Jurnal TICOM: Technology of Information and Communication
Volume 14, Nomor 2, Januari 2026 Hal: 106-114 E-ISSN: 2962-7982

Deteksi Berita Palsu Dwibahasa dengan Pemrosesan Bahasa Alami

Irfan Ramadhan[1], Eni Heni Hermaliani[2*], Zico Pratama Putra[3]

[1,3] Fakultas Teknologi Informasi, Magister Informatika, Universitas Nusa Mandiri, Jakarta, Indonesia
[2] Fakultas Teknologi Informasi, Sistem Informasi, Universitas Nusa Mandiri, Jakarta, Indonesia
Jl. Margonda Raya No. 545, RT.1/RW.7, Pondok Cina, Kec. Beji, Kota Depok, Jawa Barat 16424
E-mail: [1]14240009@nusamandiri.ac.id [2*]enie_h@nusamandiri.ac.id, [3]zico.zpp@nusamandiri.ac.id
(*: corresponding author)

Abstrak— **Mendeteksi berita palsu merupakan tantangan besar di era digital saat ini, terlebih lagi dalam konteks bilingual seperti Bahasa Indonesia dan Bahasa Inggris. Tujuan dari proyek ini adalah untuk membangun algoritma deteksi berita palsu bilateral yang akurat berdasarkan pemrosesan bahasa alami. Teknik yang digunakan adalah *Naive Bayes*, metode yang tidak rumit dan efektif untuk mengategorikan teks. Pendekatan penelitian meliputi pengumpulan dataset berita dalam bahasa Indonesia dan bahasa Inggris dan pra-pemrosesan teks yang meliputi normalisasi, tokenisasi, dan vektorisasi dengan metode TF-IDF. Teknik validasi silang kemudian digunakan untuk mengeksekusi dan mengevaluasi model *Naive Bayes* untuk menilai akurasi, presisi, dan penarikan kembali klasifikasi. Hasil eksperimen menunjukkan bahwa model ini dapat secara efektif mendeteksi berita palsu dalam kedua bahasa dengan akurasi yang kompetitif sambil menawarkan keuntungan besar yang timbul dari aspek NLP yang meningkatkan kinerja. Dalam penelitian ini, kami menunjukkan bahwa *Naive Bayes* adalah metode yang ringan dan andal untuk aplikasi deteksi berita palsu multibahasa.**

Kata Kunci— **Kecerdasan Buatan, Pemrosesan Bahasa Alami, *Naive Bayes*, Deteksi Dwibahasa Buatan dan Berita Palsu**

Abstract— ***Detecting fake news is a major challenge in today's digital era, especially in bilingual contexts such as Indonesian and English. The goal of this project is to build an accurate bilateral fake news detection algorithm based on Natural Language Processing. The technique used is Naive Bayes, a simple and effective method for categorizing text. The research approach includes collecting news datasets in Indonesian and English and pre-processing the text, including normalization, tokenization, and vectorization with the TF-IDF method. Cross-validation techniques are then used to execute and evaluate the Naive Bayes model to assess the accuracy, precision, and recall of the classification. Experimental results show that this model can effectively detect fake news in both languages with competitive accuracy while offering significant advantages arising from the NLP aspect that improves performance. In this study, we show that Naive Bayes is a lightweight and reliable method for multilingual fake news detection applications.***

Keywords— ***Artificial Intelligence, Natural Language Processing, Naive Bayes, Artificial Bilingual Detection, and Fake News***

I. PENDAHULUAN

Masalah berita palsu telah berkembang pesat dalam beberapa tahun terakhir dengan meledaknya media sosial dan saluran berita jenis tabloid. Mereka yang menyebarkan kepalsuan seperti itu melakukannya dengan kecepatan yang mengkhawatirkan dan merupakan bahaya sosial dan politik yang serius, karena mereka merusak kepercayaan masyarakat. Dalam lingkungan multibahasa, masalah ini diperburuk karena hambatan bahasa, dan juga perbedaan budaya, yang juga merupakan rintangan bagi alat deteksi. Karena perbedaan dalam bahasa-bahasa ini, dari perspektif sintaksis, semantik, dan terkadang idiomatik, pengembangan sistem deteksi universal semacam itu telah menjadi tantangan yang belum pernah terjadi sebelumnya [1][2].

Seiring dengan semakin globalnya dunia dan multikulturalnya, penting untuk mengetahui cara memilih berita palsu dalam berbagai bahasa. Sistem deteksi multibahasa ini sangat penting dalam dunia di mana kita terhubung secara global dan individu pada umumnya mengonsumsi berita dalam banyak bahasa sehingga seseorang dapat yakin akan integritas informasi [3]. Namun demikian, karena tidak ada solusi yang dapat diandalkan yang dapat bekerja dengan andal di seluruh bahasa, peningkatan di bidang ini sangat dibutuhkan [4][5][6].

Sistem deteksi berita palsu yang ada sebagian besar bekerja pada aplikasi monolingual atau spesifik wilayah. Metode yang ada untuk pengambilan dan akses informasi sebagian besar bergantung pada kumpulan data berlabel manual yang besar, yang mahal dan sulit untuk diterapkan secara luas, dan kerangka kerja identifikasi multibahasa relatif jarang. Namun, kesenjangan ini menyoroti perlunya model multibahasa yang fleksibel dan efisien [4][7][8].

Tujuan dan kontribusi: Tujuan dari penelitian ini adalah untuk mengatasi kesulitan-kesulitan ini dengan mengusulkan sistem deteksi berita palsu multibahasa menggunakan pengklasifikasi *Naïve Bayes*. Kontribusi utama meliputi: 1. membangun model yang ringan untuk dapat beradaptasi dengan berbagai bahasa ketika diperlukan untuk mengidentifikasi dalam bahasa Indonesia dan Inggris dengan akurasi tinggi. 2. Pengenalan strategi pra-pemrosesan yang disesuaikan untuk menangani teks multibahasa secara efektif. 3. Banyak pengujian telah dilakukan pada model menggunakan validasi silang untuk mengevaluasi akurasi, presisi, dan ingatannya.

Identifikasi berita palsu kini menjadi subjek penting dalam studi kecerdasan buatan dan pemrosesan bahasa alami. Beberapa teknik telah dikembangkan untuk memecahkan masalah ini dengan memanfaatkan banyak pendekatan yang memanfaatkan model pembelajaran mesin, seperti *Support Vector Machines* (SVM), *Random Forest*, dan algoritma

berbasis jaringan saraf [9]. Sebagian besar sistem ini berfokus pada berita hanya dalam satu bahasa, biasanya bahasa Inggris. Makalah oleh Shu et al. menunjukkan bahwa model berbasis sifat linguistik dan berbasis metadata sangat ampuh dalam bahasa Inggris, tetapi tidak sepenuhnya diperiksa dalam pengaturan multibahasa [10].

