

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pada zaman perkembangan teknologi yang terjadi saat ini, informasi dari segala penjuru dunia dapat kita akses melalui internet. Informasi yang didapatkan melalui internet yaitu dokumen berita yang dimana berita ini merupakan suatu informasi penting bagi kehidupan manusia. Berita banyak disebarakan melalui media cetak seperti koran dan majalah. Berita bisa dilihat melalui televisi dan bisa didengarkan melalui radio. Berita juga disebarakan media online seperti sosial media dan portal berita. Dalam portal berita terdapat banyak berbagai macam informasi yang didapatkan. Namun dengan banyaknya berita, maka diperlukan pencermatan lebih untuk mendapatkan informasi sesuai yang diinginkan oleh pengguna informasi. Berita yang sangat diminati dari berbagai kalangan maupun usia yaitu berita olahraga.

Penelitian sebelumnya memakai metode *naïve bayes* dengan *Enhanced Confix Stripping Stemmer* guna mengelompokkan menggunakan dokumen uji berita olahraga yang didapatkan dari situs *sport.detik.com*. Peneliti mengambil 151 berita latih dan 30 berita uji. Dari 30 berita uji yang disajikan, dengan acak user hanya memilih 18. Penelitian ini mampu mendapatkan hasil akurasi 77% dengan tingkat kesalahan mencapai 23%[1].

Penelitian yang lain juga memakai metode *naïve bayes* menggunakan fitur n-gram guna mengelompokkan berita lokal radar malang. Data latih berasal dari www.kompas.com dengan kategori yaitu olahraga, kesehatan, news, travel, politik, edukasi, dan ekonomi. Data uji berasal dari www.radarmalang.co.id. Peneliti telah melakukan 5 ujicoba dengan mendapatkan hasil akurasi secara berturut-turut dengan persentase 78.66%, 68.20%, 59.24%, 65.93%, 74.39% [2].

Dalam mengklasifikasi dokumen twitter guna mendapatkan karakteristik calon pegawai menggunakan metode *K-Nearest Neighbor*. Peneliti memakai empat kategori kepribadian yang ditetapkan sebagai kelas hasil klasifikasi yaitu Artisan, Guardian, Idealist dan Rasional dengan perolehan hasil keakuratan sebesar 66% dengan jumlah data benar sebanyak 53 data dan jumlah data salah 27 data[13].

Dalam penerapan algoritma *K-Nearest Neighbor* pada sistem pengelompokan sebuah kategori berita menggunakan bantuan web scrapping guna mendapatkan berita secara online dari portal berita. Penelitian ini mampu mendapatkan nilai akurasi dengan persentase 74 %, precision 35 %, recall 35 %, F-measure 35%, dan Specificity 84 % dan UAC 60% dalam klasifikasi kategori berita [4].

Dalam penelitian yang dilakukan untuk pengklasifikasian berita clickbait, diperoleh hasil paling baik dengan jumlah $k=11$ pada skenario 1 dan pembagian data sebanyak 800 data latih dan 200 data uji yang mampu mendapatkan nilai akurasi sebesar 71%, Precision 72%, dan Recall 71%. [5]

Hasil penelitian yang dilakukan untuk klasifikasi artikel wikipedia bahasa indonesia menggunakan metode *K-Nearest Neighbor* didapatkan hasil perhitungan centroid dalam setiap sub kategori terbagi atas tiga artikel yang diambil nilai tengahnya dan dilakukan perhitungan untuk jarak kedekatan pada setiap data uji. Sesuai dengan perolehan pengujian manual yang dilakukan peneliti, maka peneliti mendapatkan akurasi kebenaran sebesar 60% [6].

Berdasarkan latar belakang yang telah dijelaskan, maka judul penelitian ini adalah klasifikasi teks dalam dokumen berita olahraga dengan menggunakan metode *K-nearest neighbor (KNN)*. Diharapkan sistem dapat mengklasifikasi hasil secara akurat.

1.2 Perumusan Masalah

Ada 3 poin penting di dalam perumusan masalah, yaitu masalah penelitian, metode usulan dan pertanyaan penelitian.

1.2.1 Masalah Penelitian

Dengan banyaknya berita olahraga yang disajikan oleh penyedia situs maka dibutuhkan pencermatan lebih untuk mendapatkan informasi berita olahraga sesuai dengan yang diinginkan.

1.2.2 Metode Usulan

Metode yang diusulkan penelitian ini yaitu *K-Nearest Neighbor (KNN)* menggunakan seleksi fitur *Information Gain* dalam mengklasifikasikan teks dokumen berita olahraga.

1.2.3 Pertanyaan Penelitian

Berdasarkan permasalahan dan metode yang telah dijelaskan, maka didapatkan pertanyaan penelitian yaitu berapa tingkat akurasi untuk metode klasifikasi *K-Nearest Neighbor (KNN)* dalam mengklasifikasikan teks dokumen berita olahraga?

1.3 Tujuan dan Manfaat Penelitian

1.3.1 Tujuan Penelitian

Tujuan dari penelitian ini untuk mengetahui tingkat akurasi dalam mengklasifikasikan teks dokumen berita olahraga dengan menggunakan metode *K – Nearest Neighbor (KNN)*.

1.3.2 Manfaat Penelitian

Manfaat dari penelitian dapat dijadikan sebagai alternatif rujukan dalam membangun sistem pengklasifikasian yang lebih baik sehingga sistem yang dibangun diharapkan dapat berjalan maksimal dalam mengklasifikasikan teks dokumen berita olahraga dengan menggunakan metode *K – Nearest Neighbor (KNN)*.

1.4 Batasan Masalah

Batasan masalah dalam penelitian ini yaitu :

1. Sumber data yang akan digunakan dalam penelitian ini yaitu artikel berita olahraga pada koran online <https://www.tempo.co/>
2. Data dokumen yang digunakan yaitu 7 kelas antara lain sepakbola, bulutangkis, voli, renang, atletik, basket, dan motogp.
3. Data dokumen yang digunakan dari tahun 2017 hingga tahun 2021.
4. Dataset yang digunakan sebanyak 350 data.
5. Masing-masing kelas menggunakan 50 data.

6. Pada 1 kelas menggunakan data training sebanyak 40 data dan data yang digunakan untuk testing sebanyak 10 data.

1.5 Sistematika Penulisan

Sistematika penulisan skripsi ini adalah sebagai berikut :

BAB I PENDAHULUAN

Pada bab ini membahas mengenai Latar Belakang, Rumusan Masalah, Tujuan dan Manfaat, Batasan Masalah, serta sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini membahas mengenai tinjauan pustaka berupa teori yang digunakan dalam penelitian serta acuan penelitian berupa rangkuman penelitian sebelumnya.

BAB III METODE USULAN

Pada bab ini membahas mengenai arsitektur sistem, flowchart, dataset, tahapan penelitian dan pengujian.

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini membahas mengenai lingkungan pengembangan, pembuatan sistem dan hasil pengujian yang dilakukan.

BAB V PENUTUP

Pada bab ini membahas mengenai kesimpulan dari penelitian yang dilakukan serta saran untuk pengembangan pada penelitian selanjutnya.