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

Prediksi *Gender Username* Twitter Indonesia Terkait Otomotif Menggunakan *Hierarchical Classification* dan BERT

Syifa Ghaisani[1], Taufik Edy Sutanto[2]
[1,2]*Matematika, Universitas Islam Negeri (UIN), Syarif Hidayatullah Jakarta*
[1] syf.ghaisani@gmail.com
Corresponding author: [2]taufik.sutanto@uinjkt.ac.id

Abstract: *Preferences in the automotive industry are often influenced by gender factors, with men and women showing different tendencies in choosing the type, design, and features of vehicles. Social media platforms such as Twitter have become valuable sources of information for automotive companies to understand consumer preferences, identify market trends, and design more effective marketing strategies. This research utilizes data from Indonesian Twitter, focusing on automotive-related keywords to predict users' gender based on their usernames. The dataset consists of 14,805 labeled data points and 5,443 test data points. The methods employed in this study include Hierarchical Classification Heuristic and Bidirectional Encoder Representations from Transformers (BERT), as well as an analysis of the effectiveness of syllable cutting through feature engineering processes. Furthermore, this research introduces nameXpander, a feature engineering method that enhances username representations by expanding original usernames using commonly used Indonesian language word expansions. The research findings indicate that the combined model of BERT and nameXpander achieves the highest prediction accuracy of 90%. This confirms that social media data analysis, particularly through the combination of BERT and nameXpander methods, can be an effective tool for understanding and responding to gender preferences in the automotive market.*
Keywords: *Automotive Industry, BERT, Gender, nameXpander, Social Media, Twitter.*

Abstrak: Preferensi dalam industri otomotif seringkali dipengaruhi oleh faktor *gender*, dengan pria dan wanita menunjukkan kecenderungan yang berbeda dalam memilih jenis, desain, dan fitur kendaraan. Media sosial seperti Twitter telah menjadi sumber informasi yang berharga bagi perusahaan otomotif untuk memahami preferensi konsumen, mengidentifikasi tren pasar, dan merancang strategi pemasaran yang lebih efektif. Penelitian ini menggunakan data dari media sosial Twitter Indonesia dengan fokus pada kata kunci otomotif untuk melakukan prediksi *gender* pengguna berdasarkan nama pengguna mereka. *Dataset* yang digunakan terdiri dari 14.805 data berlabel dan 5.443 data uji. Metode yang diterapkan dalam penelitian ini meliputi *Hierarchical Classification Heuristic* dan *Bidirectional Encoder Representations from Transformers* (BERT), serta analisis efektivitas pemotongan suku kata melalui proses rekayasa fitur. Selain itu, penelitian ini juga memperkenalkan *nameXpander*, sebuah metode rekayasa fitur yang mengembangkan representasi nama pengguna dengan memperluas nama pengguna asli menggunakan ekspansi kata-kata berbahasa Indonesia yang umum digunakan. Hasil penelitian menunjukkan bahwa model kombinasi antara BERT dengan *nameXpander* mencapai akurasi prediksi tertinggi sebesar 90%. Hal ini menegaskan bahwa analisis data media sosial, khususnya melalui kombinasi metode BERT dengan *nameXpander*, dapat menjadi alat yang efektif untuk memahami dan merespons preferensi gender dalam pasar otomotif.
Kata kunci: BERT, Gender, Industri Otomotif, Media Sosial, nameXpander, Twitter.

I. PENDAHULUAN

Bayangkan jika perusahaan otomotif dapat memahami keinginan konsumen mereka bahkan sebelum konsumen itu sendiri menyadarinya. Berkat kehadiran media sosial seperti Twitter telah menjadi gudang data yang memungkinkan perusahaan menggali wawasan berharga mengenai preferensi konsumen, termasuk perbedaan yang dipengaruhi oleh *gender*. Faktor *gender* seringkali mempengaruhi preferensi dalam industri otomotif, antara pria dan wanita menunjukkan kecenderungan yang berbeda dalam memilih jenis, desain, dan fitur kendaraan. Di tengah persaingan yang semakin ketat, pemahaman mendalam mengenai preferensi *gender* dalam industri otomotif dapat menjadi kunci keberhasilan bagi perusahaan dalam memahami preferensi pelanggan, mengidentifikasi tren pasar yang dinamis, serta menyusun strategi pemasaran yang efektif dan tepat sasaran.

