Koreksi Teks untuk Kesalahan Penulisan pada Artikel Berita menggunakan Model Bahasa

1st M. Dony Samdhila Yasin
School of Computing
Telkom University
Bandung, Indonesia
samdhila@student.telkomuniversity.ac.id

2nd Dr. Ade Romadhony S.T, M.T.
School of Computing
Telkom University
Bandung, Indonesia
aderomadhony@telkomuniversity.ac.id

Abstract*— Dalam era digital ini, kualitas penulisan artikel berita sangat penting untuk menjaga integritas dan kredibilitas informasi. Namun, sering terjadi kesalahan penulisan seperti ejaan dan tata bahasa yang dapat mengurangi kualitas artikel berita dan mempengaruhi pemahaman pembaca. Kesalahan ini mengindikasikan adanya kebutuhan akan metode yang efektif untuk mendeteksi dan memperbaiki kesalahan tersebut. Penelitian ini bertujuan untuk meningkatkan kualitas artikel berita dengan mengimplementasikan teknologi koreksi teks menggunakan model bahasa IndoBERT. Tiga metode yang diterapkan adalah *pretrained error detection, finetuned error detection*, dan *dictionary-based error detection*. Dataset yang digunakan mencakup 200 artikel berita berbahasa Indonesia dengan total 55.356 kata, meliputi berbagai kategori berita. Hasil penelitian menunjukkan bahwa metode *dictionary-based error detection* menghasilkan *Overall Accuracy* sebesar 89%, yang sangat efektif dalam mendeteksi dan memperbaiki kesalahan ejaan sederhana. Di sisi lain, model IndoBERT yang di-*finetune* dengan dataset berita menunjukkan peningkatan akurasi hingga 46% dibandingkan model *pretrained*, dan lebih unggul dalam menangani kesalahan kontekstual yang lebih kompleks. Penelitian ini memberikan dampak signifikan dalam meningkatkan kualitas penulisan artikel berita dengan menggunakan teknologi koreksi teks berbasis model bahasa. Kesalahan penulisan dapat dikurangi, sehingga artikel menjadi lebih akurat dan mudah dipahami oleh pembaca. Hasil penelitian ini juga diharapkan dapat berkontribusi pada pengembangan teknologi koreksi teks otomatis di Indonesia.

Keywords*— *koreksi teks, kesalahan penulisan, artikel berita, model bahasa

I. Introduction

Dalam era informasi digital saat ini, artikel berita menjadi salah satu sumber utama bagi masyarakat untuk memperoleh informasi terkini. Kualitas penulisan artikel berita sangat mempengaruhi pemahaman dan kepercayaan pembaca terhadap informasi yang disampaikan. Namun, sering kali ditemukan kesalahan penulisan seperti kesalahan ejaan, tata bahasa, dan penempatan tanda baca yang dapat mengurangi kredibilitas dan keakuratan berita [1], [2].

Kesalahan penulisan pada artikel berita dapat disebabkan oleh berbagai faktor, termasuk kesalahan manusia saat mengetik, kurangnya waktu untuk proses penyuntingan, serta tekanan untuk menyajikan berita secepat mungkin [3], [4]. Kesalahan-kesalahan ini tidak hanya mempengaruhi kualitas berita tetapi juga dapat menyebabkan salah tafsir informasi oleh pembaca [5].

Pentingnya kualitas penulisan yang baik telah mendorong banyak penelitian dalam bidang koreksi teks otomatis. Salah satu solusi yang menjanjikan adalah penggunaan model bahasa berbasis pembelajaran mesin yang mampu mendeteksi dan mengoreksi kesalahan penulisan secara otomatis [6]. Model bahasa ini dilatih menggunakan dataset yang berisi berbagai contoh kesalahan penulisan, sehingga mampu mempelajari pola-pola kesalahan yang sering terjadi dan memberikan koreksi yang tepat [7].

