Ekstraksi Informasi Beasiswa dari Media Sosial menggunakan BiLSTM-CRF

1st Muhammad Rizki Ramadhan Setiawan
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
mrizkiramadhans@students.telkomuniversity.ac.id

2nd Ade Romadhony
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
aderomadhony@telkomuniversity.ac.id

3rd Hasmawati
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
hasmawati@telkomuniversity.ac.id

Abstrak-Sosial media merupakan tempat dimana orang-orang berkumpul dan saling bertukar informasi. Dari informasi tersebut dapat muncul berbagai macam peluang seperti beasiswa yang dikeluarkan oleh lembaga pendidikan. Peluang ini dapat banyak ditemukan pada sosial media seperti Twitter. Namun kebanyakan informasi yang dikeluarkan menggunakan format tersendiri sehingga menjadi tidak terstruktur dan menghambat upaya pengolahan informasi yang terkait. Melihat cepatnya informasi berlalu dan banyaknya kompetisi dalam meraih peluang tersebut, efisiensi menjadi faktor penting dalam mengumpulkan dan memproses informasi. Untuk mengatasi permasalahan tersebut, maka dilakukan ekstraksi informasi untuk mengubah informasi tidak terstruktur menjadi terstruktur menggunakan metode *Bidirectional Long-Short Term Memory* dengan *Conditional Random Fields* (BiLSTM-CRF). Metode ini digunakan karena dapat memberikan konteks informasi dari masa lalu dan masa depan pada teks sehingga sesuai untuk mengatasi tugas ekstraksi informasi. Tujuan penelitian ini adalah melakukan ekstraksi informasi dengan mengimplementasikan model BiLSTM-CRF untuk melakukan proses klasifikasi informasi yang diekstraksi sesuai dengan kategori pengelompokkan yang ditetapkan sehingga data yang terkumpul menjadi terstruktur dan mudah untuk dibaca. Hasil yang didapatkan dari implementasi model tersebut adalah nilai performansi dengan *Precision* 90%, *Recall* 51%, dan *F1-Score* sebesar 54%.

Kata kunci - beasiswa, twitter, sequence labelling, BiLSTM-CRF, ekstraksi informasi

Abstract-Social media is a place where people gather and exchange information. From this information, various opportunities can emerge, such as scholarships issued by educational institutions. This opportunity can be found on social media such as Twitter. However, most of the information released uses its own format so that it becomes unstructured and hampers the processing of related information. Given the speed at which information passes and there is a lot of competition for these opportunities, efficiency is an important factor in gathering and processing information. To overcome this problem, information extraction is carried out to change unstructured information into structured using the Bidirectional Long-Short Term Memory method with Conditional Random Fields (BiLSTM-CRF). This method is used because it can provide context of information from the past and future in the text so that it is suitable for solving the task of extracting information. The purpose of this research is to extract information by implementing the BiLSTM-CRF model to classify the extracted information according to the defined grouping category so that the collected data becomes structured and easy to read. The results obtained from the implementation of the model are performance values with 90% Precision, 51% Recall, and 54% F1-Score.

Keywords- scholarship, twitter, sequence labeling, BiLSTM-CRF, information extraction

I. PENDAHULUAN

A. Latar Belakang

Tidak hanya sebagai tempat bersosialisasi, media sosial seperti Twitter juga dapat menjadi sebuah platform penyebaran informasi dengan berbagai macam format seperti teks, gambar, dan video. Informasi yang mengalir pada media sosial dapat membahas topik yang beragam variasinya dari aktivitas keseharian seorang pengguna, sampai berita peristiwa besar dalam kehidupan. Sosial media juga dapat bersaing dengan outlet berita lainnya sebagai sumber berita terkini bagi beberapa orang.

Sama halnya seperti informasi lainnya yang disebarkan, lembaga pendidikan seperti universitas juga memanfaatkan platform ini sebagai tempat penyebaran informasi. Salah satunya yaitu informasi mengenai beasiswa. Sehingga Twitter dapat menjadi salah satu tempat ideal yang digunakan untuk mengumpulkan data penelitian. Namun, kebanyakan pengguna Twitter menyebarkan informasi tersebut menggunakan format tersendiri. Hal ini menyebabkan data yang dikumpulkan menjadi tidak terstruktur dan dapat memperlambat usaha pengguna dalam mengumpulkan informasi penting. Melihat cepatnya informasi berlalu serta banyaknya kompetisi dalam pencarian kesempatan seperti beasiswa, efisiensi merupakan faktor penting dalam perlombaan meraih peluang yang terbatas tersebut.

Berdasarkan permasalahan yang diuraikan, Tugas Akhir ini mengusulkan untuk melakukan

Information Extraction (IE) untuk mengambil informasi-informasi yang relevan dan mengubah data tidak terstruktur menjadi data terstruktur dengan format yang mudah dipahami. Ekstraksi informasi didefinisikan sebagai ekstraksi otomatis informasi tertentu dari teks bahasa alami. Lebih khusus lagi, IE secara otomatis memproses teks bahasa alami untuk mengekstrak informasi dari kelas entitas, hubungan, atau peristiwa tertentu[10].

Dalam penelitian ini, dilakukan implementasi ekstraksi informasi yang relevan dari kumpulan twit bahasa inggris mengenai topik beasiswa S3 yang dikirimkan oleh berbagai pihak pengguna Twitter. Implementasi dilakukan dengan melaksanakan tugas *Sequence Labelling* yaitu salah satu metode yang digunakan dalam *Natural Language Processing* (NLP). *Sequence Labelling* (SE) adalah jenis tugas pengenalan pola di cabang penting dari NLP. Pelaksanaan tugas Sequence Labelling dilakukan dengan membangun model yang menggunakan arsitektur *Bidirectional Long Short Term Memory Network* (BiLSTM) dan *Conditional Random Fields* (CRF).

BiLSTM-CRF digunakan karena BiLSTM dapat mengatasi vanishing and exploding gradient problem dan memberikan konteks informasi dua arah untuk mengatasi ketergantungan jarak jauh antar kata[12]. komponen CRF sering digunakan untuk mengatasi masalah berbasis sekuens dalam NLP. CRF memberikan informasi label tambahan pada tingkat kalimat[15]. Penggabungan CRF dengan BiLSTM juga menunjukkan peningkatan performansi dalam melakukan tugas *sequence labelling*[12].

