JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan

Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 284-289

Klasifikasi Bahasa Menggunakan FastText

Azacky Habibilah Syahlan[1], Ahmad Wildan Hisbullah[2], Muhammad Wildan[3]

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: sazackyhabibilah@gmail.com

Abstrak−Perkembangan teknologi pemrosesan bahasa alami (Natural Language Processing/NLP) mendorong meningkatnya kebutuhan akan sistem yang mampu mengidentifikasi bahasa teks secara otomatis. Klasifikasi bahasa menjadi tahap awal yang penting sebelum dilakukan proses lanjutan, seperti penerjemahan otomatis, analisis sentimen, dan klasifikasi dokumen. Penelitian ini bertujuan untuk membangun dan menguji sistem klasifikasi bahasa teks menggunakan algoritma FastText dengan pendekatan supervised learning. Dataset yang digunakan berupa teks multibahasa yang telah diberi label, dengan fokus pada Bahasa Indonesia dan Bahasa Inggris. Tahapan penelitian meliputi praproses data, pelatihan model FastText, pengujian model, serta evaluasi kinerja menggunakan metrik akurasi. Hasil pengujian menunjukkan bahwa model FastText mampu melakukan klasifikasi bahasa teks dengan tingkat akurasi sebesar 50% pada data uji. Meskipun nilai akurasi masih tergolong rendah, hasil penelitian menunjukkan bahwa FastText dapat diimplementasikan secara efektif untuk tugas klasifikasi bahasa. Keterbatasan performa model terutama dipengaruhi oleh jumlah data latih yang terbatas dan variasi data yang kurang beragam. Penelitian ini diharapkan dapat menjadi dasar pengembangan sistem identifikasi bahasa berbasis teks dengan performa yang lebih optimal di masa mendatang.

Kata Kunci: Klasifikasi Bahasa, FastText, Pemrosesan Bahasa Alami, Pembelajaran Mesin, Teks Multibahasa

Abstract−*The rapid development of Natural Language Processing (NLP) technology has increased the need for automated systems capable of identifying the language of text data. Language classification plays a crucial role as a preliminary step before further text processing tasks, such as machine translation, sentiment analysis, and document classification. This study aims to develop and evaluate a text language classification system using the FastText algorithm with a supervised learning approach. The dataset consists of labeled multilingual text, focusing on Indonesian and English languages. The research methodology includes data preprocessing, FastText model training, model testing, and performance evaluation using accuracy metrics. The experimental results show that the FastText model achieved an accuracy of 50% on the test data. Although the accuracy level is relatively low, the findings indicate that FastText can be effectively implemented for language classification tasks. The limited performance of the model is mainly influenced by the small size and limited diversity of the training data. This study is expected to serve as a foundation for further development of more robust text-based language identification systems.*

Keywords: *Language Classification, FastText, Natural Language Processing, Machine Learning, Multilingual Text*

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah mendorong peningkatan signifikan dalam penggunaan data berbasis teks pada berbagai bidang, seperti media sosial, layanan pelanggan, sistem pencarian informasi, serta aplikasi berbasis web. Data teks yang dihasilkan berasal dari beragam bahasa, sehingga diperlukan suatu mekanisme yang mampu mengidentifikasi bahasa teks secara otomatis sebelum dilakukan proses lanjutan. Tanpa adanya sistem identifikasi bahasa yang akurat, pengolahan teks multibahasa berpotensi menghasilkan kesalahan interpretasi dan menurunkan efektivitas sistem pemrosesan teks.

Klasifikasi bahasa merupakan salah satu permasalahan penting dalam bidang Pemrosesan Bahasa Alami (Natural Language Processing). Klasifikasi ini bertujuan untuk menentukan bahasa yang digunakan dalam suatu teks secara otomatis berdasarkan karakteristik linguistik yang dimilikinya. Menurut Jurafsky dan Martin (2023), Pemrosesan Bahasa Alami memungkinkan komputer untuk memahami dan menganalisis bahasa manusia dalam bentuk teks, sehingga dapat diterapkan pada berbagai tugas seperti klasifikasi teks, penerjemahan otomatis, dan analisis sentimen. Dalam konteks tersebut, klasifikasi bahasa menjadi tahap awal yang krusial sebelum dilakukan analisis teks lebih lanjut.

