

BAB I

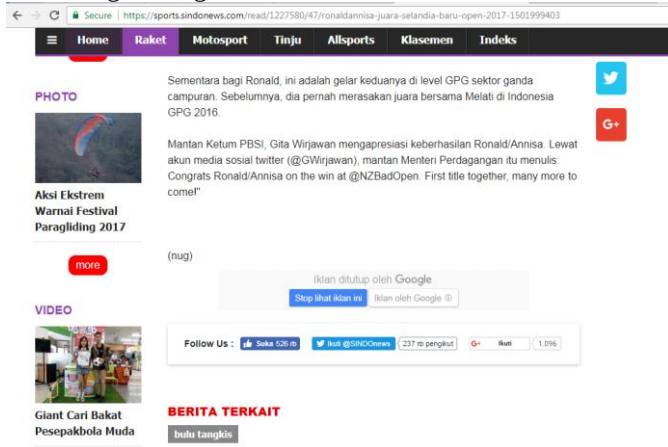
PENDAHULUAN

1.1. Latar Belakang

Dalam buku Sumadiria 2005 (hal 64-65) *Assegaf* mendefinisikan berita sebagai laporan mengenai fakta atau ide yang mampu menarik perhatian pembaca. Sedangkan menurut *Mitchel V. Charnley* dalam bukunya yang berjudul *Reporting* edisi ketiga (*Holt-Reinhart & Winston*, New York, 1975 halaman 44) menyatakan bahwa berita merupakan laporan mengenai fakta atau opini yang memiliki daya tarik atau hal penting bagi banyak orang. Kemudian *A.M. Hoeta Soehoet* dalam bukunya “Dasar-Dasar Jurnalistik” juga menuliskan bahwa berita merupakan keterangan mengenai peristiwa atau pendapat seseorang yang dianggap penting bagi pembaca untuk memenuhi kebutuhan hidupnya. Dari pengertian berita menurut tiga ahli di atas dapat disimpulkan bahwa *berita merupakan fakta ataupun opini yang membuat banyak orang merasa tertarik untuk mengetahuinya*. Berita bisa diperoleh dengan berbagai media seperti koran, surat kabar, televisi, *internet* dan lain-lain. Pada saat ini, media yang paling sering digunakan untuk memperoleh berita adalah *internet*.

Berita yang ada di *internet* memiliki berbagai macam topik. Salah satunya yaitu berita olahraga. Berita olahraga merupakan salah satu berita yang memiliki rating yang tinggi. Hal ini dibuktikan pada alexa.com di mana situs berita olahraga seperti sport.detik.com dan bola.net masuk ke dalam 25 top site di Indonesia. Sehingga berita olahraga termasuk berita favorit bagi warga Indonesia. Dalam penyajian berita, hampir semua situs memiliki cara yang sama. Berita dibagi menjadi beberapa kategori sesuai dengan cabang olahraga. Pada situs <http://sport.detik.com>, kategori yang digunakan sebanyak enam cabang olahraga yang meliputi: Basket, Raket, Formula 1, MotoGP, sepak bola dan sport lain. Pada situs <http://sport.okezone.com>, kategori yang digunakan sebanyak lima cabang olahraga yang meliputi F1, MotoGP, Netting, Basket dan Sport lain. Kemudian pada situs <http://sports.sindonews.com>, kategori yang digunakan Raket,

MotoSpot, Tinju dan all sport. Dari ketiga situs di atas, dapat disimpulkan bahwa rata-rata setiap situs menyediakan kurang lebih lima cabang olahraga.



Gambar 1.1 Screenshot Potongan berita dari situs sports.sindonew.com



Gambar 1.2 Screenshot Potongan berita dari situs sport.detik.com

Namun, masalah yang timbul di lapangan adalah cara penyedia situs mengelompokkan berita. Pengelompokan masih dilakukan secara manual sesuai dengan kategori yang sudah ditentukan sebelumnya. Pada gambar 1.1. dan gambar 1.2. dapat dilihat bahwa setiap berita memiliki satu tag kategori. Pada gambar 1.1, berita tersebut dikaitkan dengan kategori bulu tangkis. Kemudian pada gambar 1.2, berita tersebut dikaitkan dengan kategori moto-gp. Sehingga akan menjadi suatu permasalahan jika berita yang akan dikelompokkan berjumlah cukup banyak. Apabila pengelompokan dilakukan secara manual akan membutuhkan waktu yang lama. Oleh karena itu dibutuhkan cara pengelompokan secara otomatis. Hal ini bisa dilakukan dengan menggunakan *text mining*.

Text mining merupakan sebuah proses *unsupervised learning* untuk mengelompokkan kemiripan suatu dokumen dengan dokumen yang lain sehingga dapat dipisahkan menjadi beberapa kelompok[2]. Salah satu contoh *text mining* adalah *clustering*. *Clustering* dapat diartikan sebagai salah satu teknik *text mining* yang digunakan untuk pengelompokan dokumen di mana dokumen dikelompokkan dengan konten yang memiliki kemiripan berita tanpa mendefinisikan kategori sebelumnya[3]. Misalnya koleksi dokumen yang berisi berita sepak bola dan bulu tangkis dapat dikelompokkan sedemikian rupa sehingga semua berita sepak bola dikelompokkan menjadi satu *cluster* sepak bola dan semua berita bulu tangkis dikelompokkan menjadi satu *cluster* bulu tangkis.

