_auth_token = {
    'account': ******
}

data = {
    'lang': ['id', 'en', 'es', 'it', 'sk'],
    'keyword': [
        'kebakaran hutan',
        'forest fire',
        'incendio forestal',
        'incendi forestali'
        'lesné požiare',
    ],
    'limit': 500,
    'filename': 'kebakaran_hutan',
    'format': 'csv'
}

!npx --yes tweet-harvest@2.2.8 -o

"{data['filename']}_{data['lang'][1]}.{data['format']}" -s

"{data['keyword'][1]}" -l {data['limit']} --token

{twitter_auth_token['account']}
```

4.1.2. Identifikasi Data

Nergrit corpus merupakan *dataset* publik dalam Bahasa Indonesia yang dibuat oleh PT. Gria Innovation Technology (GRIT) dari kumpulan artikel situs berita di Indonesia untuk untuk tujuan umum tugas pengenalan entitas bernama dengan format anotasi BIO (*Beginning-Inside-Outside*), dimana B menunjukkan awal suatu entitas, I menunjukkan bagian dari suatu entitas dan O berarti bukan suatu entitas (non-entitas). Pelabelan data pada *dataset* ini dilakukan secara manual dengan melibatkan beberapa anotator (PT Gria Inovasi Teknologi, 2019). Selain itu, komponen data artikel berita pada *dataset* ini terdiri dari alfabet, numerik dan karakter khusus yang sesuai dengan sumber artikel berita aslinya, sedangkan *emoticon* dan tautan telah dihilangkan secara terprogram.

Pada penelitian ini, *nergrit corpus* yang dibuat untuk tujuan umum tugas pengenalan entitas bernama akan digunakan pada domain kebakaran hutan untuk mengetahui efektivitas pengunaannya, meskipun tidak mempunyai data dengan konteks kebakaran hutan.



Gambar 4.3. Contoh data pada *dataset nergrit corpus*

Pada Gambar 4.3 yang memperlihatkan salah satu contoh data pada *dataset nergrit corpus* terdapat 3 atribut data yaitu *id*, *token* dan *ner_tags*. Atribut *id* merupakan indeks data pada *dataset*, sedangkan atribut *token* merupakan pecahan kata dari suatu kalimat dan atribut *ner_tags* merupakan *id tags* atau id label entitas

pada suatu token. Setiap id label pada token merepresentasikan suatu label entitas misalnya token "CNN" memiliki id label 11 yang merupakan tag B-ORG dan token "Indonesia" memiliki id label 30 yang merupakan tag I-ORG dimana kedua tag tersebut merupakan representasi dari entitas Organisasi "CNN Indonesia". Selain itu, *dataset* ini terdiri dari 19 entitas seperti *Cardinal*, *Date*, *Event*, *Facility*, *Geopolitical Entity*, *Law Entity*, *Location*, *Money*, *Political Organization*, *Ordinal*, *Organization*, *Person*, *Percent*, *Product*, *Quantity*, *Religion*, *Time*, *Work of Art* dan *Language*. Tabel 4.2 memperlihatkan penjelasan entitas pada *dataset nergrit corpus*.

Tabel 4.2. Penjelasan entitas pada *dataset nergrit corpus*

No	Entitas	Label	Deskripsi	Contoh Data
1	Cardinal	CRD	Menunjukan suatu angka	1, 2, 3, 4, 5, dst.
2	**Date**	DAT	Format tanggal (tanggal, bulan dan tahun)	Senin, Oktober, 1997, dst.
3	Event	EVT	Nama kejadian / kegiatan	Operasi lalu lintas lodaya, pertunjukan sulap, dst.
4	Facility	FAC	Gedung, bandara, jembatan, dll.	Gedung BRI, Bandara BIJB, dst.
5	**Geopolitical Entity**	GPE	Wilayah administratif	Indonesia, jakarta, majalengka, dst.
6	Law Entity	LAW	Dokumen hukum	Undang-undang, peraturan, dst.
7	**Location**	LOC	Nama suatu tempat non administratif (non-GPE), gunung, hutan, dll.	Gunung bromo, hutan nasional ujung kulon, dst.
8	Money	MON	Jumlah uang disertai nilai mata uang	Rp. 2 miliar, Rp. 1.000, dst.
9	Political Organization	POL	Nama organisasi politik	Partai PDI-P, Golkar, dst.
10	Ordinal	ORD	Menunjukan suatu posisi dalam urutan	Kesatu, kedua, dst.
11	Organization	ORG	Perusahaan, instansi, institusi, dll.	Divisi Humas Polri, BPBD, dst.
12	Person	PER	Nama orang	Andi, budi, dst.
13	Percent	PRC	Persentase (termasuk %)	10%, 20%, dst.
14	Product	PRD	Kendaraan, senjata, makanan, dll. (Bukan jasa)	Motor beat, mobil agya, seblak, dst.

Tabel 4.2. Penjelasan entitas pada *dataset nergrit corpus* (lanjutan)

No	Entitas	Label	Deskripsi	Contoh Data
15	Quantity	QTY	Sejumlah angka disertai satuannya	1 hari, 2 bulan, 2 meter, dst.
16	Religion	REG	Nama agama	Islam, kristen, dst.
17	**Time**	TIM	Format waktu (jam, menit, detik)	08:00 WIB, 19:00 WIT, dst.
18	Work of Art	WOA	Hasil suatu pekerjaan seni	Buku data science, patung soekarno, dst.
19	Language	LAN	Bahasa	Bahasa Indonesia, Bahasa Inggris, dst.

Dataset nergrit corpus terdiri dari 17.452 data dengan proporsi data latih sebanyak 12.532 (72%), data uji sebanyak 2.399 (14%) dan data validasi sebanyak 2.521 (14%). *Dataset* ini mempunyai jumlah entitas sebanyak 60.189 data berdasarkan tag "B-". Distribusi entitas pada *dataset nergrit corpus* ditunjukkan pada Gambar 4.4.

Gambar 4.4. Grafik distribusi entitas pada *dataset nergrit corpus*

Pada Gambar 4.4 menunjukkan bahwa entitas *Person* dengan label "PER" merupakan entitas yang memiliki data paling banyak, sedangkan entitas *Language* dengan label "LAN" merupakan entitas yang memiliki data paling sedikit. Sementara itu, proses *scrapping tweet* di tahap sebelumnya menghasilkan berkas

csv (*comma separated values*) yang terdiri dari atribut *created_at, id_str, full_text, quote_count, reply_count, favorite_count, lang, user_id_str, conversation_id_str, username* dan *tweet_url*. Namun, berdasarkan Gambar 4.2 bahwa atribut yang berisi data teks *tweet* terdapat pada atribut *full_text*.

Tabel 4.3. Contoh *tweet* kebakaran hutan dengan entitas lokasi dan waktu

No	Tweet	Bahasa	Keterangan
1	Kebakaran hutan dan lahan (karhutla) terjadi di **lereng Gunung Panderman, Kota Batu, Jawa Timur**, **Selasa, 21 November 2023**. Titik api diperkirakan muncul pada pukul **15.30 WIB**. #Karhutla https://t.co/G6qLxulQjN	Indonesia	Tweet kebakaran hutan yang menampilkan lokasi, tanggal dan waktu terjadinya kebakaran hutan
2	Kaget pas lihat ke arah balkon kamar hotel keliatan ada kebakaran hutan.. ternyata kebakaran di **lereng gunung Panderman Batu Malang** sejak kemarin karena samberan petir. Semoga api nya segera padam yaa 🙏🙏 https://t.co/O3C62z9tmA	Indonesia	Tweet kebakaran hutan yang menampilkan lokasi terjadinya kebakaran hutan
3	Kebakaran hutan dan lahan (karhutla) terjadi di **Gunung Kawi, Kabupaten Malang, Jawa Timur, Selasa** malam, **7 November 2023**. #GunungKawi https://t.co/S91G2d0C5Y	Indonesia	Tweet kebakaran hutan yang menampilkan lokasi dan tanggal terjadinya kebakaran hutan
4	🔥 ***16:14*** *Incendio forestal en* ***Traslasierra****: hay una amplia zona afectada en la localidad de* ***San Lorenzo*** *con riesgo de interfase Trabaja personal de bomberos de* ***Mina Clavero****, ETAC y vecinos autoconvocados. Es EXTREMO el riesgo de incendios en* ***Córdoba****.* *https://t.co/knkPvoCx81* (🔥 **16:14** Kebakaran hutan di **Traslasierra**: ada area yang terkena dampak besar di **kota San Lorenzo** dengan risiko antarmuka Personil dari **Mina** Clavero, ETAC dan tetangga yang berkumpul sendiri sedang bekerja. Risiko kebakaran di **Cordoba** adalah EKSTRIM. https://t.co/knkPvoCx81)	Spanyol	Tweet kebakaran hutan yang menampilkan lokasi dan waktu terjadinya kebakaran hutan

Tabel 4.3. memperlihatkan contoh *tweet* kebakaran hutan hasil *scrapping* yang memiliki entitas lokasi, tanggal dan waktu terjadinya kebakaran hutan. Pada *tweet* pertama, informasi lokasi yang ditampilkan terdiri dari wilayah non-administratif dan wilayah administratif seperti "lereng Gunung Panderman" dan

"Kota Batu, Jawa Timur". Selain itu, informasi waktu pada beberapa *tweet* diatas bisa saja terdiri dari nama hari, tanggal, bulan, tahun, jam, menit maupun zona waktu seperti "Selasa, 21 November 2023", "15.30 WIB" dan "16:14".

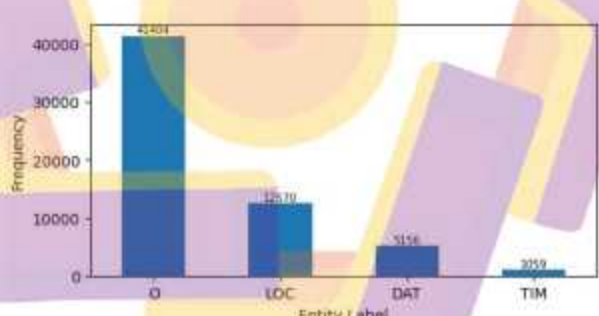
4.1.3. Analisis Data

Berdasarkan hasil identifikasi data *tweet* yang ditunjukkan pada Tabel 4.3, maka terdapat 3 data entitas yang dapat digunakan pada *dataset nergrit corpus* untuk mengekstrak lokasi dan waktu yaitu *location, date* dan *time*. Namun, ketiga entitas tersebut hanya akan mampu menangkap informasi nama tempat non-administratif dan waktunya saja. Hal itu akan menyebabkan informasi yang didapatkan dari hasil ekstraksi menjadi kurang lengkap, karena informasi kebakaran hutan di medial sosial bisa saja disertai nama negara, provinsi, kota ataupun desa seperti ditunjukkan pada Tabel 4.3. Oleh karena itu, entitas *location* diperbarui dengan menggabungkan entitas *geopolitical entity* (wilayah administratif seperti negara, provinsi, kabupaten, kecamatan, desa) dan *location* (wilayah non-administratif seperti nama gunung, hutan, lahan, dll.) untuk mendapatkan informasi secara lengkap dimana lokasi kebakaran hutan terjadi. Sementara itu, entitas *date* (format tanggal) dan *time* (format waktu) dipertahankan untuk mendapatkan informasi waktu kebakaran hutan berdasarkan nama hari, tanggal, bulan, tahun, jam, menit, detik dan zona waktu. Sisanya, 15 entitas diubah menjadi non-entitas (O) karena tidak akan digunakan.

4.2. Data Preparation

4.2.1. Pelabelan Data

Pelabelan data dilakukan karena hasil analisis data pada tahap sebelumnya menunjukkan bahwa entitas *location* diperbarui dengan menggabungkan entitas *location* dan *geopolitical entity*. Proses tersebut mengubah jumlah data pada entitas *location* yang awalnya sebanyak 3.485 data bertambah 9.085 data sehingga menjadi 12.570 data. Sementara itu, entitas *date* dan *time* dipertahankan berdasarkan hasil analisis data. Sisanya, 15 entitas dilakukan pelabelan data kembali dengan mengubahnya menjadi non-entitas sehingga menambah jumlah data non-entitas sebanyak 41.404 entitas. Gambar 4.5 menunjukkan hasil pelabelan data pada *dataset*.



Gambar 4.5. Grafik hasil pelabelan data

Proses pelabelan data dilakukan pada semua pembagian data (*train, test, validation*) sehingga proporsi jumlah data pada *dataset* masih sama dan yang berubah hanya label entitas nya saja. Proses pelabelan data dilakukan dengan menggunakan perintah sebagai berikut:

```
from datasets import load_dataset, concatenate_datasets,
Sequence, ClassLabel, Value

nergrit = load_dataset("id_nergrit_corpus", 'ner')

label_names = ['O', 'B-LOC', 'I-LOC', 'B-DAT', 'I-DAT', 'B-TIM',
'I-TIM']

id2label = {i: label for i, label in enumerate(label_names)}
label2id = {v: k for k, v in id2label.items()}

dict_nergrit = {
    7   : 1, # B-LOC
    26  : 2, # I-LOC
    4   : 1, # B-GPE
    23  : 2, # I-GPE

    1   : 3, # B-DAT
    20  : 4, # I-DAT
    17  : 5, # B-TIM
    36  : 6, # I-TIM

    38  : 0, # O (non-entitas)
}

def pelabelan_data(data):
    ner_tags = []
    ner_tokens = []

    for tags, tokens in zip(data['ner_tags'], data['tokens']):
        tag_list, token_list = [], []

        for tag, token in zip(tags, tokens):
            if tag in [*dict_nergrit][:-1]:
                tag_list.append(dict_nergrit[tag])
                token_list.append(token)
            else:
                tag_list.append(0)
                token_list.append(token)

        ner_tags.append(tag_list)
        ner_tokens.append(token_list)

    return { 'tokens': ner_tokens, 'ner_tags': ner_tags }

nergrit = nergrit.map(pelabelan_data, batched=True)
```

4.2.2. Pembersihan Data

Pembersihan data dilakukan karena setelah proses pelabelan data terdapat beberapa baris data pada *dataset* yang semua tokennya berlabel non-entitas,

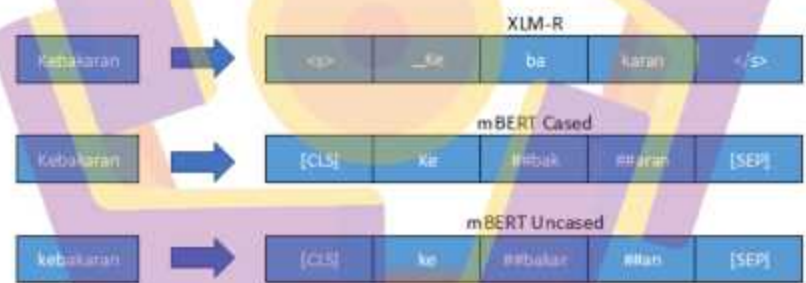
sehingga pembersihan data perlu dilakukan karena data tersebut tidak lagi memiliki nilai dan tidak relevan. Setelah proses pembersihan data, proporsi data pada *dataset* sekarang berjumlah 9.090 data dengan proporsi data *train, test* dan *validation* masing-masing sebanyak 6.611, 1.228 dan 1.251. Proses pembersihan data dilakukan dengan menggunakan fungsi *filtering* sebagai berikut:

```
dict_nergrit = {
    7   : 1, # B-LOC
    26  : 2, # I-LOC
    4   : 1, # B-GPE
    23  : 2, # I-GPE

    1   : 3, # B-DAT
    20  : 4, # I-DAT
    17  : 5, # B-TIM
    36  : 6, # I-TIM

    38  : 0, # O (non-entitas)
}


nergrit = nergrit.filter(lambda data: any(tag in
[*dict_nergrit.values()][:-1] for tag in data['ner_tags']))
```

4.2.3. Vektorisasi Data

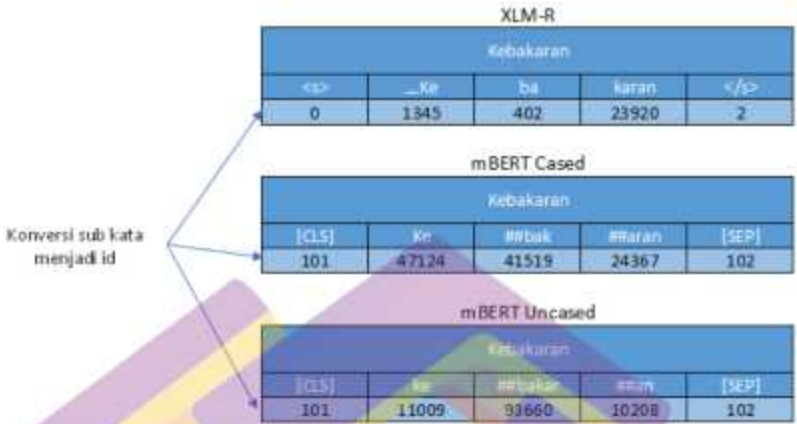
Vektorisasi data adalah tahap konversi data teks dalam *dataset* menjadi vektor numerik yang merepresentasikan data teks tersebut sehingga komputasi pembelajaran mesin dapat dilakukan. Teknik vektorisasi pada penelitian ini dilakukan pada *data training, data validation* dan *data testing* dengan menggunakan fungsi *Tokenizer* yang disediakan oleh *library transformers*. *Tokenizer* dilakukan dalam beberapa tahap yaitu tokenisasi berbasis sub kata, konversi token menjadi *id*, menandai atensi, pelabelan token, *padding* dan *trucation* (pemangkasan). Proses diawali dengan memecah token (kata) menjadi sub token (sub kata) berdasarkan teknik tokenisasi yang digunakan oleh masing-masing *pre-*

trained model. Teknik tokenisasi pada Multilingual BERT (mBERT) menggunakan *WordPiece* (Wu *dkk.*, 2016) berbasis *shared vocabulary* dengan jumlah kosakata (*vocabulary*) sebanyak 110.000 sub kata. Sedangkan, XLM-RoBERTa (XLM-R) menggunakan *SentencePiece* (Kudo & Richardson, 2018) berbasis *shared vocabulary* juga dengan jumlah kosakata sebanyak 250.000 sub kata. Selain itu, token khusus akan diberikan pada setiap awal dan akhir hasil tokenisasi untuk memformat urutan kalimat, seperti <s> dan </s> pada XLM-R serta [CLS] dan [SEP] pada mBERT. Perbedaan teknik tokenisasi, *cased* atau *uncased* dan kamus kosakata dapat mempengaruhi pemberian token khusus dan hasil tokenisasi. Contoh kata "Kebakaran" yang sudah melalui proses tokenisasi sub kata dapat dilihat pada Gambar 4.6.



Gambar 4.6. Contoh tokenisasi sub kata

Tahap kedua mengkonversi sub kata menjadi *id* berdasarkan kamus kosakata pada masing-masing *pre-trained* model. Contoh hasil konversi tokenisasi sub kata menjadi *id* dapat dilihat pada Gambar 4.7.

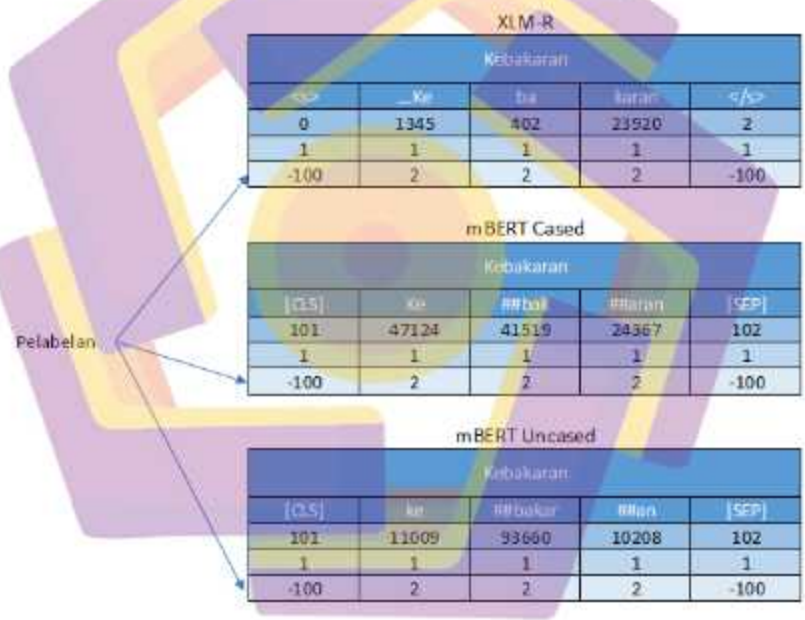


Gambar 4.7. Contoh konversi tokenisasi sub kata menjadi *id*

Tahap ketiga menandai atensi (*attention*) dengan menggunakan nilai 1. Atensi diberikan untuk menandai sub kata mana yang dipertimbangkan oleh model untuk mendapatkan konteks kalimat. Contoh hasil penandaan atensi pada sub kata dapat dilihat pada Gambar 4.8.

Gambar 4.8. Contoh penandaan atensi

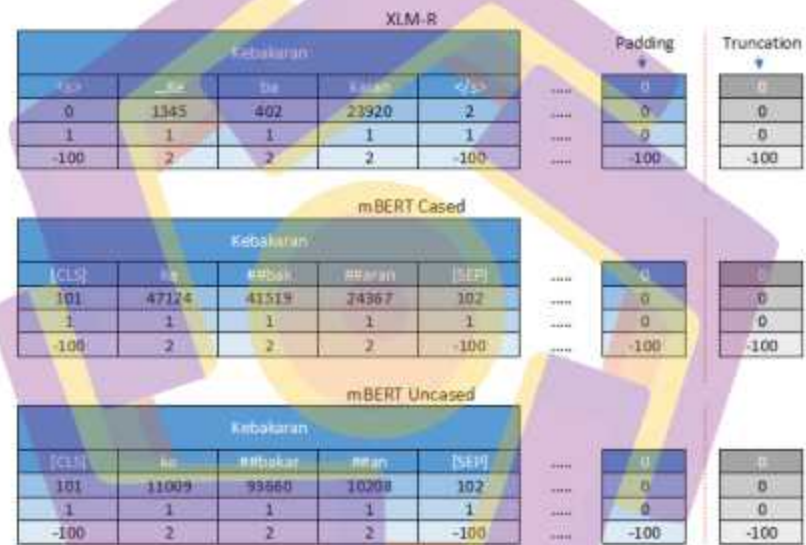
Tahap keempat adalah pelabelan dengan cara membagi label pada suatu kata kepada sub kata. Misalnya, apabila kata "Kebakaran" memiliki label entitas *event* atau suatu kejadian atau "B-EVT" dengan *id* label entitas 2, maka sub kata hasil tokenisasinya akan memiliki *id* label entitas yang sama. Selain itu, nilai -100 digunakan untuk menandai token yang tidak memiliki label entitas seperti pada token khusus dan *padding*. Contoh hasil pelabelan pada tokenisasi sub kata dapat dilihat pada Gambar 4.9.



Gambar 4.9. Contoh pelabelan sub kata

Tahap terakhir adalah *padding* dan *truncation* pada sub kata hasil tokenisasi sehingga panjangnya menjadi sama. Panjang *input* maksimal pada vektorisasi data ditentukan oleh kalimat terpanjang pada *dataset*, namun panjang maksimal bawaan dari *pre-trained* modelnya adalah sebanyak 512. *Truncation* atau pemotongan

dilakukan apabila terdapat data yang melebihi panjang *input* maksimal, sedangkan penambahan nilai *padding* dilakukan setelah tokenisasi sub kata untuk menyamakan panjang *input* maksimal. Panjang *input* yang sama mampu mempercepat proses pelatihan karena dapat memanfaatkan teknik pelatihan *batch* (berkelompok). Contoh hasil *padding* dan *truncation* dapat dilihat pada Gambar 4.10.



Gambar 4.10. Contoh *padding* dan *truncation*

Vektorisasi data dilakukan dengan menggunakan perintah sebagai berikut:

```
from transformers import AutoTokenizer
from functools import partial
from transformers import DataCollatorForTokenClassification

xlmr = "xlm-roberta-base"
mbert_cased = "bert-base-multilingual-cased"
mbert_uncased = "bert-base-multilingual-uncased"

mbert_cased_tokenizer =
AutoTokenizer.from_pretrained(mbert_cased, use_fast=True)
```

Vektorisasi data dilakukan dengan menggunakan perintah sebagai berikut (lanjutan):