B. Topik dan Batasannya

Berdasarkan uraian pada latar belakang, dapat diangkat permasalahan utama pada tugas akhir ini, yaitu bagaimana cara mengimplementasikan model ekstraksi informasi beasiswa pada twit pengguna Twitter menggunakan *BiLSTM-CRF*. Adapun batasan dari permasalahan yang diangkat, yaitu:

1. Dataset yang digunakan merupakan informasi beasiswa S3 yang diperoleh dari twit pengguna twitter.
2. Data menggunakan Bahasa Inggris
3. Data yang dimiliki menggunakan format skema BIO dimana labelnya terbagi menjadi:
    a. *Field* : Bidang yang berkait dengan beasiswa
    b. *Univ* : Universitas dikeluarkannya beasiswa tersebut
    c. *Country* : Negara tempat dikeluarkannya beasiswa tersebut
    d. *Link* : Tautan yang memiliki infor terkait dengan beasiswa yang dikeluarkan
    e. *Deadline* : Batas waktu tersedianya pendaftaran beasiswa tersebut.
    f. *Topic* : Topik yang berkait dengan beasiswa yang dikeluarkan
    g. *Lab* : Lembaga lab yang berkait dengan beasiswa yang dikeluarkan

C. Tujuan

Tujuan dari pelaksanaan tugas akhir ini adalah mengimplementasikan model yang dapat melakukan ekstraksi informasi beasiswa S3 menggunakan arsitektur *BiLSTM-CRF* pada dataset Twitter. Hasil yang didapat kemudian dianalisis dan dilakukan evaluasi terhadap performansi model yang diimplementasikan.

D. Organisasi Tulisan

Penyusunan Laporan Tugas Akhir ini terbagi menjadi 5 bagian. Dimulai pada bagian pertama, terdapat pembahasan mengenai pendahuluan yang mencakup latar belakang, topik dan batasannya, serta tujuan pelaksanaan Tugas akhir ini. bagian kedua menjelaskan studi terkait dan penelitian lainnya yang behubungan dengan topik pada Tugas akhir ini. Bagian ketika menjelaskan arsitektur yang digunakan, yaitu *BiLSTM-CRF* dan alur kerja model yang dibangun. Bagian keempat menjelaskan hasil analisis dari keluaran model dan kinerjanya. Bagian kelima membahas kesimpulan yang dapat ditarik dari pelaksanaan Tugas Akhir ini.

II. KAJIAN TEORI

A. *Natural Language Processing*

Natural Language Processing (NLP) adalah subbidang dari *Artificial Intelligence* (AI) dan ilmu linguistik, yang dikhususkan untuk membuat komputer memahami pernyataan atau kata-kata yang ditulis dalam bahasa manusia. *Natural Language* juga dikenal sebagai bahasa sederhana yang diucapkan atau ditulis oleh orang (manusia) untuk tujuan umum komunikasi. Bahasa alami muncul karena adanya keinginan pengguna untuk berkomunikasi dengan komputer. Namun, pengguna tidak dapat dipaksakan untuk mempelajari bahasa mesin tertentu. Bahasa alami ini pada dasarnya digunakan untuk orang-orang yang tidak memiliki waktu untuk mempelajari bahasa spesifik baru atau mengembangkan keahlian yang dibutuhkan. Bahasa alami yang digunakan oleh manusia diantaranya adalah seperti bahasa Hindi, Prancis, Inggris, dan Cina[5].

B. *Information Extraction*

Information extraction(IE) merupakan proses ekstraksi informasi berguna terstruktur dari data yang tidak terstruktur dalam bentuk entitas, relasi, objek, peristiwa, dan tipe lainnya. Informasi yang diekstraksi dari data yang tidak terstruktur digunakan untuk menyiapkan data untuk dianalisis[1]. Oleh karena itu, transformasi data tidak terstruktur yang efisien dan akurat dalam proses IE meningkatkan hasil analisis data. Banyak teknik yang telah diperkenalkan untuk tipe data yang berbeda seperti teks, gambar, audio, dan video[1]. Tujuan dari IE adalah untuk mengambil dan mengorganisir informasi penting dari teks dalam format yang dapat mudah dimengerti serta diproses agar dapat digunakan oleh komputer.

C. *Sequence Labelling*

Sequence Labelling (SE) adalah jenis tugas pengenalan pola di cabang penting dari NLP. Dari perspektif linguistik, unit bermakna terkecil dalam bahasa biasanya dianggap sebagai morfem, dan setiap kalimat dengan demikian dapat dilihat sebagai urutan yang terdiri dari morfem. Dengan demikian, masalah pelabelan urutan dalam domain NLP dapat dirumuskan sebagai tugas yang bertujuan untuk menetapkan label untuk kategori morfem yang umumnya memiliki peran yang sama dalam struktur gramatikal kalimat dan memiliki kesamaan sifat tata bahasa, dan arti dari yang ditugaskan label biasanya tergantung pada jenis tugas tertentu[4]. *Sequence Labelling* telah berhasil diterapkan ke sejumlah tugas NLP yang mengandalkan informasi kontekstual, seperti *named entity recognition*, *part-of-speech tagging* dan *shallow parsing*[6]. Tujuan pelabelan urutan adalah untuk menetapkan urutan label, yang diambil dari alfabet yang ditentukan, ke urutan data input. [8]

D. *Name Entity Recognition*

Named Entity Recognition (NER) merupakan salah satu tugas dalam IE dan Sequence Labeling yang terdiri dari mengidentifikasi dan mengklasifikasikan beberapa jenis elemen informasi, yang disebut Named Entity (NE)[2]. Tugas NER pada intinya terbagi menjadi 2, yaitu proses identifikasi *proper name* dalam teks dan proses klasifikasi nama-nama tersebut menjadi kategori yang telah ditentukan sebelumnya yang sesuai dengan kebutuhan, seperti nama orang, organisasi, lokasi, tanggal, dan waktu[3].