Pendekatan tradisional dalam klasifikasi bahasa umumnya menggunakan aturan linguistik atau kamus bahasa. Namun, pendekatan tersebut memiliki keterbatasan dalam menangani variasi kata, singkatan, serta teks pendek yang banyak ditemukan pada data digital modern. Oleh karena itu,

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 284-289

pendekatan berbasis pembelajaran mesin menjadi alternatif yang lebih fleksibel karena mampu mempelajari pola bahasa secara otomatis dari data. Mitchell (1997) menyatakan bahwa machine learning memungkinkan sistem untuk belajar dari data tanpa harus diprogram secara eksplisit, sehingga cocok digunakan dalam permasalahan klasifikasi teks yang bersifat kompleks dan dinamis.

Salah satu algoritma pembelajaran mesin yang banyak digunakan dalam klasifikasi teks adalah FastText. FastText merupakan algoritma yang dikembangkan oleh Facebook AI Research dan dirancang untuk menangani klasifikasi teks secara efisien dengan memanfaatkan representasi kata berbasis vektor dan informasi subword (Joulin et al., 2016). Pendekatan subword memungkinkan FastText untuk mengenali kata-kata yang jarang muncul atau tidak terdapat dalam data latih, sehingga meningkatkan kemampuan model dalam menangani variasi morfologi bahasa.

Penelitian sebelumnya menunjukkan bahwa FastText mampu memberikan performa yang baik dalam tugas klasifikasi teks multibahasa dengan waktu pelatihan yang relatif cepat. Joulin et al. (2016) membuktikan bahwa FastText memiliki efisiensi tinggi dibandingkan metode klasifikasi teks tradisional. Selain itu, Wyawhare (2023) menunjukkan bahwa FastText efektif dalam mengenali bahasa pada teks pendek, sedangkan Dewi (2025) membuktikan bahwa FastText dapat digunakan secara optimal dalam pemrosesan teks Bahasa Indonesia. Temuan-temuan tersebut menunjukkan bahwa FastText merupakan algoritma yang relevan untuk diterapkan dalam klasifikasi bahasa berbasis teks.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membangun dan menguji sistem klasifikasi bahasa teks menggunakan algoritma FastText dengan pendekatan supervised learning. Bahasa yang diklasifikasikan dalam penelitian ini dibatasi pada Bahasa Indonesia dan Bahasa Inggris dengan menggunakan dataset multibahasa yang telah diberi label. Penelitian ini diharapkan dapat memberikan gambaran mengenai penerapan FastText dalam klasifikasi bahasa serta menjadi dasar pengembangan sistem identifikasi bahasa yang lebih kompleks pada penelitian selanjutnya.

2. METODE PENELITIAN

Metode penelitian yang digunakan dalam klasifikasi bahasa teks menggunakan algoritma FastText. Pembahasan meliputi jenis dan pendekatan penelitian, karakteristik dataset, teknik pengumpulan data, tahapan penelitian, alat dan perangkat yang digunakan, serta teknik analisis data yang didukung oleh landasan teori dari para ahli.

2.1 Jenis dan Pendekatan Penelitian

Penelitian ini merupakan penelitian eksperimen dengan pendekatan kuantitatif. Penelitian eksperimen bertujuan untuk menguji hubungan sebab–akibat melalui penerapan suatu perlakuan tertentu terhadap objek penelitian dan mengamati hasilnya secara terukur. Pendekatan kuantitatif digunakan karena hasil penelitian dinyatakan dalam bentuk angka, khususnya nilai akurasi model klasifikasi.

Dalam penelitian ini digunakan pendekatan supervised learning. Menurut Mitchell (1997), supervised learning merupakan metode pembelajaran mesin di mana model dilatih menggunakan data yang telah diberi label, sehingga sistem dapat mempelajari hubungan antara data masukan dan keluaran yang diharapkan. Pendekatan ini sesuai untuk permasalahan klasifikasi bahasa karena setiap data teks telah memiliki label bahasa yang jelas.

2.2 Dataset Penelitian

Dataset yang digunakan berupa kumpulan teks multibahasa yang telah diberi label sesuai dengan bahasa masing-masing. Dataset berfungsi sebagai sumber utama dalam proses pelatihan dan pengujian model klasifikasi bahasa.

Dalam konteks pemrosesan bahasa alami, kualitas dan kuantitas dataset sangat mempengaruhi performa model. Jurafsky dan Martin (2023) menyatakan bahwa dataset yang representatif dan beragam akan membantu model NLP dalam mempelajari pola bahasa secara lebih akurat. Namun, dalam penelitian ini dataset yang digunakan berskala kecil dan bersifat contoh, karena fokus penelitian diarahkan pada pemahaman implementasi algoritma FastText, bukan pada optimasi performa model.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 284-289

Bahasa yang diklasifikasikan dibatasi pada Bahasa Indonesia dan Bahasa Inggris, sesuai dengan data yang tersedia dalam dataset.

