BAB I
PENDAHULUAN

I.1. Latar Belakang Masalah

Text mining merupakan proses ekstraksi dari sejumlah besar data yang berbentuk text atau dokumen untuk menemukan informasi yang berguna, tersembunyi, dan tidak diketahui sebelumnya [1]. Tujuan dari *text mining* adalah mendapatkan informasi yang berguna dari sekumpulan dokumen. Sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format tidak terstruktur atau minimal semi terstruktur. Menurut penelitian dari Lokesh Kumar [2], *text mining* dan *information extraction* telah menjadi area populer penelitian untuk mengekstrak informasi yang menarik dan berguna. Jadi sangat penting mengembangkan teknik dan algoritma yang lebih baik untuk mengekstrak informasi yang berguna. Salah satu area yang dapat dilakukan *text mining* adalah pada *news aggregator.*

Dalam sistem *news aggregator*, pengelompokan berita memiliki penting karena setiap kelompok berita menyatakan satu topik berita yang anggotanya merupakan artikel-artikel berita dari berbagai portal berita. Kualitas kelompok berita sangat penting karena dapat membantu pembaca untuk memilih topik berita yang diinginkan. Pengelompokan berita berbahasa Indonesia telah dilakukan oleh beberapa peneliti dengan berbagai teknik dan tujuan. Metode mengelompokan berita dengan menggunakan *partitional clustering* dengan cluster yang diinisialisasi merupakan teknik yang paling sederhana dan umum digunakan untuk berita bahasa Indonesia [3,4] karena metode ini mudah diimplementasikan [4]. Metode ini melakukan *clustering* dengan meng-inisialisasi *cluster* terlebih dahulu, setiap *cluster* diinisialisasi secara *random* sehingga pengelompokkan data yang dihasilkan dapat berbeda-beda. Jika nilai *random* untuk inisialisasi kurang baik, maka pengelompokkan yang dihasilkan pun menjadi kurang optimal. Penggunaan metode *partitional cluster*ing dengan inisialisasi

cluster pada *news aggregator* masih menghasilkan *outlier* atau dokumen yang seharusnya tidak dalam satu *cluster* [5].

Berdasarkan penelitan tentang penerapan salah satu metode *partitional clustering* dengan inisialisasi cluster [5], dengan menerapkan metode *partitional clustering* pada news aggregator terkadang menghasilkan *over cluster* jika jumlah dokumennya semakin banyak. Hasil penelitian tersebut menyarankan untuk menggunakan jenis metode *clustering* lain yang tidak sensitif terhadap inisialiasi atau menggunakan metode lain yang dapat menentukan inisialisasi secara dinamis. Oleh karena itu, maka perlu dilakukan penelitian untuk menerapakan metode *clustering* yang dinamis untuk data berita pada *news aggregator*.

I.2. Rumusan Masalah

Berdasarkan latar belakang masalah, maka rumusan masalah pada penelitian ini adalah bagaimana menerapkan metode *clustering* yang dinamis pada data berita dalam *news aggregator* untuk meningkatkan akurasi setiap *cluster*.

I.3. Maksud dan Tujuan

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, maksud dari penelitian ini adalah membuat sistem *news aggregator* untuk mengelompokkan berita berdasarkan isi beritanya.

Sedangkan tujuan yang akan dicapai adalah menerapkan metode *clustering* yang dinamis untuk meningkatkan akurasi data setiap *cluster*.

