Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Klasifikasi Produk Otomotif di Lokapasar Indonesia Menggunakan Model BERT

Rifaldi Achmad Faisal[1]
[1]*Matematika, Universitas Islam Negeri Syarif Hidayatullah Jakarta*
[1]rifaldiaf1@gmail.com

Abstract: *The use of private vehicles, especially cars, is increasing along with the development of the automotive industry. This increase is supported by the widespread development of automotive-based online stores by MSMEs and well-known automotive brands in the world. Therefore, it is interesting to classify by product type. This research aims to classify automotive product categories in marketplaces such as Tokopedia, Shopee, Blibli, Lazada, and Bukalapak, based on product titles and keywords in January and February 2023. The classification method used is BERT and compares with Random Forest and SVM methods to categorize 6 types of products. This research shows that the amount of data from product classification in the car spare parts category is very high when compared to other automotive products so that the data is not distributed evenly. This research shows that the BERT model and the other two methods are less than optimal for data with an unbalanced class distribution.*
Keywords: *Automotive, BERT, Marketplace, Product Classification*

Abstrak: Penggunaan kendaraan pribadi, khususnya mobil, semakin meningkat seiring dengan perkembangan industri otomotif. Peningkatan ini didukung oleh meluasnya perkembangan toko daring berbasis otomotif oleh UMKM dan merek otomotif ternama di dunia. Oleh karena itu, menarik untuk melakukan klasifikasi berdasarkan jenis produk. Penelitian ini bertujuan untuk mengklasifikasikan kategori produk otomotif di lokapasar seperti Tokopedia, Shopee, Blibli, Lazada, dan Bukalapak, berdasarkan judul produk dan kata kunci pada Januari dan Februari 2023. Metode klasifikasi yang digunakan adalah BERT serta membandingkan dengan metode *Random Forest* dan SVM untuk mengkategorikan 6 jenis produk. Penelitian ini menunjukkan bahwa jumlah data hasil klasifikasi produk pada kategori suku cadang mobil sangat tinggi jika dibandingkan dengan produk otomotif lainnya sehingga data tidak terdistribusi secara merata. Penelitian ini menunjukkan model BERT dan dua metode lainnya kurang optimal pada data dengan distribusi kelas yang tidak seimbang.
Kata kunci: BERT, Klasifikasi Produk, Lokapasar, Otomotif

I. PENDAHULUAN

Menjamurnya Lokapasar secara daring merupakan salah satu jalan keluar yang efisien bagi para pemasok dan distributor produk dalam menaikkan keuntungannya. Wadah jual beli produk yang beraneka ragam tersebut dapat dengan mudah diakses di situs web maupun aplikasi. Banyak dari pemasok produk ataupun distributor UMKM dan perusahaan otomotif menggunakan lokapasar untuk menjual produk-produknya. Terutama berkaitan erat dengan industri otomotif seperti mobil yang memiliki banyak komponen di dalamnya baik yang utama atau komponen pendukung. Komponen mobil tersebut dapat dikategorikan berdasarkan kegunaannya seperti aksesoris interior, aksesoris eksterior, suku cadang, serta oli dan penghemat BBM. Bahkan penjualan kendaraan seperti mobil dan motor sekalipun turut andil memperkaya produk lokapasar. Dalam lokapasar terdapat fitur yang mengkategorikan sebuah produk termasuk produk otomotif untuk memudahkan dalam pencarian berdasarkan kategori yang relevan. Namun, cukup banyak ditemui produk yang dijual di lokapasar tidak sesuai dengan kategori aslinya. Termasuk produk-produk otomotif mobil. Oleh karena itu, dibutuhkan lagi penyesuaian kategori berdasarkan produk yang dijual agar memudahkan pengguna dalam mencari barang yang diinginkan berdasarkan kategorinya. Salah satunya adalah dengan melakukan klasifikasi berdasarkan teks judul suatu produk. Dengan memanfaatkan kecenderungan teks judul produk untuk dapat masuk ke dalam kategori yang lebih spesifik.

