

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Transformasi digital yang berkembang sangat pesat telah memengaruhi cara masyarakat berinteraksi dengan informasi, mulai dari proses pencarian, penyebaran, hingga konsumsi informasi dalam kehidupan sehari-hari. Kemunculan berbagai platform media sosial dan portal berita daring telah mempermudah arus informasi lintas batas geografis secara instan. Namun di sisi lain, kemudahan tersebut juga membuka celah bagi penyebaran informasi yang tidak terverifikasi atau yang dikenal sebagai berita *Hoax* (Shu *et al.*, 2020). Fenomena ini telah menjadi permasalahan global yang memerlukan perhatian serius dari berbagai pihak, termasuk komunitas ilmiah dan teknologi.

Di Indonesia, penyebaran berita *Hoax* semakin mengkhawatirkan seiring dengan meningkatnya penetrasi internet dan penggunaan media sosial. Berdasarkan laporan Kementerian Komunikasi dan Informatika (Kominfo) Republik Indonesia, terdapat ribuan konten hoaks yang berhasil diidentifikasi dan diblokir setiap tahunnya, terutama yang berkaitan dengan isu politik, kesehatan, dan sosial (Kominfo, 2022). Penyebaran informasi palsu ini tidak hanya menimbulkan kebingungan di masyarakat, melainkan juga berpotensi memicu konflik antar kelompok, mengurangi kepercayaan publik terhadap institusi, serta mengganggu stabilitas sosial dan politik (Pratama & Widodo, 2021).

Deteksi berita *Hoax* secara manual oleh tim pemeriksa fakta (fact-checker) dinilai tidak lagi efisien mengingat volume informasi yang beredar setiap harinya terus meningkat secara eksponensial. Keterbatasan sumber daya manusia dan waktu verifikasi yang dibutuhkan menjadikan pendekatan manual tidak mampu mengimbangi laju penyebaran hoaks di era digital (Zubiaga *et al.*, 2020). Oleh karena itu, dibutuhkan solusi otomatis yang mampu mendeteksi berita *Hoax* secara cepat, akurat, dan dapat diskalakan.

Teknologi kecerdasan buatan (*Artificial Intelligence/AI*), khususnya *machine learning*, telah terbukti mampu menjadi solusi efektif dalam mengatasi permasalahan tersebut. Berbagai penelitian telah menunjukkan bahwa pendekatan

Machine Learning dapat digunakan untuk mengotomatisasi proses klasifikasi teks, termasuk identifikasi berita *Hoax*, dengan tingkat akurasi yang kompetitif dibandingkan penilaian manusia (Bondielli & Marcelloni, 2019). Dalam konteks pemrosesan teks berbahasa Indonesia, tantangan tambahan muncul karena karakteristik linguistik bahasa Indonesia yang berbeda dengan bahasa Inggris, termasuk penggunaan imbuhan yang kompleks, bahasa campuran (*code-switching*), serta berbagai jenis bahasa santai yang biasa dipakai di media sosial. (Koto *et al.*, 2020).

Salah satu cara yang sering dipakai untuk mengklasifikasikan teks ialah metode yang menggunakan *boosting ensemble*, yang menghimpun beberapa model prediksi lemah (*weak learners*) menjadi satu model yang lebih kuat (*strong learner*). Dua algoritma *boosting* yang paling banyak digunakan dan mendapat perhatian besar dalam komunitas ilmiah adalah *Extreme Gradient Boosting* (*XGBoost*) dan *Light Gradient Boosting Machine* (*LightGBM*). *XGBoost*, yang diperkenalkan oleh Chen & Guestrin (2016), dikenal sebagai algoritma dengan performa tinggi dan kemampuan generalisasi yang baik pada berbagai jenis data terstruktur. Algoritma ini telah memenangkan berbagai kompetisi *Machine Learning* di platform *Kaggle* dan banyak diterapkan dalam penelitian klasifikasi teks (Bentéjac *et al.*, 2021).