```
mbert_uncased_tokenizer =
AutoTokenizer.from_pretrained(mbert_uncased, use_fast=True)
xlmr_tokenizer = AutoTokenizer.from_pretrained(xlmr,
use_fast=True)

def align_labels_with_tokens(labels, word_ids):
    new_labels = []
    current_word = None
    for word_id in word_ids:
        if word_id != current_word:
            current_word = word_id
            label = -100 if word_id is None else labels[word_id]
            new_labels.append(label)

        elif word_id is None:
            new_labels.append(-100)

        else:

            label = labels[word_id]
            if label % 2 == 1:
                label += 1

            new_labels.append(label)

    return new_labels

def tokenize_and_align_labels(data, tokenizer):
    tokenized_inputs = tokenizer(
        data["tokens"],
        truncation=True,
        padding=True,
        is_split_into_words=True
    )

    all_labels = data["ner_tags"]
    new_labels = []

    for i, labels in enumerate(all_labels):
        word_ids = tokenized_inputs.word_ids(i)
        new_labels.append(align_labels_with_tokens(labels,
word_ids))

    tokenized_inputs["labels"] = new_labels
    return tokenized_inputs

mbert_cased_tokenized_nergrit = nergrit.map(
    partial(tokenize_and_align_labels,
tokenizer=mbert_cased_tokenizer),
    batched=True,
)
```

Vektorisasi data dilakukan dengan menggunakan perintah sebagai berikut (lanjutan):

```
mbert_uncased_tokenized_nergrit = nergrit.map(
    partial(tokenize_and_align_labels,
tokenizer=mbert_uncased_tokenizer),
    batched=True,
)

xlmr_tokenized_nergrit = nergrit.map(
    partial(tokenize_and_align_labels,
tokenizer=xlmr_tokenizer),
    batched=True,
)


mbert_cased_data_collator =
DataCollatorForTokenClassification(tokenizer=mbert_cased_tokeniz
er)
mbert_uncased_data_collator =
DataCollatorForTokenClassification(tokenizer=mbert_uncased_token
izer)
xlmr_data_collator =

DataCollatorForTokenClassification(tokenizer=xlmr_tokenizer)
```

4.3. Modelling

Setelah vektorisasi data dilakukan, langkah selanjutnya adalah *fine-tuning* pada *pre-trained* model menggunakan *data train* dan *data validation* untuk membangun model multilingual *named entity recognition*. Pemodelan yang dilakukan pada penelitian ini memanfaatkan fungsi *Trainer* yang disediakan oleh *library transformers*. Langkah pertama adalah mendefinisikan *hyperparameter* yang akan digunakan untuk *fine-tuning* model seperti ditunjukkan pada Tabel 4.4.

Tabel 4.4. *Hyperparameter* untuk *fine-tuning* model

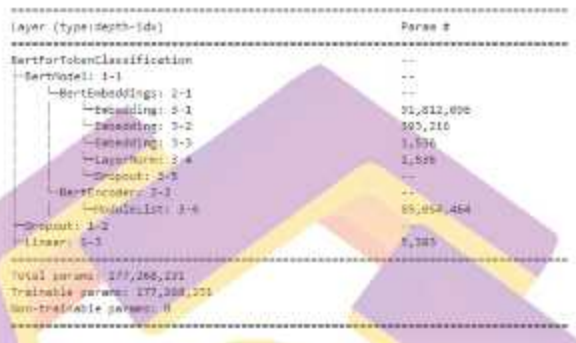
Training Arguments	Value
Learning Rate (AdamW)	2e-5
Num Epoch	3
Weight Decay	0.01
Batch Size	8

Hyperparameter yang digunakan pada penelitian ini mengacu pada prosedur *fine-tuning* yang disarankan oleh (Devlin *dkk.*, 2018) dengan jumlah *epoch* sebanyak 3 kali menggunakan fungsi optimasi *AdamW* dengan *learning rate* dan *weight decay* sebesar 0,00002 dan 0,01. Sementara itu, *batch size* yang digunakan hanya sebanyak 8 *batch* saja.

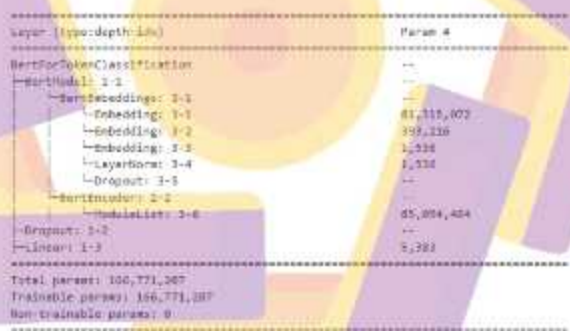
Pendefinisian *hyperparameter* dilakukan dengan menggunakan perintah sebagai berikut:

```
from transformers import TrainingArguments

mbert_cased_args = TrainingArguments(
    f"{mbert_cased}-ner-silvanus",
    evaluation_strategy="epoch",
    save_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    weight_decay=0.01,
    load_best_model_at_end=True,
)

mbert_uncased_args = TrainingArguments(
    f"{mbert_uncased}-ner-silvanus",
    evaluation_strategy="epoch",
    save_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    weight_decay=0.01,
    load_best_model_at_end=True,
)

xlmr_args = TrainingArguments(
    f"{xlmr}-ner-silvanus",
    evaluation_strategy="epoch",
    save_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    weight_decay=0.01,
    load_best_model_at_end=True,
)
```

Langkah kedua adalah mendefinisikan *pre-trained* model yang akan digunakan untuk *fine-tuning*. *Pre-trained* model yang digunakan adalah mBERT Cased, mBERT Uncased dan XLM-R.



Gambar 4.11. Arsitektur mBERT Cased



Gambar 4.12. Arsitektur mBERT Uncased

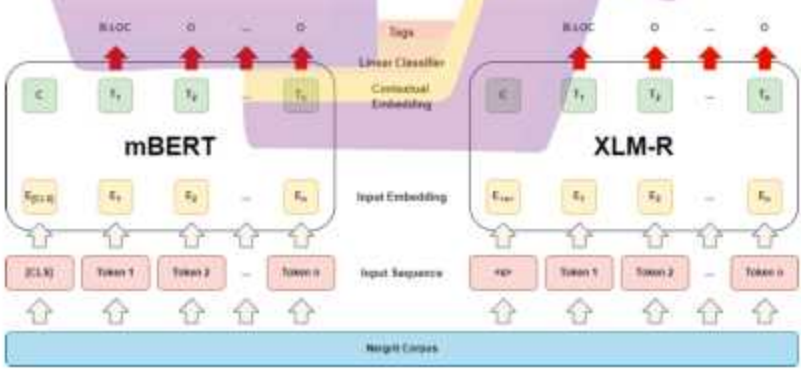
Gambar 4.13. Arsitektur XLM-R

Berdasarkan arsitektur model pada Gambar 4.11, Gambar 4.12 dan Gambar 4.13 *layer* pertama pada *input embedding* merupakan *token embedding* bawaan BERT yang memiliki jumlah parameter berdasarkan ukuran token *pre-trained* model dengan dimensi 768, kemudian pada *layer* kedua merupakan *position embedding* yang mewakili posisi setiap token dalam *input sequence* yang memiliki parameter berdasarkan panjang input maksimal 512 dengan ukuran dimensi 768. Ketiga, merupakan layer *segment embedding (token type embedding)* yang memberikan id segmen ke setiap token dengan ukuran dimensi 768. Keempat, dilakukan *layer normalization* untuk mengubah *input embedding* pada skala yang sama antara -1 dan 1 yang menghasilkan parameter bobot dan bias dengan ukuran dimensi masing-masing 768, normalisasi dilakukan sebagai strategi untuk mengurangi waktu pelatihan dan membuat model tidak bias ke fitur yang bernilai tinggi. Terakhir, *dropout* dilakukan untuk membantu mencegah *overfitting*.

Pemanggilan *pre-trained* model dilakukan dengan menggunakan perintah sebagai berikut:

```
from transformers import AutoModelForTokenClassification

xlmr_model = AutoModelForTokenClassification.from_pretrained(
    xlmr,
    id2label=id2label,
    label2id=label2id,
)

mbert_cased_model =
AutoModelForTokenClassification.from_pretrained(
    mbert_cased,
    id2label=id2label,
    label2id=label2id,
)

mbert_uncased_model =
AutoModelForTokenClassification.from_pretrained(
    mbert_uncased,
    id2label=id2label,
    label2id=label2id,
)
```

Setelah mendefinisikan model, selanjutnya adalah melakukan *fine-tuning* dengan menggunakan *Trainer* yang disediakan oleh *transformers* dengan meneruskan semua objek yang telah dibangun seperti *data train*, *data validation*, *hyperparameter*, metrik evaluasi performa dan *pre-trained* model. Terakhir, fungsi *train()* dipanggil untuk *fine-tuning* model pada *dataset* dan menghasilkan model *named entity recognition*.



Gambar 4.14. Arsitektur proses *fine-tuning* model

Pada Gambar 4.14 memperlihatkan proses *fine-tuning* model diawali dengan *input sequence* yang direpresentasikan sebagai {*Token 1, Token 2, ..., Token n*} yang berisi potongan kata dari suatu kalimat yang kemudian diberikan token khusus untuk menunjukkan awal dan akhir kalimat. Secara khusus, [CLS] dan [SEP] digunakan untuk mBERT, sedangkan <s> dan </s> digunakan untuk XLM-R. Selanjutnya, proses penyematan input yang direpresentasikan sebagai {E_1, E_2, ..., E_n} melibatkan dua metode: *token embedding* yang mengubah token menjadi id untuk kemudian diubah lagi menjadi vektor yang diambil dari matriks penyematan kata yang telah dilatih sebelumnya dan *positional embedding* yang menggabungkan informasi posisi ke dalam representasi token. Selanjutnya, masukkan hasil *input embedding* ke dalam model untuk mendapatkan penyematan kontekstual untuk setiap kata yang direpresentasikan sebagai {T_1, T_2, ..., T_n}. Penyematan kontekstual digunakan untuk menangkap konteks setiap kata melalui beberapa *multi-head attention* di setiap lapisan. Terakhir, data diteruskan ke lapisan linier *softmax* untuk mendapatkan tag label entitas {$y_1, y_2, y_3, ..., y_n$} dalam skema anotasi BIO.

Fine-tuning model dilakukan dengan menggunakan perintah sebagai berikut:

```
from transformers import Trainer
import evaluate

def compute_metrics(eval_preds):
    logits, labels = eval_preds
    predictions = np.argmax(logits, axis=-1)

    true_labels = [[label_names[l] for l in label if l != -100]
for label in labels]
    true_predictions = [
        [label_names[p] for (p, l) in zip(prediction, label) if
l != -100]
        for prediction, label in zip(predictions, labels)
    ]
```

Fine-tuning model dilakukan dengan menggunakan perintah sebagai berikut (lanjutan):

```
all_metrics = metric.compute(predictions=true_predictions,
references=true_labels)
    return {
        "precision": all_metrics["overall_precision"],
        "recall": all_metrics["overall_recall"],
        "f1": all_metrics["overall_f1"],
        "accuracy": all_metrics["overall_accuracy"],
    }

mbert_cased_trainer = Trainer(
    args=mbert_cased_args,
    model=mbert_cased_model,

    tokenizer=mbert_cased_tokenizer,
    data_collator=mbert_cased_data_collator,
    compute_metrics=compute_metrics,

    train_dataset=mbert_cased_tokenized_nergrit["train"],
    eval_dataset=mbert_cased_tokenized_nergrit["validation"]
)

mbert_uncased_trainer = Trainer(
    args=mbert_uncased_args,
    model=mbert_uncased_model,

    tokenizer=mbert_uncased_tokenizer,
    data_collator=mbert_uncased_data_collator,
    compute_metrics=compute_metrics,

    train_dataset=mbert_uncased_tokenized_nergrit["train"],
    eval_dataset=mbert_uncased_tokenized_nergrit["validation"]
)

xlmr_trainer = Trainer(
    args=xlmr_args,
    model=xlmr_model,

    tokenizer=xlmr_tokenizer,
    data_collator=xlmr_data_collator,
    compute_metrics=compute_metrics,

    train_dataset=xlmr_tokenized_nergrit["train"],
    eval_dataset=xlmr_tokenized_nergrit["validation"]
)

mbert_cased_trainer.train()
mbert_uncased_trainer.train()