Salah satu arsitektur *deep learning* terbaik dalam tugas pemrosesan bahasa alami *(Natural Language Processing)* terutama dalam tugas klasifikasi adalah *Bidirectional Encoder Representations*

from Transformers (BERT) yang dikembangkan oleh Google pada tahun 2018. BERT adalah model representasi kata kontekstual yang sudah dilatih sebelumnya pada *corpus* besar, dan menggunakan pendekatan *Masked Language Model* (MLM) dalam dua arah (*bidirectional*). *Transformer* merupakan model arsitektur yang menggunakan mekanisme perhatian (*attention mechanisms*) [2]. Struktur BERT terdiri dari beberapa lapisan *encoder-decoder transformator*, dengan koneksi *point-wise* penuh dan *self-attention* yang diurutkan untuk setiap lapisan. Metode ini memungkinkan model untuk memahami hubungan antara kata-kata dalam teks. *Pre-training* dan *fine-tuning* adalah dua langkah utama dalam pembuatan arsitektur BERT. *Pre-training* BERT menggunakan MLM. model ini menggunakan kata-kata kontekstual di sekitar token yang disembunyikan untuk memprediksi kata yang disembunyikan dalam teks. Selain itu, ada juga *Prediction of Next Sentence* (NSP), yang menentukan apakah sebuah kalimat akan mengikuti kalimat lain dalam teks. Setelah pelatihan awal, BERT dapat disesuaikan (*fine-tuning*) untuk tugas tertentu dengan mengatur parameternya menggunakan data berlabel dari tugas NLP yang lebih khusus [6].

Terdapat beberapa penelitian sebelumnya yang berkaitan dengan klasifikasi teks mengunakan BERT seperti *Comparing BERT Against Traditional Machine Learning Models in Text Classification* yang mana membandingkan model BERT dengan model klasifikasi teks tradisional seperti *voting classifier, logistic regression,* dan model tradisional lainnya pada empat skenario kasus klasifikasi yang berbeda. Hasilnya model BERT memiliki performa terbaik di antara model *NLP* tradisional pada penelitian ini [5]. *Low-Shot Classification: A Comparison of Classical and Deep Transfer Machine Learning Approaches* membahas tentang pendekatan deep transfer learning dengan model BERT dan ULMFiT. BERT menunjukkan performa yang lebih baik dibandingkan model ULMFiT dan pembelajaran mesin klasik seperti *Naive Bayes* serta SVM dalam tugas klasifikasi teks low-shot [8]. *Efficient Classification Model of Web News Documents using Machine Learning Algorithms* membandingkan *K-Nearest Neighbors* (kNN), *Decision Tree* (DT), *Support Vector Machine* (SVM), dan *Long Short-Term Memory* (LSTM) dalam melakukan klasifikasi menggunakan data artikel berita. Hasil yang didapat metode SVM memiliki performa terbaik dibanding metode lainnya [3]. *Indonesian News Classification Using IndoBert* membahas tentang tugas klasifikasi untuk kategori berita berbahasa Indonesia dengan menggunakan model IndoBERT dan dibandingkan dengan beberapa model *pre-trained* lain seperti *Indobert, XLNET, BERT-multilingual,* dan *XLMRoberta.* Kemudian menggunakan model pembelajaran mesin dengan *TF-IDF word embedding* seperti metode *XGBoost, LGB,* dan *random forest.* Hasilnya model IndoBERT mendapatkan hasil terbaik dalam tugas klasifikasi berita berbahasa Indonesia ketimbang model pre-trained dan model machine learning lain yang dibandingkan dalam penelitian [4].

Meskipun metode yang digunakan dalam penelitian sebelumnya menunjukkan hasil yang baik, metode tersebut masih memiliki kelemahan. Karena adanya regularisasi yang melekat dan cenderung menghasilkan model yang kurang cocok jika dibandingkan dengan pengklasifikasi diskriminatif, *Naive Bayes* (NB) berkinerja baik disaat fitur-fitur tidak saling bergantung (independen), dan model yang dihasilkannya mudah diinterpretasikan dan dijelaskan. Di sisi lain, representasi terbatas SVM dapat menyebabkan ketidakmampuan dalam memodelkan pola yang kompleks dalam data pelatihan. Namun, *Support Vector Machine* (SVM) mampu menangani data berukuran besar dan tidak rentan terhadap overfitting. *K-Nearest Neighbor* (KNN) memiliki biaya komputasi yang tinggi saat diterapkan pada data berukuran besar. Metode ini juga rentan terhadap overfitting dan memiliki kinerja yang buruk saat digunakan pada teks yang panjang dan dengan banyak fitur. Model turunan BERT yang lebih unggul melebihi model BERT-*Base* dalam tugas klasifikasi data berbahasa Indonesia adalah model IndoBERT

[1]. Namun, meskipun IndoBERT memiliki hasil yang lebih baik dalam melakukan tugas NLP pada data berbahasa Indonesia, dalam istilah teknis di dunia otomotif masih memiliki banyak kata yang lebih umum diucapkan dalam bahasa Inggris. Oleh karena itu, akan digunakan metode BERT-*multilingual* untuk melengkapi kekurangan-kekurangan yang ada pada metode sebelumnya dalam tugas klasifikasi. Termasuk juga dalam menangani data berbahasa Indonesia yang disisipkan banyak istilah dalam bahasa Inggris.