Penggunaan model bahasa untuk koreksi teks pada artikel berita tidak hanya bertujuan untuk memperbaiki kesalahan penulisan, tetapi juga untuk meningkatkan efisiensi proses penulisan dan penyuntingan [8]. Dengan adanya teknologi ini, jurnalis dan editor dapat lebih fokus pada konten berita itu sendiri tanpa harus khawatir tentang kesalahan penulisan yang dapat mengurangi kualitas artikel [9].

Penelitian ini bertujuan untuk mengimplementasi sistem koreksi teks menggunakan model bahasa yang mampu mendeteksi dan mengoreksi kesalahan penulisan pada artikel berita dengan tingkat akurasi yang cukup tinggi [10]. Dengan demikian, diharapkan dapat memberikan kontribusi dalam meningkatkan kualitas penulisan artikel berita dan memastikan informasi yang disampaikan kepada masyarakat lebih akurat dan mudah dipahami [11].

II. Related Works

Penelitian mengenai koreksi teks otomatis telah menjadi fokus berbagai studi dengan tujuan untuk meningkatkan akurasi dan efisiensi dalam penulisan. Salah satu penelitian yang relevan adalah oleh Naziri dan Zeinali (2024), yang menggabungkan model BERT dengan Levenshtein Distance untuk memperbaiki kesalahan ejaan. Pendekatan ini menggunakan kombinasi kekuatan representasi kontekstual dari BERT dengan algoritma Levenshtein untuk memperbaiki kesalahan ketik dengan lebih presisi, terutama pada teks yang memiliki kesalahan berulang [3].

Penelitian lain yang signifikan dilakukan oleh Hládek et al. (2020), yang menyajikan survei komprehensif tentang berbagai teknik koreksi ejaan otomatis. Mereka mengevaluasi berbagai metode, mulai dari teknik berbasis aturan hingga metode modern berbasis model pembelajaran mesin, seperti BERT dan Transformer, serta menyoroti keunggulan dan keterbatasan dari pendekatan berbasis kamus dan pembelajaran mesin [12]. Studi ini menyoroti pentingnya menggabungkan metode berbasis aturan dan pembelajaran mesin untuk mendapatkan hasil koreksi yang optimal.

Rigaud et al. (2019) dalam ICDAR 2019 Competition on Post-OCR Text Correction juga menunjukkan pentingnya pemrosesan teks setelah *Optical Character Recognition (OCR)*. Mereka mengembangkan pendekatan yang menggabungkan teknik koreksi berbasis aturan dengan model pembelajaran mesin untuk meningkatkan akurasi koreksi kesalahan ejaan pada teks hasil *OCR* [2].

Selain itu, penelitian oleh Guo et al. (2021) menggunakan pendekatan *Global Attention Decoder* untuk memperbaiki kesalahan ejaan dalam bahasa Tionghoa.

Mereka menemukan bahwa penggunaan perhatian global dalam model bahasa secara signifikan meningkatkan kemampuan model dalam mengenali dan memperbaiki kesalahan ejaan, terutama pada teks yang kompleks [4]. Penelitian ini sejalan dengan studi lain yang menunjukkan pentingnya teknik model bahasa modern dalam menangani kesalahan kontekstual.

Pendekatan berbasis kamus juga terus mendapatkan perhatian karena efektifitasnya dalam mendeteksi kesalahan ejaan sederhana. Misalnya, penelitian oleh Sun et al. (2023) menekankan pentingnya integrasi antara pendekatan berbasis kamus dengan model bahasa yang telah dioptimalkan untuk bahasa Tionghoa, dan mereka menunjukkan bahwa metode ini sangat efektif untuk mendeteksi kesalahan ejaan yang umum [5].

Dalam hal metode koreksi teks berbasis pembelajaran mesin, Liu et al. (2023) mengusulkan model Chinese Spelling Correction as Rephrasing Language Model yang memanfaatkan kekuatan model *pretrained* seperti BERT untuk memprediksi koreksi kata dalam konteks yang lebih luas. Mereka menunjukkan bahwa model ini dapat memberikan hasil yang lebih baik dibandingkan dengan metode tradisional, terutama dalam konteks kesalahan tata bahasa yang lebih rumit [6].

