

BAB I

PENDAHULUAN

1.1 Latar Belakang

Menurut situs pemeringkatan *website* dunia alexa.com situs berita di Indonesia seperti Tribunnews.com dan Detik.com menempati posisi 5 teratas untuk *top sites* setelah Google.co.id dan Google.com [1]. Untuk situs seperti Bolasport.com yang menyediakan berita bola dan olahraga terlengkap menempati urutan 41 dari 500 *website* yang dirangkum oleh alexa.com per tanggal 14 Maret 2018. Dengan fakta tersebut dapat disimpulkan bahwa orang Indonesia cukup menggemari berita olahraga.

Kegemaran orang Indonesia terhadap berita olahraga ini membuat semakin banyaknya berita olahraga. Namun cara penyajian berita yang dilakukan oleh setiap situs memiliki kesamaan yaitu disajikan berbentuk kategori sesuai dengan cabang olahraga yang ada [1]. Seperti pada situs Tribunnews.com berita olahraga memiliki 7 kategori, sport.detik.com memiliki 6 kategori, dan sports.okezone.com memiliki 5 kategori. Rata-rata setiap situs menyediakan 5 kategori untuk berita olahraga.

Permasalahan yang timbul dengan penyajian berita seperti ini adalah proses pengkategorian berita dengan jumlah besar. Apabila pengkategorian dilakukan secara manual akan membutuhkan waktu yang lama [2]. Sehingga pengkategorian yang otomatis dibutuhkan untuk menyelesaikan permasalahan tersebut. Hal ini bisa dilakukan dengan proses klasterisasi [2].

Klasterisasi adalah proses pengelompokkan dengan *unsupervised learning* pada objek yang berupa hasil pengamatan, data item, atau *feature vectors* sehingga terbentuk beberapa kelompok [3]. Sebagai contoh koleksi berita olahraga dari berbagai cabang olahraga dapat dikelompokkan berdasarkan kemiripan dokumen menjadi satu kategori tanpa menentukan kategori tersebut sebelumnya.

Banyak metode yang dapat digunakan untuk proses klasterisasi. Salah satu metode yang sering digunakan yaitu metode *k-means* [3]. Metode ini sering digunakan karena sederhana dan mudah diimplementasikan. Metode *k-means* adalah metode klasterisasi secara partisi. Hal itu memberikan keuntungan bagi *k-means* untuk menangani dataset yang besar dengan mudah [3]. Dengan keuntungannya ini *k-means* memiliki waktu komputasi yang singkat. Dari kelebihan-kelebihan tersebut, *k-means* tetap memiliki beberapa permasalahan.

Permasalahan terbesarnya yaitu cara penanganannya terhadap pembuatan titik pusat *cluster* yang kurang baik. Hasil pengelompokkan sangat dipengaruhi oleh pembuatan titik pusat *cluster* ini [3].

Titik pusat *cluster* pada *k-means* dibuat secara acak. Apabila titik pusat *cluster* terdefinisi dengan jarak yang dekat maka hasil akhir akan tidak optimal [3]. Namun apabila setiap titik pusat *cluster* tersebar merata maka hasil akhir yang didapatkan akan optimal. Hal ini dapat diatasi dengan melakukan pencegahan agar titik pusat *cluster* tidak terinisialisasi berdekatan. Pencegahan dilakukan dengan metode *cosine similarity* yang berfungsi untuk membandingkan 2 titik pusat *cluster* yang menghasilkan nilai antara 0 dan 1 [4]. Dua objek disebut identik atau berdekatan dengan yang lain apabila memiliki nilai 1 [4]. apabila tidak identik atau berjauhan dengan yang lain maka nilainya 0.

Dengan demikian diharapkan pada penelitian ini akan menghasilkan titik pusat *cluster* yang baik untuk mengelompokkan berita olahraga berbahasa Indonesia, dan diharapkan juga metode *k-means* yang dimodifikasi ini dapat memberikan hasil yang lebih baik dari *k-means* yang biasa.

