ISSN : 1411-1624
e-ISSN: 2621-8089

Contextualized Word Embedding Untuk Ekstraksi Kutipan Berita Indonesia

Syifa Khairina[1], Nayara Saffa[2], Djoko Cahyo Utomo Lieharyani[3*], Jonner Hutahaean[4]
Politeknik Negeri Bandung[1,2,3,4]
Jalan Gegerkalong Hilir, Ciwaruga, Kec. Parongpong, Kabupaten Bandung Barat, Jawa Barat 40559, Indonesia
Sur-el : syifa.khairina.tif421@polban.ac.id[1], nayara.saffa.tif421@polban.ac.id[2], djoko.c.utomo@polban.ac.id[3], jonnerh@jtk.polban.ac.id[4]
[*)] Corresponden Author

Received: 04 July 2025 Reviewed: 07 July 2025 Accepted:07 August 2025

Abstract : *This study aims to develop a Named Entity Recognition (NER) model based on Recurrent Neural Networks (RNN) to extract direct quotes from Indonesian news articles, with a focus on enhancing the Medmon system by Kabayan Group, which is used to monitor the public image of public figures and brands. The study is limited to Indonesian news articles and does not include other languages or news sources. Two models are compared in this research: one utilizing static word embedding Word2Vec and the other using contextual word embedding BERT. The experiment was conducted using PFSA-ID corpus, which consist 1,018 Indonesian news articles annotated for direct quotes using BILOU scheme. Both models were trained and evaluated using Python programming libraries such as Pytorch and Hugging Face Transformers. The results show that the BERT model outperforms Word2Vec, with an F1-Score difference of 14.03 points. The BERT model achieved a highest F1-Score of 92.28%, while Word2Vec only reached 78.05%. This research contributes to the field of online media monitoring by improving the efficiency and accuracy of direct quote extraction in Indonesian news, offering practical value for media analysts and organizations relying on automated media analysis.*

Keywords: *BERT, Direct Quotes, Indonesian News, Named Entity Recognition, Word2Vec*

Abstrak : *Penelitian ini bertujuan untuk mengembangkan model Named Entity Recognition (NER) berbasis Recurrent Neural Networks (RNN) untuk mengekstrak kutipan langsung dari artikel berita Indonesia, dengan fokus untuk meningkatkan sistem Medmon dari Kabayan Group, yang digunakan untuk memantau citra tokoh publik dan merek. Penelitian ini terbatas pada artikel berita Indonesia dan tidak mencakup bahasa atau sumber berita lainnya. Dua model dibandingkan dalam penelitian ini, yaitu model dengan word embedding statis Word2Vec dan model dengan word embedding kontekstual BERT. Eksperimen dilakukan menggunakan korpus PFSA-ID, yang terdiri dari 1.018 artikel berita berbahasa Indonesia yang telah dianotasi untuk kutipan langsung menggunakan skema BILOU. Kedua model dilatih dan dievaluasi dengan menggunakan pustaka pemrograman Python seperti Pytorch dan Hugging Face Transformers. Hasil penelitian menunjukkan bahwa model BERT memberikan kinerja yang lebih unggul dibandingkan Word2Vec, dengan selisih F1-Score sebesar 14,03 poin. Model BERT mencapai F1-Score tertinggi 92,28%, sedangkan Word2Vec hanya mencapai 78,05%. Penelitian ini memberikan kontribusi pada bidang pemantauan media online dengan meningkatkan efisiensi dan akurasi ekstraksi kutipan langsung dalam berita Indonesia, yang menawarkan nilai praktis bagi analis media dan organisasi yang bergantung pada analisis media otomatis.*

Kata kunci*: Berita Indonesia, BERT; Kutipan Langsung; Named Entity Recognition; Word2Vec*

1. PENDAHULUAN

Media berita *online* memiliki peran yang sangat penting dalam membentuk opini publik dan citra tokoh-tokoh publik, seperti politisi, selebriti, dan individu berpengaruh lainnya. Penelitian [1] menunjukkan bahwa media berita *online*, bersama dengan media sosial, tidak

ISSN : 1411-1624
e-ISSN: 2621-8089

hanya berfungsi untuk menyampaikan informasi, tetapi juga membentuk persepsi masyarakat terhadap tokoh-tokoh tersebut, baik di bidang profesional maupun pribadi. Pemberitaan dapat memperkuat atau bahkan merusak citra seseorang, bergantung pada bagaimana berita tersebut disampaikan [2]. Kabayan Group, sebuah perusahaan yang fokus pada transformasi digital dan teknologi informasi, menyediakan solusi-solusi dalam berbagai sektor, termasuk *e-government*, manajemen bisnis digital, serta pemantauan media dan pengelolaan komunikasi [3]. Salah satu produk andalan mereka, Medmon, merupakan sistem *Online Media Monitoring* yang dirancang untuk mengumpulkan informasi terkait topik atau merek dari berbagai sumber berita online dan media sosial [4], dengan tujuan untuk mengidentifikasi serta mengurangi potensi opini negatif yang mungkin muncul terkait dengan topik atau merek tersebut.

