BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan komunikasi mendorong peningkatan konsumsi berita secara signifikan. Portal berita daring seperti Detik, Kompas, CNN Indonesia, dan Liputan6 menerbitkan ratusan artikel setiap bulan pada berbagai kategori berita, termasuk pada kategori olahraga. Pada kategori ini, berita umumnya terbagi ke dalam subkategori seperti Sepak Bola, Badminton, dan MotoGP, masing-masing dengan *karakteristik linguistik*, struktur kalimat, dan istilah teknis yang berbeda. Istilah seperti gol, transfer, dan liga pada sepak bola, smash dan rally pada badminton, serta lap, sirkuit, dan pembalap pada MotoGP menunjukkan adanya variasi terminologi yang menuntut proses pengelompokan konten yang lebih akurat dan konsisten.

Selain memiliki kompleksitas linguistik, kategori olahraga juga menunjukkan tingkat konsumsi yang relatif tinggi dibandingkan kategori berita lainnya di Indonesia. Berdasarkan survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) tahun 2025, konten olahraga merupakan jenis berita daring dengan persentase akses tertinggi, yaitu sebesar 15,62%, mengungguli kategori politik, sosial hukum dan HAM yang berada pada posisi kedua dengan 14,90% (Muhamad, 2025). Data ini menegaskan bahwa berita olahraga memiliki tingkat relevansi dan keterpaparan publik yang tinggi, sehingga pemilihan domain ini sebagai objek penelitian klasifikasi teks tidak hanya tepat secara empiris tetapi juga mencerminkan kebutuhan nyata pengguna internet Indonesia. Tingginya tingkat akses ini secara langsung mendorong peningkatan produksi berita olahraga oleh portal berita daring, sehingga menambah beban pengelolaan konten yang harus dilakukan secara cepat, konsisten, dan dalam skala besar.

Peningkatan produksi berita olahraga menimbulkan tantangan baru bagi portal berita dalam pengorganisasian konten. Hingga saat ini, sebagian besar proses pengelompokan berita masih dilakukan secara manual oleh tim redaksi. Praktik ini

berpotensi menimbulkan ketidakonsistenan pelabelan, variasi standar editorial, dan keterlambatan publikasi ketika jumlah berita meningkat (Juwita et al., 2022). Temuan tersebut diperkuat oleh (Rizal et al., 2021) yang menjelaskan bahwa pengklasifikasian berita pada portal daring masih dilakukan melalui pembacaan isi secara menyeluruh sebelum menentukan kategorinya. Proses manual seperti ini tidak hanya memakan waktu, tetapi juga tidak skalabel untuk menangani *volume* berita yang terus bertambah, sehingga menegaskan urgensi penerapan sistem klasifikasi otomatis yang lebih efisien dan akurat.

Dalam penelitian *text mining* dan *Natural Language Processing* (NLP), metode *Term Frequency Inverse Document Frequency* (TF-IDF) merupakan salah satu teknik *representasi* teks yang paling umum digunakan untuk mengekstraksi fitur penting dari dokumen. TF-IDF terbukti efektif untuk teks berbahasa Indonesia, karena mampu membedakan kata bermakna tinggi dari kata yang bersifat umum dan sering muncul (Hendrawan et al., n.d.) *Representasi numerik* berbasis TF-IDF sangat sesuai untuk digunakan dalam algoritma klasifikasi seperti *K-Nearest Neighbor (KNN)* dan *Support Vector Machine (SVM)*. KNN dikenal sebagai metode *non-parametrik* yang bekerja berdasarkan kemiripan antar dokumen, dan berbagai penelitian menunjukkan bahwa kombinasi TF-IDF dan KNN menghasilkan performa kuat dalam berbagai tugas klasifikasi teks, seperti deteksi clickbait (Sagita & Primajaya, 2022) dan klasifikasi emosi (Azizah & Rahmayuda, 2025). Di sisi lain, SVM dikenal sebagai algoritma yang sangat stabil dalam memproses data berdimensi tinggi seperti hasil *representasi* TF-IDF, dan algoritma ini telah menunjukkan akurasi tinggi pada berbagai penelitian, termasuk deteksi hoaks (Desriansyah & Utna, 2025; Dewi et al., 2025) serta clickbait (Kurniawan et al., 2025)

Namun demikian, performa kedua algoritma tersebut belum dapat digeneralisasikan secara langsung ke domain berita olahraga. Berita olahraga memiliki *karakteristik linguistik* yang berbeda dibandingkan jenis teks lainnya, seperti struktur kalimat yang lebih ringkas, penggunaan istilah teknis yang berulang, serta variasi gaya penulisan antar cabang olahraga. Telaah literatur juga menunjukkan bahwa penelitian terkait klasifikasi berita olahraga berbahasa

Indonesia masih terbatas. (Rizal et al., 2021), mengusulkan klasifikasi berita olahraga menggunakan algoritma KNN berbasis *Levenshtein Distance*, namun penelitian tersebut tidak memanfaatkan *representasi* fitur modern seperti TF-IDF dan tidak melakukan perbandingan performa dengan algoritma lain seperti SVM. Sementara itu, penelitian oleh (Alif et al., 2022) telah mengembangkan model klasifikasi berita olahraga dan bukan olahraga menggunakan SVM dengan fitur TF-IDF dan n-gram, namun pendekatan tersebut masih terbatas pada klasifikasi biner dan belum mengkaji klasifikasi multikategori antar cabang olahraga maupun melakukan evaluasi komparatif antar algoritma pembelajaran mesin.