1.2 Perumusan Masalah

Rumusan masalah yang didapatkan dari latar belakang tersebut yaitu:

1. Berapa jumlah *cluster* yang dapat memberikan akurasi yang optimal dalam klasterisasi berita olahraga berbahasa Indonesia?
2. Apakah metode *k-means* yang dimodifikasi dengan metode *cosine similarity* lebih baik dari metode *k-means* biasa untuk mengklaster berita olahraga berbahasa Indonesia?

1.3 Tujuan dan Manfaat

Dari rumusan masalah tersebut maka dapat diperoleh tujuan dan manfaat sebagai berikut:

1.3.1 Tujuan

Tujuan yang akan dicapai dari penelitian ini berdasarkan perumusan masalah yang telah disebutkan yaitu:

1. Untuk mengetahui jumlah *cluster* yang dapat memberikan akurasi yang optimal dalam klasterisasi berita olahraga berbahasa Indonesia.
2. Untuk mengetahui perbandingan *cluster* terbaik antara metode *k-means* yang dimodifikasi dengan metode *k-means* biasa untuk mengklaster berita olahraga berbahasa Indonesia.

1.3.2 Manfaat

Manfaat dari penelitian ini yaitu untuk memberikan kemudahan bagi pengguna untuk mengelompokkan berita olahraga sesuai dengan kesamaan isi berita.

1.4 Batasan Masalah

Adapun agar penelitian ini tetap berjalan sesuai dengan jalur yang ditentukan dan tidak terjadi penyimpangan, maka diberlakukan batasan-batasan masalah sebagai berikut:

1. Data penelitian berupa dokumen berita olahraga berbahasa Indonesia,
2. Data penelitian yang diambil berjumlah 337 berita.
3. Aplikasi yang dibangun adalah aplikasi berbasis web.
4. Data penelitian bersumber dari beberapa website yang telah dipilih yaitu <https://sports.sindonews.com>, <https://sport.detik.com>, <https://sports.okezone.com>, dan <https://olahraga.kompas.com>.

1.5 Sistematika Penulisan

Adapun sistematika pembahasan tugas akhir ini adalah sebagai berikut :

BAB I PENDAHULUAN

Pada bab ini berisikan latar belakang tentang masalah yang diangkat pada penelitian, uraian tujuan penelitian, manfaat yang didapat setelah penelitian dilakukan, serta batasan masalah untuk ruang lingkup skripsi.

BAB II TINJAUAN PUSTAKA

Bab ini berisi tinjauan pustaka dan landasan teori. Tinjauan pustaka berisi pustaka singkat dan jelas atas pustaka yang mendasari bidang kajian sebelumnya. Sedangkan landasan teori berisi teori yang mendukung penelitian.

BAB III PERANCANGAN SISTEM

Berisi tentang tahapan-tahapan penelitian yang ditempuh dalam skripsi ini. Tahapan tersebut yaitu :

1. Studi literatur dan tinjauan pustaka untuk mengetahui dan mempelajari teknik pengelompokkan berita secara otomatis. Hal-hal yang perlu ditinjau yaitu tahapan *preprocessing* dokumen, kemudian metode *k-means* untuk mengelompokkannya
2. Setelah studi literatur dilakukan kemudian dilakukan pengumpulan data. Data yang digunakan yaitu dokumen olahraga yang diambil dari beberapa media *online*. Jumlah dokumen yang digunakan sebanyak 337 dokumen berita.



3. Kemudian dilakukan perancangan dan implementasi sistem. Mulai dari kebutuhan *software* dan *hardware* yang dibutuhkan, hingga uji coba sistem yang telah ditentukan.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Pada bab ini menjelaskan tentang hasil dari penelitian yang dilakukan. Hasil yang diperoleh berupa kekurangan, kelebihan, dan uji coba yang telah dilakukan.

BAB V PENUTUP

Bab ini berisi kesimpulan dari hasil uji coba serta saran yang dapat digunakan untuk pengembangan penelitian berikutnya.