Pada pemantauan media, kutipan langsung dari tokoh publik menjadi komponen penting karena sering kali mengandung opini, pernyataan resmi, atau sikap terhadap isu tertentu. Medmon mendukung proses ini melalui fitur ekstraksi kutipan berita, yang umumnya dilakukan dalam dua tahap, yaitu identifikasi kutipan (*quotation extraction*) dan atribusi kutipan (*quotation attribution*) [5]. Tahap pertama bertujuan menemukan bagian teks yang merupakan kutipan langsung, sedangkan tahap kedua menghubungkan kutipan tersebut dengan narasumber yang relevan. Dengan mengekstraksi kutipan secara akurat, sistem dapat mendukung analisis sentimen, pelacakan opini publik, serta identifikasi isu strategis yang berkembang di ruang media.

Namun, dalam implementasinya, model yang digunakan Medmon masih menghadapi berbagai tantangan, khususnya dalam proses identifikasi entitas seperti nama individu dan entitas penting lainnya. Salah satu masalah utama adalah ketidakakuratan dalam mengidentifikasi entitas dengan label PERSON, yang menyebabkan data selain nama individu terekstraksi ke dalam *database*. Selain itu, model juga kesulitan dalam mendeteksi entitas penting lainnya seperti organisasi, lokasi, jabatan, dan peristiwa yang sangat dibutuhkan dalam pemantauan media. Berdasarkan wawancara dengan Simanjuntak R., *Project Manager* Medmon, kesulitan ini menghambat proses ekstraksi kutipan secara otomatis dan mengharuskan analis media melakukan pengecekan manual untuk memastikan ketepatan hasilnya. Hal ini menunjukkan perlunya peningkatan pada model yang digunakan di Medmon, khususnya dalam mengenali lebih banyak entitas penting dan meningkatkan akurasi ekstraksi kutipan langsung, yang sering kali terhalang oleh struktur kalimat yang kompleks.

Salah satu masalah utama dalam ekstraksi kutipan dari teks berita adalah keragaman struktur kalimat yang digunakan dalam berita. Hal ini menjadi tantangan besar bagi model ekstraksi kutipan, terutama ketika kutipan tersebut tidak eksplisit atau terkandung dalam kalimat yang kompleks. Penelitian [5] mencatat bahwa banyak algoritma ekstraksi kutipan masih mengandalkan pendekatan berbasis aturan atau pembelajaran mesin tradisional. Meskipun

ISSN : 1411-1624
e-ISSN: 2621-8089

metode berbasis aturan dapat memberikan hasil yang akurat, metode ini memerlukan pengembangan aturan yang rumit dan pemahaman mendalam terhadap variasi struktur kalimat. Pendekatan berbasis *deep learning*, seperti *Recurrent Neural Networks* (RNN), menawarkan potensi yang lebih besar untuk mengatasi masalah ini [6], meskipun masih menghadapi tantangan dalam mendeteksi kutipan [7].

Kutipan langsung memuat pernyataan narasumber secara persis, kata demi kata, seperti yang diucapkan. Jenis kutipan ini ditandai dengan penggunaan tanda kutip untuk mengapit pernyataan tersebut [8]. Struktur kutipan dalam Bahasa Indonesia merujuk pada pedoman Ejaan yang Disempurnakan (EYD). Berdasarkan pedoman ini, tiga aturan utama terkait kutipan langsung telah didefinisikan secara rinci beserta contohnya pada Tabel 1 [9].

Tabel 1. Aturan Utama Struktur Kutipan Langsung

Struktur Kutipan	Contoh
(tanda kutip) (pernyataan) (koma | titik | tanda seru | tanda tanya) (tanda kutip) (kata kerja pengutip) (narasumber)	"MPR RI menghormati proses hukum yang berjalan, dan menyerahkan sepenuhnya kepada KPK untuk menindaklanjuti sesuai kewenangan dan ketentuan hukum yang berlaku," kata Siti
(narasumber) (kata kerja pengutip) (koma) (tanda kutip) (pernyataan) (titik | tanda seru | tanda tanya) (tanda kutip)	Presiden Prabowo mengatakan, “Bangsa dan negara kita membutuhkan kepolisian yang tangguh, unggul, bersih, dan dicintai rakyat—polisi yang berada di tengah rakyat, membela rakyat, melindungi rakyat, khususnya mereka yang paling lemah, paling tertindas, dan paling miskin.”
(tanda kutip) (pernyataan) (koma | titik | tanda seru | tanda tanya) (tanpa kutip) (kata kerja pengutip) (−nya)	"Jakarta bersyukur karena UMKM dilibatkan, 70% produk lokal. Banyak yang datang dari luar kota, menginap, sewa apartemen, dan ini mendongkrak pendapatan daerah," ucapnya.