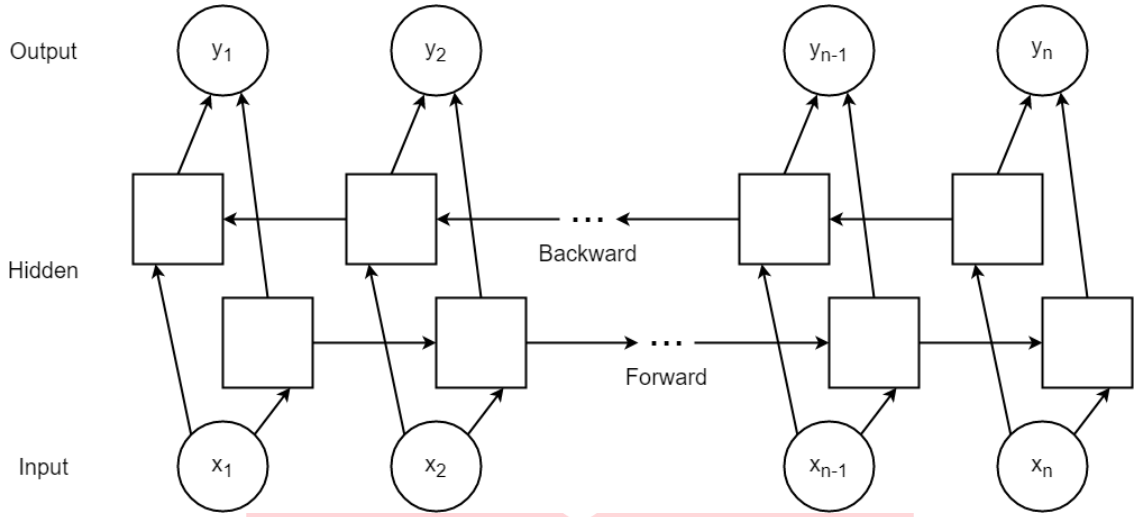
NER adalah tugas yang sangat penting di alam NLP dan merupakan teknologi dasar untuk banyak aplikasi tingkat tinggi, seperti *search engine, question and answer systems, recommendation systems,* dan *translation systems*.[4].

E. Skema *Labelling*

Umumnya, label kata dalam NER terdiri dari dua: bagian, yaitu, "X-Y", di mana "X" menunjukkan posisi kata berlabel dan "Y" mengacu pada kategori yang sesuai dalam taksonomi yang telah ditentukan sebelumnya. Secara khusus, mungkin pelabelan dilakukan dengan pemberian label khusus (misalnya, "*none*"), jika sebuah kata tidak dapat diklasifikasikan ke dalam kategori yang telah ditentukan sebelumnya. Umumnya, skema penandaan yang diadopsi secara luas di industri adalah sistem BIOES, yang adalah, kata berlabel "B" (*Begin*), "I" (*Inside*) dan "E" (*End*) berarti bahwa itu adalah kata pertama, tengah atau terakhir dari entitas bernama frase, masing-masing. Kata berlabel "0-" (*Outside*) berarti itu bukan milik frase entitas bernama dan "S-" (Tunggal) menunjukkan itu adalah satu-satunya kata yang mewakili entitas[4]. Namun pada penelitian ini digunakan skema lain yaitu BIO dimana *Beginning* (B) menandakan awal dari suatu frase label, *Inside* (I) menandakan bagian dari suatu frase label, dan *Outside* (O) menandakan tidak termasuk frase label yang telah ditentukan.

F. *Bidirectional Long Short Term Memory Network* (BiLSTM)

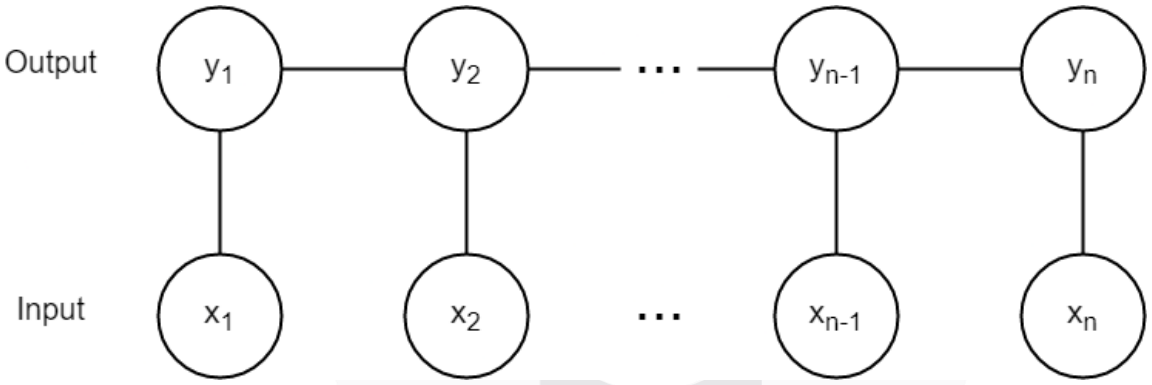
Long Short Term Memory Networks (LSTM) adalah jenis khusus dari *Recurrent Neural Networks* (RNNs) dan dirancang untuk mengatasi *gradient vanishing problem* dari RNNs. Secara khusus, LSTM memiliki sel memori tambahan, yang dapat menyimpan memori dari *long distance terms* [7]. Mayoritas tugas *sequence labelling*, akan terbantu pelaksanaannya dengan adanya akses terhadap konteks masa depan dan masa lalu. Dengan menggunakan LSTM sebagai arsitektur *network* pada *bidirectional* RNN, maka dihasilkan *Bidirectional* LSTM yang menyediakan akses terhadap konteks jarak jauh pada kedua arah tersebut[8].



GAMBAR 1
CONTOH ARSITEKTUR LSTM

G. Conditional Random Field (CRF)

Conditional Random Field (CRF) adalah kerangka kerja probabilistik untuk pelabelan dan segmentasi data. CRF merupakan bentuk model grafis tidak berarah yang mendefinisikan distribusi log-linier tunggal atas urutan label yang diberikan urutan pengamatan tertentu. CRF mendefinisikan distribusi probabilitas bersyarat
$P(\ Y \mid X\ )$ dari urutan label yang diberikan urutan input[9].