Dalam penelitian ini, model BERT digunakan untuk klasifikasi produk otomotif di berbagai lokapasar di Indonesia, yang mana pendekatan ini menunjukkan efisiensi dan keakuratan dalam mengolah data teks yang besar, serupa dengan metode clustering via ranking yang diusulkan oleh Sutanto dan Nayak dalam studi mereka mengenai penemuan pengetahuan cepat dari data media sosial melalui pendekatan FCAR (Fast Clustering Approximations via Ranking) [11].

II. METODE PENELITIAN

II.1. Strategi Pengumpulan Data

Dalam paper ini, penulis menggunakan analitik berbasis relevansi untuk mendapatkan pemahaman tentang persepsi merek dan dinamika pasar dalam konteks peraturan data media sosial yang ketat. "Pengambilan Sampel Relevan" dan "Pembelajaran Mesin Berbasis Pengambilan Sampel Relevan" adalah dua komponen utama metodologi inovatif ini. Setiap komponen dibuat untuk mengatasi masalah etis dalam pengumpulan data dan memastikan analisis menyeluruh terhadap konten yang dibuat oleh pengguna.

Akan digunakan "Pengambilan Sampel Relevan" untuk mengatasi masalah terkait privasi dan batas rekam. Metode ini menggunakan data mesin pencari yang diindeks untuk menghasilkan sampel representatif konten media sosial yang berkaitan dengan kehadiran merek otomotif global di Indonesia. penulis memastikan kepatuhan terhadap kebijakan perlindungan data sambil mempertahankan keanekaragaman dan kekayaan dataset dengan berfokus pada data yang tersedia melalui indeks mesin pencari publik.

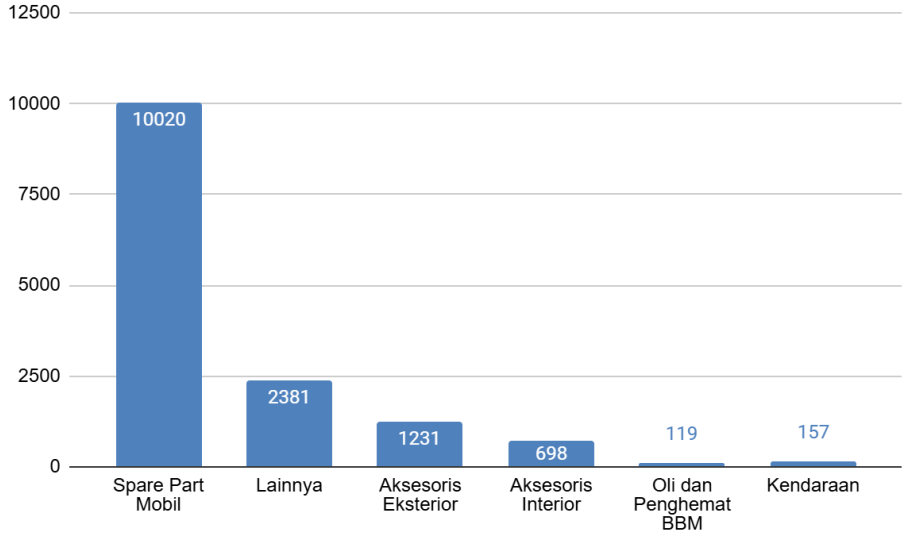
Penulis menggunakan metode "Pengambilan Sampel Relevan" untuk mengatasi masalah privasi dan batasan data. Dengan menggunakan data yang diindeks oleh mesin pencari publik, metodologi ini menghasilkan sampel representatif konten media sosial yang terkait dengan merek otomotif internasional di Indonesia. Penulis memastikan kepatuhan terhadap kebijakan perlindungan data sekaligus mempertahankan keanekaragaman dan kekayaan dataset.

Metodologi ini menggunakan pendekatan strategis pengambilan sampel untuk menekankan relevansi. Relevansi didefinisikan berdasarkan kriteria yang telah ditetapkan sebelumnya, seperti informasi tentang merek otomotif, pengguna Indonesia, dan periode waktu tertentu. Metode pengambilan sampel ini memungkinkan penulis untuk fokus pada data yang sesuai dengan undang-undang privasi dan relevan. Kata kunci yang digunakan dalam pengambilan sampel melibatkan istilah-istilah yang memastikan data yang terambil adalah data otomotif, seperti: "toyota", "daihatsu", "honda", "hyundai", "suzuki", "mitsubishi", dan "wuling".