Telah ada aplikasi BERT, yang pada dasarnya merupakan teknik berbasis pembelajaran transfer, untuk mengidentifikasi berita palsu dalam berbagai bahasa dalam beberapa situasi. Meskipun model ini memerlukan terlalu banyak komputasi, yang mungkin tidak cocok untuk penggunaan praktis [3]. Meskipun demikian, metode yang lebih sederhana seperti *Naive Bayes* lebih sering cocok untuk mengorganisasi teks ketika Anda menggabungkannya dengan beberapa teknik NLP seperti TF-IDF.

Karena tidak ada standar umum untuk struktur informasi dan daya tarik budaya, teknik deteksi berita palsu multibahasa menghadapi rintangan khusus. Dalam kasus orang Indonesia dan penutur bahasa Inggris, kata-kata yang diucapkan dalam ide yang sama tetapi dalam bahasa yang berbeda dapat mengurangi kemampuan mesin untuk mengidentifikasi konteks dan pola bahasa [11].

Lebih lanjut, penelitian menunjukkan bahwa sumber interpretasi yang salah paling sering terletak pada cacat dalam proses penerjemahan mesin yang cenderung memengaruhi efisiensi sistem deteksi berita palsu [12]. Kelangkaan kumpulan data multibahasa berkualitas tinggi untuk melatih dan menguji algoritma deteksi juga ditangani.

Masalah ini telah dicoba dipecahkan oleh beberapa penelitian dengan cara mengintegrasikan kumpulan data monolingual yang ada [13]. Namun, harmonisasi dan normalisasi data lintas bahasa masih menjadi penghalang. Untuk mengatasi kesenjangan ini, karya ini mengusulkan sistem deteksi berita palsu yang secara khusus terdiri dari data Indonesia dan Inggris, dan berdasarkan metode *Naive Bayes* yang ringan dan dapat beradaptasi lintas bahasa.

II. METODOLOGI PENELITIAN

Tujuan dari penelitian ini adalah untuk membangun sistem otomatis deteksi berita palsu dalam dua bahasa menggunakan metode NLP dan algoritma Naive Bayes.

A. *Arsitektur Sistem*

Sistem deteksi berita palsu yang dirangkum dalam makalah ini dirancang berdasarkan pendekatan metodis yang terstruktur dan sistematis untuk klasifikasi artikel berita dalam bahasa Inggris dan Indonesia. Tahap awal, berita dikumpulkan dari berbagai sumber daring, kemudian dibedakan antara berita asli dan berita palsu, dan akhirnya menjamin cakupan artikel yang inklusif dalam kedua bahasa. Dataset ini sangat berguna untuk melatih model, karena memungkinkan sistem untuk mendefinisikan pola dengan berbagai jenis konten dari berita. Tren terkini analisis *Big Data* memperkenalkan fase pra-pemrosesan setelah pengumpulan data, yang dilakukan di sini: teks diproses dan distandardisasi, semua karakter khusus dihapus, semua kata diubah menjadi huruf kecil, dan semua kata non-standar ditransliterasi. Teknik pemrosesan di atas mengurangi tingkat kebisingan dan mempersiapkan data untuk langkah selanjutnya.

Inti dari sistem ini adalah pengklasifikasi *Naive Bayes*, yang menggunakan data yang telah diproses untuk menentukan apakah artikel tersebut termasuk dalam pola teks tertentu.

Algoritma Deteksi Berita Palsu Dwibahasa:

A. **Input:** Artikel berita (Teks).
B. **Preprocessing:**
   1. Identifikasi Bahasa (Indonesia/Inggris).
   2. *Case Folding*, *Filtering*, dan *Tokenization*.
   3. *Stopword Removal* (menggunakan NLTK untuk Inggris atau Sastrawi untuk Indonesia).
C. **Feature Extraction:** Transformasi teks ke bentuk numerik menggunakan TfidfVectorizer.
D. **Classification:**
   1. Hitung probabilitas artikel masuk ke kelas 'Palsu' vs 'Nyata' menggunakan model MultinomialNB.
   2. Tentukan label kelas berdasarkan probabilitas tertinggi.
E. **Output:** Prediksi label (Asli/Palsu) dan nilai akurasi.

Dalam studi ini, digunakan *Count Vectorizer* untuk mengonversi data teks sehingga pengklasifikasi dapat dengan mudah menemukan fitur unik dalam berita palsu dan berita asli. Terakhir, namun tidak kalah pentingnya, sistem ini menawarkan antarmuka pengguna yang ramah yang menerima artikel berita dan langsung menampilkan hasilnya, apakah berita tersebut asli atau palsu. Pendekatan ini membuat sistem ini efisien waktu dan mudah digunakan, dan mereka yang menggunakan sistem ini di lingkungan multibahasa mengatakan bahwa sistem ini akan menjadi alat yang andal untuk mengetahui kebenaran berita secara *real-time* [14].

B. Teknik NLP Multibahasa

Untuk menangani keberagaman bahasa, teknik NLP yang digunakan meliputi:

Ada beberapa langkah untuk menyiapkan teks guna mendeteksi berita palsu. Tokenisasi merupakan langkah utama dalam model untuk mengelompokkan teks menjadi kata atau frasa baru yang dapat dibedah lebih lanjut. Bergantung pada bahasa yang dipilih, program menggunakan *tokenizer* yang berbeda sehingga memungkinkan untuk menggunakan teks dalam bahasa Indonesia dan Inggris. Kemudian, penghilangan kata henti dalam pemrosesan kata mengecualikan kata-kata yang tidak perlu karena kata tersebut, dan/atau dalam konteks produksi informasi yang bermakna dalam klasifikasi. Kata henti juga bergantung pada bahasa, dan setiap bahasa dikecualikan dari pencarian untuk mempertahankan hanya istilah yang paling penting. Lebih jauh, praproses yang bergantung pada bahasa dilakukan di mana aksara non-Latin diubah menjadi bahasa Latin; dan untuk beberapa bahasa,

terutama bahasa Indonesia, fenomena bahasa tertentu diproses setelahnya, di mana tujuan utamanya adalah untuk mengatur bahasa tersebut sehingga memperoleh hasil yang lebih baik untuk model tersebut. Terakhir, ide penambahan data digunakan untuk meningkatkan ukuran kumpulan data melalui proses yang didasarkan pada produksi *synset* atau *parafrase*. Hal ini juga membantu meningkatkan tingkat variasi kata dalam kalimat, jumlah kalimat, variasi struktural, dan distribusi kata untuk meningkatkan kemampuan generalisasi model dan mengurangi overfitting. Setiap langkah praproses ini secara kolektif membuat data siap untuk klasifikasi yang lebih efektif dan membantu model meningkatkan kemampuan deteksi berita palsu [15][16].