GAMBAR 2
CONTOH LAPISAN CONDITIONAL RANDOM FIELD (CRF)

H. Penelitian Terkait

Penelitian mengenai ekstraksi informasi sudah dilakukan dalam penelitian sebelumnya, seperti ekstraksi kata kunci pada artikel ilmiah[12] berdasarkan abstrak dari 2000 artikel ilmiah yang disertai dengan dua set kata kunci beranotasi secara manual oleh manusia, yaitu *controlled* yang diberikan oleh penulis dan *uncontrolled* yang diberikan oleh pembaca. Permasalahan ini diangkat sebagai tugas *sequence labelling* dan memanfaatkan *contextual embedding* sebagai ekstraksi kata kunci dengan menggunakan dan membandingkan arsitektur *BiLSTM-CRF* dan *BiLSTM* [12]. Hasil yang didapatkan pada penelitian ini menunjukkan bahwa model dengan arsitektur BiLSTM-CRF merupakan arsitektur terbaik dan konsisten di semua dataset.

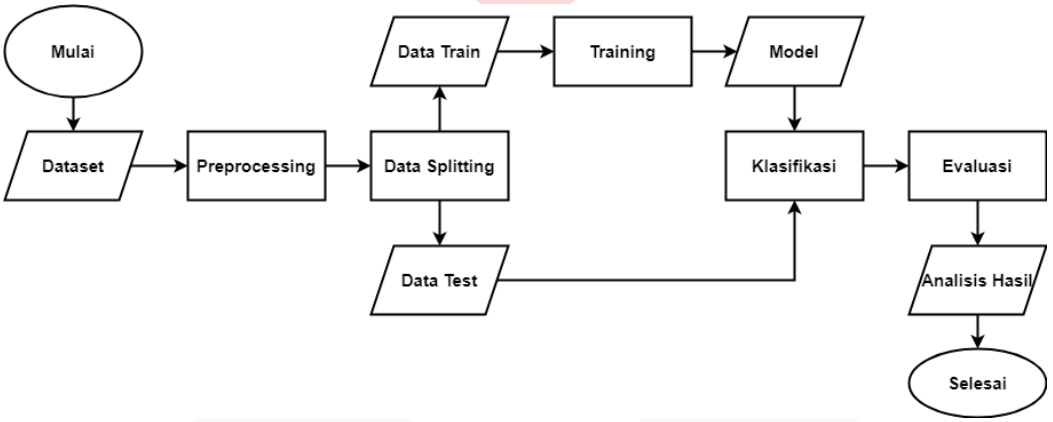
Penelitian lainnya yaitu ekstraksi informasi dari pesan-pesan pada Twitter berbasis ontologi (OBIE) yang menggunakan pendekatan rule-based[13]. Penelitian tersebut menggabungkan Named Entity Recognition (NER) dan modul yang disambiguasi terhadap pesan-pesan singkat yang berasal dari BBC News, New York Times, dan akun Twitter The Times[13]. Hasil yang didapat menunjukkan bahwa sistem OBIE mendapatkan peningkatan ketika menggunakan proses diasmbiguation dengan skor F-Measures traditional mencapai 86% dan Augmented F-Measures 90%.

Selanjutnya penelitan yang menggunakan NER [14] untuk melakukan ekstraksi informasi seperti keahlian dan sejarah karir pada 543 dokumen resume dalam bahasa inggris dan 142 dokumen resume dalam bahasa jepang dengan format PDF. Pelaksanaan tugas NER dilakukan dengan menggunakan metode *neural network* yang memanfaatkan arsitektur BiLSTM-LSTM-CRF dengan beberapa variasi *word/character embedding* pada pelaksanaan eksperimen. Hasil yang didapat adalah untuk dataset dokumen inggris, penggunaan *pre-trained word embeddings* membantu meningkatkan hasil yang didapatkan dengan *Precision* 76.40, *Recall* 75.74, dan *F1-Score* 76,07. Namun, apabila dibandingkan dengan dataset dokumen jepang, maka didapatkan tren yang berbeda dimana *character embedding* tanpa *pre-trained word embedding* menunjukkan hasil yang lebih baik dari yang lain dengan *Precision* 74.70, *Recall* 77.17, dan *F1-Score* 74.81.

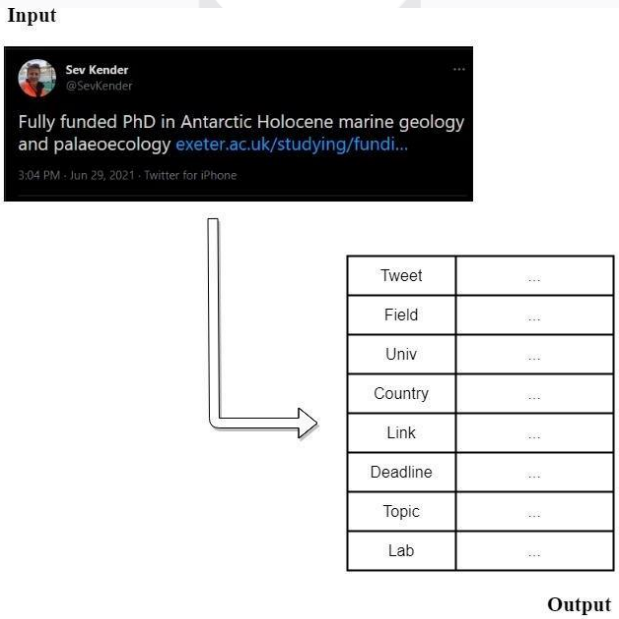
III. METODE

A. Gambaran umum Sistem

Secara umum gambaran sistem memiliki beberapa tahapan dalam memproses teks untuk melakukan ekstraksi informasi. Pertama, sistem melakukan *preprocessing* terhadap dataset yang didapatkan. Kemudian dilakukan pembagian dataset yang telah diolah menjadi *train* data dan *test* data. Data *train* digunakan untuk melatih model mengklasifikasi dan memberikan label yang sesuai terhadap input yang diterima model. Kemudian data *test* digunakan untuk melakukan ekstraksi informasi dengan memberikan label terhadap input model dan dilakukan pengelompokan output sesuai dengan kategori yang ditentukan dan disimpan dalam format yang mudah dipahami. Hasil klasifikasi yang dikeluarkan oleh model kemudian dievaluasi dari performansi dan akurasi pelabelan yang dilakukan model.