II.2. Dataset

Penelitian ini menggunakan data yang telah diberi label sebanyak 14.606 data. Data ini diambil dari beberapa platform lokapasar di Indonesia yaitu Tokopedia, Shopee, Blibli, Lazada, dan Bukalapak. Terdapat dua kolom dalam data yaitu kolom *Title* yang berisi teks judul dari produk atau kata kunci dalam *listing* pencarian dan kolom *Label* yang berisi kategori dari isi kolom *Title*. Berikut di bawah ini ditampilkan distribusi label pada produk otomotif dan lima data awal dalam dataset.

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 1. Grafik Distribusi Kategori Produk Otomotif

Tabel 1. Dataset

Title	Label
Jual Tongkat transmisi handle pindah gigi Toyota New dyna - Kota ...	Suku Cadang Mobil
Jual Kaos Toyota baju Toyota kaos kerah Toyota polo shirt toyota ...	Lainnya
Jual 76625-0K370/76626-0K370 MUDGUARD BELAKANG RH/LH ...	Aksesoris Eksterior
Jual TERBARU Soket head unit sansui SA5200i android pnp Toyota ...	Aksesoris Interior
Jual Stopper Gardan Empuk INNOVA / FORTUNER Anti Jedug ...	Suku Cadang Mobil

II.3. Prepocessing Data

Proses *prepocessing* data teks merupakan tahap krusial dalam analisis teks yang bertujuan untuk membersihkan dan mempersiapkan data agar siap digunakan dalam tahap analisis selanjutnya. Dalam penelitian ini terdapat fungsi *cleanText* yang sudah dibuat sebelumnya di dalam library *tauData*. *cleanText* melakukan banyak langkah untuk membersihkan dan memproses teks. Termasuk di dalamnya menghapus tag HTML, entitas HTML, aksen, URL, karakter non-alfanumerik, huruf tunggal, dan spasi berlebih. Fungsi ini juga dapat menghapus *stopwords*, melakukan lemmatisasi, mengonversi *slang*, dan memfilter kata berdasarkan panjangnya. Dalam kaitannya dengan data, kami melakukan *prepocessing* berupa membersihkan teks dari berbagai elemen yang tidak diinginkan seperti karakter non-alfanumerik, menghapus angka, menghapus *stopwords*, menerapkan *lower case* dan memfilter kata berdasarkan panjang yang ditentukan. Kemudian menyesuaikan enam kategori yang ada dengan label numerik. Berikut lebih jelasnya ada pada tabel di bawah ini.

Tabel 2. Kategori dan Label

Kategori	Label
Aksesoris Eksterior	0
Oli dan Penghemat BBM	1
Lainnya	2
Aksesoris Interior	3
Kendaraan	4
Suku Cadang Mobil	5

Tabel 3. Dataset Hasil Prepocessing

Title	Label
jual tongkat transmisi handle pindah gigi toyota dyna kota	5
jual kaos toyota baju toyota kaos kerah toyota polo shirt toyota	2
jual mudguard rhlh	0
jual terbaru soket head unit sansui android pnp toyota	3
jual stopper gardan empuk innova fortuner anti jedug	5

II.4. Pemberian Label pada Teks Menggunakan Model BERT

Setelah data melalui tahap *prepocessing*, langkah berikutnya adalah penyesuaian pada data *input*. Beberapa model pembelajaran yang efisien dan otomatis tidak bisa menerima masukan dalam bentuk teks [9], sehingga diperlukan penyesuaian input lebih lanjut. Oleh karena itu butuh beberapa tahapan sebelum memasukkan *input* ke dalam model BERT. Pada langkah pertama, *input* data diubah terlebih dahulu ke dalam bentuk *input* yang dapat dibaca oleh model BERT. Untuk membuat BERT dapat membaca *input*, maka harus menambahkan token [CLS] di awal kalimat dan token [SEP] di akhir kalimat, serta menentukan panjang kalimat untuk menambahkan token [PAD] sebanyak sisa token. Setelah menambahkan token ini, BERT akan mengubah setiap token kata menjadi id token, dan mengembalikan hasil input_ids dan attention_mask, yang kemudian dapat diteruskan ke dalam model BERT [7].