Penelitian terkait lainnya dilakukan oleh Jayanthi et al. (2020) dengan memperkenalkan NeuSpell, sebuah *toolkit neural* untuk koreksi ejaan yang menggunakan model pembelajaran mendalam untuk menangani berbagai kesalahan dalam teks [9]. *Toolkit* ini terbukti sangat efektif dalam memperbaiki kesalahan ketik dan kesalahan tata bahasa sederhana, namun juga menunjukkan kelemahan dalam konteks kesalahan yang lebih kompleks.

Secara keseluruhan, penelitian-penelitian ini menunjukkan bahwa penggunaan model pembelajaran mesin modern, seperti BERT dan Transformer, memberikan peningkatan yang signifikan dalam tugas koreksi teks. Namun, integrasi dengan pendekatan berbasis aturan, seperti yang digunakan dalam metode berbasis kamus, tetap memberikan kontribusi yang penting untuk mengatasi kesalahan ejaan yang sederhana.

III. DATASETS

Penelitian ini menggunakan dataset artikel berita berbahasa Indonesia yang berasal dari *repository* HuggingFace di indonesian-nlp/id_newspapers_2018, atau dapat juga diakses melalui *repository* GitHub di feryandi/Dataset-Artikel. Dataset ini terdiri dari berbagai artikel berita berbahasa Indonesia dari berbagai media jurnalistik yang penting untuk analisis dalam penelitian ini. Dataset ini mencakup berbagai kategori berita, mulai dari keseharian, ekonomi, olahraga, politik, dan berbagai kategori lainnya. Diversitas kategori ini memastikan bahwa model bahasa yang dilatih memiliki kemampuan generalisasi yang baik terhadap berbagai jenis artikel berita.

Proses persiapan data melibatkan beberapa tahap:

- **Pengumpulan**: Dataset yang didapatkan dari *repository* HuggingFace ataupun GitHub berisi berita dari berbagai sumber tepercaya, yang kemudian digabungkan menjadi dataset yang lengkap. Dataset ini memastikan keberagaman sumber berita dan representasi yang baik dari penggunaan bahasa Indonesia dalam berbagai konteks.
- **Pembersihan**: Data yang tidak relevan atau mengandung noise dihapus, seperti karakter non-standar atau format yang tidak sesuai. Pembersihan ini bertujuan untuk meningkatkan kualitas data, sehingga model dapat dilatih tanpa terganggu oleh data yang tidak perlu.
- ***Formatting***: Dataset asli terdiri dari ribuan *file* JSON, dengan satu *file* per artikel. Oleh karena itu, diperlukan tahap *formatting* untuk menggabungkan data dari ribuan *file* ini menjadi satu format yang terstruktur. Tahap ini bertujuan agar data dari berbagai *file* dapat diakses dan diproses secara efisien oleh model, sehingga mempercepat proses pelatihan dan analisis.
- **Pelabelan**: Kesalahan ejaan dan bentuk teks yang benar diberi label untuk digunakan dalam proses pelatihan model. Tahap pelabelan dilakukan pada seluruh artikel berita yang digunakan untuk melakukan pelatihan. Tahap ini memastikan bahwa model mendapatkan data yang sudah terstruktur untuk mengidentifikasi pola kesalahan dan koreksi yang diperlukan.

Tahapan ini memastikan bahwa dataset yang digunakan berkualitas tinggi dan representatif untuk tugas koreksi teks, sehingga model dapat bekerja secara efektif di berbagai kategori artikel berita.

IV. PROPOSED METHOD

Berikut adalah arsitektur dari sistem *Text Correction* yang digunakan pada penelitian ini

A. Error Detection

Tahap *error detection* bertujuan untuk mendeteksi kata atau frasa yang dianggap salah dalam teks. Penelitian ini menggunakan tiga variasi metode *error detection* untuk mendeteksi kesalahan ejaan, dan modul *text correction* yang digunakan dalam ketiga variasi ini tetaplah sama, yaitu model IndoBERT. Tahap ini bertujuan untuk mengidentifikasi secara tepat bagian teks yang salah, sehingga koreksi dapat diarahkan secara akurat dan efisien.