Penelitian ini dilakukan untuk menjawab tantangan tersebut dengan memanfaatkan media sosial, khususnya Twitter, dalam memahami preferensi *gender* konsumen otomotif di Indonesia. Dengan menggunakan *dataset* yang berisi ribuan data berlabel dan data uji, penelitian ini menerapkan metode seperti *Hierarchical Classification Heuristic* dan *Bidirectional Encoder Representations from Transformers* (BERT). Selain itu, inovasi dalam rekayasa fitur melalui metode *nameXpander* juga diperkenalkan. Metode *Hierarchical Classification Heuristic* memungkinkan analisis yang terstruktur, sementara BERT yang dikembangkan oleh *Google AI Language* dilatih untuk memiliki kemampuan memahami konteks bahasa dengan lebih mendalam [1]. Inovasi dalam rekayasa fitur melalui metode *nameXpander* mengembangkan representasi nama pengguna dengan memperluas kata-kata umum berbahasa Indonesia.

Supitcha Yuenyong and Sukree Sinthupinyo [2] melakukan penelitian untuk mengklasifikasi *gender* berdasarkan nama pengguna Facebook Thailand. Data yang digunakan dalam penelitian ini dikumpulkan dari bulan Januari hingga Maret 2019 yang bersumber dari pustaka *selenium*. Pengguna yang dipilih adalah yang memiliki nama pengguna dalam karakter Thailand dan memiliki profil *gender* yang terbuka, total data yang berhasil dikumpulkan sebanyak 4.317 data. Dalam penelitian ini terdapat 5 model yang digunakan yaitu *K-Nearest Neighbor, Super Vector Machine, Random Forest, Multinomial Naïve Bayes*, dan *Neural Network*. Hasil eksperimen menunjukkan bahwa menggunakan tokenisasi kata untuk semua nama pengguna mencapai tingkat akurasi dasar sebesar 65,81%, tetapi model gabungan mencapai peningkatan kinerja dengan tingkat akurasi sebesar 91,75%.

Babatunde Onikoyi, Nonso Nnamoko, dan Ioannis Korkontzelos [3] dalam penelitiannya melakukan prediksi *gender* dengan data tekstual deskriptif menggunakan Pendekatan *machine learning*. *Dataset* yang digunakan dalam penelitian ini sebanyak 20.050 pengguna Twitter dengan satu *tweet* acak per pengguna. Penelitian ini memiliki empat model untuk dibandingkan yaitu BOW, TF-IDF, W2Vec, dan GloVE serta memiliki lima metode yaitu *Logistic Regression* (LR), *Support Vector Machine* (SVM), *Naive Bayes* (NB), *Random Forest* (RF), dan *XGBoost* (XGB) . Hasil yang didapatkan ditemukan bahwa model yang memiliki kinerja terbaik adalah GloVE. Algoritma ML yang berkinerja terbaik saat dikombinasikan dengan GloVE adalah *Random Forest*, yang mencapai akurasi sebesar 70%.

Annisa Selma Zakia, Indriati, dan Marji [4] melakukan klasifikasi jenis kelamin pengguna Twitter dengan menggunakan metode BM25 dan *K-Nearest Neighbor* (KNN). Data yang digunakan dalam penelitian ini didapatkan dari dua sumber yaitu laman Twitter dan penyebaran kuesioner, total data yang dimiliki sebanyak 1.000 data. Proses klasifikasi dilakukan dengan memasukan nilai *k* dan data uji, kemudian sistem akan melakukan proses klasifikasi secara otomatis. Nilai optimal KNN dalam penelitian adalah *k*=3. Nilai pengujian *k*=3 adalah 68,6%, 67,63%, 71,52% dan 69,34% untuk nilai akurasi sebesar 68,6%, untuk *precision* sebesar 67,63%, untuk *recall* sebesar 71,52% dan *f-score*nya sebesar 69,34%. Saat nilai *k* lebih dari 3 hingga 10 hasil yang dikeluarkan mengalami penurunan, kemudian nilai tersebut mengalami kenaikan pada saat *k* lebih dari 10 tetapi hasil yang didapatkan tidak melebihi nilai *k*=3. Hasil akhir dalam penelitian ini menyimpulkan bahwa metode BM25 dan KNN sudah bisa melakukan klasifikasi jenis kelamin pada pengguna Twitter.