II.5. Model BERT untuk Tugas Klasifikasi

Model klasifikasi yang digunakan dalam penelitian ini adalah BERT-*Multilingual.* BERT-*Multilingual* merupakan model multibahasa yang baik dalam mentransfer informasi antar bahasa dengan skrip yang berbeda dan bahasa yang serupa secara tipologis. Kemampuan BERT-*Multilingual* untuk generalisasi tidak hanya bergantung pada menghafal kosakata, tetapi juga pada representasi multibahasa yang lebih mendalam [10]. Data yang digunakan pada penelitian ini merupakan data otomotif yang masih mengadopsi beberapa kata dalam bahasa Inggris, sehingga data ini cocok digunakan menggunakan model BERT-*multilingual* karena model ini dapat menangani berbagai bahasa sekaligus. Model *fine-tuned* BERT yang dipakai ini memiliki 12 *layer*, 768 *hidden size,* dan 12 *attention heads.*

II.6. Evaluasi dan Uji Model

Data yang bersih kemudian dibagi menjadi 80% untuk data latih dan 20% untuk data uji. Data di proses model dalam tiga kali epoch dan jumlah batch sebanyak 32. Namun, sebelum melakukan evaluasi pada model, terlebih dahulu akan ditentukan hyperparameter yang sesuai. Berikut merupakan konfigurasi hyperparameter pada model.

Tabel 4. Konfigurasi Hyperparameter

Jenis	Nilai
Dropout	0.1
Learning Rate	2e-5
Batch Size	32
Epoch	3

II.7. Analisis Prediksi Data Baru

Setelah dilakukan evaluasi pada model, maka akan diprediksi menggunakan data baru. Data ini masih merupakan satu dataset pada data evaluasi namun tidak sama. Sebanyak 2000 data tanpa label akan diprediksi dan akan diperiksa kesalahan prediksi tersebut. Sebelum diprediksi, data terlebih dahulu dibersihkan menggunakan fungsi *cleanText*. Hasil dari prediksi akan dianalisis lebih lanjut terkait kesalahan-kesalahan pada model dalam melakukan tugas klasifikasi.

II.8. Random Forest dan Support Vector Machine sebagai Algoritma Perbandingan

Random Forest (RF) dan *Support Vector Machine* (SVM) dipilih sebagai algoritma perbandingan dengan metode BERT dalam penelitian klasifikasi ini. Random Forest dipilih karena kemampuannya dalam menangani kompleksitas data dan kecenderungan untuk tidak overfitting, berkat penggunaan ensemble trees yang mampu memodelkan hubungan non-linear dan menangani variasi dalam distribusi kelas. SVM dipilih karena sifatnya yang memungkinkan untuk menemukan hyperplane terbaik yang memisahkan kelas-kelas berbeda dengan baik dalam ruang fitur yang kompleks. Kedua metode ini dikenal karena kemampuan mereka dalam mempertahankan performa yang baik dengan dataset berdimensi tinggi dan kompleksitas yang beragam, serta memberikan interpretasi yang lebih mudah untuk pemahaman faktor-faktor yang mempengaruhi prediksi kelas.

III. HASIL DAN PEMBAHASAN

III.1. Hasil Evaluasi Model

Penulis menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-Score* sebagai tolak ukur performa pada model. Didapatkan *accuracy* sebesar 90.83%, *precision* sebesar 61%, *recall* sebesar 59%, dan *F1-Score* sebesar 60% dan rata-rata training loss sebesar 0.28. Dari hasil evaluasi tersebut hanya *accuracy* yang memiliki nilai yang baik, namun metrik lain justru kurang baik. Hal ini mengindikasikan adanya ketidakseimbangan data atau kesulitan model dalam mengklasifikasikan beberapa kategori dengan benar.

Kemudian berikut merupakan hasil dari *training loss*, *validation loss*, dan *validation accuracy* dari tiap *epoch*.

Tabel 5. Hasil Evaluasi Model

Epoch	Training Loss	Validation Loss	Validation Accuracy
1	0.739295	0.486406	0.852196
2	0.400906	0.402908	0.890203
3	0.279544	0.340855	0.908361

Berdasarkan **tabel 5**, dapat dilihat bahwa di tiap epoch, nilai *training loss* semakin kecil, yang menunjukkan bahwa model semakin belajar mengurangi kesalahan pada data pelatihan. Selain itu *validation loss* juga semakin kecil, yang menunjukkan bahwa model mengurangi overfitting dan belajar untuk menggeneralisasi. Kebalikannya, *Validation accuracy* justru semakin besar yang menunjukkan bahwa kinerja model pada data validasi meningkat.