Penelitian [10] menggunakan arsitektur *Recurrent Neural Networks* (RNN) berbasis *static word embedding* dalam sistem ekstraksi kutipan untuk teks berita bahasa Indonesia. Meskipun hasilnya menunjukkan performa terbaik dalam hal akurasi, model tersebut masih memiliki ruang untuk pengembangan lebih lanjut. Salah satu saran yang diberikan adalah untuk menggunakan *Large Language Models* (LLM) seperti BERT, ELMo dan XLM-RoBerta dalam model tersebut. Penelitian ini bertujuan untuk menguji penggunaan *contextualized embeddings* dari LLM dalam arsitektur RNN untuk tugas *Named Entity Recognition* (NER), guna melihat sejauh mana perubahan ini dapat meningkatkan kinerja model dalam mengenali entitas, khususnya dalam mendeteksi kutipan langsung dan entitas penting lainnya seperti

ISSN : 1411-1624
e-ISSN: 2621-8089

organisasi dan jabatan. Dengan pemahaman konteks yang lebih mendalam yang diberikan oleh model berbasis konteks, diharapkan model NER ini akan lebih mampu mengenali struktur kalimat yang lebih kompleks dalam teks berita Indonesia.

Penelitian ini bertujuan untuk mengeksplorasi apakah peralihan dari *static word embedding* ke *contextualized word embedding* dapat meningkatkan akurasi model NER dalam ekstraksi kutipan langsung, yang akan dievaluasi menggunakan matriks F1-*Score*. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem Medmon, khususnya dalam meningkatkan akurasi ekstraksi kutipan langsung, yang pada akhirnya dapat meningkatkan efisiensi proses pemantauan media *online*.

2. METODOLOGI PENELITIAN

2.1 Persiapan Data

Sumber data utama yang digunakan dalam penelitian ini adalah korpus PFSA-ID (*Public Figure Statement Attribution for the Indonesian Language*) yang dikembangkan oleh Purnomo W.P. et al. [11]. Korpus ini dirancang untuk tugas atribusi pernyataan *public figure* dalam berita berbahasa Indonesia, dengan fokus utama pada anotasi kutipan langsung. Setiap kutipan dalam korpus telah dianotasi secara rinci, mencakup elemen-elemen penting seperti narasumber, isi pernyataan, dan penandaan kutipan yang muncul dalam konteks berita.

Setiap token dalam korpus dianotasi menggunakan skema BILOU (*Begin, Inside, Last, Outside, Unit*), yang merupakan skema umum dalam tugas *sequence labeling* seperti *Named Entity Recognition* (NER) [12]. Anotasi dilakukan terhadap sebelas jenis label entitas, yaitu PERSON, PERSONCOREF, ROLE, AFFILIATION, CUE, CUECOREF, STATEMENT, ISSUE, DATETIME, LOCATION, dan EVENT. Tabel 2 menyajikan penjelasan masing-masing label entitas yang digunakan dalam korpus PFSA-ID.

Tabel 2. Label Entitas Korpus PFSA-ID

Label Entitas	Deskripsi
PERSON	Nama orang yang disebutkan.
PERSONCOREF	Rujukan kepada orang yang telah disebutkan sebelumnya.
ROLE	Peran atau jabatan seseorang.
AFFILIATION	Keterkaitan dengan organisasi atau institusi tertentu.
CUE	Penanda kutipan langsung (misalnya, "kata").
CUECOREF	Menandai token yang menggabungkan kata kerja atribusi dengan referensi narasumber (misalnya, “katanya”)
STATEMENT	Isi dari kutipan langsung.

ISSN : 1411-1624
e-ISSN: 2621-8089

Label Entitas	Deskripsi
ISSUE	Masalah atau topik utama yang dibahas.
DATETIME	Informasi waktu dan tanggal terkait berita.
LOCATION	Lokasi kejadian yang dilaporkan.
EVENT	Peristiwa yang dijelaskan dalam berita.

Penelitian [10] menggunakan Korpus PFSA-ID dibagi menjadi dua set data. Set pertama terdiri dari 1.018 artikel yang digunakan untuk proses pelatihan dan validasi model, sedangkan set kedua merupakan data uji yang terdiri dari 144 artikel. Seluruh data disimpan dalam format berkas teks (.txt), di mana setiap baris mewakili satu token beserta label entitasnya.

2.2 Pra-pemrosesan Data

Sebelum data digunakan dalam proses pelatihan model, diperlukan tahapan pra-pemrosesan untuk memastikan format input sesuai dengan kebutuhan arsitektur model. Langkah-langkah utama dalam tahap ini mencakup tokenisasi, *encoding* karakter, serta proses *padding* dan *masking*.

Proses tokenisasi disesuaikan dengan jenis *word embedding* yang digunakan. Pada model berbasis Word2Vec, tokenisasi dilakukan langsung terhadap token asli dalam dataset tanpa pemecahan menjadi sub-kata. Sebaliknya, model berbasis BERT menggunakan tokenisasi sub-kata dengan menggunakan *WordPiece*. Metode ini memungkinkan sebuah kata dipecah menjadi beberapa unit sub-kata, yang membantu dalam menangani kata-kata yang jarang muncul atau belum pernah dilihat sebelumnya. sebagai bagian dari format input BERT, token khusus [CLS] ditambahkan di awal kalimat untuk merepresentasikan keseluruhan kalimat, dan token [SEP] di akhir kalimat sebagai penanda batas [13].