GAMBAR 3
GAMBARAN UMUM SISTEM



GAMBAR 4.
ILUSTRASI INPUT DAN OUTPUT YANG DIHARAPKAN

B. Dataset

Dataset yang digunakan merupakan kumpulan twit yang diambil dari Twitter mengenai informasi beasiswa S3 yang telah dilabeli dengan menggunakan skema BIO. Skema tersebut digunakan karena terdapat entitas atau label yang memiliki panjang token lebih dari satu. "B-" menandakan awalan suatu entitas atau label baru, "I-" menandakan bahwa token tersebut merupakan bagian dari entitas atau label tersebut, dan "O" menandakan bahwa token tersebut tidak termasuk label manapun yang telah ditentukan. Dataset diperoleh dari hasil pengerjaan kerja praktek mahasiswa pada tahun 2020-2021. Terdapat 165 twit atau 5690 token dengan label BIO yang sesuai. Berikut merupakan contoh representasi dataset yang digunakan.

TABEL 1
CONTOH REPRESENTASI DATA YANG DIGUNAKAN

Token	**BIO Tag**
Fully	O
funded	O
PhD	O
in	O
Antarctic	B-B-Field
Holocene	B-B-Field
marine	B-I-Field
geology	B-I-Field
and	B-I-Field
palaeoecology	B-I-Field
https://t.co/C4OWhWMCCf	B-B-Link

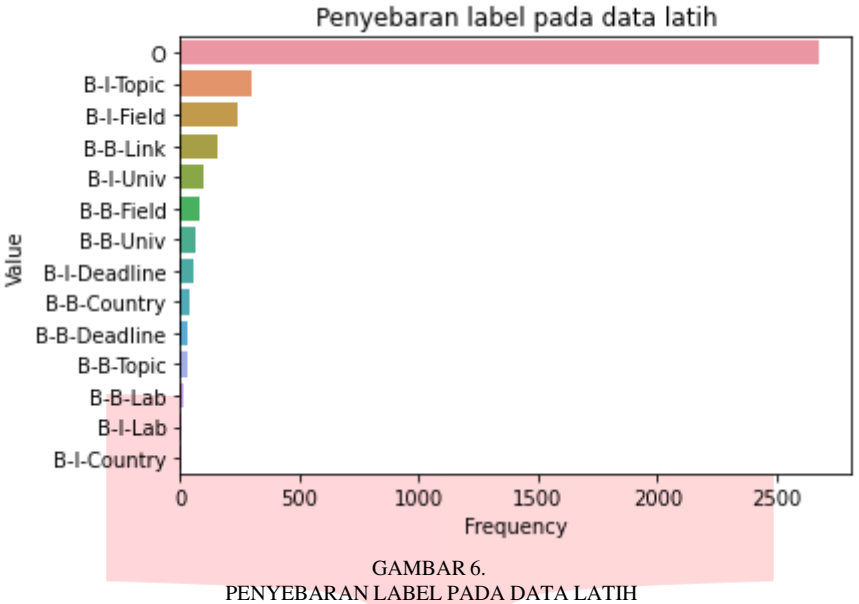
C. *Preprocessing*

Pada tahap ini dataset yang dimiliki akan melalui beberapa tahap pengolahan agar dapat dipahami oleh komputer. Tahapan yang dilalui yaitu:

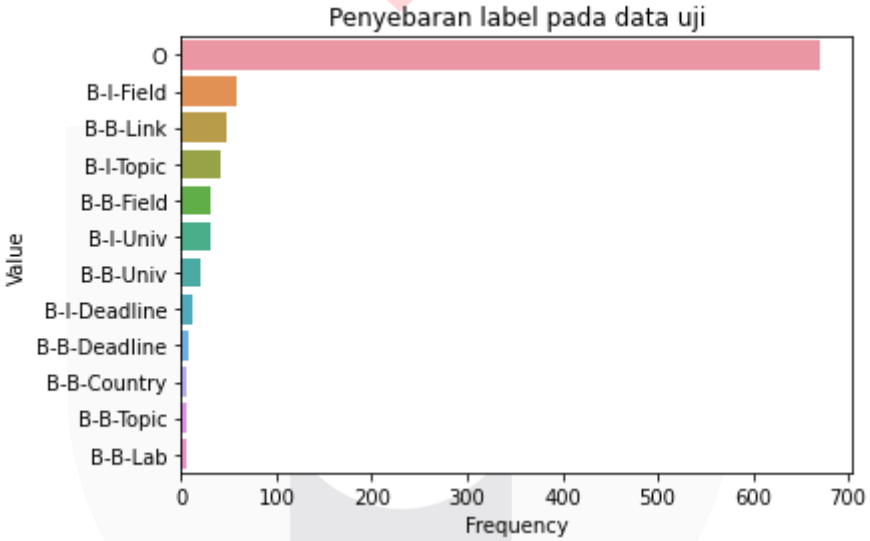
1. *Tokenizing*
   Setiap twit yang ada pada dataset dipecah menjadi sekumpulan kata dengan label BIO yang sesuai.
2. *Filtering*
   merupakan tahap dimana dilakukan penghapusan karakter spesial seperti tanda baca, sehingga dataset hanya mengandung karakter alfabet dan angka.
3. *Case Folding*
   Ukuran teks pada dataset diubah menjadi ukuran yang sama, yaitu *lower case* untuk setiap twit pada dataset.
4. *Encoding*
   Setiap token dan label diubah menjadi suatu bilangan bulat (*integer*) unik sesuai dengan kemunculan token dan label yang unik tersendiri. Label diubah dalam bentuk *one-hot encoding* dimana setiap kelas dikategorikan dalam bentuk *binary*.
5. *Padding*
   Model yang dibangun hanya dapat menerima input dengan panjang yang serupa. Oleh karena itu, setiap twit dengan panjang yang bervariasi diberikan *padding* sesuai dengan twit yang paling panjang, yaitu 46 token/kata.

D. Data *Splitting*

Dataset yang telah diproses dibagi menjadi data latih (70%) untuk pelatihan dan pembetukan model, data uji (20%) sebagai input pada tahap klasifikasi, serta data validasi (10%) yang digunakan untuk menyesuaikan *hyperparameter* model yang dibangun.