LightGBM, yang dikembangkan oleh *Microsoft Research* (Ke *et al.*, 2017), merupakan varian *boosting* modern yang dirancang khusus untuk mengatasi keterbatasan *XGBoost* dalam hal efisiensi komputasi. *LightGBM* menggunakan teknik *Gradient-based One-Side Sampling* (GOSS) dan *Exclusive Feature Bundling* (EFB) yang memungkinkan proses pelatihan model berjalan lebih cepat dengan konsumsi memori yang lebih rendah tanpa mengorbankan akurasi secara signifikan (Ke *et al.*, 2017). Keunggulan ini menjadikan *LightGBM* sebagai pilihan utama dalam pengolahan dataset berskala besar, termasuk data teks yang memiliki dimensi fitur yang tinggi setelah proses ekstraksi fitur.

XGBoost dan *LightGBM* dipilih dalam studi ini karena keduanya adalah algoritma ensemble *boosting* yang menunjukkan performa superior dalam beragam tugas klasifikasi data, termasuk klasifikasi teks. *XGBoost* terkenal karena kemampuannya yang baik dan stabil dalam memberikan prediksi yang tepat,

sementara *LightGBM* lebih unggul dalam efisiensi komputasi dan kecepatan pelatihan untuk data berdimensi tinggi. Ciri-ciri tersebut membuat kedua algoritma menarik untuk diperbandingkan dalam pengklasifikasian berita *hoax* berbahasa Indonesia yang memiliki kompleksitas fitur teks yang tinggi setelah dilakukan proses ekstraksi fitur TF-IDF.

Dalam klasifikasi teks, representasi fitur memegang peranan penting dalam menentukan kualitas model. *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah salah satu metode ekstraksi fitur yang paling umum digunakan, dimana metode ini dapat secara akurat menggambarkan pentingnya suatu kata dalam dokumen subjektif (Ramos, 2003). Metode TF-IDF ini terbukti efisien dalam berbagai tugas klasifikasi teks berbahasa Indonesia dan sering dikombinasikan dengan algoritma *Machine Learning* konvensional (Tho *et al.*, 2021).

Meskipun *XGBoost* dan *LightGBM* telah terbukti unggul dalam berbagai domain, performa keduanya sangat dipengaruhi oleh konfigurasi *Hyperparameter* yang digunakan. Parameter yang tidak dipelajari secara otomatis oleh model dan harus ditentukan sebelum proses tersebut dimulai disebut *hyperparameter* (Bergstra & Bengio, 2012). Pemilihan *Hyperparameter* yang tidak tepat dapat mengakibatkan *underfitting* atau *overfitting*, yang berujung pada penurunan performa model secara signifikan. Oleh karena itu, proses *Hyperparameter tuning* sangat penting untuk membangun model *Machine Learning* yang terbaik.

Berbagai teknik *Hyperparameter tuning* telah diterapkan, seperti *Grid Search* dan *Random Search*, sampai pendekatan lebih canggih berbasis optimasi *Bayesian* (Bergstra *et al.*, 2011). Penerapan *Hyperparameter tuning* telah terbukti secara konsisten mampu meningkatkan performa model secara signifikan pada berbagai tugas klasifikasi, termasuk klasifikasi teks (Probst *et al.*, 2019). Dengan demikian, perbandingan antara *XGBoost* dan *LightGBM* yang dilakukan setelah masing-masing model dioptimalkan melalui *Hyperparameter tuning* akan menghasilkan kesimpulan yang lebih valid dan dapat dipertanggungjawabkan secara ilmiah.

Walaupun sejumlah penelitian telah mengkaji klasifikasi berita *hoax* berbahasa Indonesia menggunakan *machine learning*, penelitian sebelumnya umumnya menggunakan algoritma konvensional atau belum melakukan tuning

hyperparameter secara sistematis, sehingga perbandingan kinerja antar algoritma belum optimal dan kurang objektif. Selain itu, studi yang secara khusus membandingkan *XGBoost* dan *LightGBM* pada klasifikasi berita *hoax* berbahasa Indonesia masih sangat terbatas. Penelitian ini berupaya mengisi celah tersebut dengan menyajikan perbandingan komprehensif yang didasarkan pada konfigurasi optimal masing-masing algoritma.