- ***Pretrained Error Detection***: Pada variasi ini, digunakan model IndoBERT versi *pretrained* yang belum di-*finetune*. Model ini digunakan untuk mendeteksi kesalahan tanpa penyesuaian khusus pada dataset artikel berita. Model *pretrained* ini mampu mendeteksi kesalahan yang lebih kompleks terkait konteks dan tata bahasa, namun akurasinya masih lebih rendah dibandingkan model yang sudah di-*finetune* karena belum disesuaikan untuk tugas spesifik koreksi teks.
- ***Finetuned Error Detection***: Variasi kedua menggunakan model IndoBERT yang telah di-*finetune* menggunakan dataset artikel berita berbahasa Indonesia yang disiapkan. Proses finetuning membuat model lebih spesifik dalam mendeteksi kesalahan pada teks berita, terutama kesalahan ejaan dan kata yang salah ketik. Pendekatan ini lebih efektif untuk tugas spesifik koreksi teks.

- ***Dictionary Error Detection***: Metode ini menggunakan Kamus Besar Bahasa Indonesia (KBBI) Edisi V yang berisikan 110.000 kata baku bahasa Indonesia sebagai basis untuk mendeteksi kesalahan ejaan. Setiap kata yang muncul dalam teks berita dibandingkan dengan daftar kata yang ada di KBBI. Jika kata tidak ditemukan dalam kamus, kata tersebut dianggap sebagai kesalahan dan akan di-masking untuk proses pengoreksian. Pendekatan ini berfokus pada penanganan kesalahan ejaan yang sederhana dan umum dalam teks.

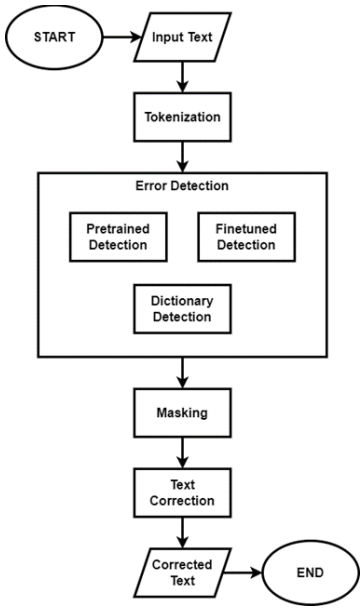
Setelah kesalahan terdeteksi menggunakan salah satu metode di atas, proses koreksi dilakukan menggunakan model IndoBERT yang sama pada setiap variasi. Tahap error detection berfungsi sebagai landasan untuk menentukan bagian mana dari teks yang perlu diperbaiki, sehingga model tidak mencoba memperbaiki bagian teks yang sebenarnya sudah benar.

B. Sistem Utama Text Correction

Sistem koreksi teks yang dibangun dalam penelitian ini terdiri dari beberapa komponen yang berfungsi secara berurutan untuk mendeteksi dan mengoreksi kesalahan penulisan pada artikel berita. Setiap komponen memiliki peran khusus untuk memastikan bahwa teks yang diproses dapat diperbaiki dengan akurat dan efisien.