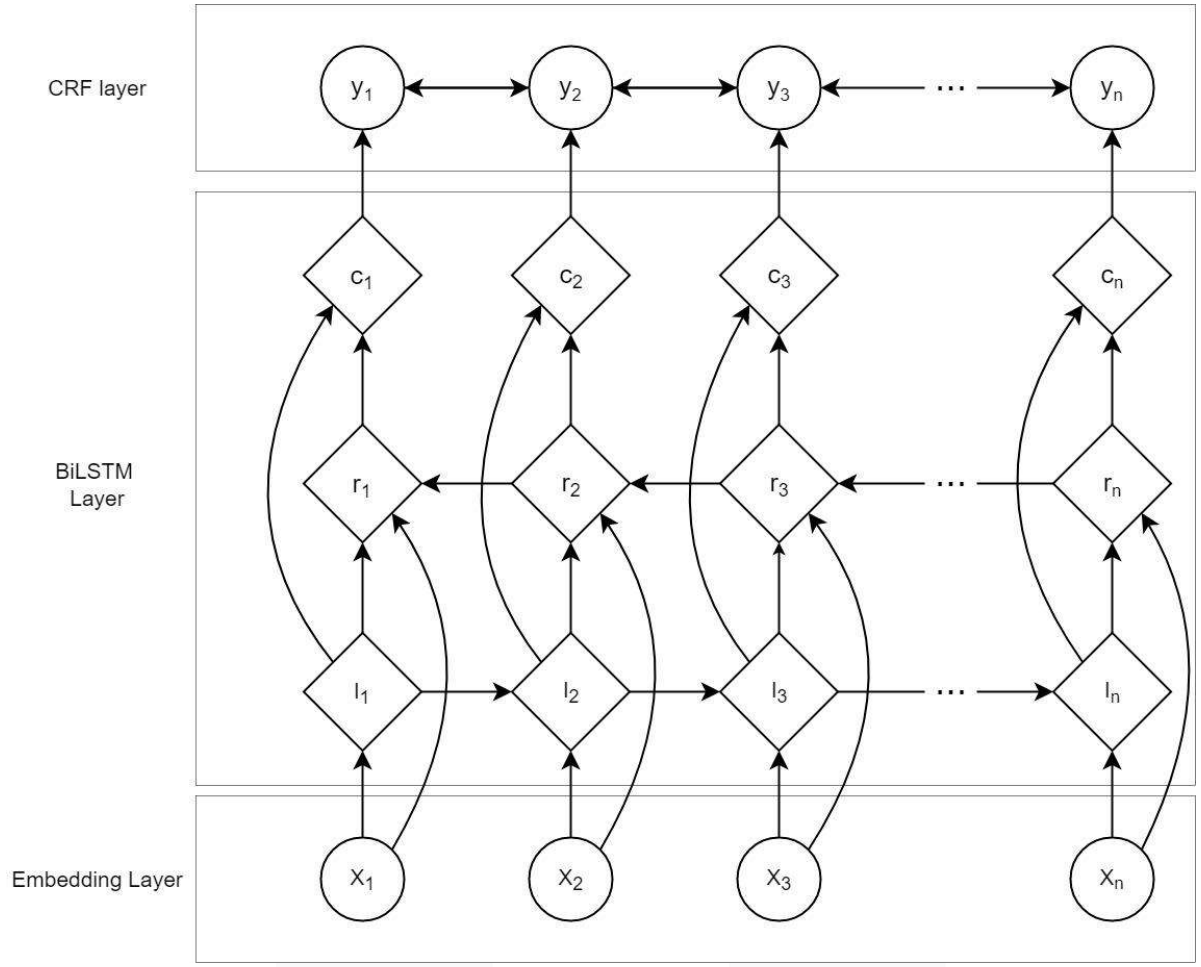
GAMBAR 6.
PENYEBARAN LABEL PADA DATA LATIH



GAMBAR 7.
PENYEBARAN LABEL PADA DATA UJI

E. BiLSTM-CRF

Arsitektur yang digunakan untuk membangun model adalah dengan menggunakan Bidirectional LSTM (BiLSTM) yang ditambahkan dengan lapisan *Conditional Random Field* (CRF).



GAMBAR 8.
CONTOH ARSITEKTUR BILSTM UNTUK *SEQUENCE LABELLING*

Model tersebut mengambil setiap token dari sequence sebagai input yang kemudian dipetakan menjadi vektor kata pada lapisan *embeddings*. Dalam implementasinya, input model tersebut merupakan vektor *integer* dengan panjang yang sama.

Kemudian pada lapisan BiLSTM, diambil informasi yang relevan dengan menggabungkan informasi 2 arah yang didapatkan dari I_n (kiri ke kanan) dan r_n (kanan ke kiri) yang kemudian disimpan pada c_n sebagai probabilitas label setiap token sebagai output dari lapisan tersebut.

Pada lapisan CRF, dilakukan prediksi dengan mengambil nilai probabilitas terbesar dari keluaran lapisan sebelumnya, dengan mempertimbangkan rangkaian sekuens output secara keseluruhan.

TABEL 2.
CONTOH INPUT DAN OUTPUT MODEL

Input	['phd', 'opportunity', 'estimating', 'sources', 'and', 'population', 'exposure', 'to', 'voc', 'pollution', 'indoors', 'at', '@atmoschemyork', 'fully', 'funded', 'for', '3.5', 'yrs', 'by', '@atmosscience', 'closes', 'mon', '19', 'july', 'https://t.co/jpnyzj99mq"]
Output	['O', 'O', 'B-B-Topic', 'B-I-Topic', 'B-I-Topic', 'B-I-Topic', 'B-I-Topic', 'B-I-Topic', 'B-I-Topic', 'B-I-Topic', 'B-I-Topic', 'O', 'B-B-Lab', 'O', 'O', 'O', 'O', 'O', 'O', 'B-B-Lab', 'O', 'B-B-Deadline', 'B-I-Deadline', 'B-I-Deadline', 'B-B-Link']

F. Klasifikasi

Model diberikan input data latih untuk melakukan klasifikasi dengan memprediksi rangkaian label yang tepat untuk input yang diterima model. hasil klasifikasi kemudian diubah menjadi bentuk tabular sehingga mudah untuk dibaca.