Ridho Akbar [5] dalam penelitiannya melakukan klasifikasi *gender* pada nama-nama orang indonesia menggunakan *multinomial naive bayes* dan *random forest classifiers*. Data yang digunakan dalam penelitian ini sebanyak 50.000 nama indonesia. Metode yang digunakan sebanyak dua metode untuk dibandingkan yaitu *Naive Bayes* dan *Random Forest*. Hasil yang dimiliki dalam penelitian ini menunjukkan bahwa pengklasifikasi *Random Forest* (RF) bekerja lebih baik bila dibandingkan dengan

Naive Bayes (NB) dalam memprediksi *gender* berdasarkan nama. Distribusi keyakinan menunjukkan bahwa pengklasifikasi RF cenderung yakin ketika memprediksi nilai yang benar. Selain itu. Kedua pengklasifikasi menunjukkan akurasi lebih dari 70% pada *dataset* pengujian yang lebih besar dibandingkan dengan *dataset* pelatihan. Menambah ukuran *dataset* pelatihan mungkin akan meningkatkan akurasi. Namun, dalam hal efisiensi, RF lebih unggul dari NB dengan akurasi lebih dari 80% setelah dilatih menggunakan *dataset* pelatihan.

Parth Aggarwal dan Rhea Mahajan [6] melakukan pendeteksian dan klasifikasi perundungan siber dengan menggabungkan metode *Bidirectional Encoder Representations from Transformers* (BERT) dan *Support Vector Machine* (SVM) dalam teks media sosial. Data yang digunakan dalam penelitian ini berasal dari IEEE *data port*, data yang digunakan sebanyak 2.140 data tweet yang dikategorikan menjadi lima kategori yaitu *Sexual Harassment, Doxing, Cyberstalking, Revenge porn* dan *Slut Shaming*. Hasil yang didapatkan dari penelitian ini melalui serangkaian eksperimen dan evaluasi yang komprehensif, penelitian ini berhasil menunjukkan bahwa model *ensemble* yang digunakan sangat efektif, mencapai tingkat akurasi yang tinggi mencapai 90% pada data uji. Keberhasilan ini diperoleh dengan memanfaatkan keunggulan teknologi deep learning dan pembelajaran mesin konvensional.

Hal yang menjadi pembeda dalam penelitian ini bukan hanya dari metode yang digunakan yaitu *Hierarchical Classification Heuristic*, BERT serta metode *nameXpander* tetapi juga untuk data yang kami dapatkan dalam penelitian ini memenuhi standar perlindungan data dan juga mematuhi protokol hukum yang berlaku.

Penelitian ini bertujuan untuk membandingkan efektivitas ketiga metode tersebut dalam memprediksi *gender* pengguna berdasarkan nama pengguna mereka, dimana pemrosesan data yang efektif dan pemilihan model yang tepat dapat menghasilkan prediksi yang akurat [7] . Dengan melakukan perbandingan ini, diharapkan dapat ditemukan metode yang paling efektif dalam memahami dan merespons preferensi *gender* di pasar otomotif.

II. METODE PENELITIAN

Penelitian ini dilakukan dengan memanfaatkan data dari media sosial Twitter untuk memprediksi *gender* konsumen otomotif di Indonesia. Berikut tahapan dan metode yang digunakan dalam penelitian ini:

a. Pengumpulan Data

Data yang dikumpulkan berasal dari Twitter menggunakan kata kunci terkait otomotif menghasilkan *dataset* yang terdiri dari 14.805 data berlabel dan 5.443 data uji. Untuk memahami persepsi merek dan dinamika pasar dalam konteks regulasi ketat media sosial, penelitian ini memperkenalkan Analitik Berbasis Relevansi. Metodologi ini mencakup dua komponen utama: Pengambilan Sampel Relevan dan Pembelajaran Mesin Berbasis Pengambilan Sampel Relevan, yang dirancang untuk mengatasi tantangan etis dalam pengumpulan data sambil memastikan analisis mendalam terhadap konten pengguna.

Untuk mengatasi tantangan privasi dan batasan regulasi, kami menggunakan pendekatan pengambilan sampel relevansi. Metode ini menggunakan data yang diindeks oleh mesin pencari untuk menghasilkan sampel representatif dari konten media sosial terkait kehadiran merek otomotif internasional di Indonesia. Dengan berfokus pada data yang tersedia melalui indeks mesin pencari publik, kami menjaga integritas dan keragaman kumpulan data sekaligus memastikan kepatuhan terhadap kebijakan privasi.