- ***Tokenization***: Proses tokenisasi memecah teks menjadi unit-unit kecil seperti kata atau sub-kata (token). Tokenisasi berfungsi untuk memastikan bahwa model bahasa dapat memproses teks dalam bentuk yang terstruktur. IndoBERT menggunakan token ini dalam proses masking dan prediksi koreksi. Tahap ini diperlukan karena model bekerja dengan cara mengolah unit-unit kecil dari teks (token), sehingga pemecahan teks menjadi token memastikan pemrosesan lebih terarah dan efisien.
- ***Error Detection***: Proses ini merupakan langkah awal dalam sistem, di mana teks diperiksa untuk mendeteksi kesalahan. *Error detection* dilakukan menggunakan salah satu dari tiga variasi yang telah dijelaskan pada bagian sebelumnya (*pretrained, finetuned,* atau *dictionary-based*). Deteksi kesalahan ini mencakup kesalahan ejaan dan kata yang tidak dikenal. Tahap ini diperlukan untuk memisahkan kata atau frasa yang salah dari yang benar, sehingga hanya bagian yang salah yang akan diperbaiki.
- ***Masking***: Setelah kesalahan terdeteksi, token kata-kata yang salah akan diganti dengan token [MASK]. *Masking* adalah teknik yang digunakan untuk menandai tempat kesalahan dalam teks, sehingga model bahasa dapat memprediksi kata yang benar berdasarkan konteks sekitarnya. Tahap ini berfungsi untuk memberikan sinyal kepada model untuk fokus pada kata yang salah saja dan menghindari memperbaiki bagian teks yang sudah benar.
- ***Text Correction***: Modul koreksi teks yang digunakan adalah model IndoBERT. Setelah kata yang salah diberi token [MASK], model IndoBERT memprediksi kata yang benar berdasarkan konteks sekitar token. Proses prediksi ini menghasilkan beberapa kandidat kata, dan kata dengan skor tertinggi akan dipilih sebagai koreksi. Modul *text correction* yang digunakan tetap konsisten di seluruh variasi *error detection*, memastikan bahwa kesalahan yang dideteksi akan dikoreksi menggunakan pendekatan berbasis konteks yang sama, yaitu model IndoBERT.

Proses ini memastikan bahwa setiap kesalahan yang terdeteksi dikoreksi secara akurat, dengan mempertimbangkan konteks keseluruhan dari artikel berita, sehingga hasil akhir dari sistem koreksi teks ini dapat dipertanggungjawabkan dan relevan. Gambar 1 di bawah ini mengilustrasikan alur dari sistem *Text Correction*



Gambar 1. *Flowchart* Sistem Koreksi Teks

V. EXPERIMENTS AND RESULTS

Eksperimen dilakukan dengan menjalankan masing-masing versi sistem koreksi teks pada dataset yang sama. Setiap versi diuji berdasarkan kemampuan untuk mendeteksi dan mengoreksi kesalahan penulisan pada artikel berita.

Tabel 2. Hasil Eksperimen

System Variation	Detection Accuracy	Correction Accuracy	Overall Accuracy
Pretrained	0,32	0,33	0,28
Finetuned	0,48	0,53	0,46
Dictionary	0,90	0,90	0,89

Hasil eksperimen menunjukkan performa yang berbeda di antara ketiga versi sistem koreksi teks:

***Pretrained Error Detection*:**

Variasi *pretrained* menunjukkan hasil yang cukup rendah dalam hal *Detection Accuracy* (32%) dan *Correction Accuracy* (33%). Hasil ini mengindikasikan bahwa model pretrained sering kali gagal mendeteksi kesalahan penulisan yang umum terjadi dalam teks berita, dan juga sering kali gagal mengoreksi dengan tepat kesalahan yang terdeteksi. Model *pretrained* sering kali menandai kosakata yang sudah benar sebagai kesalahan, dan juga sering kali melewatkan kata yang sebenarnya salah akan tetapi tidak terdeteksi sebagai kesalahan.

Hal ini disebabkan karena model belum di-*finetune*, sehingga belum terfokus pada pola kesalahan yang spesifik dalam teks artikel berita. Model ini mengandalkan pengetahuan umum dari data pelatihan awal, tanpa kemampuan untuk menangani kesalahan yang lebih spesifik dalam artikel berita berbahasa Indonesia, sehingga performanya terbatas.

***Finetuned Error Detection*:**

Variasi *finetuned* menunjukkan hasil yang cukup bagus dalam hal *Detection Accuracy* (48%) dan *Correction Accuracy* (53%). Hasil ini mengindikasikan bahwa model yang telah di-*finetune* mengalami peningkatan performa dalam mendeteksi dan mengoreksi kesalahan teks dibandingkan dengan variasi *pretrained*. Proses *finetuning* memungkinkan model untuk memahami kesalahan penulisan yang lebih spesifik terhadap artikel berita berbahasa Indonesia.