xlmr_trainer.train()
```

Selama *fine-tuning*, performa model diperoleh dengan menggunakan *seqeval* sebagai evaluasi pengenalan entitas. Hasil evaluasi performa *fine-tuning* ketiga model dapat dilihat pada Tabel 4.5, Tabel 4.6 dan Tabel 4.7.

Tabel 4.5. Performa *fine-tuning* model mBERT Cased

Epoch	Train. Loss	Val. Loss	Precision	Recall	F1-score	Accuracy
1	0,1336	0,0551	0,9034	0,9130	0,9082	0,9844
2	0,0461	0,0604	0,9098	0,9134	0,9116	0,9842
3	0,0299	0,0621	0,9099	0,9202	0,9135	0,9852

Tabel 4.6. Performa *fine-tuning* model mBERT Uncased

Epoch	Train. Loss	Val. Loss	Precision	Recall	F1-score	Accuracy
1	0,1429	0,0587	0,8885	0,9075	0,8979	0,9829
2	0,0464	0,0609	0,9081	0,9103	0,9092	0,9846
3	0,0288	0,0662	0,9022	0,9190	0,9105	0,9838

Tabel 4.7. Performa *fine-tuning* model XLM-R

Epoch	Train. Loss	Val. Loss	Precision	Recall	F1-score	Accuracy
1	0,1394	0,0559	0,8808	0,9257	0,9027	0,9842
2	0,0468	0,0575	0,9107	0,9190	0,9148	0,9849
3	0,0279	0,0567	0,9189	0,9273	0,9231	0,9859

Pada Tabel 4.5 memperlihatkan bahwa selama 3 kali *epoch* model mBERT Cased mengalami penurunan *training loss* sedangkan *validation loss* meningkat di *epoch* 3, selain itu *precision* juga mengalami penurunan di *epoch* 3 bersamaan dengan meningkatnya *recall*, *f1-score* dan *accuracy*. Pada Tabel 4.6 memperlihatkan *accuracy* model mBERT Uncased mengalami penurunan dan peningkatan nilai *validation loss* di *epoch* 3. Pada Tabel 4.7 memperlihatkan bahwa performa model XLM-R selama proses *fine-tuning* nilai *training loss* dan *validation loss* cukup stabil mengalami penurunan selama pelatihan bersamaan dengan *precision*, *recall*, *f1-score* dan *accuracy* yang meningkat.

Gambar 4.15. Grafik perbandingan *loss fine-tuning*

Gambar 4.16. Grafik perbandingan performa selama *fine-tuning* model

Pada Gambar 4.16 memperlihatkan bahwa XLM-R memperoleh performa paling baik diantara model lainnya selama *fine-tuning* model menggunakan *data train* dan *data validation*. Selain itu, tipe model *cased* seperti XLM-R dan mBERT

Cased yang memperhatikan huruf kapital pada *dataset* memperoleh performa lebih baik dibandingkan tipe model *uncased* yang hanya menggunakan huruf kecil saja.

Gambar 4.17. Grafik perbandingan waktu *fine-tuning* model

Pada Gambar 4.17 memperlihatkan bahwa model yang memiliki durasi paling rendah dalam melakukan *fine-tuning* model dengan ukuran *batch* 8 adalah mBERT Uncased dengan waktu *fine-tuning* selama 747 detik sedangkan XLM-R memiliki durasi paling lama dengan waktu *fine-tuning* selama 861 detik. Selain itu, mBERT Cased memiliki durasi waktu *fine-tuning* selama 754 detik.

Tabel 4.8. Perbandingan model yang dilatih menggunakan *dataset nergrit corpus*

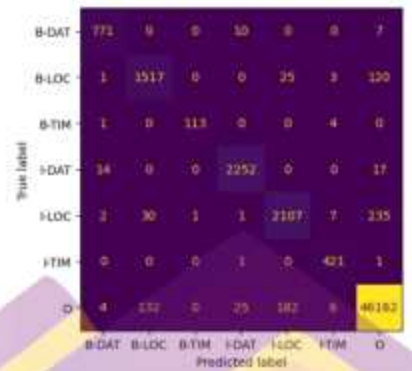
Model	Ukuran Parameter	Akurasi	Validation Loss	Waktu Fine-Tuning
mBERT Cased	177.268.231	98,52%	0,0621	754 detik
mBERT Uncased	166.771.207	98,38%	0,0662	**747 detik**
XLM-R	277.458.439	**98,59%**	**0,0567**	861 detik

Pada Tabel 4.8 menunjukkan bahwa mBERT Uncased memiliki ukuran parameter sebanyak 166 juta parameter, yang mana merupakan model dengan ukuran parameter terkecil dibandingkan model lainnya. Hal itu mengakibatkan

waktu *fine-tuning* model mBERT Uncased juga yang tercepat diantara model lainnya dengan waktu pelatihan selama 747 detik. Namun, akurasi terbaik dicapai oleh model XLM-R dengan akurasi sebesar 98,59% yang diikuti dengan *validation loss* terkecil sebesar 0,0567. Akurasi yang tinggi dengan *validation loss* yang kecil menunjukkan bahwa model XLM-R dapat mengenali entitas lokasi dan waktu di *dataset* dengan baik. Selain itu, model XLM-R memiliki ukuran parameter terbesar sebanyak 277 juta parameter. Hal itu mengakibatkan waku *fine-tuning* model XLM-R menjadi yang terlama dengan waktu pelatihan selama 861 detik. Performa yang tinggi pada ketiga model menunjukkan bahwa teknik pengolahan data pada *dataset* efektif dan menghasilkan bias yang rendah (data tidak seimbang) pada proporsi data, mengingat hasil pelabelan data pada Gambar 4.5 menunjukkan bahwa jumlah data pada masing-masing entitas label berbeda lebih dari 50%.

4.4. Evaluation

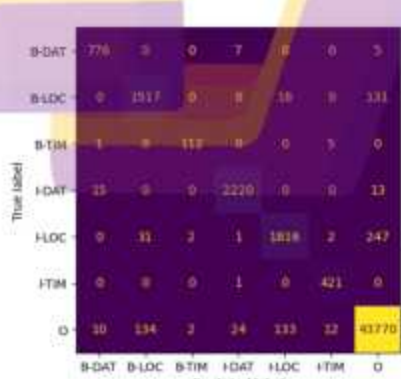
Tiga model multilingual *named entity recognition* yang telah terbentuk dari proses *fine-tuning*, selanjutnya akan dievaluasi untuk memastikan keakuratan model. Evaluasi dilakukan dengan cara menguji ketiga model dalam memprediksi label suatu token pada 1.228 *data testing* yang sudah divektorisasi ditahap sebelumnya sesuai dengan format masing-masing model. Selanjutnya, *confusion matrix* digunakan untuk menghitung akurasi.



Gambar 4.18. Grafik *confusion matrix* pengujian model mBERT Cased

Gambar 4.18 memperlihatkan hasil pengujian klasifikasi token menggunakan 54.172 token pada *data testing* sehingga diketahui nilai *True Positive* (TP) atau jumlah token entitas yang berhasil diklasifikasi adalah sebanyak 7.181 token dan *True Negative* (TN) atau jumlah token non-entitas yang berhasil diklasifikasi sebanyak 46.162 token untuk kemudian dihitung nilai *accuracy*.

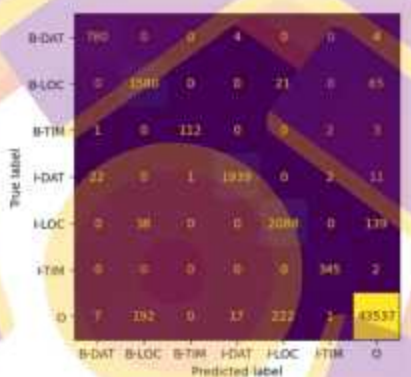
$$accuracy = \frac{7.181 + 46.162}{54.172} = 0{,}9847 \text{ atau } 98{,}47\%$$



Gambar 4.19. Grafik *confusion matrix* pengujian model mBERT Uncased

Gambar 4.19 memperlihatkan hasil pengujian klasifikasi token menggunakan 51.426 token pada *data testing* sehingga diketahui nilai *True Positive* (TP) atau jumlah token entitas yang berhasil diklasifikasi adalah sebanyak 6.862 token dan *True Negative* (TN) atau jumlah token non-entitas yang berhasil diklasifikasi sebanyak 43.770 token untuk kemudian dihitung nilai *accuracy*.

$$accuracy = \frac{6.862 + 43.770}{51.426} = 0{,}9846 \text{ atau } 98{,}46\%$$



Gambar 4.20. Grafik *confusion matrix* pengujian model XLM-RoBERTa

Gambar 4.20 memperlihatkan hasil pengujian klasifikasi token menggunakan 51.135 token pada *data testing* sehingga diketahui nilai *True Positive* (TP) atau jumlah token entitas yang berhasil diklasifikasi adalah sebanyak 6.844 token dan *True Negative* (TN) atau jumlah token non-entitas yang berhasil diklasifikasi sebanyak 43.537 token untuk kemudian dihitung nilai *accuracy*.

$$accuracy = \frac{6.844 + 43.537}{51.135} = 0{,}9853 \text{ atau } 98{,}53\%$$

Setelah dilakukan pengujian pada ketiga model dengan menggunakan *data testing* diketahui jumlah token, TP, TN dan akurasi masing-masing model dalam melakukan prediksi label suatu token yang ditunjukkan pada Tabel 4.9.

Tabel 4.9. Perbandingan hasil pengujian prediksi label suatu token

Model NER	Jumlah Token	TP	TN	Akurasi
mBERT Cased	54.172	7.181	46.162	98,47%
mBERT Uncased	51.426	6.862	43.770	98,46%
XLM-R	51.135	6.844	43.537	98,53%

Pada Tabel 4.9 memperlihatkan bahwa model mBERT Cased menghasilkan token terbanyak yaitu 54.172 token, karena memperhatikan huruf kapital dan menggunakan model tokenisasi *WordPiece* dengan ukuran *vocabulary* sebanyak 110.000 sub kata. Meskipun menghasilkan jumlah token terbanyak, akurasi pengujian model mBERT Cased bukan yang tertinggi yaitu sebesar 98,47%. Model tokenisasi *WordPiece* juga digunakan pada model mBERT Uncased, namun menghasilkan jumlah token yang lebih sedikit dari mBERT Cased yaitu 51.426 token, karena hanya memperhatikan huruf kecil yang mengakibatkan ukuran *vocabulary* nya lebih sedikit yaitu 105.879 sub kata. Akurasi pengujian yang diperoleh model mBERT Uncased merupakan yang terkecil yaitu sebesar 98,46%. Selain itu, model XLM-R menghasilkan token paling sedikit diantara kedua model lainnya yaitu 51.135 token, karena menggunakan model tokenisasi *SentencePiece* dengan ukuran *vocabulary* terbanyak yaitu 250.000 sub kata. Meskipun jumlah token paling sedikit, namun akurasi pengujian yang diperoleh model XLM-R merupakan yang tertinggi dibandingkan kedua model lainnya yaitu sebesar 98,53%. Jumlah token yang sedikit dengan akurasi pengujian yang tinggi menunjukkan bahwa model XLM-R efisien dan efektif dalam memprediksi label suatu token

ketika diberikan data baru dari *dataset*. Selain itu, akurasi pengujian yang tinggi pada ketiga model menunjukkan bahwa teknik pengolahan data yang dilakukan pada *dataset* juga terbilang efektif.

Pengujian pada ketiga model menggunakan *data testing* yang divektorisasi dilakukan menggunakan perintah sebagai berikut:

```
from transformers import pipeline, AutoTokenizer,
AutoModelForTokenClassification, Trainer, TrainingArguments
from sklearn.metrics import confusion_matrix,
ConfusionMatrixDisplay

def eval_confusion_matrix(eval_preds):
    logits, labels = eval_preds
    predictions = np.argmax(logits, axis=-1)

    true_labels = [[label_names[l] for l in label if l != -100]
for label in labels]
    true_predictions = [
        [label_names[p] for (p, l) in zip(prediction, label) if
l != -100]
        for prediction, label in zip(predictions, labels)
    ]

    x_labels = np.concatenate(true_labels)
    y_labels = np.concatenate(true_predictions)

    cm = confusion_matrix(x_labels, y_labels)
    cm_display = ConfusionMatrixDisplay( confusion_matrix = cm,
display_labels = sorted(list(set(label_names))) )
    cm_display.plot()

    return {}

mbert_cased_trainer = Trainer(
                model =
AutoModelForTokenClassification.from_pretrained("rollerhafeezh-
amikom/bert-base-multilingual-cased-ner-silvanus"),
                compute_metrics = eval_confusion_matrix )

mbert_uncased_trainer = Trainer(
                model =
AutoModelForTokenClassification.from_pretrained("rollerhafeezh-
amikom/bert-base-multilingual-uncased-ner-silvanus"),
                compute_metrics = eval_confusion_matrix )

xlmr_trainer = Trainer(
                model =
AutoModelForTokenClassification.from_pretrained("rollerhafeezh-
amikom/xlm-roberta-base-ner-silvanus"),
                compute_metrics = eval_confusion_matrix )
```

Pengujian pada ketiga model menggunakan *data testing* yang divektorisasi dilakukan menggunakan perintah sebagai berikut (lanjutan):