Untuk mematuhi praktik penelitian yang etis dan meningkatkan keandalan pengumpulan data, kami mengintegrasikan API protokol hukum ke dalam proses pengambilan sampel yang relevan. API ini memfasilitasi ekstraksi data sesuai dengan pedoman hukum yang berlaku, memberikan jaminan tambahan bahwa metodologi kami sejalan dengan kerangka kerja regulasi. Integrasi ini memastikan bahwa pengumpulan data kami tidak hanya memenuhi standar perlindungan data, namun juga mematuhi protokol hukum.

Istilah "Relevan" dalam metodologi kami menggambarkan pendekatan strategis terhadap pengambilan sampel. Kami mendefinisikan relevansi berdasarkan kriteria tertentu, seperti konten yang berkaitan dengan merek otomotif, berasal dari pengguna di Indonesia, dan mencakup periode waktu tertentu. Pendekatan pengambilan sampel yang terfokus ini memungkinkan kami untuk mengumpulkan data yang tidak hanya sesuai dengan regulasi privasi tetapi juga relevan dengan tujuan penelitian kami, memastikan analisis yang terarah dan bermakna.

Kata kunci yang kami gunakan dalam pengambilan sampel dirancang untuk memastikan bahwa data yang diambil adalah dalam bahasa Indonesia, seperti: "toyota indonesia","daihatsu indonesia","honda indonesia","hyundai indonesia","suzuki indonesia","mitsubishi indonesia","wuling indonesia", "mobil toyota","mobil daihatsu","mobil honda","mobil hyundai","mobil suzuki","mobil mitsubishi", dan "mobil wuling".

b. Pra Pemprosesan Data

Data yang dikumpulkan dari Twitter kemudian diproses untuk memastikan relevansi dan kualitas data. Pembersihan data merupakan langkah awal yang sangat penting dalam proses analisis data. Proses ini bertujuan untuk memastikan bahwa data yang digunakan untuk analisis terlepas dari kesalahan atau anomali yang dapat mempengaruhi hasil akhir. Pembersihan data dalam penelitian ini melibatkan beberapa hal seperti mengidentifikasi serta menghapus data duplikat untuk menghindari bias analisis dan untuk memastikan setiap entri berbentuk unik, menginisialisasi dan menghapus nilai yang hilang, memfilter data yang tidak relevan atau tidak valid, untuk memfilter data yang tidak relevan atau tidak valid, kami menerapkan beberapa aturan yaitu jika panjang nama atau *username* kurang dari 5 karakter, data tersebut dianggap tidak valid dan diberi label *"Unknown/Hidden"*. Jika *username* hanya terdiri dari angka, data tersebut akan dianggap tidak valid dan diberi label *"Unknown/Hidden"*. *Username* yang hanya terdiri dari angka biasanya tidak memberikan informasi yang berguna untuk prediksi *gender*. Jika nama atau *username* mengandung kata dari daftar *'unknown'* yang sudah kami buat, data tersebut dianggap tidak relevan dan ditandai sebagai *"Unknown/Hidden"*. Kata-kata ini dianggap tidak bermakna atau umum dan tidak memberikan informasi berguna untuk memprediksi *gender*. Kami memberikan daftar kata-kata yang berhubungan dengan media. Jika nama atau *username* mengandung kata yang termasuk dalam daftar 'media', maka data tersebut dianggap sebagai entitas atau institusi media massa dan ditandai sebagai "MediaMasa/Institusi". Hal ini membantu memisahkan unit-unit organisasi dan institusi yang tidak relevan dengan prediksi *gender* individu. Nama atau *username* yang masih valid kemudian diperiksa apakah mengandung kata-kata yang berada dalam daftar nama pria atau wanita yang telah kami buat untuk memberikan label "Pria" atau "Wanita".

Selanjutnya ada normalisasi data, normalisasi data merupakan proses penting yang melakukan standarisasi data sehingga setiap fitur atau entri memiliki skala yang sama. Dalam penelitian ini, normalisasi terutama berfokus pada pra pemrosesan teks untuk digunakan dalam model BERT. Normalisasi yang dilakukan yaitu *preprocessing* nama, fungsi *nameXpander* digunakan untuk membersihkan dan menstandarisasi teks nama. Langkah-langkah *preprocessing*nya adalah dengan menghapus karakter non-abjad dan angka, menghapus karakter berulang, mengubah teks menjadi huruf

kecil, dan menghapus tanda baca dan karakter khusus. Normalisasi berikutnya yaitu *syllabification* atau pemenggalan suku kata Proses ini membagi kata menjadi suku kata-suku kata yang lebih kecil. Hal ini membantu dalam memperluas nama untuk analisis lebih lanjut.Contohnya seperti nama "risa" menjadi ri-sa.