C. *Model Pembelajaran Mesin*

Model terpilih adalah *Naive Bayes* untuk klasifikasi teks karena kesederhanaan dan efisiensi komputasinya. Model tersebut beroperasi dengan memperkirakan probabilitas salah satu kelas (palsu atau nyata) berdasarkan teks yang diberikan. Model divalidasi dengan validasi silang sehingga hasil yang serupa diperoleh dalam data pelatihan dan data tambahan untuk menguji model. Persamaan dasar dari algoritma *Naive Bayes* yang digunakan adalah sebagai berikut:

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)}$$

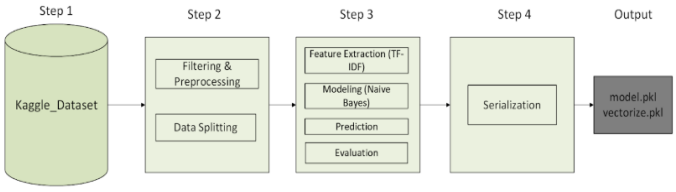
Keterangan:

a. P(C|X): Probabilitas kelas C (Palsu/Nyata) diberikan fitur teks X (Posterior).
b. P(X|C): Probabilitas fitur X muncul pada kelas C (Likelihood).
c. P(C): Probabilitas awal kelas C (Prior).
d. P(X): Probabilitas awal fitur X (*Predictor Prior*).

D. *Pertimbangan Model Multilingual*

Saat mengerjakan kedua bahasa tersebut, penyesuaian berikut dilakukan, di mana algoritma *Naive Bayes* yang sama dilatih untuk data tweet bahasa Indonesia dan Inggris guna membandingkan pola-pola khusus bahasa [17][18]. Arsitektur independen bahasa juga disertakan, di mana fitur-fitur mulai dari n-gram hingga pemodelan bahasa disertakan karena lintas bahasa [19]. Untuk mencegah bias tersebut, distribusi keseluruhan kumpulan data lintas bahasa juga dijaga agar sedapat mungkin merata, atau sedekat mungkin dengannya [20].

Metodologi ini diharapkan tidak hanya menghasilkan model yang akurat, tetapi juga ringan dan dapat diimplementasikan secara praktis di berbagai *platform*.



Gambar 1. Diagram Alir Model dan Vektorisasi pkl

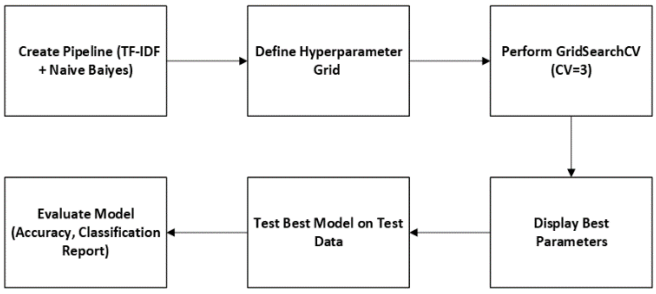
Diagram alir pada Gambar 1. Memetakan metodologi untuk menggunakan model pembelajaran mesin dalam mengidentifikasi detektor berita palsu dan untuk menghasilkan file serial untuk penggunaan lebih lanjut. Proses dimulai dengan akumulasi sekumpulan data, misalnya, *Kaggle Dataset*, yang sebagian besar digunakan sebagai input. Langkah berikutnya adalah penyaringan dan praproses untuk membersihkan data dan tidak memasukkan data yang tidak perlu, kemudian data dibagi menjadi set pelatihan dan pengujian. Bobot kata dihitung menggunakan persamaan TF-IDF:

$$W_{dt} = tf_{dt} \cdot log\left(\frac{N}{dft}\right)$$

Keterangan:

a. Wdt: Bobot kata t dalam dokumen d.
b. tfdt: Frekuensi kemunculan kata t dalam dokumen d.
c. N: Total seluruh dokumen dalam dataset.
d. dft: Jumlah dokumen yang mengandung kata t.

Kedua, data teks diubah menjadi vektor numerik dengan menerapkan metodologi ekstraksi fitur TF-IDF. Akhirnya, data yang diproses yang merupakan fitur yang diekstraksi dari video dimasukkan ke dalam model *Naive Bayes*. Model diuji untuk menentukan kinerjanya dalam konteks aplikasinya dan kemampuannya untuk membuat prediksi yang wajar secara efektif. Setelah itu, proses pelatihan model diikuti oleh fase serialisasi dalam sistem. Di sini hanya model yang dilatih dan vektorisator yang mengubah teks menjadi nilai numerik yang disimpan dalam format .pkl, yaitu model.pkl dan *vectorize*.pkl. Berkas-berkas ini memungkinkan seseorang untuk mengimplementasikan model dalam penggunaan praktis dan tepat waktu dengan jaminan untuk mengukur model realitas yang benar dan tepat.



Gambar 2. Pelatihan dan Pengujian Model Diagram Alur

Diagram alir pada Gambar 2 menggambarkan panduan untuk melatih dan menguji model pembelajaran mesin yang mencontohkan proses pelatihan dan pengujian, terutama untuk pengoptimalan akurasi keluaran. Proses ini dimulai dengan konstruksi pipa yang mencakup TF-IDF untuk mengonversi fitur teks dan *Naive Bayes* sebagai pengklasifikasi untuk praproses dan pelatihan teks. Selanjutnya, kisi hiperparameter karakteristik dibuat untuk merinci nilai-nilai untuk menyesuaikan model *Naive Bayes*. Mengikuti *GridSearchCV* dan validasi silang 3 kali lipat, model dinilai dengan semua

kemungkinan kombinasi parameter ini. Ini memungkinkan penentuan kondisi khusus yang memungkinkan produsen untuk mencapai hasil terbaik. Setelah itu, pencarian kisi hiperparameter selesai, yang terbaik ditunjukkan untuk memberikan hiperparameter optimal untuk model. Dari model yang dioptimalkan tersebut, kami menguji kinerjanya pada set data uji lain yang tidak terlihat untuk menentukan kemampuan generalisasinya. Terakhir, kinerja model disimpulkan dalam hal akurasi dan laporan klasifikasi untuk memberikan evaluasi yang lebih deskriptif tentang kelayakannya. Langkah demi langkah ini membangun model pembelajaran mesin yang kuat, positif, dan efisien yang secara tepat disesuaikan dengan tugasnya.