Setelah tokenisasi, setiap token yang telah ditokenisasi juga dipecah menjadi karakter individual untuk diproses melalui lapisan *characater-level embedding* berbasis *Convolutional Neural Network* (CNN). Representasi karakter ini digunakan untuk menangkap informasi morfologis yang sering kali terlewat pada level token.

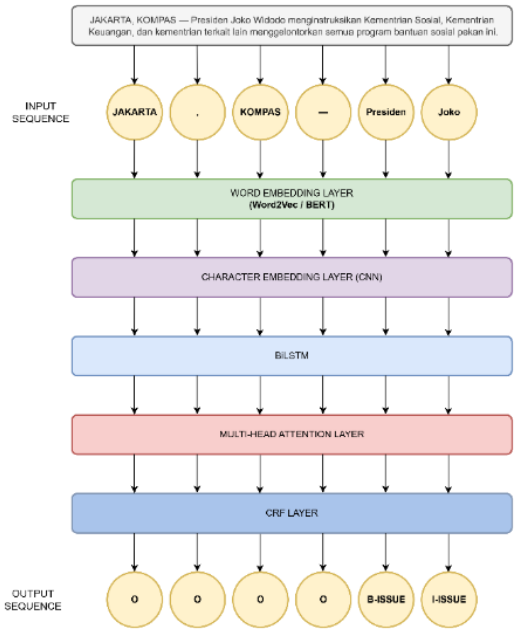
Langkah selanjutnya adalah proses *padding* dan *masking* untuk menyamakan panjang input. Pada model Word2Vec, digunakan teknik BucketIterator, di mana setiap *batch* dipadatkan berdasarkan panjang kalimat dan kata terpanjang dalam *batch* tersebut. Sementara itu, pada model berbasis BERT, panjang input ditetapkan secara tetap, yaitu 256 token per kalimat dan maksimal 32 karakter per token. Jika panjang token atau karakter lebih pendek dari batas tersebut, maka *padding* ditambahkan. Untuk membedakan token asli dan *padding*, diterapkan *masking* yang menandai bagian mana yang valid untuk diproses oleh model.

2.3 Arsitektur Model

Arsitektur model dalam penelitian ini terdiri dari lima komponen utama yang bekerja secara berurutan, seperti ditunjukkan pada Gambar 1.

ISSN : 1411-1624
e-ISSN: 2621-8089

Langkah awal dalam arsitektur model adalah membentuk *word embedding*, yaitu merepresentasikan setiap kata dalam bentuk vektor numerik agar dapat diproses oleh model. Penelitian ini menggunakan dua pendekatan *embedding*, yakni Word2Vec dan BERT. Word2Vec menghasilkan representasi kata berbasis vektor dari *vocabulary* yang telah dilatih sebelumnya [14]. Dengan memanfaatkan model Word2Vec untuk bahasa Indonesia, setiap kata dalam *vocabulary* dapat langsung dipetakan ke vektor yang sesuai. Jika sebuah kata tidak ditemukan, vektor nol diberikan agar seluruh kata tetap memiliki representasi yang seragam.



Gambar 1. Arsitektur Model

Berbeda dengan Word2Vec yang menghasilkan representasi kata secara statis, pendekatan BERT menggunakan *embedding* kontekstual. Penelitian ini menggunakan varian IndoBERT dari model indobenchmark/indobert-base-p1, yang dikembangkan khusus untuk bahasa Indonesia [15]. Setiap token diubah menjadi indeks numerik, lalu dipetakan ke tabel *embedding* dan diperkaya dengan *positional encoding* untuk mempertimbangkan urutan kata dalam kalimat. Dengan demikian, makna vektor tiap token dapat menyesuaikan konteks kata-kata di sekitarnya.

Selain representasi kata melalui *word embedding*, model juga menggunakan *character embedding* untuk menangkap informasi morfologis dari bentuk kata. Pada tahap ini, setiap karakter dikonversi menjadi indeks numerik dan dipetakan ke dalam vektor berdimensi tetap melalui *embedding layer*. Vektor karakter tersebut kemudian diproses oleh lapisan CNN untuk mengekstraksi pola-pola antar karakter. Hasil konvolusi selanjutnya diproses oleh *max pooling* untuk menyaring fitur paling relevan. Representasi karakter ini kemudian digabungkan dengan *word embedding*.

Gabungan representasi kata dan karakter tersebut kemudian diproses oleh *Bidirectional Long Short-Term Memory* (BiLSTM) untuk memahami konteks dari dua arah dalam kalimat. BiLSTM memungkinkan model membaca kata secara maju dan mundur, sehingga makna kata dipahami secara menyeluruh [16]. Mekanisme LSTM, seperti *forget gate*, *input gate*, dan *output gate*, membantu mengelola informasi penting dalam urutan data. Representasi ini lalu diperkuat dengan *multihead attention*, yang menghitung relevansi antar-token agar model lebih fokus pada kata-kata yang penting dalam kalimat.