Tabel 1. Dataset Penelitian

No	Bahasa	Jumlah Data	Keterangan
1	Bahasa Indonesia	Beberapa kalimat	Data contoh
2	Bahasa Inggris	Beberapa kalimat	Data contoh

Tabel 1 menunjukkan komposisi dataset yang digunakan dalam penelitian ini, yang terdiri dari dua kelas bahasa, yaitu Bahasa Indonesia dan Bahasa Inggris. Dataset berskala kecil dan digunakan sebagai data contoh untuk mendemonstrasikan proses klasifikasi bahasa menggunakan algoritma FastText. Penggunaan dataset dengan jumlah data yang terbatas dilakukan karena penelitian ini difokuskan pada pemahaman alur kerja dan implementasi algoritma, bukan pada pencapaian performa klasifikasi yang optimal.
Dataset dibagi menjadi data latih dan data uji. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan bahasa teks yang belum pernah dilihat sebelumnya.

2.3 Teknik Pengumpulan Data

Teknik pengumpulan data dalam penelitian ini dilakukan melalui studi pustaka dan penggunaan dataset sekunder. Studi pustaka bertujuan untuk memperoleh dasar teori yang berkaitan dengan klasifikasi bahasa, pemrosesan bahasa alami, dan algoritma FastText.
Penggunaan dataset sekunder dilakukan tanpa pengumpulan data secara langsung. Pendekatan ini umum digunakan dalam penelitian pembelajaran mesin, terutama ketika fokus penelitian adalah evaluasi metode atau algoritma tertentu, bukan eksplorasi fenomena lapangan.

2.4 Tahapan Penelitian

Tahapan penelitian disusun secara sistematis untuk memastikan proses klasifikasi bahasa berjalan sesuai dengan alur metodologis yang benar. Menurut Jurafsky dan Martin (2023), proses pemodelan dalam NLP umumnya meliputi praproses data, pelatihan model, pengujian, dan evaluasi.

Tabel 2. Tahapan Penelitian

No	Tahap	Deskripsi
1	Praproses Data	Membersihkan dan menyiapkan data teks
2	Pelatihan Model	Melatih model FastText dengan data latih
3	Pengujian Model	Memprediksi bahasa pada data uji
4	Evaluasi	Menghitung akurasi dan kinerja model

Tabel 2 menggambarkan tahapan penelitian yang dilakukan secara berurutan dan sistematis. Proses dimulai dari praproses data untuk memastikan data teks sesuai dengan format yang dibutuhkan oleh algoritma FastText. Selanjutnya, model dilatih menggunakan data latih dan diuji menggunakan data uji untuk memperoleh hasil prediksi bahasa. Tahap evaluasi dilakukan untuk menilai kinerja model berdasarkan hasil klasifikasi yang dihasilkan.
Pada tahap praproses, data teks disesuaikan dengan format input FastText. Selanjutnya, model dilatih menggunakan pendekatan supervised learning. Model yang telah dilatih kemudian diuji

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 284-289

menggunakan data uji untuk memperoleh hasil prediksi yang akan dianalisis pada tahap evaluasi.

2.5 Alat dan Perangkat yang Digunakan

Alat dan perangkat yang digunakan dalam penelitian ini meliputi perangkat lunak dan perangkat keras yang mendukung proses klasifikasi bahasa.

Tabel 3. Alat dan Perangkat Penelitian

No	Alat / Perangkat	Fungsi
1	Python	Implementasi program
2	FastText	Klasifikasi bahasa
3	Laptop	Pengolahan data
4	Sistem Operasi	Windows

Tabel 3 menyajikan alat dan perangkat yang digunakan untuk mendukung proses penelitian. Bahasa pemrograman Python digunakan sebagai media implementasi program karena fleksibilitasnya dalam pengolahan data teks. FastText berperan sebagai algoritma utama dalam proses klasifikasi bahasa, sedangkan perangkat komputer dengan sistem operasi Windows digunakan sebagai sarana pengolahan dan pengujian data.
FastText dipilih karena kemampuannya dalam memanfaatkan representasi subword untuk meningkatkan akurasi klasifikasi teks pendek. Joulin et al. (2016) menyatakan bahwa penggunaan subword memungkinkan FastText menangani kata yang jarang muncul atau tidak terdapat dalam data latih.