Guna mengimplementasikan alur logika yang telah dirancang, proses pembangunan model dilakukan dengan mengintegrasikan ekstraksi fitur dan algoritma klasifikasi ke dalam satu kesatuan sistem. Implementasi teknis dari tahapan *pipeline* menggunakan bahasa pemrograman *Python* adalah sebagai berikut:

```
# Implementasi Pipeline Naive Bayes Dwibahasa
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import
TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB

# Inisialisasi pipeline untuk ekstraksi fitur dan klasifikasi
pipeline = Pipeline([
   ('tfidf', TfidfVectorizer()), # Mengonversi teks ke vektor
numerik
   ('clf', MultinomialNB())       # Algoritma Naive Bayes
Classifier
])

# Melatih model dengan data latih
pipeline.fit(X_train, y_train)

# Melakukan prediksi pada data uji
y_pred = pipeline.predict(X_test)
```

Untuk membangun sistem deteksi berita palsu multibahasa yang akurat, tahap pengumpulan dan penyiapan data memegang peranan yang sangat penting. Tahap ini meliputi identifikasi sumber data, penambahan data untuk meningkatkan keragaman bahasa, dan praproses data untuk memastikan data siap digunakan dalam pelatihan model.

Untuk memperkuat basis data dalam penelitian ini, khusunya pada klasifikasi teks dwibahasa dalam Bahasa Inggris dan Bahasa Indonesia, diterapkan Teknik augmentasi data melalui strategi penerjemahan silang. Dengan alat seperti berita palsu dalam bahasa Inggris atau bahasa Indonesia yang diterjemahkan dengan bantuan *Google*, peningkatan tersebut dapat dilakukan. Strategi ini tidak hanya memperkaya kumpulan data, tetapi juga membantu model mengenali varian bahasa yang berasal dari proses penerjemahan yang biasanya terjadi di lingkungan multibahasa. Selain itu, augmentasi ini menyediakan sinonim, varian frasa, dan frasa linguistik unik yang meningkatkan ketahanan model terhadap varian dalam pola bahasa.

Prapemrosesan data merupakan langkah penting untuk memastikan kualitas dan konsistensi kumpulan data sebelum digunakan dalam pelatihan model. Langkah-langkah utama meliputi:

Pemrosesan data untuk model deteksi berita palsu melibatkan serangkaian proses rutin untuk membersihkan kumpulan data atau membuatnya dalam format yang tepat yang dapat digunakan untuk analisis. Langkah pertama terdiri dari identifikasi bahasa yang melibatkan penggunaan alat (termasuk *langdetect*) untuk memeriksa apakah dokumen terkait berbahasa Inggris atau Indonesia. Dokumen resmi yang bukan berbahasa Inggris tidak disertakan dalam dataset untuk menjaga cakupan penelitian tetap sesuai. Opini dan analisis sentimen yang akan diterapkan pada teks biasanya muncul berikutnya setelah tokenisasi memecah teks menjadi bentuk terstruktur dari kata-kata atau frasa individual menggunakan alat khusus bahasa. Untuk bagian bahasa Inggris, Tokenisasi dilakukan dengan menggunakan *tokenizer* NLTK untuk teks berbahasa Inggris, sedangkan untuk teks berbahasa Indonesia digunakan tokenizer dari pustaka Sastrawi agar sesuai dengan karakteristik linguistik masing-masing bahasa. Selanjutnya, penghapusan stopword diterapkan untuk menghilangkan kata-kata umum yang tidak memberikan kontribusi signifikan terhadap makna, seperti “dan”, “the” pada bahasa Inggris, serta “dan”, “yang” pada bahasa Indonesia. Daftar *stopword* disesuaikan secara spesifik per bahasa guna mengurangi noise dan meningkatkan bobot pada istilah-istilah yang lebih relevan. Tahap normalisasi meliputi konversi seluruh teks ke huruf kecil, penghapusan tanda baca, serta penanganan bentuk nonstandar dan bahasa gaul (contoh: “4nda” diubah menjadi “anda”). Langkah ini memastikan konsistensi data di seluruh dataset, khususnya dalam mengurangi variasi informal yang umum pada teks berbahasa Indonesia. Akhirnya, teks yang telah diproses diubah menjadi representasi numerik menggunakan metode TF-IDF atau *Count Vectorizer*. Transformasi ini mempersiapkan data agar siap digunakan dalam pelatihan model *Naive Bayes* dan proses klasifikasi.

Karena langkah-langkah praproses ini membentuk tulang punggung model yang baik, penting untuk memastikan bahwa data yang digunakan untuk pemodelan cukup baik untuk mendukung tujuan deteksi berita palsu. Langkah-langkah ini ditujukan terutama untuk memastikan penyediaan kualitas data dalam konteks multibahasa. Dengan menggunakan segmentasi berbasis bahasa, model dapat membagi data menjadi data yang bergantung pada bahasa yang lebih terarah untuk diproses. Prosedur untuk tokenisasi dan penghapusan *stopword* disesuaikan dengan perbedaan linguistik struktural antara bahasa Inggris dan bahasa Indonesia. Augmentasi berbasis terjemahan meningkatkan kemampuan penanganan keragaman

teks pada model, sehingga sistem tidak terlalu sensitif terhadap perbedaan pola teks. Dengan demikian, dalam skenario ini, data pelatihan yang digunakan dalam pelatihan adalah variasi bahasa yang seragam dan realistis, yang merupakan tantangan utama dalam pendeteksian berita palsu multibahasa.

Langkah awal yang menjelaskan cara melatih model berfokus pada stratifikasi kumpulan data menjadi dua bagian: kumpulan pelatihan dan pengujian. Kumpulan pelatihan terdiri dari data yang akan digunakan untuk membangun model, dan kumpulan data pengujian adalah kumpulan data yang akan diterapkan pada model setelah pelatihan dilakukan. Pemisahan ini dicapai melalui penggunaan teknik pemisahan data, misalnya, 80 persen data dapat diterapkan pada model, sementara 20 persen dapat dicadangkan untuk pengujian atau, dalam hal yang lebih kompleks, metode validasi silang seperti teknik validasi silang *K-fold*, di mana data dibagi menjadi sejumlah subset untuk memastikan bahwa model diuji dengan berbagai kombinasi data pelatihan dan pengujian. Kemudian, tahap berikutnya, hiperparameter model dioptimalkan dengan tujuan untuk meningkatkan kinerjanya. Dalam model *Naive Bayes*, parameter spesifik yang dianggap sebagai hiperparameter meliputi distribusi yang digunakan untuk model *Naive Bayes* (Gaussian atau Multinomial) serta parameter penghalusan yang mengontrol kehalusan distribusi probabilitas. Hal ini dilakukan dengan menggunakan teknik pencarian grid atau pencarian acak untuk mendapatkan parameter optimal.