Lalu terdapat *Labeling*, proses ini merupakan pemberian label atau kategori pada data yang tidak berlabel untuk tujuan klasifikasi atau pengelompokan. Dalam penelitian ini, dilakukan *Manual Labeling* yang berarti kami melakukan klasifikasikan data secara manual. Misalnya, menentukan apakah *username* tertentu berhubungan dengan *gender* pria atau wanita berdasarkan aturan atau pengetahuan yang kami miliki.

c. Rekayasa Fitur

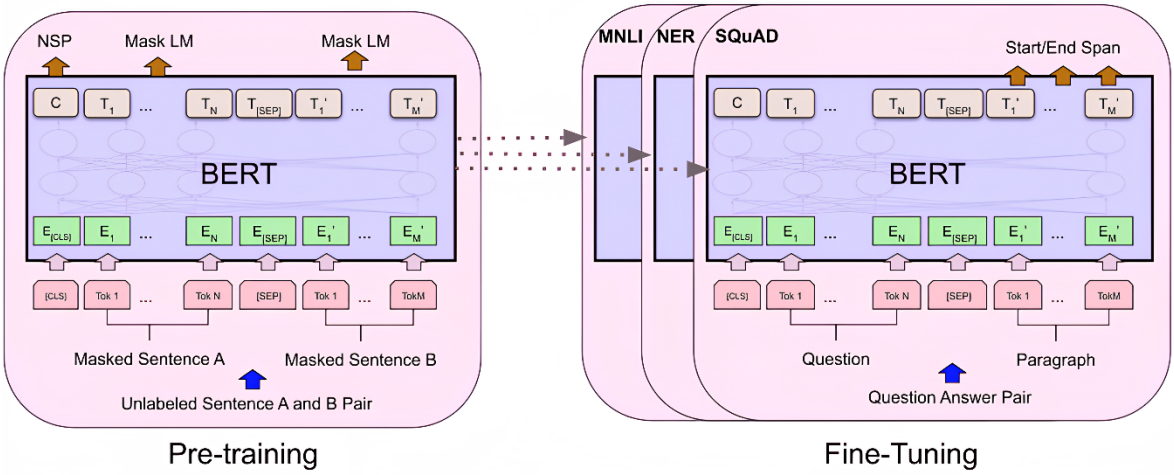
NameXpander merupakan metode rekayasa fitur inovatif yang bertujuan untuk meningkatkan akurasi prediksi *gender* berdasarkan nama pengguna. Metode ini memperluas representasi nama pengguna hingga mencakup suku kata dan kata-kata umum yang umum ditemukan dalam bahasa Indonesia. *NameXpander* memanfaatkan prinsip-prinsip dalam jurnal terdahulu yang memperkenalkan model *transformer* dan mekanisme *self-attention* yang menjadi dasar untuk model NLP *modern* seperti BERT memungkinkan model untuk memperhatikan detail-detail penting dalam nama yang mungkin mengindikasikan *gender*, menghasilkan prediksi yang lebih akurat dan informatif [8]. *NameXpander* dibuat untuk memperbanyak informasi yang terkandung dalam nama pengguna dengan cara memperluas dan memecah nama menjadi elemen-elemen yang lebih granular, seperti suku kata atau bentuk kata-kata yang umum diketahui. Oleh karena itu, metode ini memungkinkan membantu model NLP, khususnya model BERT, dalam menangkap pola-pola linguistik yang lebih halus yang mungkin tidak terlihat pada analisis nama secara keseluruhan. Contoh penerapan nameXpander seperti nama "Rahmawati Abdullah" hasilnya akan menjadi "rah ma wa ti ab dul lah", lalu nama "Budi56 Santoso" hasilnya menjadi "bu di san to so".

d. Metode Klasifikasi

Metode klasifikasi digunakan dalam penelitian ini untuk memprediksi *gender* pengguna berdasarkan nama pengguna mereka, metode yang digunakan ada dua yaitu *Hierarchical Classification Heuristic* (HCH), pendekatan dalam klasifikasi yang menggunakan struktur hierarkis untuk mengelompokkan dan mengkategorikan data secara lebih efisien. Metode ini cocok digunakan untuk *dataset* yang kompleks dengan banyak kelas yang memiliki hubungan hierarkis atau berjenjang. Dalam HCH, output dari algoritma klasifikasi didasarkan pada struktur taksonomi kelas, di mana kelas-kelas dibagi berdasarkan hierarki tertentu. Pendekatan ini membantu meningkatkan efisiensi serta akurasi dalam proses klasifikasi dengan memanfaatkan informasi struktur hierarkis yang ada [9].