```
mbert_cased_results =
mbert_cased_trainer.predict(mbert_cased_tokenized_nergrit['test'
])

mbert_uncased_results =
mbert_uncased_trainer.predict(mbert_uncased_tokenized_nergrit['t
est'])

xlmr_results                                                    =

xlmr_trainer.predict(xlmr_tokenized_nergrit['test'])
```

4.5. Deployment

Setelah model multilingual *named entity recognition* dievaluasi menggunakan *data testing* dari *dataset nergrit corpus*, selanjutnya pada tahap ini model digunakan pada skenario nyata untuk mengekstrak lokasi dan waktu dari *tweet* kebakaran hutan dalam lima bahasa dan divalidasi hasil ekstraksinya untuk mengetahui tingkat akurasi model dalam mengenali entitas lokasi dan waktu selain Bahasa Indonesia. Pada penelitian ini, *tweet* yang digunakan untuk validasi model tidak dilakukan *text preprocessing* sehingga *tweet* masih terdiri dari *hashtag* (#), *user mention (@)*, *emoticon*, karakter khusus maupun tautan. Contoh *tweet* kebakaran hutan untuk ekstraksi lokasi dan waktu dapat dilihat pada Tabel 4.10.

Tabel 4.10. Contoh *tweet* kebakaran hutan untuk ekstraksi lokasi dan waktu

No	Tweet	Bahasa
1	**Kebakaran hutan** dan lahan (karhutla) terjadi di lereng Gunung Panderman, Kota Batu, Jawa Timur, Selasa, 21 November 2023. Titik api diperkirakan muncul pada pukul 15.30 WIB. #Karhutla https://t.co/G6qLxuIQiN	Indonesia
2	Bolivia **Forest Fire**: International aids respond as fires raze swathes o... https://t.co/VY9i7OiniT via @YouTube	Inggris

Tabel 4.10. Contoh tweet kebakaran hutan untuk ekstraksi lokasi dan waktu (lanjutan)

No	Tweet	Bahasa
3	🔴 DECLARADO **incendio forestal** en Alhaurín de La Torre. MEDIOS: 🚒 1 técnico de Operaciones, 1 agente de Medio Ambiente, 1 Brica y 1 grupo de bomberos forestales 🚁 1 helicóptero pesado #axarquia #AxarquiaPlus #infoca #incendio #malaga #alhaurindelatorre https://t.co/6CJmCVsHM2	Spanyol
4	🔴 **Incendi forestal** a Sant Julià de Ramis: les flames han saltat a la carretera empeses pel fort vent Una vintena de dotacions dels Bombers treballen per aturar el foc, que ha generat focus secundaris https://t.co/Db2iMzPCSE	Italia
5	**Lesné požiare** v Austrálii si vyžiadali evakuáciu stoviek obyvateľov https://t.co/99QWAS0Y7I https://t.co/dFal3Vaotf	Slovakia

Tabel 4.10 memperlihatkan bahwa semua bahasa menggunakan alfabet latin, namun berbeda dengan Bahasa Spanyol, Italia dan Slovakia yang menambahkan beberapa huruf aksen seperti pada kata "*Alhaurín de La Torre*", "*Sant Julià de Ramis*" dan "*Austrálii*".

Berikut merupakan tahapan-tahapan yang digunakan dalam mengekstrak lokasi dan waktu dari *tweet* kebakaran hutan untuk validasi model:

4.5.1. Pengenalan Entitas Lokasi dan Waktu

Tahapan pertama dalam mengekstrak lokasi dan waktu dari *tweet* kebakaran hutan adalan dengan memanggil model multilingual *named entity recognition* beserta teknik tokenisasinya untuk dimasukkan ke *pipeline* tugas pengenalan entitas. Selain itu, pengaturan strategi agregasi (*aggregation strategy*) pada tokenisasi sub token juga digunakan untuk menggabungkan sub token hasil pengenalan entitas menjadi token. Hasil pengenalan entitas menggunakan model multilingual *named entity recognition* pada *tweet* kebakaran hutan akan menghasilkan *array* yang berisi *dictionary* dengan atribut-atribut sebagai berikut:

a. *entity group:* merupakan entitas yang diprediksi untuk suatu token (kata) berdasarkan penggabungan (agregasi) beberapa sub token (sub kata).

b. *score:* nilai probabilitas entitas yang sesuai pada suatu kata.

c. *word*: token atau kata yang dikenali atau diklasifikasi

d. *start:* indeks awal token entitas yang dikenali dalam kalimat

e. *end:* indeks akhir token entitas yang dikenali dalam kalimat

Contoh hasil pengenalan entitas model multilingual *named entity recognition* pada contoh *tweet* kebakaran hutan di Tabel 4.10 dapat dilihat pada Tabel 4.11, Tabel 4.12, Tabel 4.13, Tabel 4.14 dan Tabel 4.15.

Tabel 4.11. Contoh hasil pengenalan entitas lokasi dan waktu pada *tweet* kebakaran hutan dalam Bahasa Indonesia

mBERT Cased	mBERT Uncased	XLM-R
[{ "entity_group": "LOC", "score": 0.97621083, "word": "lereng Gunung Panderman", "start": 48, "end": 71, }, { "entity_group": "LOC", "score": 0.99475217, "word": "Kota Batu", "start": 73, "end": 82, }, { "entity_group": "LOC", "score": 0.99490625, "word": "Jawa Timur", "start": 84,	[{ "entity_group": "LOC", "score": 0.79765606, "word": "lereng", "start": 48, "end": 54, }, { "entity_group": "LOC", "score": 0.8781987, "word": "gunung panderman", "start": 55, "end": 71, }, { "entity_group": "LOC", "score": 0.9945995, "word": "kota batu", "start": 73,	[{ "entity_group": "LOC", "score": 0.92402005, "word": "lereng Gunung Panderman", "start": 48, "end": 71, }, { "entity_group": "LOC", "score": 0.99571913, "word": "Kota Batu", "start": 73, "end": 82, }, { "entity_group": "LOC", "score": 0.9953073, "word": "Jawa Timur", "start": 84,

Tabel 4.11. Contoh hasil pengenalan entitas lokasi dan waktu pada *tweet* kebakaran hutan dalam Bahasa Indonesia (lanjutan)

mBERT Cased	mBERT Uncased	XLM-R
"end": 94, }, { "entity_group": "DAT", "score": 0.9965285, "word": "Selasa, 21 November 2023", "start": 96, "end": 120, }, { "entity_group": "TIM", "score": 0.9987256, "word": "15. 30 WIB", "start": 163, "end": 172, },]	"end": 82, }, { "entity_group": "LOC", "score": 0.99332273, "word": "jawa timur", "start": 84, "end": 94, }, { "entity_group": "DAT", "score": 0.9924701, "word": "selasa, 21 november 2023", "start": 96, "end": 120, }, { "entity_group": "TIM", "score": 0.9977441, "word": "15. 30 wib", "start": 163, "end": 172, },]	"end": 94, }, { "entity_group": "DAT", "score": 0.9949247, "word": "Selasa, 21 November 2023", "start": 96, "end": 120, }, { "entity_group": "TIM", "score": 0.9928413, "word": "15.30 WIB", "start": 163, "end": 172, },]

Pada Tabel 4.11 memperlihatkan model mBERT Cased mendapatkan hasil pengenalan entitas yang sama dengan XLM-R, namun berbeda pada nilai probabilitasnya (*score*). Sementara itu, mBERT Uncased mendapatkan hasil pengenalan entitas yang berbeda karena melakukan tokenisasi pada kata pertama, selain itu mengubah bentuk semua huruf menjadi huruf kecil semua. Ketiga model menghasilkan pengenalan entitas dengan kata (*word*) yang sama dengan nilai kebenaran, namun berbeda pada tokenisasi, nilai probabilitas (*score*) dan bentuk huruf.

Tabel 4.12. Contoh hasil pengenalan entitas lokasi dan waktu pada *tweet* kebakaran hutan dalam Bahasa Inggris

mBERT Cased	mBERT Uncased	XLM-R

[
{
"entity_group": "LOC",
"score": 0.9301841,
"word": "Bolivia",
"start": 0,
"end": 7
}
]
	[
	{
	"entity_group": "LOC",
	"score": 0.9757205,
	"word": "bolivia",
	"start": 0,
	"end": 7
	}
	]
		[
		{
		"entity_group": "LOC",
		"score": 0.9905618,
		"word": "Bolivia",
		"start": 0,
		"end": 7
		}
		]

Pada Tabel 4.12 memperlihatkan model mBERT Cased mendapatkan hasil pengenalan entitas yang sama dengan XLM-R, namun berbeda pada nilai probabilitasnya (*score*). Sementara itu, mBERT Uncased mendapatkan hasil pengenalan entitas yang berbeda karena mengubah bentuk semua huruf menjadi huruf kecil semua. Ketiga model menghasilkan pengenalan entitas dengan kata (*word*) yang sama dengan nilai kebenaran, namun berbeda pada nilai probabilitas (*score*) dan bentuk huruf.

Tabel 4.13. Contoh hasil pengenalan entitas lokasi dan waktu pada *tweet* kebakaran hutan dalam Bahasa Spanyol

mBERT Cased	mBERT Uncased	XLM-R

[
{
"entity_group": "LOC",
"score": 0.9815024,
"word": "Alhaurin de La Torre",
"start": 33,
"end": 53
}
]
	[
	{
	"entity_group": "LOC",
	"score": 0.93453556,
	"word": "alhaurin",
	"start": 33,
	"end": 41
	},
	{
	"entity_group": "LOC",
	"score": 0.5015954,
	"word": "torre",
	"start": 48,
	"end": 53
	},
	{
		[
		{
		"entity_group": "LOC",
		"score": 0.98847574,
		"word": "Alhaurin de La Torre",
		"start": 33,
		"end": 53
		},
		{
		"entity_group": "LOC",
		"score": 0.7300759,
		"word": "malaga",
		"start": 230,
		"end": 236
		},

Tabel 4.13. Contoh hasil pengenalan entitas lokasi dan waktu pada tweet kebakaran hutan dalam Bahasa Spanyol (lanjutan)

mBERT Cased	mBERT Uncased	XLM-R

	"entity_group": "LOC",
	"score": 0.9795688,
	"word": "malaga",
	"start": 230,
	"end": 236
	}
	]
		{
		"entity_group": "LOC",
		"score": 0.93473524,
		"word": "alhaurindelatorre",
		"start": 238,
		"end": 255
		}
		]

Pada Tabel 4.13 memperlihatkan model mBERT Cased hanya mengenali lokasi pada teks saja, berbeda dengan mBERT Uncased dan XLM-R yang bisa mengenali lokasi pada *hashtag* tanpa mengenali simbol *hashtag*-nya. Namun, mBERT Uncased mendapatkan hasil pengenalan entitas yang berbeda karena mengubah bentuk semua huruf menjadi huruf kecil semua dan mengganti jenis huruf aksen menjadi huruf latin. Ketiga model menghasilkan pengenalan entitas yang baik dalam mengenali lokasi dan sesuai dengan nilai kebenaran, namun berbeda pada nilai probabilitas (*score*), bentuk huruf, jenis huruf dan kemampuan dalam mengenali lokasi pada teks ataupun *hashtag*.

Tabel 4.14. Contoh hasil pengenalan entitas lokasi dan waktu pada *tweet* kebakaran hutan dalam Bahasa Italia

mBERT Cased	mBERT Uncased	XLM-R

[
{
"entity_group": "LOC",
"score": 0.9737876,
"word": "Sant Julià de Ramis",
"start": 21,
"end": 40
}
]
	[
	{
	"entity_group": "LOC",
	"score": 0.96996534,
	"word": "sant julia de ramis",
	"start": 21,
	"end": 40
	}
	]
		[
		{
		"entity_group": "LOC",
		"score": 0.9669585,
		"word": "Sant Julià de Ramis",
		"start": 21,
		"end": 40
		}
		]

Pada Tabel 4.14 memperlihatkan model mBERT Cased mendapatkan hasil pengenalan entitas yang sama dengan XLM-R, namun berbeda pada nilai probabilitasnya (*score*). Sementara itu, mBERT Uncased mendapatkan hasil pengenalan entitas yang berbeda karena mengubah bentuk semua huruf menjadi huruf kecil semua dan mengganti jenis huruf aksen menjadi huruf latin. Ketiga model menghasilkan pengenalan entitas dengan kata (*word*) yang sama dengan nilai kebenaran, namun berbeda pada nilai probabilitas (*score*), bentuk huruf dan jenis huruf.

Tabel 4.15. Contoh hasil pengenalan entitas lokasi dan waktu pada *tweet* kebakaran hutan dalam Bahasa Slovakia

mBERT Cased	mBERT Uncased	XLM-R

[
{
"entity_group": "LOC",
"score": 0.9857793,
"word": "Austrálii",
"start": 16,
"end": 25
},
{
"entity_group": "LOC",
"score": 0.5633619,
"word": "##S0Y",
"start": 86,
"end": 89
}
]
	[
	{
	"entity_group": "LOC",
	"score": 0.9854446,
	"word": "australii",
	"start": 16,
	"end": 25
	}
	]
		[
		{
		"entity_group": "LOC",
		"score": 0.98909366,
		"word": "Austrálii",
		"start": 16,
		"end": 25
		}
		]

Pada Tabel 4.15 memperlihatkan model mBERT Uncased mendapatkan hasil pengenalan entitas yang sama dengan XLM-R, namun berbeda pada bentuk huruf, jenis huruf dan nilai probabilitasnya (*score*), dimana mBERT Uncased mendapatkan hasil pengenalan entitas yang berbeda karena mengubah bentuk semua huruf menjadi huruf kecil semua dan mengganti jenis huruf aksen menjadi

huruf latin. Sementara itu, model mBERT Cased mampu mengenali lokasi pada teks seperti kedua model lainnya, namun memiliki kelemahan karena bisa menghasilkan *noise* dengan mengenali sub token yang belum diagregasi menjadi token. Model mBERT Uncased dan XLM-R menghasilkan pengenalan entitas dengan kata (*word*) yang sama, namun berbeda pada nilai probabilitas (*score*), bentuk huruf dan jenis huruf.

Ketiga model mampu untuk mengenali entitas lokasi dan waktu dari *tweet* kebakaran hutan pada Bahasa Indonesia, Inggris, Spanyol, Italia dan Slovakia. Namun, terdapat beberapa perbedaan hasil ekstraksi. Dimana, model mBERT Cased tetap mempertahankan teks aslinya tanpa mengubah bentuk huruf dan jenis huruf, namun bisa menghasilkan *noise* dengan mengenali sub token yang belum diagregasi menjadi token. Sedangkan, model mBERT Uncased mengubah bentuk semua huruf menjadi huruf kecil semua dan mengganti jenis huruf aksen menjadi huruf latin serta mampu mengenali lokasi pada *hashtag*. Selain itu, pada model XLM-R tetap mempertahankan teks aslinya tanpa mengubah bentuk huruf dan jenis huruf serta mampu mengenali lokasi pada *hashtag*.

4.5.2. Pengelompokan Token Entitas

Entitas lokasi dan waktu yang dikenali di tahapan sebelumnya masih berbentuk *raw data* yang terdiri dari beberapa token, sehingga di tahapan ini akan dilakukan pengelompokan token yang dikenali sebagai entitas bertujuan untuk mendapatkan seluruh informasi pada masing-masing entitas tersebut. Pengelompokan token entitas dilakukan dengan cara menggabungkan token (*word*) berdasarkan masing-masing entitas (*entity_group*) yang dipisahkan menggunakan

spasi. Contoh hasil pengelompokan token berdasarkan masing-masing entitas dari *tweet* kebakaran hutan di Tabel 4.10 dapat dilihat pada Tabel 4.16, Tabel 4.17, Tabel 4.18, Tabel 4.19 dan Tabel 4.20.

Tabel 4.16. Contoh hasil pengelompokan token entitas pada *tweet* kebakaran hutan dalam Bahasa Indonesia

Model	Token Entitas		
	Lokasi (LOC)	Tanggal (DAT)	Waktu (TIM)
mBERT Cased	lereng Gunung Panderman Kota Batu Jawa Timur	Selasa, 21 November 2023	15.30 WIB
mBERT Uncased	lereng gunung panderman kota batu jawa timur	selasa, 21 november 2023	15.30 wib
XLM-R	lereng Gunung Panderman Kota Batu Jawa Timur	Selasa, 21 November 2023	15.30 WIB

Tabel 4.17. Contoh hasil pengelompokan token entitas pada *tweet* kebakaran hutan dalam Bahasa Inggris

Model	Token Entitas		
	Lokasi (LOC)	Tanggal (DAT)	Waktu (TIM)
mBERT Cased	Bolivia	-	-
mBERT Uncased	bolivia	-	-
XLM-R	Bolivias	-	-

Tabel 4.18. Contoh hasil pengelompokan token entitas pada *tweet* kebakaran hutan dalam Bahasa Spanyol

Model	Token Entitas		
	Lokasi (LOC)	Tanggal (DAT)	Waktu (TIM)
mBERT Cased	Alhaurín de La Torre	-	-
mBERT Uncased	alhaurin torre malaga	-	-
XLM-R	Alhaurín de La Torre malaga alhaurindelatorre	-	-

Tabel 4.19. Contoh hasil pengelompokan token entitas pada *tweet* kebakaran hutan dalam Bahasa Italia

Model	Token Entitas		
	Lokasi (LOC)	Tanggal (DAT)	Waktu (TIM)
mBERT Cased	Sant Julià de Ramis	-	-
mBERT Uncased	sant julia de ramis	-	-
XLM-R	Sant Julià de Ramis	-	-

Tabel 4.20. Contoh hasil pengelompokan token entitas pada *tweet* kebakaran hutan dalam Bahasa Italia

Model	Token Entitas		
	Lokasi (LOC)	Tanggal (DAT)	Waktu (TIM)
mBERT Cased	Australii ##S0Y	-	-
mBERT Uncased	australii	-	-
XLM-R	Australii	-	-

4.5.3. Penyimpanan Hasil Ekstraksi

Setelah pengelompokan token entitas dilakukan untuk mendapatkan keseluruhan informasi pada masing-masing entitas, selanjutnya adalah menyimpannya dalam format .csv (*comma separated value*) sebagai hasil ekstraksi lokasi dan waktu menggunakan model multilingual *named entity recognition* dari *tweet* kebakaran hutan. Hasil ekstraksi yang disimpan pada tahapan ini selanjutnya akan digunakan untuk validasi model.

Proses ekstraksi lokasi dan waktu dari *tweet* kebakaran hutan dilakukan dengan menggunakan perintah sebagai berikut:

```
models = [
    'rollerhafeezh-amikom/xlm-roberta-base-ner-silvanus',
    'rollerhafeezh-amikom/bert-base-multilingual-cased-ner-
silvanus',
    'rollerhafeezh-amikom/bert-base-multilingual-uncased-ner-
silvanus'
]

file_path = {
    'id': 'tweets/id_tweets.csv',
    'en': 'tweets/en_tweets.csv',
    'es': 'tweets/es_tweets.csv',
    'it': 'tweets/it_tweets.csv',
    'sk': 'tweets/sk_tweets.csv'
}

id = pd.read_csv(file_path['id'], sep=';')
en = pd.read_csv(file_path['en'], sep=';')
es = pd.read_csv(file_path['es'], sep=';')
it = pd.read_csv(file_path['it'], sep=';')
sk = pd.read_csv(file_path['sk'], sep=';')
```

Proses ekstraksi lokasi dan waktu dari *tweet* kebakaran hutan dilakukan dengan menggunakan perintah sebagai berikut (lanjutan):

```
df = pd.concat([id, en, es, it, sk])
df = df[['created_at', 'full_text', 'location_detect',
'location_true']]

def ner_extraction(models, text):
    for i, ner_model in enumerate(models):
        ner_result = []
        location = []
        date = []
        time = []

        tokenizer = AutoTokenizer.from_pretrained(ner_model)
        model =
AutoModelForTokenClassification.from_pretrained(ner_model)

        ner_pipeline = pipeline("ner",
                                model=model,
                                tokenizer=tokenizer,
                                aggregation_strategy="simple")

        for t in tqdm(text, desc=ner_model):
            results = ner_pipeline(t)
            ner_result.append(results)

            LOC = ' '.join( [ t['word'] for t in results if
t['entity_group'] == 'LOC' ] )
            DAT = ' '.join([ t['word'] for t in results if
t['entity_group'] == 'DAT' ])
            TIM = ' '.join([ t['word'] for t in results if
t['entity_group'] == 'TIM' ])

            location.append(LOC)
            date.append(DAT)
            time.append(TIM)

        df[f'model_{i}_location'] = location
        df[f'model_{i}_date'] = date
        df[f'model_{i}_time'] = time
        df[f'model_{i}_ner'] = ner_result

        df.to_csv('validasi tweet.csv', encoding='utf-8',
sep=';', index=False)