TABEL 3.
CONTOH TABEL HASIL KLASIFIKASI BILSTM-CRF

Kategori	Deskripsi
Twit	fully funded phd positions at the university of nebraska at omaha usa starting january 2022 if interested send your cv undergraduate and graduate if applicable transcripts and toefl ielts gre scores to dr mastorakis s.mastorakis@unomaha.edu https://t.co/ne0we7uy8b https://t.co/owhlkeskrc
Field	-
Univ	university of nebraska at omaha
Country	usa
Link	https://t.co/ne0we7uy8b https://t.co/owhlkeskrc
Deadline	-
Topic	-
Lab	-

G. Metode Evaluasi

Dalam pembuatan model , dibutuhkan metode pengukuran yang digunakan untuk menakar kualitas model yang dibangun tersebut. Penelitian ini menggunakan metode evaluasi yang umum digunakan dalam algoritma klasifikasi, yaitu *F1-score, precision* dan *recall*. Nilai-nilai tersebut didapatkan dengan menghitung Confusion matrix. Confusion matrix merupakan hasil klasifikasi yang merupakan prediksi model terhadap data yang dimiliki dibandingkan dengan data aslinya. Confusion matrix dapat digambarkan

TABEL 4.
CONFUSION MATRIX

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Precision merepresentasikan proporsi hasil yang didapatkan dari sistem yang merupakan akurat dan benar terhadap data yang diuji [11].

$$P = \frac{|TP|}{|TP|+|FP|} \quad (1)$$

oleh sistem dan sebaliknya [11].

Recall mengindikasikan jumlah data yang seharusnya positif, diklasifikasikan sebagai positif

$$R = \frac{|TP|}{|TP|+|FP|} \quad (2)$$

Sehingga didapatkan *F1-score*:

$$F_1 = 2 \times \frac{P \cdot R}{P+R} \quad (3)$$

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengujian

Setelah dibangun model BiLSTM-CRF, dilakukan pengujian performansi model dan ketepatan hasil klasifikasi yang dilakukan. Metode evaluasi dilakukan menggunakan *Precision, Recall, dan F1-score* untuk setiap klasifikasi. Hyperparameter yang digunakan untuk membangun model adalah *batch* sebesar 32, *epoch* sebanyak 150, dan 128 dimensi untuk *embedding layer.* Berikut merupakan hasil yang didapatkan.

TABEL 5.
HASIL PENGUJIAN MODEL BILSTM-CRF

Label	Precision	Recall	F1-Score
B-B-Country	1.00	0.00	0.00
B-B-Deadline	1.00	0.29	0.44
B-B-Field	0.96	0.71	0.81
B-B-Lab	1.00	0.00	0.00
B-B-Link	0.70	0.92	0.79
B-B-Topic	1.00	0.00	0.00
B-B-Univ	0.85	0.55	0.67
B-I-Deadline	1.00	0.58	0.74
B-I-Field	0.93	0.71	0.80
B-I-Topic	0.85	1.00	0.92
B-I-Univ	0.67	0.87	0.75

Setelah dibentuknya model BiLSTM-CRF, dilakukan perbandingan terhadap variasi arsitektur LSTM lainnya menggunakan dataset dan parameter yang sama. Tabel 6 menunjukkan perbandingan hasil kinerja model dengan arsitektur LSTM yang berbeda.

TABEL 6.
PERBANDINGAN HASIL PREDIKSI MODEL

No.	Model	Precision	Recall	F1-Score
1	BiLSTM-CRF	0.90	0.51	**0.54**
2	BiLSTM	0.82	0.53	0.49
3	Forward LSTM	0.83	0.48	0.43
4	Backward LSTM	0.81	0.09	0.07

Berdasarkan Tabel 6, rata-rata nilai yang didapatkan untuk model BiLSTM-CRF merupakan 90% untuk *precision*, 51% untuk *recall* dan 54% untuk *f1-score*

TABEL 7.
HASIL PENGUJIAN MODEL BILSTM-CRF UNTUK EKSTRAKSI INFORMASI

	Precision	Recall	F1-Score
Country	1.00	0.00	0.00
Deadline	0.93	0.68	0.79
Field	0.99	0.74	0.85
Lab	1.00	0.00	0.00
Link	0.73	0.94	0.82
Topic	0.98	1.00	0.99
University	0.82	0.84	0.83

Tabel 7 menunjukkan performansi model BiLSTM-CRF dalam melakukan ekstraksi informasi penting tanpa mempertimbangkan posisi token dengan Skema BIO.

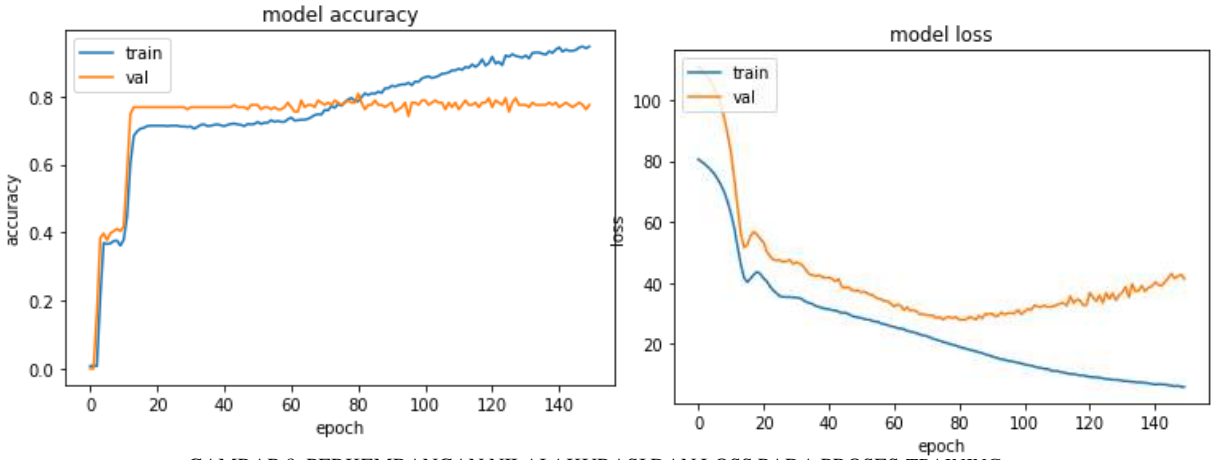
TABEL 8.
PERBANDINGAN HASIL PREDIKSI MODEL