Sebagai tahap akhir, model menggunakan *Conditional Random Field* (CRF) untuk memprediksi label entitas secara sekuensial. Sebelum masuk ke CRF, representasi dari

ISSN : 1411-1624
e-ISSN: 2621-8089

multihead attention terlebih dahulu diproses oleh lapisan *fully-connected* untuk menghasilkan skor awal setiap label. CRF kemudian menghitung kemungkinan transisi antar label, sekaligus menentukan urutan label terbaik. Selain menghasilkan prediksi, CRF juga menghitung nilai *loss* yang digunakan dalam proses pelatihan untuk mengukur akurasi model terhadap label sebenarnya [17].

2.4 Pelatihan Model

Pelatihan model dilakukan secara berulang dengan menguji berbagai kombinasi *hyperparameter* untuk memperoleh konfigurasi terbaik. Kombinasi yang digunakan mencakup variasi yang umum digunakan pada penelitian NLP dengan rasio pembagian data pelatihan dan validasi (60:40, 70:30, dan 80:20) [10], *batch size* (16, 32, 128) [13] [18], serta jumlah *epoch* (10, 20, 30, 40, 50). Sementara itu, nilai *learning rate* ditentukan secara otomatis dengan memanfaatkan LR Finder [19]. Untuk mencegah *overfitting*, diterapkan mekanisme *early stopping*. Seluruh proses pelatihan dilakukan di Kaggle Notebook yang dilengkapi dua GPU NVIDIA T4 dan RAM sebesar 29 GB, sehingga proses pelatihan dapat berjalan lebih efisien dan cepat.

2.5 Evaluasi Model

Evaluasi model dilakukan menggunakan data validasi untuk mengukur performa model dalam mengenali dan mengklasifikasikan entitas dalam teks. Metrik evaluasi yang digunakan adalah F1-*Score*, yang mengukur keseimbangan antara *precision* dan *recall*. Metrik ini dipilih karena telah menjadi standar umum dalam berbagai penelitian sebelumnya terkait tugas NER [20][21]. Melalui F1-*Score*, efektivitas model dapat dinilai secara kuantitatif dan konsisten sepanjang eksperimen yang dilakukan.

3. HASIL DAN PEMBAHASAN

Hasil evaluasi yang diperoleh pada eksperimen ini mengukur kinerja model dalam mengekstraksi kutipan langsung dengan menggunakan F1-*Score*, yang merupakan metrik yang menggabungkan *precision* dan *recall* untuk memberikan gambaran yang lebih lengkap tentang performa model. Gambar 2 menunjukkan grafik perbandingan rata-rata F1-*Score* untuk model dengan *word embedding* menggunakan Word2Vec dan BERT pada variasi pembagian dataset dan *batch size*.

Grafik tersebut menunjukkan rata-rata F1-*Score* yang dihitung untuk masing-masing model (Word2Vec dan BERT) dengan pembagian dataset dan *batch size* yang berbeda. Model berbasis BERT selalu menunjukkan performa yang lebih baik dibandingkan dengan model Word2Vec, dengan nilai tertinggi mencapai 92% pada *batch size* 128 dan *split* dataset 80:20.

Lebih lanjut, analisis terhadap rasio pembagian data menunjukkan bahwa model Word2Vec lebih sensitif terhadap proporsi data latih dan validasi. Performa terbaik dicapai pada skema 70:30 dengan rata-rata F1-*Score* sebesar 74,53%, sementara pada skema 60:40 dan 80:20 performanya menurun. Sebaliknya, model BERT menunjukkan kecenderungan yang berbeda, di mana peningkatan proporsi data pelatihan