2.6 Teknik Analisis Data

Teknik analisis data dilakukan dengan membandingkan hasil prediksi model dengan label bahasa yang sebenarnya. Evaluasi difokuskan pada nilai akurasi sebagai indikator utama keberhasilan klasifikasi.
Akurasi digunakan untuk mengukur persentase prediksi yang benar terhadap keseluruhan data uji. Menurut Mitchell (1997), evaluasi kuantitatif seperti ini penting untuk menilai sejauh mana model pembelajaran mesin mampu melakukan generalisasi terhadap data baru. Hasil analisis menjadi dasar dalam pembahasan serta penarikan kesimpulan penelitian.

3. ANALISA DAN PEMBAHASAN

Hasil pelatihan dan pengujian model klasifikasi bahasa menggunakan algoritma FastText serta pembahasan terhadap hasil yang diperoleh. Analisis difokuskan pada kinerja model dalam mengklasifikasikan bahasa teks serta faktor-faktor yang mempengaruhi hasil klasifikasi.

3.1 Hasil Pelatihan Model

Pelatihan model dilakukan menggunakan dataset teks multibahasa yang telah melalui tahap praproses. Proses pelatihan menghasilkan sebuah model FastText yang disimpan dalam format file .bin. Model yang dibangun mampu mengenali dua kelas bahasa, yaitu Bahasa Indonesia dan Bahasa Inggris, sesuai dengan label pada data latih.
FastText memanfaatkan representasi kata berbasis vektor serta informasi subword, sehingga model mampu mengenali karakteristik bahasa meskipun menggunakan data dengan skala terbatas. Joulin et al. (2016) menyatakan bahwa pendekatan subword pada FastText memungkinkan model tetap bekerja secara efektif pada teks pendek dan kata yang jarang muncul.

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 284-289

3.2 Hasil Pengujian Model

Pengujian model dilakukan dengan memberikan sejumlah teks uji berbahasa Indonesia dan Bahasa Inggris yang tidak digunakan dalam proses pelatihan. Model menghasilkan prediksi berupa label bahasa untuk setiap teks uji.
Hasil pengujian menunjukkan bahwa model FastText mampu melakukan klasifikasi bahasa pada sebagian data uji, namun masih ditemukan kesalahan prediksi pada teks tertentu. Hal ini menunjukkan bahwa model telah mempelajari pola dasar bahasa, tetapi belum mampu membedakan karakteristik bahasa secara konsisten.

3.3 Evaluasi Kinerja Model

Evaluasi kinerja model dilakukan dengan membandingkan hasil prediksi model dengan label bahasa yang sebenarnya. Berdasarkan pengujian yang dilakukan, diperoleh nilai akurasi sebesar 50%.

Tabel 4. Hasil Evaluasi Kinerja Model

No	Parameter Evaluasi	Hasil
1	Jumlah kelas bahasa	2
2	Bahasa yang diuji	Indonesia & Inggris
3	Akurasi model	50%

Tabel 4 menunjukkan bahwa model FastText mampu mengklasifikasikan bahasa teks dengan tingkat akurasi sebesar 50%. Nilai ini menunjukkan bahwa setengah dari data uji berhasil diklasifikasikan dengan benar. Hasil ini mencerminkan kemampuan dasar model dalam mengenali bahasa, meskipun performanya masih terbatas.

3.4 Pembahasan

Hasil evaluasi menunjukkan bahwa model FastText yang dibangun dalam penelitian ini memperoleh nilai akurasi sebesar 50%. Nilai tersebut menunjukkan bahwa model mampu mengklasifikasikan bahasa teks secara dasar, namun performanya belum optimal. Capaian ini perlu dipahami dengan mempertimbangkan karakteristik dataset, metode pelatihan, serta pendekatan algoritma yang digunakan dalam penelitian.
Salah satu faktor utama yang mempengaruhi kinerja model adalah jumlah data latih yang sangat terbatas. Dataset berskala kecil menyebabkan model tidak memiliki cukup variasi contoh bahasa untuk mempelajari pola linguistik secara mendalam. Dalam konteks pemrosesan bahasa alami, Jurafsky dan Martin (2023) menegaskan bahwa kualitas dan kuantitas data merupakan faktor kunci dalam menentukan performa model NLP. Dataset yang kurang representatif akan membatasi kemampuan model dalam mengenali struktur bahasa yang lebih kompleks.
Selain jumlah data, keberagaman teks juga berpengaruh terhadap kemampuan model dalam melakukan generalisasi. Teks yang digunakan dalam penelitian ini bersifat sederhana dan memiliki struktur kalimat yang relatif seragam. Akibatnya, model kesulitan membedakan karakteristik bahasa pada teks yang memiliki kemiripan kosakata antarbahasa. Mitchell (1997) menyatakan bahwa model supervised learning dengan data latih terbatas cenderung mengalami keterbatasan dalam melakukan generalisasi terhadap data baru, sehingga performa prediksi menjadi kurang stabil.
Meskipun demikian, hasil penelitian menunjukkan bahwa FastText tetap mampu menjalankan fungsi klasifikasi bahasa dengan memanfaatkan representasi subword. Pendekatan subword memungkinkan FastText mengenali pola bahasa berdasarkan potongan karakter, bukan hanya kata utuh. Joulin et al. (2016) menjelaskan bahwa mekanisme ini sangat efektif dalam menangani kata yang jarang muncul atau tidak terdapat dalam data latih. Hal ini menjelaskan