Dengan demikian, penelitian ini berfokus pada judul "PERBANDINGAN KINERJA *XGBOOST* DAN *LIGHTGBM* DENGAN *HYPERPARAMETER TUNING* PADA KLASIFIKASI BERITA *HOAX* INDONESIA". Diharapkan penelitian ini dapat membantu dalam pengembangan sistem deteksi hoaks berbahasa Indonesia yang efisien dan akurat, sekaligus menyajikan rekomendasi algoritma terbaik yang dapat dijadikan acuan bagi peneliti dan praktisi di bidang *Data Mining* dan natural language processing.

1.2 Rumusan Masalah

Rumusan masalah penelitian ini adalah sebagai berikut, berdasarkan latar belakang yang telah dijelaskan:

1. Bagaimana kinerja algoritma *XGBoost* dalam melakukan klasifikasi berita *Hoax* berbahasa Indonesia setelah diterapkan proses *Hyperparameter tuning*?
2. Bagaimana kinerja algoritma *LightGBM* dalam melakukan klasifikasi berita *Hoax* berbahasa Indonesia setelah diterapkan proses *Hyperparameter tuning*?
3. Algoritma manakah, antara *XGBoost* dan *LightGBM*, yang menghasilkan kinerja klasifikasi berita *Hoax* berbahasa Indonesia yang lebih optimal berdasarkan metrik akurasi, presisi, *Recall*, dan *F1-Score*?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan, tujuan penelitian ini adalah sebagai berikut:

1. Menganalisis dan mengukur kinerja algoritma *XGBoost* dalam klasifikasi berita *Hoax* berbahasa Indonesia setelah diterapkan proses *Hyperparameter tuning*.
2. Menganalisis dan mengukur kinerja algoritma *LightGBM* dalam klasifikasi berita *Hoax* berbahasa Indonesia setelah diterapkan proses *Hyperparameter tuning*.
3. Membandingkan kinerja algoritma *XGBoost* dan *LightGBM* secara objektif guna menentukan algoritma yang paling optimal untuk klasifikasi berita *Hoax* berbahasa Indonesia.

1.4 Batasan Masalah

Penelitian ini dibatasi pada hal-hal berikut untuk tetap fokus dan menghindari diskusi yang terlalu luas.

1. Jumlah data yang digunakan terdiri dari 31.353 data berita berbahasa Indonesia yang dikategorikan sebagai *Hoax* dan non-*Hoax*, yang diakses melalui Kaggle: <https://www.kaggle.com/datasets/linkgish/indonesian-fact-and-Hoax-political-news>
2. Algoritma *Machine Learning* yang dibandingkan dibatasi hanya pada dua algoritma, yaitu *XGBoost* dan *LightGBM*, tanpa membandingkan dengan algoritma lain di luar keduanya.
3. *Term Frequency-Inverse Document Frequency* (TF-IDF) digunakan untuk mengekstraksi fitur dan tidak mencakup metode representasi teks berbasis deep learning seperti word embedding atau transformer.
4. Proses *Hyperparameter tuning* yang diterapkan mencakup metode yang sesuai dengan kapasitas komputasi yang tersedia, yaitu *Grid Search*.
5. Model dievaluasi dengan metrik akurasi, presisi, *Recall*, dan *F1-Score*.
6. Penelitian ini tidak mencakup pengembangan sistem aplikasi deteksi hoaks secara langsung, melainkan berfokus pada eksperimen dan analisis perbandingan performa kedua algoritma.
7. *Tools* yang dipakai *Google Colaboratory* (*Google Colab*) dengan bahasa pemrograman *Python*.

1.5 Manfaat Penelitian

Diharapkan penelitian ini dapat memberi manfaat bagi banyak pihak, antara lain:

- a. Memberikan dedikasi ilmiah mengenai perbandingan kinerja algoritma *XGBoost* dan *LightGBM* dalam klasifikasi berita *Hoax* berbahasa Indonesia.
- b. Menambah referensi di bidang *Data Mining* dan pemrosesan bahasa alami (NLP).
- c. Memberikan pemahaman tentang pengaruh *Hyperparameter tuning* terhadap kinerja model.
- d. Memberikan rekomendasi algoritma yang lebih optimal untuk deteksi berita *Hoax*.
- e. Menjadi panduan bagi peneliti atau pengembang dalam penerapan klasifikasi teks.
- f. Meningkatkan pemahaman dan keterampilan peneliti dalam *Machine Learning* dan eksperimen model.