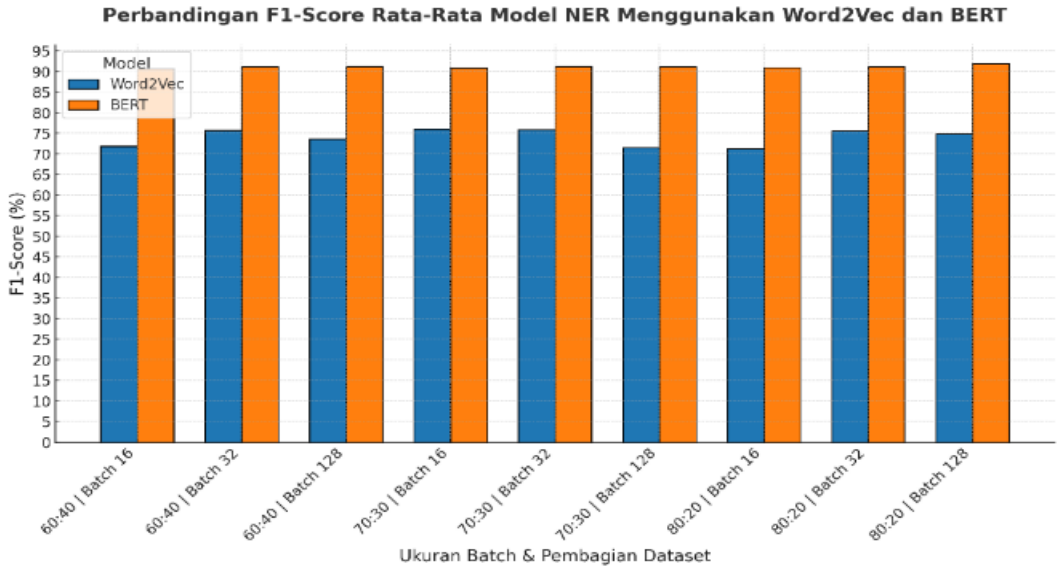
ISSN : 1411-1624
e-ISSN: 2621-8089

memberikan kontribusi positif terhadap kinerjanya. Konfigurasi 80:20 menghasilkan rata-rata F1-*Score* tertinggi sebesar 91,41%.

Selain pembagian data, variasi *batch size* juga memberikan dampak terhadap performa model. Pada model Word2Vec, peningkatan *batch size* dari 16 ke 32 menghasilkan kenaikan rata-rata F1-*Score* dari 73,76% menjadi 75,78%. Namun, ketika *batch size* ditingkatkan menjadi 128, performa justru menurun. Sebaliknya, model BERT menunjukkan peningkatan performa yang lebih stabil seiring penambahan *batch size*, di mana nilai rata-rata *F1-Score* tertinggi tercapai pada *batch size* 128, yaitu sebesar 91,51%.

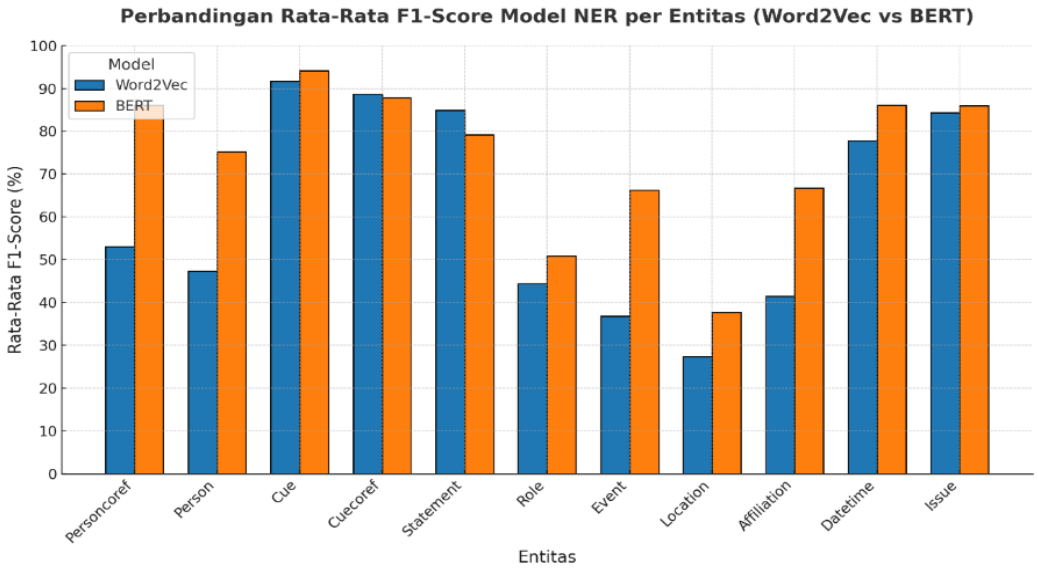
Dari segi jumlah *epoch*, meskipun eksperimen pelatihan model diatur hingga maksimum 50 *epoch*, seluruh proses dihentikan lebih awal oleh mekanisme *early stopping*. Model Word2Vec menunjukkan rentang *stopping epoch* yang cukup luas, antara 6 hingga 23 *epoch*, dengan frekuensi tertinggi pada *epoch* ke-10. Sebaliknya, model BERT cenderung lebih cepat mencapai konvergensi, dengan *stopping epoch* yang lebih sempit, yaitu antara 6 hingga 11 *epoch*, dan paling sering berhenti pada *epoch* ke-7.

Gambar 3 menunjukkan perbandingan performa model Word2Vec dan BERT pada masing-masing entitas yang diuji. BERT unggul pada sebagian besar entitas, terutama pada PERSONCOREF, PERSON, dan EVENT. Sebagai contoh, F1-*Score* untuk PERSONCOREF pada BERT mencapai 86.09%, sementara Word2Vec hanya mencatat 53.08%. BERT lebih efektif dalam mengenali referensi antar individu dalam teks yang menggunakan kata ganti atau istilah yang mengarah pada tokoh yang sudah disebutkan sebelumnya. Pemahaman konteks ini penting karena referensi antar entitas dalam teks bisa tidak eksplisit, sehingga model yang dapat menangkap hubungan ini dengan lebih baik akan memberikan hasil yang lebih akurat.