Metrik evaluasi utama yang digunakan untuk mengukur kinerja model dalam pendeteksian berita palsu multibahasa adalah:

Pengukuran yang digunakan dalam menilai model pendeteksian berita palsu mencakup pengukuran berikut.

1. **Akurasi (*Accuracy*)**:

Digunakan untuk menilai kemampuan model dalam mengategorikan berita secara benar secara keseluruhan.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **Presisi (*Precision*)**:

Menghitung ketepatan prediksi positif untuk memastikan model jarang salah mengira berita nyata sebagai palsu.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Presisi = \frac{TP}{TP + FP}$$

3. ***Recall***:

Menghitung kapasitas model untuk mengidentifikasi seluruh berita palsu yang ada dalam kumpulan data.

$$Recall = \frac{TP}{TP + FN}$$

4. ***F1-Score***:

Merupakan skor gabungan yang menyeimbangkan nilai presisi dan *recall*.

$$F1 - Score = 2 \cdot \frac{Presisi \cdot Recall}{Presisi + Recall}$$

Keterangan untuk naskah Anda:

a. **TP (*True Positive*):** Berita palsu yang terdeteksi benar sebagai palsu.
b. **TN (*True Negative*):** Berita asli yang terdeteksi benar sebagai asli.
c. **FP (*False Positive*):** Berita asli yang salah terdeteksi sebagai palsu.
d. **FN (*False Negative*):** Berita palsu yang salah terdeteksi sebagai asli

Akurasi menilai kemampuan model dalam mengategorikan berita dengan benar, yang diberikan oleh total data uji dan berguna dalam kumpulan data yang mencakup jumlah berita palsu dan nyata yang sama. Skor F1 menawarkan skor gabungan yang jauh lebih dapat digunakan dari tampilan presisi dan recall. Hal ini terutama penting selama pendeteksian berita palsu; berita nyata dapat digolongkan sebagai berita palsu, atau berita palsu tidak dapat digolongkan sebagai berita palsu sama sekali. Presisi, menganggap prediksi positif sebagai tepat, dan nilai tepat yang besar berarti model jarang salah mengira berita nyata sebagai berita palsu. Di sisi lain, recall menghitung kapasitas model untuk mengidentifikasi semua prediksi positif dari total data positif yang diuji model. *Recall* yang tinggi menunjukkan efisiensi model dalam mendeteksi semua berita palsu dalam kumpulan data. Terakhir, AUC (*Area Under the Curve*) bekerja berdasarkan hipotesis untuk menentukan efektivitas model dalam memisahkan kelas (berita palsu dan nyata) dengan ambang batas diferensiasi yang beragam. Skor tinggi dalam AUC berarti model tersebut memiliki kemampuan untuk menentukan perbedaan antara kedua kategori dan memang dapat berguna di area yang sensitif. Evaluasi model menggunakan parameter ini memberikan pandangan holistik tentang kinerja model dengan menawarkan hasil yang akurat serta menjaga sensitivitas dan spesifisitas yang sangat penting dalam aplikasi seperti deteksi berita palsu.

III. HASIL DAN PEMBAHASAN

A. Deskripsi Dataset

Kumpulan data utama yang digunakan dalam penelitian ini bersumber dari situs *web Kaggle*, yang menawarkan kumpulan data berita palsu berbahasa Inggris. Kumpulan data ini memiliki ratusan berita yang dikategorikan sebagai "palsu" atau "asli". Untuk memenuhi persyaratan multibahasa, kumpulan data tambahan untuk berita Indonesia diperoleh dari sumber-sumber lokal, yang meliputi situs web berita dan platform terkemuka yang didedikasikan untuk investigasi berita palsu.

1) Kumpulan Data Bahasa Inggris: Kumpulan data *WELFake* yang berisi data elaboratif yang dikumpulkan dari berbagai sumber terkemuka seperti *Kaggle, McIntire, Reuters,* dan

BuzzFeed. Ada sekitar 72.000 item dalam kumpulan data ini yang diidentifikasi sebagai berita benar atau salah. Kumpulan data ini menyediakan titik referensi untuk solusi berbasis ML untuk mendeteksi berita palsu yang kuat karena seimbang dengan volume artikel berita asli dan palsu, dengan kedua kategori tersebar di seluruh kumpulan data dalam proporsi yang kurang lebih sama.

2) Kumpulan Data Bahasa Indonesia: Kumpulan data bahasa Indonesia yang disebut sebagai "Kumpulan Data Berita Hoax 2023" diperoleh melalui Kaggle. Kumpulan data ini ditujukan untuk item berita bahasa Indonesia dan diberi tag untuk mengidentifikasi berita asli dan berita palsu. Kumpulan data ini dikompilasi dan dirancang dengan tujuan untuk mewujudkan bahasa dan konteks budaya setempat sehingga dapat digunakan untuk mengembangkan model untuk mendeteksi misinformasi dalam bahasa Indonesia.

Kumpulan data gabungan ini dipilih untuk memastikan sistem dapat bekerja secara efektif pada dua bahasa yang berbeda. Dengan mengintegrasikan kumpulan data *WELFake* untuk bahasa Inggris dengan kumpulan data *Kaggle* untuk bahasa Indonesia, pekerjaan ini memberikan dasar yang baik untuk identifikasi berita palsu multibahasa. Proses praproses menjamin bahwa kehalusan bahasa dan kontekstual terekam dengan memadai, yang mendukung kategorisasi yang benar. Selain itu, penggabungan augmentasi berbasis terjemahan meningkatkan kapasitas sistem untuk berfungsi dalam situasi multibahasa

	Unnamed: 0	title	text	label
0	0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	No comment is expected from Barack Obama Membe...	1
1	1	NaN	Did they post their votes for Hillary already?	1
2	2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	Now, most of the demonstrators gathered last ...	1
3	3	Bobby Jindal, raised Hindu, uses story of Chri...	A dozen politically active pastors came here f...	0
4	4	SATAN 2: Russia unvelis an image of its terrif...	The RS-28 Sarmat missile, dubbed Satan 2, will...	1

Gambar 3. Tampilan dataset WELFake sebelum tokenisasi dan stopword bahasa Inggris