III.2. Hasil Prediksi Data Baru

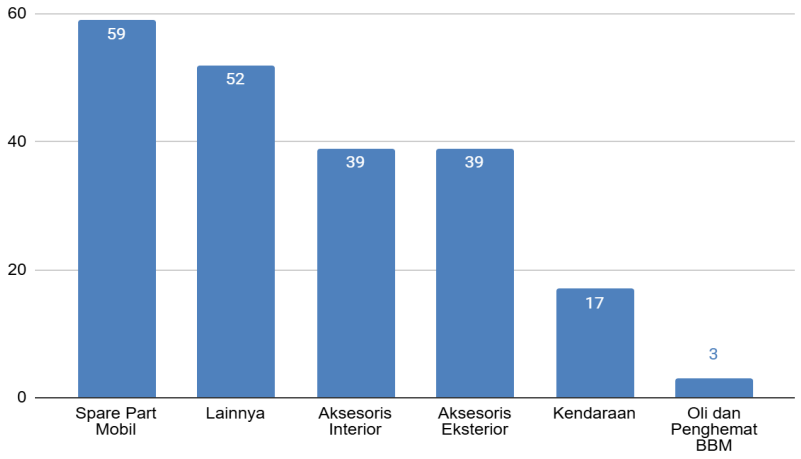
Berikut ini merupakan kategori hasil dari prediksi menggunakan data tidak berlabel.

Tabel 6. Kategori Hasil Prediksi

Jenis	Nilai
Aksesoris Eksterior	328
Oli dan Penghemat BBM	19
Lainnya	372

Jenis	Nilai
Aksesoris Interior	96
Kendaraan	2
Suku Cadang Mobil	1183

Dari **tabel 6** dapat dilihat bahwa kategori paling sedikit adalah "Kendaraan" yaitu hanya dua data. Angka ini sangat jauh jika dibandingkan dengan jumlah prediksi tertinggi yaitu "Suku Cadang Mobil". Kemudian akan ditampilkan grafik jumlah kesalahan prediksi pada masing-masing kategori.



Gambar 2. Grafik Distribusi Kesalahan Prediksi

Seperti yang terlihat di **gambar 2**, terdapat perbedaan jumlah kesalahan prediksi dari masing-masing kategori. Jika dibandingkan dengan jumlah hasil prediksi, kategori "Kendaraan" lebih banyak memiliki kesalahan dalam prediksi. Berbeda dari kategori-kategori lain yang cenderung memiliki jumlah kesalahan prediksi lebih kecil dari jumlah hasil prediksinya. Berikut merupakan gambaran sampel dari lima data kategori "Kendaraan" yang salah diprediksi

Tabel 7. Hasil Prediksi pada Kategori "Kendaraan"

Data Asli	Data Prepocessing	Hasil Prediksi
Jual sx4 modifikasi Harga Terbaik & Termurah Mei 2024 | Shopee ...	jual sx modifikasi harga terbaik termurah mei shopee	Lainnya
Jual jual nissan teana Harga Terbaik & Termurah Mei 2024 ...	jual jual nissan teana harga terbaik termurah mei	Suku Cadang Mobil
Jual jual toyota cressida Harga Terbaik & Termurah Mei 2024 ...	jual jual toyota cressida harga terbaik termurah mei	Suku Cadang Mobil
Jual Mobil Pintu Geser Terlengkap & Harga Terbaru Mei 2024 ...	jual mobil pintu geser terlengkap harga terbaru mei	Suku Cadang Mobil
Jual jual nissan silvia Harga Terbaik & Termurah Mei 2024 | Shopee ...	jual jual nissan silvia harga terbaik termurah mei shopee	Suku Cadang Mobil

Dari **tabel 7** terlihat bahwa satu data diprediksi ke kategori "Lainnya" dan empat lainnya ke kategori "Suku Cadang Mobil". Data *prepocessing* pada penelitian ini salah satunya menghapus karakter angka, sehingga dapat dilihat di data pertama kata "sx4" yang merupakan sebuah tipe mobil berubah menjadi "sx" setelah *prepocessing*. Hal ini sangat berpengaruh dalam keefektifan pengkategorian pada tipe-tipe mobil.