Gambar 2. Grafik Perbandingan Rata-Rata *F1-Score* Model NER Menggunakan Word2Vec dan BERT

ISSN : 1411-1624
e-ISSN: 2621-8089



Gambar 3. Grafik Perbandingan Rata-Rata *F1-score* Model NER per Entitas

Pada PERSON, BERT kembali menunjukkan keunggulannya dengan F1-*Score* 75.19%, sedangkan Word2Vec hanya mencatatkan 47.31%. Ini mengindikasikan bahwa BERT lebih mampu mengenali individu dalam kalimat yang lebih kompleks, terutama ketika individu tersebut disebutkan dalam konteks yang lebih luas atau dengan informasi tambahan. Word2Vec, dengan representasi kata yang statis, tidak mampu menyesuaikan konteks seperti BERT yang menggunakan *contextual embeddings*, sehingga kemampuan BERT dalam memahami hubungan antar kata dalam kalimat lebih mendalam.

Entitas CUE dan CUECOREF, yang berkaitan dengan petunjuk atau referensi dalam kutipan, juga menunjukkan hasil yang lebih baik pada BERT, dengan F1-*Score* 94.12% dan 87.77% dibandingkan dengan Word2Vec yang mencatat 91.69% dan 88.71%. CUE berfungsi sebagai penanda awal kutipan langsung, dan BERT lebih efektif dalam mengenali tanda kutipan yang digunakan dalam teks yang kompleks. Sementara itu, CUECOREF mengindikasikan referensi ulang terhadap kutipan yang telah disebutkan sebelumnya. BERT lebih unggul dalam mengidentifikasi petunjuk ini, terutama dalam kalimat yang memiliki banyak elemen yang saling bergantung, seperti dalam kutipan panjang atau yang melibatkan penggantian kata.

Namun, pada entitas STATEMENT, Word2Vec sedikit lebih unggul dengan F1-*Score* 84.86%, sementara BERT mencatatkan 79.13%. STATEMENT merujuk pada isi kutipan langsung, yang umumnya lebih eksplisit dan tidak terlalu bergantung pada konteks yang rumit. Word2Vec, meskipun tidak sefleksibel BERT, efektif dalam mengenali pernyataan langsung yang tidak membutuhkan pemahaman kalimat yang dalam. BERT, yang lebih memfokuskan pada konteks kalimat secara keseluruhan, mungkin sedikit kehilangan akurasi pada entitas yang lebih langsung dan sederhana seperti STATEMENT.

Pada entitas lainnya, seperti ROLE, EVENT, dan AFFILIATION, BERT menunjukkan hasil yang lebih konsisten dan

ISSN : 1411-1624
e-ISSN: 2621-8089

lebih tinggi. Sebagai contoh, EVENT pada BERT mencapai 66.25%, jauh lebih baik dibandingkan dengan Word2Vec yang hanya mencatat 36.84%. BERT lebih mampu mengenali entitas EVENT yang berhubungan dengan peristiwa atau kejadian yang terjadi dalam teks. Karena EVENT sering kali melibatkan hubungan antara berbagai elemen dalam cerita (misalnya waktu, lokasi, dan orang-orang yang terlibat), BERT lebih efektif dalam menangkap hubungan tersebut dibandingkan Word2Vec, yang tidak dapat menyesuaikan representasi kata dengan konteks yang lebih luas. AFFILIATION juga menunjukkan hasil yang lebih baik pada BERT (66.75%) dibandingkan dengan Word2Vec (41.49%), mengindikasikan kemampuan BERT dalam mengenali hubungan antara individu dan organisasi atau institusi yang relevan dengan lebih efektif.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penggunaan *contextualized word embedding* BERT secara signifikan meningkatkan akurasi model *Named Entity Recognition* (NER) berbasis *Recurrent Neural Networks* (RNN) untuk ekstraksi kutipan langsung dari berita berbahasa Indonesia. Model berbasis BERT mampu mencapai F1-*Score* tertinggi sebesar 92,28%, jauh lebih tinggi dibandingkan model dengan Word2Vec yang hanya memperoleh skor tertinggi 78,05%.

Penelitian ini telah berhasil melakukan pelatihan dan evaluasi model untuk tugas ekstraksi kutipan langsung pada teks berita berbahasa Indonesia. Sebagai pengembangan lebih lanjut, disarankan agar penelitian mencakup pelatihan model pada dataset untuk kutipan tidak langsung. Selain itu, karena model berbasis BERT menunjukkan performa yang lebih unggul dibandingkan Word2Vec, maka penggunaan *word embedding* kontekstual lainnya seperti RoBERTa dapat dijadikan alternatif dibandingkan. Penggunaan model *pre-trained* multibahasa seperti mBERT dan XLM-RoBERTa juga dapat dipertimbangkan, karena kemampuannya dalam memahami konteks lintas bahasa yang lebih luas dan beragam. Diharapkan, eksplorasi terhadap model-model tersebut dapat memberikan kontribusi lebih lanjut dalam meningkatkan akurasi ekstraksi kutipan, baik kutipan langsung maupun tidak langsung.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Kabayan Group atas dukungan yang diberikan dalam penelitian ini, khususnya terkait Medmon, dan kepada semua pihak yang telah berkontribusi dalam penelitian ini, baik dalam hal sumber data maupun masukan teknis yang berharga. Penulis juga mengucapkan terima kasih kepada Purnomo W.P. yang telah menyediakan dataset PFSA-ID yang digunakan dalam pengembangan model ini. Ucapan terima kasih juga disampaikan kepada Politeknik Negeri Bandung atas dukungan materiil yang diberikan melalui pendanaan tugas akhir, sehingga penelitian ini dapat terlaksana dengan baik.