JURNAL AI dan SPK
Jurnal Artificial Intelligent dan Sistem Penunjang Keputusan
Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan
Volume 3, No. 2, Tahun 2025
ISSN 3025-0927 (media online)
Hal 284-289

mengapa model masih mampu memberikan prediksi bahasa meskipun menggunakan dataset yang terbatas.

Dari sisi implementasi, penelitian ini berhasil menunjukkan bahwa algoritma FastText dapat diterapkan secara praktis untuk membangun sistem klasifikasi bahasa teks. Proses pelatihan dan pengujian dapat dilakukan dengan relatif cepat dan sederhana, sehingga FastText cocok digunakan sebagai pendekatan awal dalam pengembangan sistem identifikasi bahasa. Hasil ini sejalan dengan tujuan penelitian yang lebih menekankan pada pemahaman alur kerja dan implementasi algoritma, bukan pada pencapaian performa maksimal.

Dengan demikian, meskipun nilai akurasi yang diperoleh masih rendah, penelitian ini tetap memberikan kontribusi dalam menunjukkan potensi FastText sebagai metode klasifikasi bahasa berbasis teks. Untuk meningkatkan kinerja model, penelitian selanjutnya disarankan menggunakan dataset dengan jumlah dan variasi yang lebih besar, menambah jumlah kelas bahasa, serta melakukan penyesuaian parameter pelatihan. Upaya tersebut diharapkan dapat meningkatkan kemampuan model dalam mengenali pola bahasa secara lebih akurat dan konsisten.

4. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah diuraikan, dapat disimpulkan bahwa algoritma FastText dapat diterapkan untuk melakukan klasifikasi bahasa teks berbasis supervised learning. Model yang dibangun mampu mengidentifikasi dua kelas bahasa, yaitu Bahasa Indonesia dan Bahasa Inggris, dengan memanfaatkan representasi kata berbasis vektor serta informasi subword.

Hasil evaluasi menunjukkan bahwa model FastText memperoleh nilai akurasi sebesar 50% pada data uji. Nilai ini menunjukkan bahwa model telah mampu melakukan klasifikasi bahasa secara dasar, namun performanya belum optimal. Keterbatasan kinerja model terutama dipengaruhi oleh jumlah data latih yang terbatas serta kurangnya variasi teks dalam dataset, sehingga kemampuan model dalam melakukan generalisasi terhadap data baru masih rendah.

Meskipun demikian, penelitian ini berhasil menunjukkan bahwa FastText dapat diimplementasikan secara efektif sebagai pendekatan awal dalam pengembangan sistem klasifikasi bahasa teks. Proses pelatihan dan pengujian model dapat dilakukan secara sederhana dan efisien, sehingga FastText berpotensi digunakan sebagai dasar pengembangan sistem identifikasi bahasa yang lebih kompleks.

Sebagai tindak lanjut, penelitian selanjutnya disarankan untuk menggunakan dataset dengan jumlah dan variasi yang lebih besar, menambah jumlah kelas bahasa yang diklasifikasikan, serta melakukan penyesuaian parameter pelatihan guna meningkatkan akurasi dan kestabilan model. Dengan pengembangan tersebut, diharapkan sistem klasifikasi bahasa berbasis FastText dapat memberikan performa yang lebih optimal dan aplikatif.

REFERENCES

Dewi, B. E. S. (2025). Pengukuran kemiripan kalimat bahasa Indonesia menggunakan representasi word embedding FastText. Jurnal Teknologi Informasi dan Digital, 3(1), 20–29.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed.). Stanford University.

Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

Wyawhare, A. (2023). Comparative analysis of multilingual text classification and identification through deep learning and embedding visualization. arXiv preprint arXiv:2312.03789.