Meskipun model *finetuned* menunjukkan peningkatan performa yang signifikan dibandingkan dengan variasi *pretrained*, model ini masih tetap menunjukkan keterbatasan performa dalam mendeteksi dan mengoreksi kesalahan teks yang lebih kompleks. Variasi ini masih membutuhkan proses *finetuning* yang lebih jauh lagi untuk mencapai performa yang optimal.

***Dictionary Error Detection*:**

Metode berbasis kamus (KBBI) memberikan hasil terbaik dengan *Detection Accuracy* mencapai 90% dan *Correction Accuracy* sebesar 90%. Hasil ini menunjukkan bahwa sistem sangat efektif dalam mendeteksi kesalahan yang berkaitan dengan ejaan, terutama kesalahan yang melibatkan kata yang tidak baku atau salah ketik. Karena sistem ini memeriksa setiap kata dalam teks dan membandingkannya dengan daftar kata baku di KBBI, kata-kata yang tidak sesuai dengan kamus akan langsung terdeteksi sebagai kesalahan.

Namun, sistem ini memiliki keterbatasan dalam mendeteksi kesalahan yang lebih kompleks, terutama yang melibatkan konteks kalimat. Sistem ini tidak memperhitungkan hubungan antar kata dalam kalimat, sehingga kesalahan yang terjadi karena kosakata yang tidak sesuai dengan konteks kalimat akan tetapi kata tersebut benar berdasarkan dari kamus, tidak akan terdeteksi sebagai kesalahan. Contohnya, pada kalimat “para petani memanen tadi di siang hari.” Kosakata “tadi” seharusnya adalah “padi”, akan tetapi tidak terdeteksi sebagai kesalahan karena kosakata “tadi” adalah kosakata yang valid dalam KBBI. Sistem ini juga menganggap kosakata tidak baku seperti *slang*/plesetan kata sebagai kesalahan, dan akan menganggap salah kosakata baku yang lebih baru, yang masih belum ada di KBBI.

VI. Discussion

Evaluasi dari hasil eksperimen menunjukkan bahwa proses *finetuning* model bahasa dengan dataset yang relevan secara signifikan meningkatkan kinerja model dalam tugas koreksi teks. Model IndoBERT yang di-*finetune* mampu mengenali kesalahan yang lebih spesifik dan kontekstual dalam artikel berita berbahasa Indonesia, sehingga hasil deteksi dan koreksi kesalahan meningkat dibandingkan dengan model *pretrained*. *Finetuning* membantu model memahami pola kesalahan penulisan yang umum terjadi, seperti salah ketik atau kesalahan penggunaan kata, yang tidak dapat dikenali oleh model yang dilatih secara umum.

Namun, meskipun ada peningkatan performa setelah *finetuning*, model ini masih menunjukkan keterbatasan dalam menangani kesalahan yang lebih kompleks. Beberapa kesalahan, terutama yang terkait dengan tata bahasa atau pemilihan kata yang tepat dalam konteks yang lebih rumit, sering kali terlewat atau dikoreksi dengan tidak tepat. Misalnya, model *finetuned* sering kali tidak mendeteksi perbedaan antara kata-kata yang benar secara ejaan tetapi salah dalam konteks penggunaannya. Hal ini menunjukkan bahwa model masih memerlukan lebih banyak penyesuaian atau penambahan data pelatihan untuk meningkatkan pemahaman terhadap konteks yang lebih mendalam.

Di sisi lain, variasi *Dictionary Error Detection* menghasilkan kinerja terbaik dalam hal *Overall Accuracy* dengan skor mencapai 90%. Pendekatan berbasis kamus ini sangat efektif dalam mendeteksi kesalahan ejaan yang sederhana, terutama kata-kata yang salah ketik atau tidak sesuai dengan kata baku yang tercantum dalam Kamus Besar Bahasa Indonesia (KBBI). Sistem ini secara konsisten menunjukkan kemampuan yang sangat baik dalam mengidentifikasi kata yang tidak baku dan memberikan koreksi yang tepat berdasarkan daftar kata dalam kamus. Hal ini menunjukkan bahwa pendekatan berbasis aturan, seperti penggunaan kamus, masih menjadi metode yang sangat andal dalam mendeteksi dan mengoreksi kesalahan ejaan yang sederhana, terutama di lingkungan yang terstruktur dengan baik seperti artikel berita.