Berdasarkan Gambar 3, tampilan dataset menunjukkan struktur awal sebelum pra-pemrosesan.

	content	label
0	law enforcement on high alert following threat...	1
2	unbelievable obama's attorney general says mos...	1
3	bobby jindal raised hindu uses story of christ...	0
4	satan russia unvelis an image of its terrifyi...	1
5	about time christian group sues amazon and spl...	1

Gambar 4. Tampilan dataset WELFake setelah tokenisasi dan penghapusan kata-kata penghenti dalam bahasa Inggris

Berdasarkan Gambar 4, tampilan dataset setelah pra-pemrosesan menunjukkan teks yang lebih bersih. Tabel Kumpulan data yang disajikan ditujukan untuk melatih dan menguji model bahasa Inggris, menghitung distribusi umum data. Ukuran sampel data keseluruhan dari kumpulan data adalah 14308; dari jumlah tersebut 7081 termasuk dalam kategori "BENAR" (berita) dan 7227 termasuk dalam kategori "SALAH" (menyesatkan). Distribusi ini bermanfaat untuk melatih model yang memungkinkan pembedaan berita palsu dan nyata dalam bahasa Inggris, karena menargetkan kedua kategori tersebut. Tabel tersebut juga memberikan ukuran penting dari kecenderungan sentral dan penyebaran, yang terbukti penting dalam mengevaluasi kesesuaian kumpulan data yang digunakan untuk mendefinisikan dan melatih model deteksi setiap kelas.

Struktur kumpulan data seperti itu penting untuk tujuan kalibrasi model yang akurat karena jika tidak, angka-angka seperti akurasi dan *F1_score* dapat menjadi tidak akurat karena perbedaan besar dalam volume data di kelas.

	Judul	Konten	Unnamed: 2	Label
0	Jokowi: Terlalu banyak peraturan kita pusing s...	Jokowi: Terlalu banyak peraturan kita pusing s...	NaN	1
1	Elon Musk Menghapus Akun Twitter Greta Thunber...	Ini adalah salah satu hal terbaik dan luar bia...	NaN	1
2	Jenazah Ferdy Sambo Dipulangkan ke Jakarta Usa...	Pecah Tangis Keluarga dan Ibu PC saat Jenazah ...	NaN	1
3	Ayah Brigadir J Ditahan Usai Hina Kapolri	GegerAyah Brigadir J Ditahan !? Usai Hina Kapo...	NaN	1
4	Gelombang Tinggi Lenyapkan Kota di Indonesia	LENYAP SEKETIKA… INDONESIA HARI INI MENANGIS G...	NaN	1

Gambar 5. Tampilan Dataset Berita Palsu Indonesia Sebelum Tokenisasi dan Stopword Diterapkan

Berdasarkan Gambar 5, tampilan dataset menunjukkan struktur awal.

	content	Label
0	jokowi peraturan pusing stres jokowi peraturan...	1
1	elon musk menghapus akun twitter greta thunber...	1
2	jenazah ferdy sambo dipulangkan jakarta jalani...	1
3	ayah brigadir j ditahan hina kapolri gegerayah...	1
4	gelombang lenyapkan kota indonesia lenyap seke...	1

Gambar 6. Tampilan Dataset Berita Palsu Indonesia Setelah Tokenisasi dan Penghapusan *Stopword*

Berdasarkan Gambar 6, tampilan dataset setelah pra-pemrosesan menunjukkan perbaikan signifikan dalam bentuk teks yang lebih bersih dan terstandarisasi. Tabel distribusi dataset bahasa Indonesia yang digunakan untuk pelatihan dan pengujian model menunjukkan bahwa kategori **BENAR** (berita asli) berjumlah 305 sampel, sedangkan kategori **SALAH** (berita palsu) berjumlah 425 sampel, dengan total keseluruhan 730 catatan. Distribusi ini relatif tidak seimbang (*imbalanced*), di mana jumlah entri berita palsu lebih banyak daripada berita asli. Ketidakseimbangan ini perlu ditangani selama proses pelatihan model, misalnya melalui teknik oversampling, undersampling, atau *class weighting*, guna menghindari bias dan memastikan model dapat menggeneralisasi kedua kelas secara merata.

B. Hasil Evaluasi

Durasi Cross-Validation: 0.64 detik

Laporan Klasifikasi:

	precision	recall	f1-score	support
0.0	0.89	0.96	0.92	305
1.0	0.97	0.92	0.94	425
accuracy			0.93	730
macro avg	0.93	0.94	0.93	730
weighted avg	0.94	0.93	0.93	730

Gambar 7. Hasil Evaluasi Model Naive Bayes Indonesia

Berdasarkan Gambar 7, hasil evaluasi menunjukkan metrik kinerja untuk model Indonesia.

Akurasi Model: 0.8605675146771037

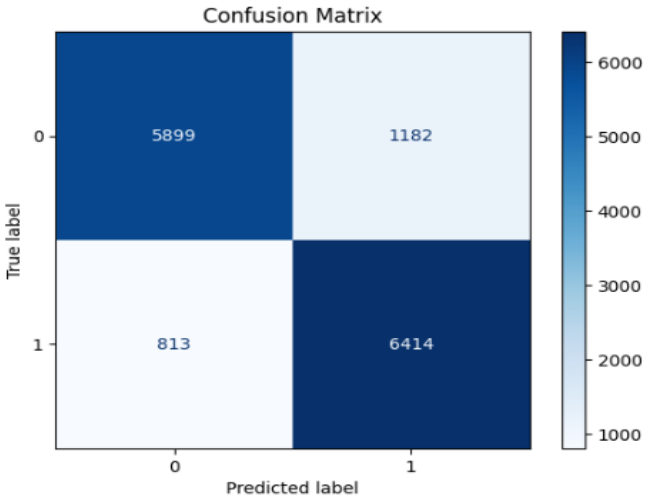
Laporan Klasifikasi:

	precision	recall	f1-score	support
0	0.88	0.83	0.86	7081
1	0.84	0.89	0.87	7227
accuracy			0.86	14308
macro avg	0.86	0.86	0.86	14308
weighted avg	0.86	0.86	0.86	14308

Gambar 8. Hasil Evaluasi Model Naive Bayes Dalam Bahasa Inggris

Berdasarkan Gambar 8, hasil evaluasi menunjukkan metrik kinerja untuk model Inggris.