ner_extraction( models, df['full_text'].to_list() )
```

4.5.4. Validasi Model

Validasi model dilakukan secara manual dengan melibatkan beberapa validator untuk mengetahui tingkat akurasi model dalam mengekstrak lokasi dan waktu dari *tweet* kebakaran hutan. Daftar validator yang melakukan validasi model dapat dilihat pada Tabel 4.21.

Tabel 4.21. Daftar validator model

No	Bahasa Tweet	Validator
1	Indonesia	Amikom
2	Inggris	Amikom
3	Spanyol	Atos
4	Italia	EAI
5	Slovakia	3Mon

Validasi model yang dilakukan pada penelitian ini adalah menilai kesesuaian lokasi dan waktu yang dideteksi dengan nilai kebenaran secara subjektif. Kriteria yang digunakan untuk menilai kesesuaian pengenalan entitas lokasi dan waktu pada suatu *tweet* adalah sebagai berikut:

a. Bernilai 1 atau benar, apabila:
    1. Hasil pengenalan entitas lokasi atau waktu pada *tweet* sesuai dengan nilai kebenaran secara subjektif.
    2. Model tidak mengenali entitas lokasi atau waktu dari suatu *tweet* yang memang tidak ada entitas tersebut.
b. Bernilai 0 atau salah, apabila model salah mengenali entitas lokasi atau waktu pada suatu *tweet*.

Tabel 4.22, Tabel 4.23, Tabel 4.24, Tabel 4.25 dan Tabel 4.26 memperlihatkan contoh hasil validasi model secara manual pada Bahasa Indonesia,

Inggris, Spanyol, Italia dan Slovakia berdasarkan kriteria kesesuaian pada hasil ekstraksi yang telah disimpan sebelumnya.

Tabel 4.22. Contoh hasil validasi model pada hasil ekstraksi lokasi dan waktu dalam Bahasa Indonesia

Model	Hasil Ekstraksi			Hasil Validasi	
	Lokasi (LOC)	Waktu (DAT)	Waktu (TIM)	Lokasi	Waktu
mBERT Cased	lereng Gunung Panderman Kota Batu Jawa Timur	Selasa, 21 November 2023	15. 30 WIB	Benar	Benar
mBERT Uncased	lereng gunung panderman kota batu jawa timur	selasa, 21 november 2023	15. 30 wib	Benar	Benar
XLM-R	lereng Gunung Panderman Kota Batu Jawa Timur	Selasa, 21 November 2023	15.30 WIB	Benar	Benar

Tabel 4.23. Contoh hasil validasi model pada hasil ekstraksi lokasi dan waktu dalam Bahasa Inggris

Model	Hasil Ekstraksi			Hasil Validasi	
	Lokasi (LOC)	Waktu (DAT)	Waktu (TIM)	Lokasi	Waktu
mBERT Cased	Bolivia	-	-	Benar	Benar
mBERT Uncased	bolivia	-	-	Benar	Benar
XLM-R	Bolivia	-	-	Benar	Benar

Tabel 4.24. Contoh hasil validasi model pada hasil ekstraksi lokasi dan waktu dalam Bahasa Spanyol

Model	Hasil Ekstraksi			Hasil Validasi	
	Lokasi (LOC)	Waktu (DAT)	Waktu (TIM)	Lokasi	Waktu
mBERT Cased	Alhaurín de La Torre	-	-	Benar	Benar
mBERT Uncased	alhaurin torre malaga	-	-	Benar	Benar
XLM-R	Alhaurín de La Torre malaga alhaurindelatorre	-	-	Benar	Benar

Tabel 4.25. Contoh hasil validasi model pada hasil ekstraksi lokasi dan waktu dalam Bahasa Italia

Model	Hasil Ekstraksi			Hasil Validasi	
	Lokasi (LOC)	Waktu (DAT)	Waktu (TIM)	Lokasi	Waktu
mBERT Cased	Sant Julià de Ramis	-	-	Benar	Benar

Tabel 4.25. Contoh hasil validasi model pada hasil ekstraksi lokasi dan waktu dalam Bahasa Italia (lanjutan)

Model	Hasil Ekstraksi			Hasil Validasi	
	Lokasi (LOC)	Waktu (DAT)	Waktu (TIM)	Lokasi	Waktu
mBERT Uncased	sant julia de ramis	-	-	Benar	Benar
XLM-R	Sant Julià de Ramis	-	-	Benar	Benar

Tabel 4.26. Contoh hasil validasi model pada hasil ekstraksi lokasi dan waktu dalam Bahasa Slovakia

Model	Hasil Ekstraksi			Hasil Validasi	
	Lokasi (LOC)	Waktu (DAT)	Waktu (TIM)	Lokasi	Waktu
mBERT Cased	Austrálii ##S0Y	-	-	Salah	Benar
mBERT Uncased	australii	-	-	Benar	Benar
XLM-R	Austrálii	-	-	Benar	Benar

Selanjutnya jumlah nilai benar hasil validasi lokasi dan waktu dibandingkan dengan jumlah data pada masing-masing bahasa untuk menghitung akurasi masing-masing model dalam mengekstrak lokasi dan waktu. Berikut persamaan yang digunakan untuk menghitung akurasi validasi pada masing-masing entitas dapat dilihat pada Persamaan 4.1 dan Persamaan 4.2.

$$\text{Akurasi}(lokasi) = \frac{\text{Jumlah Benar}_{\text{lokasi}}}{Jumlah\ Tweet_{bahasa}} \times 100\% \quad (4.1)$$

$$\text{Akurasi}(waktu) = \frac{\text{Jumlah Benar}_{\text{waktu}}}{Jumlah\ Tweet_{bahasa}} \times 100\% \quad (4.2)$$

Selain menghitung akurasi validasi masing-masing entitas, penelitian ini juga menghitung akurasi validasi dalam mengenali kedua entitas secara bersamaan dengan membandingkan jumlah nilai benar dalam mengenali entitas lokasi dan waktu dengan jumlah data dari masing-masing bahasa. Berikut persamaan yang

digunakan untuk menghitung akurasi validasi lokasi dan waktu dapat dilihat pada Persamaan 4.3.

$$\text{Akurasi}(lokasi\ \&\ waktu) = \frac{\text{Jumlah Benar}_{\text{lokasi \& waktu}}}{Jumlah\ Tweet_{bahasa}} \times 100\% \quad (4.3)$$

Dengan menggunakan Persamaan 4.1, Persamaan 4.2 dan Persamaan 4.3, maka diperoleh hasil validasi dari model multilingual *named entity recognition* berbasis mBERT Cased, mBERT Uncased dan XLM-R dengan hasil validasi dapat dilihat pada Tabel 4.27, Tabel 4.28 dan Tabel 4.29.

Tabel 4.27. Hasil validasi model mBERT Cased

No	Bahasa Tweet	Jumlah Tweet	Lokasi			Waktu			Lokasi & Waktu		
			Benar	Salah	Akurasi	Benar	Salah	Akurasi	Benar	Salah	Akurasi
1	Indonesia	482	344	138	71,37%	**480**	**2**	**99,59%**	343	139	71,16%
2	Inggris	511	259	252	50,68%	503	8	98,43%	256	255	50,10%
3	Spanyol	510	315	195	61,76%	495	15	97,06%	305	205	59,80%
4	Italia	509	346	163	67,98%	487	22	95,68%	335	174	65,82%
5	Slovakia	314	**246**	**68**	**78,34%**	301	13	95,86%	**238**	**76**	**75,80%**

Tabel 4.27 memperlihatkan hasil validasi *tweet* dalam mengenali entitas lokasi menggunakan model multilingual *named entity recognition* berbasis mBERT Cased dimana Bahasa Slovakia memperoleh akurasi tertinggi sebesar 78,34% dan Bahasa Inggris memperoleh akurasi terendah sebesar 50,68%. Perbandingan akurasi validasi dalam mengenali entitas lokasi terlihat lumayan jauh berbeda. Sedangkan, dalam mengenali entitas waktu dimana Bahasa Indonesia memperoleh akurasi tertinggi sebesar 99,59% dan Bahasa Italia memperoleh akurasi terendah sebesar 95,68%. Perbandingan akurasi validasi dalam mengenali entitas waktu terlihat tidak jauh berbeda. Sementara itu, dalam mengenali entitas lokasi dan waktu Bahasa Slovakia memperoleh akurasi tertinggi sebesar 75,80% dan Bahasa Inggris

memperoleh akurasi terendah sebesar 50,10%. Perbandingan akurasi validasi dalam mengenali entitas dan waktu terlihat lumayan jauh berbeda.

Tabel 4.28. Hasil validasi model mBERT Uncased

No	Bahasa Tweet	Jumlah Data	Lokasi			Waktu			Lokasi & Waktu		
			Benar	Salah	Akurasi	Benar	Salah	Akurasi	Benar	Salah	Akurasi
1	Indonesia	482	369	113	76,56%	**480**	**2**	**99,59%**	368	114	76,35%
2	Inggris	511	365	146	71,43%	500	11	97,85%	358	153	70,06%
3	Spanyol	510	389	121	76,27%	499	11	97,84%	383	127	75,10%
4	Italia	509	397	112	78,00%	495	14	97,25%	385	124	75,64%
5	Slovakia	314	**294**	**20**	**93,63%**	307	7	97,77%	**287**	**27**	**91,40%**

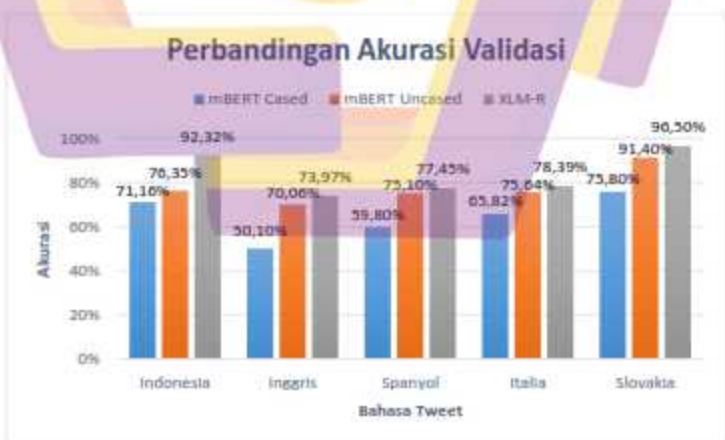
Tabel 4.28 memperlihatkan hasil validasi *tweet* dalam mengenali entitas lokasi menggunakan model multilingual *named entity recognition* berbasis mBERT Uncased dimana Bahasa Slovakia memperoleh akurasi tertinggi sebesar 93,63% dan Bahasa Inggris memperoleh akurasi terendah sebesar 71,43%. Perbandingan akurasi validasi dalam mengenali entitas lokasi terlihat lumayan jauh berbeda. Sedangkan, dalam mengenali entitas waktu dimana Bahasa Indonesia memperoleh akurasi tertinggi sebesar 99,59% dan Bahasa Italia memperoleh akurasi terendah sebesar 97,25%. Perbandingan akurasi validasi dalam mengenali entitas waktu terlihat tidak jauh berbeda. Sementara itu, dalam mengenali entitas lokasi dan waktu Bahasa Slovakia memperoleh akurasi tertinggi sebesar 91,40% dan Bahasa Inggris memperoleh akurasi terendah sebesar 70,06%. Perbandingan akurasi validasi dalam mengenali entitas dan waktu terlihat lumayan jauh berbeda.

Tabel 4.29. Hasil validasi model XLM-R

No	Bahasa Tweet	Jumlah Data	Lokasi			Waktu			Lokasi & Waktu		
			Benar	Salah	Akurasi	Benar	Salah	Akurasi	Benar	Salah	Akurasi
1	Indonesia	482	447	35	92,74%	480	2	99,59%	445	37	92,32%
2	Inggris	511	379	132	74,17%	510	1	99,80%	378	133	73,97%
3	Spanyol	510	396	114	77,65%	509	1	99,80%	395	115	77,45%
4	Italia	509	403	106	79,17%	500	9	98,23%	399	110	78,39%
5	Slovakia	314	**303**	**11**	**96,50%**	**314**	**0**	**100%**	**303**	**11**	**96,50%**

Tabel 4.29 memperlihatkan hasil validasi *tweet* dalam mengenali entitas lokasi menggunakan model multilingual *named entity recognition* berbasis XLM-R dimana Bahasa Slovakia memperoleh akurasi tertinggi sebesar 96,50% dan Bahasa Inggris memperoleh akurasi terendah sebesar 74,17%. Perbandingan akurasi validasi dalam mengenali entitas lokasi terlihat lumayan jauh berbeda. Sedangkan, dalam mengenali entitas waktu dimana Bahasa Slovakia juga memperoleh akurasi tertinggi sebesar 100% dan Bahasa Italia memperoleh akurasi terendah sebesar 98,23%. Perbandingan akurasi validasi dalam mengenali entitas waktu terlihat tidak jauh berbeda. Sementara itu, dalam mengenali entitas lokasi dan waktu Bahasa Slovakia juga memperoleh akurasi tertinggi sebesar 96,50% dan Bahasa Inggris memperoleh akurasi terendah sebesar 73,97%. Perbandingan akurasi validasi dalam mengenali entitas dan waktu terlihat lumayan jauh berbeda.

Perbandingan akurasi validasi ketiga model multilingual *named entity recognition* dapat dilihat pada Gambar 4.21.



Gambar 4.21. Grafik perbandingan akurasi validasi

Pada Gambar 4.21 menunjukkan bahwa model multilingual *named entity recognition* berbasis XLM-R yang divalidasi secara manual dalam mengekstrak lokasi dan waktu dari *tweet* kebakaran hutan memperoleh akurasi tertinggi pada Bahasa Indonesia 92,32%, Bahasa Inggris 73,97%, Bahasa Spanyol 77,45%, Bahasa Italia 78,39% dan Bahasa Slovakia 96,50% jika dibandingkan Multilingual BERT (mBERT) Cased dengan akurasi validasi pada Bahasa Indonesia 71,16%, Bahasa Inggris 50,10%, Bahasa Spanyol 59,80%, Bahasa Italia 65,82% dan Bahasa Slovakia 75,80% dan Multilingual BERT (mBERT) Uncased dengan akurasi validasi pada Bahasa Indonesia 76,35%, Bahasa Inggris 70,06%, Bahasa Spanyol 75,10%, Bahasa Italia 75,64% dan Bahasa Slovakia sebesar 91,40%. Hasil validasi manual pada model XLM-R menunjukkan bahwa suatu *dataset* publik untuk tujuan umum tugas pengenalan entitas bernama dalam Bahasa Indonesia bisa digunakan secara efektif untuk domain kebakaran hutan dengan akurasi validasi yang dapat diterima lebih dari 74% pada Bahasa Indonesia, Inggris, Spanyol, Italia dan Slovakia. Selain itu, akurasi validasi tertinggi diperoleh Bahasa Slovakia karena banyak *tweet* kebakaran hutan pada bahasa itu yang penulisannya baku berisi lokasi dan waktu dan memiliki sedikit *noise* seperti singkatan, salah penulisan dan penggunaan kata tidak baku.

Terdapat temuan yang menyebabkan perubahan akurasi validasi dalam mengenali entitas lokasi, waktu maupun lokasi dan waktu yaitu dikarenakan dalam proses validasi model terdapat nilai benar di entitas lokasi namun ada nilai salah di entitas waktu atau sebaliknya dan juga adanya nilai salah di kedua entitas.

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil analisa pada penelitian yang telah dilakukan terhadap tiga *pre-trained* multilingual model berbasis BERT dapat disimpulkan bahwa:

a. Hasil penelitian menunjukkan bahwa *pre-trained* model multibahasa XLM-RoBERTa (XLM-R) yang disempurnakan (*fine-tuning*) menggunakan *data train* dan *data validation nergrit corpus* memperoleh performa tertinggi dengan *accuracy* 98,59%, *precision* 91,89%, *recall* 92,73% dan *f1-score* 92,31% jika dibandingkan Multilingual BERT (mBERT) Cased dengan performa *accuracy* 98,52%, *precision* 90,69%, *recall* 92,02% dan *f1-score* 91,35% maupun Multilingual BERT (mBERT) Uncased dengan performa *accuracy* 98,38%, *precision* 90,22%, *recall* 91,90% dan *f1-score* 91,05%. Performa yang tinggi pada ketiga model menunjukkan bahwa *dataset* yang digunakan berkualitas tinggi dan teknik pengolahan data yang dilakukan pada *dataset* efektif sehingga model sangat baik dalam mengenali label entitas lokasi dan waktu.