Selain itu, data dari evaluasi model **(gambar 2)** hanya memiliki kategori "Kendaraan" sebanyak 157 data, sangat tidak seimbang pada data keseluruhan sebanyak 14.606. Di lain sisi, kategori "Suku Cadang Mobil" sangat mendominasi data. Ada sebanyak 10.020 data berlabel "Suku Cadang Mobil". Semakin banyak data maka semakin banyak variasinya, sehingga semakin baik dalam melakukan klasifikasi. Terlebih dengan komponen mobil yang dijual di lokapasar sangat beragam jumlahnya. Hal ini dapat membuat kategori "Suku Cadang Mobil" mengambil bagian dari kategori "Kendaraan" yang sudah melalui tahap *prepocessing* ketika proses pengkategorian data berlangsung.

Dari **gambar 2** dapat dilihat untuk kesalahan prediksi kategori "Oli dan Penghemat BBM" tidak sebanyak kategori "Kendaraan". Meskipun dari **gambar 1,** jumlah kategori keduanya tidak jauh berbeda. Berikut merupakan contoh lima data hasil prediksi "Oli dan Penghemat BBM".

Tabel 8. Hasil Prediksi pada Kategori "Oli dan Penghemat BBM"

Data Asli	Data Prepocessing	Hasil Prediksi
Jual oli daihatsu 10w40 Harga Terbaik & Termurah Mei 2024 ...	jual oli daihatsu w harga terbaik termurah mei	Oli dan Penghemat BBM
Jual OLI MESIN MOBIL TMO 10W-40 4 LITER GALON AVANZA ...	jual oli mesin mobil tmo w liter galon avanza	Oli dan Penghemat BBM
Paket oli Shell HX6 10-40 SUZUKI KARIMUN WAGON R ESTILO ...	paket oli shell hx suzuki karimun wagon r estilo	Oli dan Penghemat BBM
Jual Oli Gardan TMO GL4 75WL-90 Original 08885-81624 di lapak ...	jual oli gardan tmo gl wl original di lapak	Oli dan Penghemat BBM
Jual Oli Petronas Syntium 3000 5w-40 1liter Oil Minyak Pelumas ...	jual oli petronas syntium w liter oil minyak pelumas	Oli dan Penghemat BBM

Dari **tabel 8** di atas, terlihat bahwa semua produk oli tersebut berhasil diprediksi ke kategori "Oli dan Penghemat BBM". Dari data *prepocessing,* kata "oli" tidak terhapuskan dan tidak banyak nama varian oli di dalam data. Sehingga model cukup mudah memprediksi produk oli dengan akurat. Berbeda halnya dengan kendaraan, terutama pada mobil yang memiliki tipe sangat beragam dan memiliki nomor-nomor unik antar produk.

III.3. Perbandingan dengan Hasil Algoritma Lain

Penulis kemudian membandingkan metode BERT dengan algoritma machine learning untuk tugas klasifikasi seperti *Random Forest* (RF) dan *Support Vector Machine* (SVM). Berikut ditampilkan tabel perbandingan hasil evaluasi dari ketiga algoritma

Tabel 9. Perbandingan Evaluasi Algoritma

Algoritma	Presisi (%)	Recall (%)	F1-Score (%)	Akurasi (%)
BERT	61	59	60	92
RF	91	71	78	90
SVM	91	79	84	92

Dari **tabel 9** terlihat bahwa ketiga model yang dievaluasi menunjukkan karakteristik yang berbeda dalam kinerja msaing-masing. Meskipun BERT mencapai akurasi yang tinggi sebesar 92%, namun menunjukkan kesulitan dalam menangani ketidakseimbangan kelas dibuktikan dengan *precision* sebesar 61% dan *recall* sebesar 59%. Hal ini mengindikasikan bahwa meskipun BERT mampu secara umum memprediksi mayoritas kategori dengan baik, namun kemampuannya dalam mengenali dan mengklasifikasikan kasus-kasus dari kategori minoritas seperti "Oli dan penghemat BBM" atau "Kendaraan" masih perlu diperbaiki.

Di sisi lain, Random Forest dan SVM menunjukkan presisi yang sangat tinggi yaitu sebesar 91%, tetapi *recall* hanya mencapai 71%. Meskipun akurasinya cukup baik yaitu sebesar 90%, model ini cenderung lebih fokus pada presisi dalam mengidentifikasi kelas yang dominan dan mengabaikan sebagian identifikasi kasus dari kelas minoritas. Sementara itu, Support Vector Machine (SVM) menunjukkan *recall* yang sedikit lebih baik yaitu sebesar 79%. Begitupun dengan akurasi yang baik yaitu sebesar 92%. Namun hal ini belumlah cukup jika melibatkan semua metrik evaluasi dalam menggambarkan kinerja model.