Untuk mengelompokkan berita dengan *clustering* dapat menggunakan beberapa metode. Salah satu metode yang sering digunakan adalah *k-means*. Metode *k-means* merupakan salah satu algoritma *cluster* yang sering digunakan karena algoritma ini cukup sederhana. Metode *k-means* mampu mengelompokkan data dengan jumlah yang cukup besar dan waktu komputasi yang cukup singkat. Namun, *k-means* memiliki kelemahan pada penentuan pusat *cluster* awal. Karena *cluster-cluster* akhir yang dihasilkan sangat dipengaruhi oleh *cluster* awal yang telah didefinisikan.[4]

K-medoid merupakan salah satu metode *clustering* yang memiliki efisiensi untuk menangani data set yang kecil. *K-medoid* memiliki kemiripan dengan *k-means*. Karena kedua

metode ini tergolong pada metode *partitional clustering*. Namun, *K-meoid* lebih unggul dibandingkan dengan *k-means* karena *k-means* digunakan untuk *data set* dengan jumlah yang besar.[5] Ketika *data set* yang digunakan adalah data yang kecil, maka data yang dihasilkan kurang optimal. Sehingga penelitian ini menggunakan metode *K-meoid* untuk menangani masalah tersebut.

K-medoid memiliki kekurangan yang sama dengan *k-means*. Di mana pada saat penginisialisasian *cluster* awal sangat mempengaruhi hasil *cluster*. Karena metode ini langsung mengambil secara acak sebanyak *cluster* yang diinginkan oleh pengguna. Untuk menangani hal itu, penelitian ini menambahkan proses pencegahan kesamaan atau tingkat kemiripan yang tinggi pada saat penginisialisasian *cluster* awal. Sehingga inisialisasi *cluster* awal memiliki tingkat kemiripan yang rendah. Selain itu, jumlah *cluster* yang dimasukkan pengguna juga berpengaruh dalam pengujian akurasi hasil *cluster*. Sehingga penelitian ini diharapkan menghasilkan jumlah *cluster* yang optimal untuk mengelompokkan berita olahraga berbahasa Indonesia dengan akurasi yang tinggi. Selain itu, penelitian ini juga diharapkan pengelompokkan berita olahraga dengan metode *k-medoid* bersyarat mendapatkan hasil lebih baik dibandingkan dengan menggunakan metode *k-medoid* biasa.

1.2. Perumusan Masalah

Berdasarkan latar belakang di atas, maka timbul rumusan masalah yaitu:

1. Berapa jumlah *cluster* yang dapat digunakan untuk menghasilkan akurasi yang paling optimal dalam pengelompokan berita olahraga berbahasa Indonesia?
2. Apakah metode *k-medoid* bersyarat lebih baik dibandingkan dengan metode *k-medoid* dalam mengelompokkan berita olahraga?

1.3. Tujuan dan Manfaat

Berdasarkan rumusan masalah di atas maka tujuan dan manfaat yang dapat diambil dari penelitian ini adalah sebagai berikut:

1.3.1. Tujuan

Berdasarkan perumusan masalah di atas, maka tujuan yang ingin dicapai dalam penelitian ini yaitu:

1. Untuk mengetahui jumlah *cluster* yang digunakan untuk menghasilkan akurasi yang optimal dalam pengelompokan berita olahraga berbahasa Indonesia.
2. Untuk mengetahui cluster terbaik dari hasil perbandingan menggunakan metode k-medoid bersyarat dengan menggunakan metode k-medoid.

1.3.2. Manfaat

Manfaat dari penelitian ini yaitu untuk mempermudah pengguna untuk mengelompokkan berita olahraga sesuai dengan kemiripan konten dalam berita tersebut.

1.4. Batasan Masalah

Adapun batasan masalah dalam penelitian ini agar tetap terarah dan tidak menyimpang dari apa yang sudah direncanakan adalah sebagai berikut:

1. Dokumen berupa dokumen olahraga berbahasa Indonesia.
2. Dokumen yang diuji sebanyak 25 dokumen.
3. Aplikasi yang dibangun merupakan aplikasi berbasis web.
4. Dokumen didapat dari website (<http://sport.detik.com>, <http://sports.okezone.com>, <http://sports.sindonews.com>, dan <http://sport.idntimes.com>).
5. Data yang digunakan merupakan data yang sudah disediakan dalam sistem dan tidak ada proses pembaruan berita.