b. Hasil penelitian menunjukkan bahwa model multilingual *named entity recognition* berbasis XLM-RoBERTa (XLM-R) yang diuji menggunakan *data testing* pada *nergrit corpus* memperoleh akurasi tertinggi sebesar 98,53% jika dibandingkan Multilingual BERT (mBERT) Cased dengan akurasi pengujian sebesar 98,47% dan Multilingual BERT (mBERT) Uncased dengan akurasi pengujian sebesar 98,46%. Akurasi pengujian yang tinggi pada ketiga model

menunjukkan bahwa model sangat akurat dalam memprediksi label entitas ketika diberikan data baru dari *dataset*. Selain itu, akurasi pengujian yang tinggi pada ketiga model menunjukkan bahwa teknik pengolahan data yang dilakukan pada *dataset* efektif, sehingga model sangat baik dalam mengenali label entitas lokasi dan waktu.

c. Hasil penelitian menunjukkan bahwa model multilingual *named entity recognition* berbasis XLM-RoBERTa (XLM-R) yang divalidasi secara manual dalam mengekstrak lokasi dan waktu dari *tweet* kebakaran hutan memperoleh akurasi tertinggi pada Bahasa Indonesia 92,32%, Bahasa Inggris 73,97%, Bahasa Spanyol 77,45%, Bahasa Italia 78,39% dan Bahasa Slovakia 96,50% jika dibandingkan Multilingual BERT (mBERT) Cased dengan akurasi validasi pada Bahasa Indonesia 71,16%, Bahasa Inggris 50,10%, Bahasa Spanyol 59,80%, Bahasa Italia 65,82% dan Bahasa Slovakia 75,80% dan Multilingual BERT (mBERT) Uncased dengan akurasi validasi pada Bahasa Indonesia 76,35%, Bahasa Inggris 70,06%, Bahasa Spanyol 75,10%, Bahasa Italia 75,64% dan Bahasa Slovakia sebesar 91,40%. Hasil validasi pada model XLM-R menunjukkan bahwa suatu *dataset* publik untuk tujuan umum tugas pengenalan entitas bernama dalam Bahasa Indonesia bisa digunakan secara efektif untuk domain kebakaran hutan dengan akurasi validasi yang dapat diterima lebih dari 74% pada Bahasa Indonesia, Inggris, Spanyol, Italia dan Slovakia.

d. Hasil penelitian menunjukkan bahwa model mBERT Cased, mBERT Uncased dan XLM-R mampu untuk mengenali entitas lokasi dan waktu dari *tweet* kebakaran hutan pada Bahasa Indonesia, Inggris, Spanyol, Italia dan Slovakia.

Namun, terdapat beberapa perbedaan hasil ekstraksi. Dimana, model mBERT Cased tetap mempertahankan teks aslinya tanpa mengubah bentuk huruf dan jenis huruf, namun bisa menghasilkan *noise* dengan mengenali sub token yang belum diagregasi menjadi token. Sedangkan, model mBERT Uncased mengubah bentuk semua huruf menjadi huruf kecil semua dan mengganti jenis huruf aksen menjadi huruf latin serta mampu mengenali lokasi pada *hashtag*. Selain itu, model XLM-R tetap mempertahankan teks aslinya tanpa mengubah bentuk huruf dan jenis huruf serta mampu mengenali lokasi pada *hashtag*.

e. Hasil penelitian menunjukkan bahwa model multilingual *named entity recognition* berbasis Multilingual BERT (mBERT) Cased, Multilingual BERT (mBERT) Uncased dan XLM-RoBERTa (XLM-R) memperoleh akurasi validasi tertinggi dalam mengekstrak lokasi dan waktu pada Bahasa Slovakia yang menggunakan huruf alfabet-aksen, meskipun model dilatih menggunakan *dataset nergrit corpus* dalam Bahasa Indonesia yang menggunakan huruf alfabet saja. Hal itu dipengaruhi karena banyaknya *tweet* kebakaran hutan pada Bahasa Slovakia yang penulisannya baku berisi lokasi atau waktu dan memiliki sedikit *noise* seperti singkatan, salah penulisan dan penggunaan kata tidak baku.

f. Hasil penelitian menunjukkan bahwa jenis model *cased* yang memperhatikan huruf kapital di *dataset* tidak selalu lebih baik dari model *uncased* yang hanya menggunakan huruf kecil dalam melakukan ekstraksi entitas lokasi dan waktu kebakaran hutan dari *tweet* secara multibahasa, ditunjukkan dengan akurasi validasi yang lebih tinggi pada model Multilingual BERT (mBERT) Uncased jika dibandingkan dengan Multilingual BERT (mBERT) Cased.

5.2. Saran

Berikut adalah beberapa saran yang dapat dijadikan pedoman untuk melakukan pengembangan penelitian ini, diantaranya:

a. Sebaiknya dilakukan penambahan *dataset* pada bahasa lain dengan kualitas *dataset* yang sama ataupun dibuatkan *dataset* khusus pada domain kebakaran hutan supaya hasil pengenalan entitas lebih optimal.
b. Sebaiknya melakukan *hyperparameter tuning* untuk mencari *hyperparameter* model terbaik sehingga mendapatkan performa yang lebih optimal.
c. Sebaiknya menambah metode evaluasi lain seperti evaluasi pengenalan entitas yang menampilkan performa masing-masing entitas.
d. Sebaiknya kriteria validasi model dijelaskan lebih detail untuk setiap kondisi sehingga mendapatkan analisis data yang lebih lengkap.
e. Sebaiknya menambah data validasi pada *platform* media sosial dan bahasa lain untuk menambah daftar performa validasi model.
f. Menambahkan skenario penggunaan pendekatan pemrosesan bahasa alami pada teks media sosial untuk menambah daftar performa validasi model.
g. Mengembangkan hasil ekstraksi lokasi dan waktu untuk mendeteksi lokasi berdasarkan teknik geolokasi (*geocoding*).
h. Dikembangkan sistem manajemen bencana kebakaran hutan untuk sistem pendukung keputusan terintegrasi big data yang terdiri dari data teknologi *remote sensing*, *social media sensing*, *wireless sensing*, dan lainnya.

DAFTAR PUSTAKA

Agustiar, A.B., Mustajib, M., Amin, F. & Hidayatullah, A.F. 2020. Kebakaran hutan dan lahan perspektif etika lingkungan. *Profetika: Jurnal Studi Islam*, 20(2): 124–132.

Aini, N.N. & Basuki, R.S. 2020. Pengaruh Electronic Word of Mouth Media Sosial Instagram@ Gartenhaus_Co dan Store Atmosphere Terhadap Minat Beli di Cafe Gartenhaus Malang Jawa Timur. *JAB: Jurnal Aplikasi Bisnis*, 6(1): 25–28.

Arisanty, D., Anis, M.Z.A., Putro, H.P.N., Muhaimin, M. & Syarifuddin 2020. *Kebakaran Lahan Gambut: Faktor Penyebab dan Mitigasinya*. Banjarmasin: Program Studi Pendidikan IPS, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Lambung Mangkurat.

Berragan, C., Singleton, A., Calafiore, A. & Morley, J. 2023. Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science*, 37(4): 747–766.

Budi, I. & Suryono, R.R. 2023. Application of named entity recognition method for Indonesian datasets: a review. *Bulletin of Electrical Engineering and Informatics*, 12(2): 969–978.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eligüzel, N., Çetinkaya, C. & Dereli, T. 2022. Application of named entity recognition on tweets during earthquake disaster: a deep learning-based approach. *Soft Computing*, 26(1): 395–421.

Finkel, J.R., Grenager, T. & Manning, C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 363–370.

Indradjad, A., Purwanto, J. & Sunarmodo, W. 2019. Analisis tingkat akurasi titik hotspot dari S-NPP VIIRS dan TERRA/AQUA MODIS terhadap kejadian

kebakaran. *Jurnal Penginderaan Jauh dan Pengolahan Data Citra Digital*, 16(1): 53–60.

Jati, B.S., Widyawan, S.T. & Muhammad Nur Rizal, S.T. 2020. Multilingual Named Entity Recognition Model for Indonesian Health Insurance Question Answering System. *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*. hlm.180–184.

Kaku, K. 2019. Satellite remote sensing for disaster management support: A holistic and staged approach based on case studies in Sentinel Asia. *International Journal of Disaster Risk Reduction*, 33: 417–432.

Khalid, H., Murtaza, G. & Abbas, Q. 2023. Using Data Augmentation and Bidirectional Encoder Representations from Transformers for Improving Punjabi Named Entity Recognition. *ACM Transactions on Computing Education*.

Li, J., He, Z., Plaza, J., Li, S., Chen, J., Wu, H., Wang, Y. & Liu, Y. 2017. Social Media: New Perspectives to Improve Remote Sensing for Emergency Response. *Proceedings of the IEEE*, 105(10): 1900–1912.

Li, L., Ma, Z. & Cao, T. 2021. Data-driven investigations of using social media to aid evacuations amid Western United States wildfire season. *Fire Safety Journal*, 126: 103480.

MacCarthy, J., Richter, J., Tyukavina, S., Mikaela, W. & Harris, N. 2023. *The Latest Data Confirms: Forest Fires Are Getting Worse*. World Resources Institute.

Nugraha, F.A., Harani, N.H. & Habibi, R. 2020. *Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning*. Kreatif.

Phengsuwan, J., Shah, T., Thekkummal, N.B., Wen, Z., Sun, R., Pullarkatt, D., Thirugnanam, H., Ramesh, M.V., Morgan, G., James, P. & Ranjan, R. 2021. Use of Social Media Data in Disaster Management: A Survey. *Future Internet*, 13(2).

Pires, T., Schlinger, E. & Garrette, D. 2019. How multilingual is Multilingual BERT? *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 4996–5001.

PT Gria Inovasi Teknologi 2019. *Nergrit Corpus*. Huggingface. Tersedia di https://huggingface.co/datasets/id_nergrit_corpus [Accessed 15 Januari 2024].

Raaijmakers, S. 2022. *Deep Learning for Natural Language Processing*. Manning.

Salas, E.B. 2023. *Number of deaths due to wildfires worldwide from 1990 to 2023*. Statita. Tersedia di https://www.statista.com/statistics/1293254/global-number-of-deaths-due-to-wildfires/ [Accessed 11 November 2023].

Salminen, J., Hopf, M., Chowdhury, S.A., Jung, S., Almerekhi, H. & Jansen, B.J. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10: 1–34.

Shah, S.A., Yahia, S. Ben, McBride, K., Jamil, A. & Draheim, D. 2021. Twitter Streaming Data Analytics for Disaster Alerts. *2021 2nd International Informatics and Software Engineering Conference (IISEC)*. hlm.1–6.

Sharma, A., Amrita, Chakraborty, S. & Kumar, S. 2022. Named entity recognition in natural language processing: A systematic review. *Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021*. hlm.817–828.

Shi, K., Peng, X., Lu, H., Zhu, Y. & Niu, Z. 2022. Application of Social Sensors in Natural Disasters Emergency Management: A Review. *IEEE Transactions on Computational Social Systems*.

Silvanus Amikom 2022. *SILVANUS Press Release (Indonesia) - Silvanus & Amikom Project*. Silvanus & Amikom Project. Tersedia di https://silvanus.amikom.ac.id/index.php/2022/03/01/silvanus-press-release-indonesia/ [Accessed 7 Juni 2024].

Suganda Girsang, A. & Noveta, B.K. 2022. Location Prediction Using Named Entity Recognition for Indonesia Natural Disasters in Data Twitter. *SSRN Electronic Journal*.

Sun, J., Liu, Y., Cui, J. & He, H. 2022. Deep learning-based methods for natural hazard named entity recognition. *Scientific Reports 2022 12:1*, 12(1): 1–15.

Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P. & Xiong, C. 2020. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *arXiv preprint arXiv:2003.04985*.

Suwaileh, R., Elsayed, T., Imran, M. & Sajjad, H. 2022. When a disaster happens, we are ready: Location mention recognition from crisis tweets. *International Journal of Disaster Risk Reduction*, 78: 103107.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wahyudi, M. 2021. Analisis Kebijakan Pencegahan Dan Penanganan Kebakaran Hutan Dan Lahan Di Provinsi Kalimantan Tengah: Policy Analysis Of Forest

And Land Fire Prevention And Management In Central Kalimantan Province. *Anterior Jurnal*, 20(2): 153–159.

Wirth, R. & Hipp, J. 2000. CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. hlm.29–39.

Yang, Y., Wu, Z., Yang, Y., Lian, S., Guo, F. & Wang, Z. 2022. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19): 9691.

Young, T., Hazarika, D., Poria, S. & Cambria, E. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3): 55–75.

Zhang, D., Wang, D., Vance, N., Zhang, Y. & Mike, S. 2019. On Scalable and Robust Truth Discovery in Big Data Social Media Sensing Applications. *IEEE Transactions on Big Data*, 5(2): 195–208.