Matriks kebingungan untuk model Bahasa Inggris ditunjukkan di bawah ini:

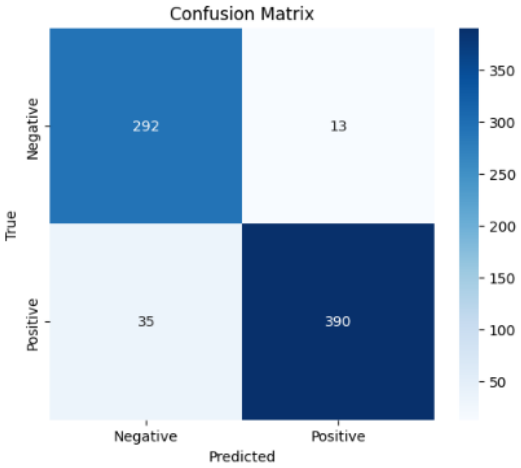


Gambar 9. Matriks Kebingungan

Berdasarkan Gambar 9, matriks kebingungan menunjukkan nilai TP dan TN tinggi, dengan FP dan FN rendah, mengonfirmasi kemampuan klasifikasi yang baik.
Dari matriks kebingungan, kita dapat melihat bahwa model tersebut berkinerja baik, dengan nilai positif benar (TP) dan negatif benar (TN) yang tinggi, dan jumlah positif salah (FP) dan negatif salah (FN) yang relatif rendah. Hal ini menegaskan kemampuan model untuk mengklasifikasikan kasus positif dan negatif dengan benar.

Matriks kebingungan untuk model Indonesia ditunjukkan di pada Gambar 10 berikut:

Confusion Matrix:
[[292 13]
[35 390]]



Gambar 10. Matriks Kebingungan

Berdasarkan Gambar 10, matriks kebingungan menunjukkan performa serupa dengan model Inggris, dengan TP dan TN relatif tinggi. Mirip dengan model Inggris, matriks kebingungan memperlihatkan bahwa model Indonesia memiliki jumlah positif benar dan negatif benar yang relatif tinggi, dengan jumlah positif salah dan negatif salah yang dapat dikelola.

Seperti dalam studi bahasa Indonesia dan Inggris ini, deteksi berita palsu multibahasa lintas bahasa menimbulkan tantangan karena akurasi dalam bahasa rentan terhadap perbedaan struktural, tata bahasa, dan kosa kata di antara bahasa-bahasa tersebut. Meskipun model *Naive Bayes* mampu bekerja dengan memuaskan di kedua dialek. Namun, mengingat konteks lokal dan bahasa gaul informal, keandalan untuk klasifikasi yang efektif terganggu. Aksesibilitas data juga menjadi masalah, khususnya bahasa Indonesia yang memiliki kumpulan data yang relatif sedikit dibandingkan dengan kumpulan data bahasa Inggris, yang mengakibatkan model yang kurang terlatih untuk menangani kefasihan bahasa yang bervariasi secara efisien. Juga tidak terkait dengan bahasa Indonesia, augmentasi data berbasis terjemahan otomatis sering kali tidak akurat karena mengorbankan kualitas kumpulan data. Area yang juga memerlukan penjelasan dalam teknik khusus ini adalah kenyataan bahwa pelatihan model multibahasa

membutuhkan lebih banyak daya komputasi karena terdiri dari 2 bahasa tidak perlu penjelasan karena jelas lebih banyak daya dan memori yang dibutuhkan. Meskipun *Naive Bayes* cukup efisien, menjalankan dua bahasa secara bersamaan memerlukan lebih banyak pengoptimalan untuk mengurangi waktu pelatihan yang diperlukan dalam model dan meningkatkan efisiensi model. Terlepas dari kekurangan yang terkait dengan data dan komputasi di atas, metode ini telah terbukti berhasil dalam mengidentifikasi berita palsu dalam beberapa bahasa.

Model *Naive Bayes* yang diterapkan pada bahasa Indonesia dan Inggris menunjukkan kinerja yang luar biasa, dengan akurasi sekitar 86% dalam kedua bahasa. Meskipun model tersebut secara signifikan mengungguli bahasa Inggris karena kumpulan data dan referensi linguistiknya yang lebih luas, tidak adanya data untuk bahasa Indonesia menghambat kinerja model. *Naive Bayes* juga memiliki keunggulan dalam hal kecepatan dan efisiensi komputasi, meskipun terkadang kurang tepat dalam menangani nuansa linguistik yang lebih kompleks dibandingkan dengan model berbasis pembelajaran mendalam.

Secara keseluruhan, meskipun ada tantangan terkait ketidakakuratan terjemahan dan data terbatas untuk bahasa Indonesia, model *Naive Bayes* terbukti efektif dan efisien dalam pendeteksian berita palsu multibahasa. Meningkatkan jumlah data dan mengembangkan teknik NLP untuk bahasa dengan sumber daya rendah dapat meningkatkan kinerja model dalam aplikasi dunia nyata.

Model *Naive Bayes* menunjukkan hasil yang sangat baik dalam mendeteksi berita palsu dalam dua bahasa, yaitu bahasa Indonesia dan bahasa Inggris. Berdasarkan evaluasi menggunakan metrik akurasi, presisi, recall, dan skor F1, model ini mencatat akurasi rata-rata 86,06% untuk kedua bahasa. Untuk bahasa Indonesia, skor F1 mencapai 0,86, sedangkan untuk bahasa Inggris sedikit lebih tinggi yaitu 0,87, menunjukkan kinerja yang konsisten di kedua bahasa. Hal ini menunjukkan efisiensi sistem dalam lingkungan multibahasa, meskipun terdapat perbedaan dalam struktur bahasa dan kosakata antara bahasa Indonesia dan bahasa Inggris.

Dalam pengujian model ini, ditemukan beberapa ketidakkonsistenan dalam penerjemahan, khususnya dengan pemanfaatan teknik augmentasi data berbasis penerjemahan, antara bahasa Indonesia dan bahasa Inggris. Meskipun memperluas kumpulan data, penerjemahan mesin dapat gagal dan mungkin mengandung masalah seperti pemilihan kata yang salah atau perubahan makna yang dangkal yang akan berdampak negatif pada kemampuan model dalam mendeteksi berita palsu. Jadi, untuk kasus bahasa Indonesia, ada juga masalah karena bahasa tersebut merupakan bahasa dengan sumber daya yang rendah. Model tersebut dapat mengalami kesulitan dengan bahasa sehari-hari setempat dan konstruksi yang tidak diketahui dalam kumpulan data yang luas, yang pada gilirannya dapat berdampak buruk pada kemampuan untuk menentukan terjadinya berita palsu dalam lingkungan budaya tersebut.