ISSN : 1411-1624
e-ISSN: 2621-8089

DAFTAR PUSTAKA

[1] S. Hong, "Shaping Public Opinion in the Digital Age: The Role of Online News and Social Media in Forming Political Leaders' Image," *Public Relat Rev*, vol. 46, no. 2, pp. 101–111, 2020.

[2] C. J. Vargo, L. Guo, and M. A. Amazeen, "Media Coverage and Public Perception: How Online News Influences Public Image of Political Figures," *Digital Journalism*, vol. 7, no. 3, pp. 348–365, 2019.

[3] Kabayan Group, "Kabayan Group." [Online]. Available: https://kabayan.id

[4] Kabayan Group, "Company Profile CV. Kabayan Consulting," 2022.

[5] Y. S. Purnomo W.P., Y. J. Kumar, and N. Z. Zulkarnain, "Understanding quotation extraction and attribution: towards automatic extraction of public figure's statements for journalism in Indonesia," *Global Knowledge, Memory and Communication*, vol. 70, no. 6–7, pp. 655–671, 2020, doi: 10.1108/GKMC-07-2020-0098.

[6] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *Neural Comput Appl*, vol. 36, no. 16, pp. 8995–9022, 2022, doi: 10.1007/s00521-024-09646-6.

[7] J. Papay and S. Pado, "Quotation Extraction Using Deep Learning," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, 2019, pp. 1431–1440.

[8] L. Ishwara, *Jurnalisme dasar*. Penerbit Buku Kompas, 2011.

[9] Y. Syaifudin and A. Nurwidyantoro, "Quotations identification from Indonesian online news using rule-based method," *Proceeding - 2016 International Seminar on Intelligent Technology and Its Application, ISITIA 2016: Recent Trends in Intelligent Computational Technologies for Sustainable Energy*, pp. 187–194, 2016, doi: 10.1109/ISITIA.2016.7828656.

[10] Y. S. Purnomo W.P., Y. J. Kumar, N. Z. Zulkarnain, and B. Raza, "Extraction and attribution of public figures statements for journalism in Indonesia using deep learning," *Knowl Based Syst*, vol. 289, no. February, 2024, doi: 10.1016/j.knosys.2024.111558.

[11] Y. S. Purnomo W.P., Y. J. Kumar, and N. Z. Zulkarnain, "PFSA-ID: an annotated Indonesian corpus and baseline model of public figures statements attributions," *Global Knowledge, Memory and Communication*, vol. 73, no. 6–7, pp. 853–870, 2024, doi: 10.1108/GKMC-04-2022-0091.

[12] X. Zhong and E. Cambria, *Time Expression and Named Entity Analysis and Recognition*. 2021.

[13] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Naacl-Hlt 2019*, no. Mlm, pp. 4171–4186, 2019, [Online]. Available: https://aclanthology.org/N19-1423.pdf

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pp. 1–12, 2013.

[15] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 8875–8898, 2021, doi: 10.18653/v1/2021.emnlp-main.699.

[16] D. C. Wintaka, M. A. Bijaksana, and I. Asror, "Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF," *Procedia Comput Sci*, vol. 157, pp. 221–228, 2019, doi: 10.1016/j.procs.2019.08.161.

[17] C. Che, C. Zhou, H. Zhao, B. Jin, and Z. Gao, "Fast and effective biomedical named entity recognition using temporal convolutional network with conditional random field," *Mathematical Biosciences and Engineering*, vol. 17, no. 4, pp. 3553–3566, 2020, doi: 10.3934/MBE.2020200.

[18] S. Khan *et al.*, "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection," *Journal of King Saud University - Computer and*

ISSN : 1411-1624
e-ISSN: 2621-8089

Information Sciences, vol. 34, no. 7, pp. 4335–4344, 2022, doi: 10.1016/j.jksuci.2022.05.006.

[19] R. An, J. M. Perez-Cruet, X. Wang, and Y. Yang, "Build Deep Neural Network Models to Detect Common Edible Nuts from Photos and Estimate Nutrient Portfolio," *Nutrients*, vol. 16, no. 9, pp. 1–9, 2024, doi: 10.3390/nu16091294.

[20] G. Popovski, B. K. Seljak, and T. Eftimov, "A Survey of Named-Entity Recognition Methods for Food Information Extraction," *IEEE Access*, vol. 8, pp. 31586–31594, 2020, doi: 10.1109/ACCESS.2020.2973502.

[21] Warto, Muljono, Purwanto, and E. Noersasongko, "Improving Named Entity Recognition in Bahasa Indonesia with Transformer-Word2Vec-CNN-Attention Model," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 4, pp. 655–668, 2023, doi: 10.22266/ijies2023.0831.53.