Metode yang kedua yaitu *Bidirectional Encoder Representations from Transformers* (BERT) yang merupakan salah satu model dalam bagian *transformer* yang memiliki beberapa kemampuan untuk melakukan pemrosesan bahasa alami, melakukan tugas seperti pemahaman kata, pemahaman dokumen, dan pemahaman kalimat. BERT menggunakan arsitektur *transformer*, yang memiliki beberapa lapisan *encoder* yang digunakan untuk mempelajari representasi kata [10]. Dalam pengerjaannya BERT memiliki dua langkah yaitu *pre-training* dan *fine-tuning*, dalam *pre-training* model dilatih kepada data yang tidak berlabel melalui tugas *pre-training* yang berbeda. dan untuk *fine-tuning* model BERT pertama kali akan di inisialisasi dengan parameter yang sudah dilatih di *pre-training*, dan semua parameter *fine-tuning* menggunakan data berlabel dari tugas turunan (*downstream*). setiap tugas turunan (*downstream*) memiliki model *fine-tuning* yang terpisah, meskipun telah diinisialisasi dengan parameter yang sama yang telah dilatih [11]. Berikut dua langkah yang digunakan dalam BERT:

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



Gambar 1. Prosedur pre-training dan fine-tuning pada arsitektur BERT [11]

Dalam penelitian ini memiliki beberapa tantangan dalam menggunakan metode BERT berdasarkan data yang dimiliki, tantangan yang dihadapi seperti banyaknya *username* di Twitter tidak secara jelas mencerminkan informasi tentang gender pengguna. Misalnya, *username* yang hanya berisi kombinasi angka atau karakter acak, atau *username* dengan kata-kata yang tidak spesifik seperti kombinasi huruf acak. Ini dapat membuat tugas prediksi *gender* menjadi lebih sulit karena kurangnya indikasi langsung tentang jenis kelamin dari *username* itu sendiri. Meskipun BERT sangat baik dalam memahami konteks dan makna teks, namun seringkali sulit untuk BERT membuat prediksi berdasarkan *username* tertentu, terutama tanpa alat tambahan atau teknik interpretasi yang canggih. Sehingga dibutuhkan penggunaan metode rekayasa fitur seperti *nameXpander* yang penting untuk meningkatkan representasi *username*. Hal ini membantu BERT dalam memahami informasi yang lebih halus dalam *username*, yang dapat mempengaruhi prediksi *gender* secara signifikan. Tanpa adanya modifikasi atau rekayasa fitur tambahan seperti *nameXpander*, BERT mungkin tidak dapat menangkap informasi yang diperlukan untuk memprediksi *gender* dengan akurasi tinggi dari *username* yang kurang jelas atau non-indikatif. BERT sangat bergantung pada seberapa baik teks masukan direpresentasikan, dan tanpa penyesuaian fitur yang tepat, performanya tidak akan memiliki akurasi yang tinggi pada penelitian ini.

e. Evaluasi Model

Untuk mengetahui seberapa baik masing-masing metode memprediksi *gender* pengguna, evaluasi model yang dilakukan dalam penelitian ini mencakup dua model evaluasi, yang pertama metode *stratified sampling* yang merupakan metode pengambilan sampel dengan cara pembagian populasi ke dalam strata, lalu memilih sampel acak setiap stratum, dan menggabungkannya untuk menaksir parameter populasi [12]. Metode ini digunakan untuk memilih sampel yang mewakili setiap kategori (*Hierarchical Classification Heuristic*, BERT, kombinasi *Hierarchical Classification Heuristic* dengan BERT, kombinasi *Hierarchical Classification Heuristic* dengan BERT beserta *nameXpander,* dan yang terakhir kombinasi BERT dengan *nameXpander*) secara proporsional. Sepuluh data diambil dari setiap kategori, sehingga total sampel adalah 50 data. Selanjutnya model evaluasi yang kedua yaitu matrik evaluasi, dalam penelitian ini evaluasi kinerja model prediksi *gender* menggunakan metrik *F1-Score* dan Akurasi. *F1-Score* digunakan karena menggabungkan presisi (*precision*) dan *recall* dalam satu metrik, memberikan gambaran holistik tentang kemampuan model dalam mengklasifikasikan data. Sedangkan Akurasi digunakan untuk mengevaluasi seberapa sering model memberikan prediksi yang benar secara keseluruhan. Kedua metrik ini memberikan *insight* yang penting mengenai efektivitas model dalam tugas prediksi *gender* berdasarkan *username* Twitter.