Antarmuka pengguna (UI) yang dibuat untuk tujuan penggunaan aplikasi ini untuk mendeteksi berita palsu sangat membantu karena melibatkan pengguna secara efektif. Hasil klasifikasi berita dalam kedua bahasa ini ditampilkan dengan mudah kepada pengguna karena tata letak antarmuka yang jelas yang cukup sederhana dan lancar. Menanggapi ulasan pengguna, mereka yang bekerja dengan antarmuka tersebut tidak menghadapi kesulitan apa pun karena adanya berbagai informasi tentang jenis konten yang diklasifikasikan dan apa metriknya – berita palsu atau nyata. Hasil pencarian untuk konten yang diberi tag juga dapat dilacak, karena detail relevan tambahan mengenai sistem presisi memang merupakan faktor ekspektasi yang dapat diandalkan selamanya pada saat menggunakan aplikasi ini dari hari ke hari. Singkatnya, meskipun ada masalah penerjemahan dan bahasa yang minim sumber daya, sistem deteksi berita palsu berbasis *Naive Bayes* multibahasa telah berhasil dan dapat diandalkan, dengan bahasa yang baik dan UI yang ramah interaksi.

IV. PENUTUP

Dalam penelitian ini, sistem deteksi berita palsu dwibahasa (Bahasa Indonesia dan Inggris) berbasis algoritma *Naive Bayes* dan pemrosesan bahasa alami multibahasa telah berhasil dikembangkan. Hasil eksperimen menunjukkan performa yang kompetitif dengan akurasi rata-rata sekitar 86% serta *F1-score* 0,86–0,87 pada kedua bahasa. Kontribusi signifikan terhadap peningkatan presisi dan recall diberikan oleh strategi pra-pemrosesan yang disesuaikan per bahasa, khususnya tokenisasi dan eliminasi stopword. Sistem ini juga unggul dalam hal efisiensi komputasi karena mengonsumsi daya yang lebih rendah dibandingkan pendekatan berbasis deep learning yang lebih kompleks, sehingga mudah diimplementasikan pada perangkat dengan sumber daya terbatas. Meskipun demikian, tantangan utama yang masih ada meliputi ketidakseimbangan dataset (khususnya pada bahasa Indonesia) serta keterbatasan dalam menangani variasi bahasa informal dan konteks budaya lokal. Untuk pengembangan lebih lanjut, disarankan agar penelitian selanjutnya: (1) memperluas dan menyeimbangkan dataset multibahasa dengan sumber data yang lebih representatif, (2) mengintegrasikan teknik penanganan imbalance seperti SMOTE atau *class weighting*, serta (3) mengeksplorasi kombinasi dengan model multilingual modern (seperti mBERT atau XLM-R) untuk meningkatkan ketahanan terhadap nuansa bahasa dan konteks real-time.

REFERENSI

[1] N. Ahuja and S. Kumar, "Mul-FaD: attention based detection of multiLingual fake news," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 3, pp. 2481–2491, 2023, doi: 10.1007/s12652-022-04499-0.

[2] K. Shu, S. Wang, and H. Liu, "Understanding User Profiles on Social Media for Fake News Detection," *Proc. - IEEE 1st Conf. Multimed. Inf. Process. Retrieval, MIPR 2018*, pp. 430–435, 2018, doi: 10.1109/MIPR.2018.00092.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist.*

Hum. Lang. Technol. - Proc. Conf., vol. 1, no. Mlm, pp. 4171–4186, 2019.
[4] A. Mishra and H. Sadia, "A Comprehensive Analysis of Fake News Detection Models: A Systematic Literature Review and Current Challenges †," *Eng. Proc.*, vol. 59, no. 1, 2023, doi: 10.3390/engproc2023059028.
[5] E. Papageorgiou, C. Chronis, I. Varlamis, and Y. Himeur, "A Survey on the Use of Large Language Models (LLMs) in Fake News," *Futur. Internet*, vol. 16, no. 8, pp. 1–29, 2024, doi: 10.3390/fi16080298.
[6] X. Zhou, J. Wu, and R. Zafarani, "SAFE : Similarity-Aware Multi-Modal Fake News Detection," no. 1, pp. 1–12.
[7] K. Kowsari, K. J. Meimandi, M. Heidarysafa, and S. Mendu, "Text Classification Algorithms : A Survey," pp. 1–68, 2019, doi: 10.3390/info10040150.
[8] J. Su, C. Cardie, and P. Nakov, "Adapting Fake News Detection to the Era of Large Language Models," *Find. Assoc. Comput. Linguist. NAACL 2024 - Find.*, pp. 1473–1490, 2024, doi: 10.18653/v1/2024.findings-naacl.95.
[9] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, 2017, doi: 10.1257/jep.31.2.211.
[10] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," no. i, 2017, [Online]. Available: http://arxiv.org/abs/1708.01967
[11] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online. Science, 359(6380), 1146–1151 | 10.1126/science.aap9559," vol. 1151, no. March, pp. 1146–1151, 2018, [Online]. Available: 10.1126/science.aap9559
[12] A. Conneau *et al.*, "XNLI: Evaluating cross-lingual sentence representations," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 2475–2485, 2018, doi: 10.18653/v1/d18-1269.
[13] S. Ruder, M. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing tutorial," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Tutor. Abstr.*, no. 2010, pp. 15–18, 2019.
[14] F. K. Alarfaj, "Deep Dive into Fake News Detection : Feature-Centric Classification with Ensemble and Deep Learning Methods," 2023.
[15] D. Mimno, "Pulling Out the Stops : Rethinking Stopword Removal for Topic Models," vol. 2, pp. 432–436, 2017.
[16] A. S. M, L. Thompson, and D. Mimno, "Understanding Text Pre-Processing for Latent Dirichlet Allocation," 2017.
[17] C. Padurariu and M. Elena, "ScienceDirect Dealing with Data Imbalance in Text Classification Dealing with Data Imbalance in Text Classification," *Procedia Comput. Sci.*, vol. 159, pp. 736–745, 2019, doi: 10.1016/j.procs.2019.09.229.
[18] M. Zhang, J. M. Peña, and V. Robles, "Feature selection for multi-label naive Bayes classification," *Inf. Sci. (Ny).*, vol. 179, no. 19, pp. 3218–3229, 2009, doi: 10.1016/j.ins.2009.06.010.
[19] N. L. Models, "N-gram Language Models," 2024.
[20] C. Zhao, M. Wu, X. Yang, and W. Zhang, *A Systematic Review of Cross-Lingual Sentiment Analysis : Tasks , Strategies , and Prospects*, vol. 56, no. 7. 2024. doi: 10.1145/3645106.