1.5. Metodologi Penelitian

Secara garis besar tahapan metodologi penelitian yang digunakan dalam penelitian skripsi ini adalah sebagai berikut :

1. Studi literatur dan tinjauan pustaka

Studi literatur perlu dilakukan oleh peneliti untuk mengetahui bagaimana cara mengelompokkan berita dengan benar. Peneliti perlu mempelajari beberapa pengetahuan sebagai dasar dalam pengerjaan penelitian ini. Pengetahuan yang perlu dipelajari sebagai berikut:

a. *Clustering* beserta tahapan-tahapannya

Clustering merupakan salah satu cara untuk mengelompokkan dokumen atau berita. Tahap-tahap melakukan *clustering* yaitu pengambilan berita, *preprocessing*, dan memilih metode yang tepat. Pengambilan data bisa dilakukan dengan dua cara. Cara pertama yaitu dengan cara langsung mengambil berita dari situs berita olahraga. Sedangkan cara yang kedua yaitu dengan cara memanfaatkan *crawler*. *Crawler* merupakan salah satu cara mendapatkan dokumen dengan mengunduh secara otomatis menggunakan aplikasi. Setelah berita didapatkan, maka akan dilanjutkan ke proses *preprocessing*, yaitu tahap mengubah dokumen ke dalam format yang sesuai agar dapat diproses pada klusterisasi. Hasil yang didapatkan dari tahap *preprocessing* berupa kumpulan kata atau istilah yang memiliki bobot tertentu. Setelah berita diubah menjadi kumpulan kata, maka perlu dilanjutkan dengan memilih metode *clustering*.

b. Metode *K-medoid*

K-medoid merupakan salah satu metode *clustering* yang dilakukan dengan cara pengambilan beberapa berita sebagai *medoid*. Beberapa *medoid* tersebut akan mencari jarak terhadap semua berita yang ada. Berita akan dikelompokkan sesuai dengan *medoid* yang memiliki jarak terdekat.

2. Pengumpulan Data

Data yang digunakan sebagai sampel adalah dokumen berita olahraga yang berhasil diambil dari media *online* seperti <http://sport.detik.com>, <http://sports.okezone.com>, <http://sports.sindonews.com>, dan <http://sport.idntimes.com>. Adapun dokumen yang dikumpulkan sebanyak 25 dokumen. Analisis dan perancangan sistem menggunakan algoritma *k-medoid*.

3. Implementasi

a. Ekstraksi dokumen

Proses ekstraksi ini bertujuan untuk menghasilkan term-term yang akan digunakan sebagai *prototype* bagi setiap dokumen yang berisi bobot *term* pada kumpulan dokumen. Ekstraksi dokumen dilakukan pada tahap *preprocessing*. Pada tahap ini yang dilakukan adalah *case folding*, *stopwords*, *stemming* dan *term weighting*. Metode *term weighting* yang digunakan adalah TF-IDF yang merupakan kombinasi antara TF (*Term Frequency*) dengan IDF (*Invers Document Frequency*).

b. Perhitungan jarak

Perhitungan jarak yang digunakan pada penelitian ini adalah menggunakan metode *Euclidean Distance*.

c. Pengelompokan dokumen

Untuk melakukan pengelompokan dokumen algoritma yang digunakan adalah algoritma *k-medoid*. Pengelompokan dokumen dengan melakukan perbandingan jarak antara *closest medoid* dengan berita yang ada.

4. Uji Coba Sistem

Untuk mengetahui kinerja algoritma *k-medoid* pada tahap uji coba, dilakukan pengukuran berdasarkan kualitas *cluster*. Sistem akan menerima input berupa jumlah *cluster* yang akan dihasilkan.

5. Dokumentasi

Pada tahap ini dilakukan penyusunan laporan dan dokumentasi dari pelaksanaan secara keseluruhan, mulai dari tahap awal sampai tahap analisa dan evaluasi.

1.6. Sistematika Penulisan

Secara garis besar, Sistematika penulisan buku Skripsi ini adalah sebagai berikut :

BAB I : PENDAHULUAN

Pada bab ini menjelaskan tentang latar belakang penelitian, perumusan masalah, tujuan, manfaat, batasan

masalah, metode penelitian, serta sistematika penulisan dalam laporan tugas akhir ini.

BAB II : STUDI PUSTAKA

Bab ini membahas mengenai landasan teori-teori pokok yang digunakan dalam menyusun penelitian ini. Landasan teori yang dibahas meliputi teori umum dan teori khusus.

BAB III : PERANCANGAN SISTEM

Pada bab ini menjelaskan tentang rancangan sistem dan keseluruhan sistem yang dibuat beserta rancangan alur flowchart pada tiap-tiap proses pada sistem.

BAB IV : IMPLEMENTASI SISTEM

Pada bab ini menjelaskan tentang implementasi sistem berdasarkan rancangan pada BAB III. Dan menjelaskan hasil analisa dari sistem yang telah dibuat.

BAB V : PENUTUP

Pada bab penutup ini menjelaskan tentang kesimpulan dari hasil uji coba yang telah dilakukan serta saran yang menjadi masukan dalam pengembangan penelitian berikutnya.