Namun, meskipun sistem berbasis kamus menunjukkan keunggulan dalam mendeteksi kesalahan ejaan, sistem ini masih memiliki keterbatasan dalam menangani kesalahan yang lebih kompleks, seperti kesalahan tata bahasa atau penggunaan kata yang salah dalam konteks tertentu. Sistem ini tidak memperhitungkan hubungan antar kata dalam kalimat, sehingga kesalahan yang melibatkan kata yang benar menurut KBBI tetapi salah dalam konteks kalimat tidak akan terdeteksi. Sebagai contoh, pada kalimat “para petani memanen tadi di siang hari,” sistem berbasis kamus tidak akan mendeteksi kata "tadi" sebagai kesalahan karena kata tersebut ada dalam kamus, meskipun konteks kalimat menunjukkan bahwa kata yang benar seharusnya adalah "padi". Keterbatasan ini menunjukkan bahwa metode berbasis kamus kurang mampu dalam memahami konteks semantik dan struktur kalimat yang lebih kompleks, sehingga dalam situasi yang membutuhkan pemahaman konteks, model berbasis kamus kurang efektif.

Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa kedua pendekatan memiliki keunggulan dan keterbatasan masing-masing. *Finetuning* model bahasa meningkatkan kemampuan model dalam menangani konteks yang lebih rumit dan kesalahan penulisan yang spesifik, tetapi masih perlu pengembangan lebih lanjut untuk mencapai kinerja yang optimal. Di sisi lain, *Dictionary Error Detection* memberikan hasil yang sangat baik dalam mendeteksi kesalahan ejaan sederhana, namun metode ini tidak fleksibel dalam menangani konteks kalimat yang lebih kompleks. Kombinasi dari pendekatan berbasis kamus dengan model bahasa yang di-*finetune* dapat menjadi solusi yang lebih kuat untuk menangani berbagai jenis kesalahan dalam teks artikel berita.

VII. Future Work

Penelitian di masa depan akan berfokus pada pengembangan sistem koreksi teks yang lebih canggih untuk menangani kesalahan tata bahasa dan sintaksis yang lebih kompleks. Selain itu, sistem ini juga akan diuji di lingkungan *real-time* seperti *newsroom*, di mana artikel berita dipublikasikan dalam waktu singkat. Pengembangan model untuk bahasa lain atau domain teks yang lebih spesifik, seperti teks akademik atau hukum, juga akan menjadi fokus pengembangan di masa depan.

VIII. Conclusion

Penelitian ini telah mengimplementasikan dan mengevaluasi tiga variasi sistem koreksi teks berbasis model bahasa IndoBERT untuk mendeteksi dan mengoreksi kesalahan penulisan pada artikel berita berbahasa Indonesia. Hasil evaluasi menunjukkan bahwa meskipun model yang telah di-*finetune* menghasilkan peningkatan kinerja dibandingkan dengan model yang belum di-*finetune*, pendekatan berbasis kamus (KBBI) tetap memberikan hasil terbaik dalam hal deteksi dan koreksi kesalahan ejaan.

Pendekatan berbasis kamus berhasil mendeteksi dan mengoreksi lebih banyak kesalahan dengan tingkat kesalahan koreksi yang lebih rendah, menjadikannya lebih efektif untuk tugas koreksi teks di lingkungan yang terstruktur seperti artikel berita. Sementara itu, model berbasis pembelajaran mesin, meskipun memiliki potensi besar, memerlukan lebih banyak data dan proses *finetuning* yang lebih intensif agar mampu menyaingi keakuratan metode berbasis aturan. Ini menunjukkan bahwa penggunaan metode berbasis kamus lebih efisien untuk kasus-kasus kesalahan yang umum dan berstruktur, meskipun model berbasis pembelajaran mesin memiliki kemampuan yang lebih baik dalam memahami konteks kesalahan yang lebih kompleks.