f. Analisis Hasil

Hasil dari setiap metode dianalisis dan dibandingkan untuk menentukan metode yang paling efektif dalam memprediksi *gender* pengguna. Penelitian ini diharapkan dapat memberikan metode paling efektif untuk memprediksi preferensi tersebut berdasarkan analisis data media sosial.

III. HASIL DAN PEMBAHASAN

a. Metode Klasifikasi

Penelitian ini mengevaluasi lima metode klasifikasi untuk memprediksi *gender* pengguna Twitter berdasarkan nama pengguna dan konten tweet terkait otomotif, kelima metode klasifikasinya yaitu *Hierarchical Classification Heuristic,* BERT, kombinasi *Hierarchical Classification Heuristic* dengan BERT, kombinasi *Hierarchical Classification Heuristic* dengan BERT beserta *nameXpander,* kombinasi BERT dengan *nameXpander.*

b. Evaluasi Model

Berikut adalah hasil dari masing-masing metode klasifikasi:

Tabel 1. Hasil Akurasi dan F1-Score

	Akurasi	***F1-Score***
Hierarchical Classification Heuristic	40%	0,17
BERT	60%	0,38
Kombinasi *Hierarchical Classification Heuristic* dengan BERT	80%	0,65
Kombinasi *Hierarchical Classification Heuristic* dengan BERT beserta *nameXpander*	80%	0,83
Kombinasi BERT dengan *nameXpander*	90%	0,86

c. Analisis Hasil

Model kombinasi antara BERT dengan *nameXpander* memiliki kinerja terbaik yaitu dengan akurasi model sebesar 90% dan *F1-Score* sebesar 0,86 yang berarti bahwa menggabungkan metode dengan teknik rekayasa fitur seperti *nameXpander* dapat membuat peningkatan kinerja model dengan signifikan, *nameXpander* disini membantu BERT dalam memperbaiki pemahaman konteks dengan lebih baik, sehingga dapat meningkatkan kemampuan yang dimiliki oleh model dalam memprediksi *gender* pengguna Twitter. Sedangkan model *Hierarchical Classification Heuristic* memiliki hasil kinerja terendah yang terlihat dari *F1-Score* dan juga akurasinya. Model *Hierarchical Classification Heuristic* memiliki *F1-Score* sebesar 0,17 dan akurasi sebesar 40%, pendekatan ini terlihat tidak mampu secara efektif untuk menangani kompleksitas dan variasi dalam data nama pengguna Twitter untuk memprediksi *gender,* dan model ini juga tampak terlalu sederhana serta kurang memiliki kemampuan untuk menangkap pola-pola yang lebih halus dalam data. Model BERT menunjukkan peningkatan yang cukup signifikan dalam kinerja nya bila dibandingkan dengan model *Hierarchical Classification Heuristic*, dengan *F1-Score* nya memiliki nilai sebesar 0,38 dan akurasinya sebesar 60%, hal ini menunjukkan bahwa model BERT lebih mampu memahami konteks linguistik dalam nama pengguna Twitter, meskipun masih perlu beberapa perbaikan. Ketika menggabungkan metode *Hierarchical Classification Heuristic* dengan metode BERT menghasilkan kinerja model dengan *F1-Score* sebesar 0,65 dan akurasi sebesar 80%, setelah ditambahkan *nameXpander* dalam kombinasi tersebut ternyata memiliki peningkatan hasil pada *F1-Score* nya menjadi sebesar 0,83, tetapi akurasinya tetap berada di

angka 80%. *nameXpander* yang mampu memperluas representasi nama pengguna hingga berbagai suku kata, dapat membantu model dalam menangkap lebih banyak informasi yang relevan untuk prediksi *gender* ternyata peningkatan akurasi tidak terlihat signifikan dalam hal ini.