IV. KESIMPULAN

Ditribusi kelas pada data menunjukkan ketidakseimbangan yang menyebabkan model yang dipakai, dalam hal ini BERT, dapat mencapai akurasi yang tinggi, namun dengan hanya memprediksi kelas mayoritas yang benar dan mengabaikan kelas minoritas. Semakin jauh perbandingan proporsi pada data, maka kelas minoritas relatif sulit untuk dikategorikan dengan tepat. Model akan cenderung mengkategorikan data baru ke dalam kelas mayoritas.

Pada data otomotif sangat perlu memerhatikan *prepocessing* pada data. Penghilangan angka pada data sangat berpengaruh dalam model. Dikarenakan banyak istilah dalam dunia otomotif, termasuk mobil yang menggunakan angka sebagai tipe unik dari produknya. Jika angka dihilangkan pada tipe produk, maka bisa mengakibatkan salah tafsir oleh model. Dari hasil penelitian terlihat bahwa meskipun kelas "Kendaraan" dan "Oli dan Penghemat BBM" merupakan kelas minoritas, namun yang paling sering salah prediksi adalah kategori "Kendaraan". Hal ini dikarenakan *prepocessing* yang kurang baik sehingga menghilangkan angka pada produk mobil sebagai informasi penting dalam tugas klasifikasi. Oleh karena itu, kelas "Kendaraan" menjadi salah satu kelas minoritas yang memiliki proporsi salah prediksi yang besar dibandingkan kelas lainnya.

Dari hasil *accuracy*, ketiga model memang baik. Namun dari *recall* menunjukkan hasil yang masih kurang bagus. Ini berarti model belum cukup bisa mendeteksi prediksi positif. Hal ini dapat terjadi karena data yang dievaluasi tidak seimbang dan juga prepocessing yang kurang baik. Selain itu, hasil *presicion* dan *F1-score* dari BERT lebih kecil dari dua model lainnya. Hal ini dikarenakan model BERT kurang baik dalam mengidentifikasi data pada kelas minoritas sehingga justru memprediksinya masuk ke dalam kelas mayoritas.

REFERENSI

1. A. Faridhah, "Perbandingan Kinerja Klasifikasi Pesan Gejala Covid-19 dari Pesan Sosial Media dengan BERT dan IndoBERT," *Repo-Mhsulmacid*, [Online]. Available: https://repo-mhs.ulm.ac.id/handle/123456789/41862
2. A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
3. A. M. K. G. K. Zrar, S. Ghafoorc Halgurd, and M. D. B. Rawate, "Efficient Classification Model of Web News Documents using Machine Learning Algorithms".
4. B. Juarto, "Indonesian news classification using indobert," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 2, pp. 454–460, 2023.
5. E. C. Garrido-Merchan, "Comparing BERT Against Traditional Machine Learning Models in Text Classification," *J. Comput. Cogn. Eng.*, vol. 2, no. 4, pp. 352–356, 2023, doi: 10.47852/bonviewJCCE3202838.
6. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. Accessed: Jun. 16, 2024. [Online]. Available: http://arxiv.org/abs/1810.04805
7. J. H. Tandijaya, L. Liliana, and I. Sugiarto, "Klasifikasi dalam Pembuatan Portal Berita Online dengan Menggunakan Metode BERT," *J. Infra*, 2021, [Online]. Available: http://publication.petra.ac.id/index.php/teknik-informatika/article/view/11467

8. P. Usherwood and S. Smit, "Low-Shot Classification: A Comparison of Classical and Deep Transfer Machine Learning Approaches." arXiv, Jul. 17, 2019. Accessed: Jun. 16, 2024. [Online]. Available: http://arxiv.org/abs/1907.07543
9. R. Qasim, "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification," *J. Healthc. Eng.*, vol. 2022, no. Query date: 2024-06-11 14:05:55, 2022, doi: 10.1155/2022/3498123.
10. T. Pires, E. Schlinger, and D. Garrette, "How multilingual is Multilingual BERT?" arXiv, Jun. 04, 2019. Accessed: Jun. 16, 2024. [Online]. Available: http://arxiv.org/abs/1906.0150
11. T. Sutanto and R. Nayak, "Fast Knowledge Discovery in Social Media Data using Clustering via Ranking," in *2021 9th International Conference on Cyber and IT Service Management (CITSM)*, Bengkulu, Indonesia: IEEE, Sep. 2021, pp. 1–8. doi: 10.1109/CITSM52892.2021.9588866.