Dengan implementasi sistem ini, diharapkan kualitas penulisan artikel berita di Indonesia dapat meningkat, sehingga informasi yang disampaikan menjadi lebih akurat dan dapat dipercaya oleh masyarakat. Penggunaan teknologi koreksi teks otomatis ini juga dapat membantu mengurangi beban kerja jurnalis dan editor, memungkinkan mereka untuk lebih fokus pada pengembangan konten dan penyajian berita yang berkualitas.

Penelitian ini juga membuka peluang untuk pengembangan lebih lanjut, termasuk pengujian pada domain atau bahasa lain, serta integrasi dengan alat bantu penulisan dan penyuntingan yang sudah ada. Dengan demikian, teknologi 1ini dapat terus dikembangkan dan dioptimalkan, sehingga mendukung kualitas penulisan yang lebih baik secara lebih luas dan menyeluruh.

References

[1] H. T. Ngo, D. T. Ham, T. Huynh, and K. Hoang, "A Combination of BERT and Transformer for Vietnamese Spelling Correction," *The Saigon International University, Vietnam*, 2024. [Online]. Available: https://arxiv.org/abs/2405.02573v1. [Accessed: Aug. 28, 2024].

[2] C. Rigaud, A. Doucet, M. Coustaty, and J.-P. Moreux, "ICDAR 2019 Competition on Post-OCR Text Correction," in *15th International Conference on Document Analysis and Recognition*, Sydney, Australia, 2019, pp. 1588-1593. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02304334. [Accessed: Aug. 28, 2024].

[3] A. Naziri and H. Zeinali, "A Comprehensive Approach to Misspelling Correction with BERT and Levenshtein Distance," arXiv, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.17383.

[4] Z. Guo, Y. Ni, K. Wang, W. Zhu, and G. Xie, "Global Attention Decoder for Chinese Spelling Error Correction," *in Proceedings of the ACL-IJCNLP 2021*, 2021, pp. 209-219.

[5] R. Sun, X. Wu, and Y. Wu, "An Error-Guided Correction Model for Chinese Spelling Error Correction," Peking University, MOE Key Laboratory, 2023.

[6] L. Liu, H. Wu, and H. Zhao, "Chinese Spelling Correction as Rephrasing Language Model," Shanghai Jiao Tong University, 2023.

[7] Z. Liang, X. Quan, and Q. Wang, "Disentangled Phonetic Representation for Chinese Spelling Correction," Sun Yat-sen University, Meta AI, 2023.

[8] Y. Hu, X. Jing, Y. Ko, and J. T. Rayz, "Misspelling Correction with Pre-trained Contextual Language Model," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 158-164, 2021.

[9] S. M. Jayanthi, D. Pruthi, and G. Neubig, "NeuSpell: A Neural Spelling Correction Toolkit," in *Proceedings of the 2020 EMNLP (Systems Demonstrations)*, 2020, pp. 158-164.

[10] H. Wu, S. Zhang, Y. Zhang, and H. Zhao, "Rethinking Masked Language Modeling for Chinese Spelling Correction," *in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, Toronto, Canada, 2023, pp. 10743-10756.

[11] S. Zhang, H. Huang, J. Liu, and H. Li, "Spelling Error Correction with Soft-Masked BERT," *in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 882-890.

[12] D. Hládek, J. Staš, and M. Pleva, "Survey of Automatic Spelling Correction," Electronics, vol. 9, no. 10, p. 1670, Oct. 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/10/1670.

[13] C. Wei, S. Huang, R. Li, N. Yan, and R. Wang, "Training a Better Chinese Spelling Correction Model via Prior-knowledge Guided Teacher," *in Proceedings of the 2024 Association for Computational Linguistics (ACL)*, 2024, pp. 13578-13589.

[14] D. Nguyen Tien, T. T. M. Tuoi, L. Le Vu, and T. D. Minh, "Vietnamese Spelling Error Detection and Correction using BERT and N-gram Language Model," CMC Institute of Science Technology, 2023.