Keterbatasan tersebut menunjukkan adanya kesenjangan penelitian yang signifikan, yaitu belum adanya studi *komprehensif* yang menerapkan *representasi* TF-IDF untuk klasifikasi multikategori berita olahraga sekaligus membandingkan performa algoritma KNN dan SVM dalam konteks tersebut. Oleh karena itu, penelitian ini dilakukan untuk mengisi kekosongan tersebut dengan membangun dan membandingkan model klasifikasi berita olahraga ke dalam tiga subkategori utama Sepak Bola, Badminton, dan MotoGP menggunakan kombinasi TF-IDF dengan algoritma KNN dan SVM sebagai pendekatan pembelajaran mesin.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang di atas, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana algoritma *K-Nearest Neighbor (KNN)* dan *Support Vector Machine (SVM)* dapat diimplementasikan untuk mengklasifikasikan berita olahraga ke dalam subkategori Sepak Bola, Badminton, dan MotoGP?
2. Bagaimana metode *Term Frequency Inverse Document Frequency (TF-IDF)* digunakan untuk merepresentasikan teks berita olahraga dalam bentuk fitur numerik yang dapat diproses oleh model klasifikasi?
3. Bagaimana hasil perbandingan kinerja antara model KNN dan SVM berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-score* yang diperoleh dari hasil pengujian sistem?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Menerapkan algoritma *K-Nearest Neighbor (KNN)* dan *Support Vector Machine (SVM)* untuk mengklasifikasikan berita olahraga ke dalam subkategori Sepak Bola, Badminton, dan MotoGP secara otomatis
2. Menerapkan metode *Term Frequency Inverse Document Frequency* (TF-IDF) sebagai teknik *representasi* teks untuk mengubah berita olahraga menjadi *vektor numerik* yang mencerminkan tingkat kepentingan kata dalam setiap dokumen.
3. Menganalisis dan membandingkan kinerja model KNN dan SVM berdasarkan metrik evaluasi *accuracy*, *Precision*, *recall*, dan *F1-score* guna menentukan algoritma yang paling efektif dalam mengklasifikasikan berita olahraga berbahasa Indonesia

1.4 Manfaat Penelitian

a. Manfaat Teoretis

Penelitian ini berkontribusi dalam memperkaya kajian di bidang text mining dan *Natural Language Processing (NLP)* berbahasa Indonesia, khususnya dalam penerapan metode pembobotan *Term Frequency Inverse Document Frequency* (TF-IDF) serta algoritma *K-Nearest Neighbor (KNN)* dan *Support Vector Machine (SVM)* untuk klasifikasi berita. Melalui hasil pengujian yang dilakukan, penelitian ini dapat menjadi referensi empiris bagi studi-studi selanjutnya yang ingin melakukan analisis komparatif terhadap performa berbagai algoritma *machine learning* pada teks berbahasa Indonesia. Selain itu, penelitian ini juga memperkuat literatur terkait *efektivitas* metode pembobotan TF-IDF dalam menangani teks berdimensi tinggi, sekaligus memberikan dasar teoretis mengenai keunggulan dan keterbatasan model KNN dan SVM dalam konteks klasifikasi berita digital.

b. Manfaat praktis

Penelitian ini diharapkan dapat menjadi dasar ilmiah bagi pengembangan sistem klasifikasi berita otomatis yang dapat diterapkan oleh industri media daring di Indonesia. Dengan adanya sistem klasifikasi berbasis kecerdasan buatan, proses pengelompokan berita ke dalam subkategori seperti Sepak Bola, Badminton, dan MotoGP dapat dilakukan secara lebih cepat, akurat, dan konsisten dibandingkan proses manual oleh redaksi. Hasil penelitian

ini juga diharapkan dapat membantu meningkatkan efisiensi kerja redaksi dengan mempercepat proses pengelolaan dan penandaan berita, serta memastikan distribusi informasi kepada pembaca menjadi lebih relevan dan tepat sasaran. Dengan demikian, penelitian ini memberikan kontribusi nyata terhadap penerapan teknologi kecerdasan buatan dalam peningkatan kualitas manajemen konten media digital di Indonesia.

1.5 Batasan Masalah

Agar penelitian ini lebih terarah dan fokus, maka ruang lingkup penelitian dibatasi pada hal-hal berikut:

1. Penelitian yang dilakukan hanya berfokus pada berita olahraga berbahasa Indonesia yang didapatkan melalui proses *Web Scraping* dari portal berita daring, dengan tiga subkategori utama yaitu Sepak Bola, Badminton, dan MotoGP.
2. Proses *representasi* teks dilakukan menggunakan metode *Term Frequency Inverse Document Frequency (TF-IDF)* sebagai teknik pembobotan kata, tanpa melibatkan metode *representasi* lain seperti *word embedding* (Word2Vec, FastText, atau BERT).
3. Algoritma klasifikasi yang digunakan terbatas *pada K-Nearest Neighbor (KNN)* dan *Support Vector Machine (SVM)*, dengan evaluasi kinerja model dilakukan berdasarkan empat metrik utama, yaitu *accuracy*, *Precision*, *recall*, dan *F1-score*.
4. Dataset yang digunakan dibatasi pada hasil *Web Scraping* yang telah melalui tahap preprocessing (pembersihan teks, tokenisasi, *stopword removal*, dan *stemming*), tanpa mempertimbangkan aspek analisis sentimen atau topik lain di luar klasifikasi berita olahraga.