No.	Model	Precision	Recall	F1-Score
1	BiLSTM-CRF	0.92	0.60	**0.61**
2	BiLSTM	0.83	0.58	0.49
3	Forward LSTM	0.72	0.62	0.45
4	Backward LSTM	0.71	0.14	0.11

Tabel 8 menunjukkan perbandingan performansi model dengan variasi LSTM yang berbeda tanpa mempertimbangkan posisi token dengan Skema BIO. Hasil yang didapatkan menunjukkan model BiLSTM-CRF lebih unggul dengan rata-rata nilai *precision* 0.92, *recall* 0.60, dan *f1-score* 0.61

B. Analisis Hasil Pengujian

Berdasarkan hasil yang didapat, BiLSTM-CRF lebih unggul dari model lainnya dalam kategori *precision*, *recall*, dan *f1-score*. Akan tetapi, model memiliki nilai *precision* yang lebih tinggi dibandingkan dengan nilai *recall* yang didapatkan. Rendahnya nilai recall yang dapat dilihat pada tabel 5 untuk label seperti B-B-Country, B-B-Topic, dan B-B-Lab kemungkinan dikarenakan oleh penyebaran label yang tidak merata pada dataset yang digunakan seperti yang ditunjukkan pada gambar 5. Hal ini menyebabkan model yang dibangun jarang atau bahkan tidak pernah bertemu data dengan label tersebut pada proses *training* sehingga tidak dapat memberikan prediksi yang baik.



GAMBAR 9. PERKEMBANGAN NILAI AKURASI DAN LOSS PADA PROSES *TRAINING*

Berdasarkan perbandingan nilai akurasi dan *loss* yang didapatkan ketika proses training pada gambar 9, nilai akurasi dan *loss* pada data validasi yang lebih rendah dibandingkan dengan *training* data menunjukkan kemungkinan model mengalami *overfitting*, dimana model tidak dapat melakukan generalisasi dengan baik sehingga mempengaruhi hasil prediksi yang dikeluarkan.

V. KESIMPULAN

Dalam penelitian ini, telah dilakukan implementasi ekstraksi informasi menggunakan BiLSTM-CRF. Model tersebut menunjukkan performansi dengan rata-rata nilai F1-Score 54% dan 61% apabila hanya mempertimbangkan performansi ekstraksi informasi penting. Implementasi model menggunakan hyperparameter dengan dimensi vektor sebesar 128 pada embedding layer, 32 *batch size* ketika *training* yang dilakukan sebanyak 150 *epoch* . Hasil yang didapat melebihi performansi model dengan variasi arsitektur LSTM yang berbeda dengan nilai rata-rata f1-score 49% untuk BiLSTM tanpa CRF, 43% untuk Forward LSTM, dan 7% untuk Backward LSTM. Namun, tingginya nilai *precision* (90%) dan rendahnya *recall* (51%) pada model BiLSTM-CRF dapat disebabkan oleh penyebaran label pada dataset yang tidak merata, yang kemungkinan menyebabkan model tidak bertemu dengan label yang memiliki jumlah data yang sedikit saat proses pelatihan sehingga model tidak dapat memberikan prediksi yang baik. Dalam proses training, model juga menunjukkan tanda-tanda *overfitting* dari perbandingan nilai akurasi dan *loss* antara data validasi dan data pelatihan yang mempengaruhi hasil prediksi akhir lebih lanjut.

Adapun saran untuk penelitian selanjutnya, yaitu diperlukannya jumlah dataset yang lebih banyak dengan penyebaran label yang lebih merata. Penyesuaian hyperparameter lebih lanjut dapat dilakukan untuk mendapatkan hasil klasifikasi yang lebih baik dan akurat. Dapat dilakukan modifikasi terhadap *layer* pada arsitektur yang digunakan untuk meningkatkan efisiensi kinerja model yang dibangun. Diperlukan observasi dan penyesuaian lebih lanjut dalam tahapan pembangunan model untuk menghindari terjadinya *overfitting* pada model.

REFERENSI

[1] Adnan, K., & Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data, 6(1), 1-38.

[2] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. Named Entity Recognition: Fallacies, Challenges and Opportunities.

[3] Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. International Journal of Computer Science and Network Security, 8(2), 339-344.

[4] He, Z., Wang, Z., Wei, W., Feng, S., Mao, X., & Jiang, S. (2020). A Survey on Recent Advances in Sequence Labeling from Deep Learning Models. arXiv preprint arXiv:2011.06727.

[5] Chopra, A., Prashar, A., & Sain, C. (2013). Natural language processing. International journal of technology enhancements and emerging engineering research, 1(4), 131-134.

[6] Gooding, S., & Kochmar, E. (2019, July). Complex word identification as a sequence labelling task. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1148-1153).

[7] Al-Zaidy, R. A., Caragea, C., & Giles, C. L. (2019). Bi-LSTM-CRF Sequence Labeling for Keyphrase Extraction from

Scholarly Documents.
[8] Kawakami, K. (2008). Supervised sequence labelling with recurrent neural networks. Ph. D. thesis.
[9] Sangal, R., Bendre, S., Sharma, D. M., & Mannem, P. R. (2007). Shallow Parsing for South Asian Languages.
[10] Liu, K., & El-Gohary, N. (2016). Ontology-based sequence labelling for automated information extraction for supporting bridge data analytics. Procedia Engineering, 145, 504-510.
[11] Derczynski, L. (2016, May). Complementarity, F-score, and NLP Evaluation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 261-266).
[12] D. Sahrawat et al., 'Keyphrase Extraction from Scholarly Articles as Sequence Labeling using Contextualized Embeddings', 2019
[13] K. Nebhi, 'Ontology-based information extraction from twitter', στο Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, 2012, σσ. 17–22.
[14] A. Katsuta et al., 'Infomation extraction from english & japanese résumé with neural sequence labelling methods', 2018.
[15] Z. Huang, W. Xu, και K. Yu, 'Bidirectional LSTM-CRF models for sequence tagging'.

.