IV. KESIMPULAN

Berdasarkan penelitian yang sudah dilakukan dapat disimpulkan bahwa model dengan kombinasi metode BERT dan *nameXpander* menunjukkan kinerja terbaik dengan akurasi sebesar 90% dan *F1-Score* sebesar 0,86 dalam memprediksi *gender* pengguna Twitter, yang berarti kombinasi ini tidak hanya meningkatkan kemampuan model dalam memberikan prediksi yang benar secara keseluruhan, tetapi juga dapat memiliki keseimbangan antara presisi dengan *recall* yang penting dalam klasifikasi *gender.*

Berdasarkan peningkatan kinerja model diatas menunjukkan bahwa rekayasa fitur yang tepat sangat penting dalam upaya meningkatkan kemampuan model. Kombinasi metode yang berbeda memungkinkan model untuk mengatasi kelemahan dari masing-masing metode dan memanfaatkan kekuatan yang mereka miliki, sehingga menghasilkan model yang lebih akurat. Berdasarkan pendekatan ini menegaskan juga bahwa pentingnya memilih dan mengkombinasikan metode yang tepat dalam membuat kasus prediksi yang kompleks. Penelitian ini juga memberikan kontribusi penting dalam analisis data media sosial Twitter, khususnya dalam memprediksi *gender* pengguna berdasarkan *username*. Dengan fokus pada industri otomotif di Indonesia, penelitian ini juga dapat memberikan wawasan berharga bagi perusahaan dalam memahami dan merespons preferensi *gender* di pasar otomotif Indonesia.

DAFTAR PUSTAKA

1. S. Alaparthi and and M. Mishra, "Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey," *Anal. Biochem.*, no. 1, pp. 1–15, 2020.
2. S. Yuenyong and S. Sinthupinyo, "Gender Classification of Thai Facebook Usernames," *Int. J. Mach. Learn. Comput.*, vol. 10, no. 5, pp. 618–623, 2020, doi: 10.18178/ijmlc.2020.10.5.982.
3. B. Onikoyi, N. Nnamoko, and I. Korkontzelos, "Gender prediction with descriptive textual data using a Machine Learning approach," *Nat. Lang. Process. J.*, vol. 4, no. February, p. 100018, 2023, doi: 10.1016/j.nlp.2023.100018.
4. A. S. Zakia, Indriati, and Marji, "Klasifikasi Jenis Kelamin Pengguna Twitter dengan menggunakan Metode BM25 dan K-Nearest Neighbor (KNN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 10, pp. 3331–3337, 2020, [Online]. Available: http://j-ptiik.ub.ac.id
5. R. Akbar, "Gender Classification of Indonesian Names Using Multinomial Naive Bayes and Random Forrest Classifiers," no. November, 2016, [Online]. Available: https://www.researchgate.net/publication/308785770
6. P. Aggarwal and R. Mahajan, "Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification," *J. Inf. Syst. Informatics*, vol. 6, no. 2, pp. 607–623, 2024, doi: 10.51519/journalisi.v6i2.692.
7. F. C. Asyuraa, S. Abdullah, and T. E. Sutanto, "Empirical evaluation on discounted Thompson sampling for multi-armed bandit problem with piecewise-stationary Bernoulli arms," *J. Phys. Conf. Ser.*, vol. 1722, no. 1, 2021, doi: 10.1088/1742-6596/1722/1/012096.
8. K. Mohiuddin *et al.*, "Attention Is All You Need,Advances in Neural Information Processing Systems," *Int. Conf. Inf. Knowl. Manag. Proc.*, no. Nips, pp. 4752–4758, 2017, doi: 10.1145/3583780.3615497.
9. C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Min. Knowl. Discov.*, vol. 22, no. 1–2, pp. 31–72, 2011, doi: 10.1007/s10618-010-0175-9.
10. R. Yunanto, E. P. Wibowo, and R. Rianto, "a Bert Model To Detect Provocative Hoax," *J. Eng. Sci. Technol.*, vol. 18, no. 5, pp. 2281–2297, 2023.
11. 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, 2018, [Online]. Available: https://aclanthology.org/N19-1423.pdf
12. S. F. Ulya, Y. Sukestiyarno, and P. Hendikawati, "Random Sampling Confidence Interval," *UNNES J. Math.*, vol. 7, no. 1, pp. 108–119, 2018.