I.4. Batasan Masalah

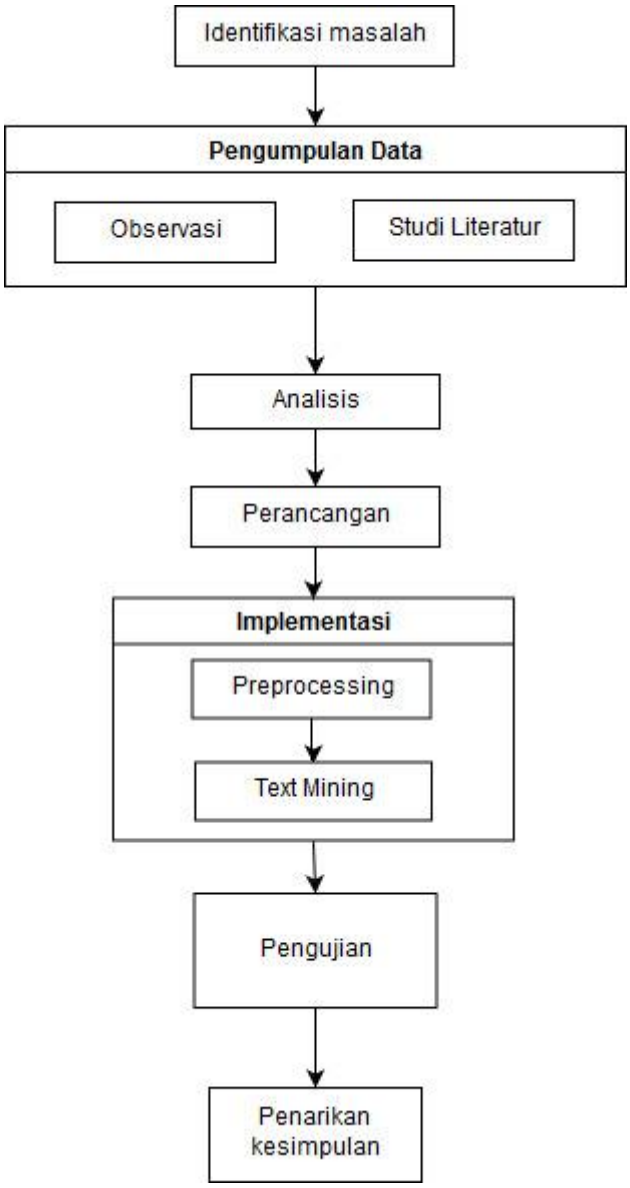
Adapun batasan masalah dalam *text mining* pada *news aggregator* ini adalah sebagai berikut:

a. Berita yang dikumpulkan berasal dari media berita yang berbahasa Indonesia.
b. Data masukan berupa teks dalam format html.
c. Data masukan langsung disimpan ke dalam database.

d. Tahap *preprocessing* akan dilakukan pada data masukan sehingga data masukan siap untuk diolah.
e. Tahap processing yang dilakukan adalah *cleaning, case folding, tokenizing,* dan *stopword removal.*
f. Metode pembobotan yang digunakan adalah *TF-IDF*.
g. Metode *text mining* yang digunakan adalah *cosine similarity* dan *single pass clustering.*
h. Data yang dihasilkan dari proses *text mining* berupa data yang sudah di *cluster.*
i. Data keluaran dalam bentuk teks berita yang berisi sumber, kategori, judul, isi, tanggal dan penulis.

I.5. Metode Penelitian

Metode penelitian yang digunakan dalam membangun ssitem ini adalah metode penelitian *deskriptif* [7]. Metode ini digunakan karena pada penelitian ini data sumber berasal dari internet dan tidak dilakukan manipulasi variabel pada data yang akan digunakan. Data yang digunakan merupakan data yang diperoleh apa adanya. Oleh karena itu, metode penelitian deskriptif dirasa cocok untuk digunakan pada penelitian ini. Tahapan penelitian yang dilakukan dapat dilihat pada Gambar I.1



Gambar I.1 Tahap Penelitian

I.5.1. Identifikasi Masalah

Tahap ini merupakan tahap yang pertama dilakukan, yaitu mengidentifikasi atau mengenali masalah yang terjadi pada sistem *news aggregator*.

I.5.2. Pengumpulan Data

Pengumpulan data pada penelitian ini dilakukan dengan dua tahap, diantaranya adalah sebagai berikut.

a. Observasi

   Tapahan ini dilakukan dengan cara mengumpulkan data dari beberapa portal berita bahasa Indonesia.

b. Studi Literatur

   Tahapan ini dilakukan dengan cara membaca beberapa jurnal dan buku yang sesuai dengan penelitian sebagai referensi.

I.5.3. Analisis

Pada tahap ini peneliti akan melakukan analisis dan perancangan sistem yang akan dibangun. Tahap ini meliputi dua bagian, diantaranya adalah sebagai berikut.

a. Analisis Masalah

   Melakukan analisis permasalahan yang menyebakan penelitian ini dilakukan, yaitu menganalisa masalah yang terjadi pada pengolahan data berita dan *news aggregator*.

b. Analisis Sistem

   Melakukan analisis pada sistem yang akan dibangun, pada tahap ini akan dilakukan analisis data masukan dan analisis proses.

I.5.4. Perancangan

Pada tahap ini peneliti akan melakukan perancangan sistem yang akan dibangun termasuk perancangan sistem, basis data, dan antarmuka.

I.5.5. Implementasi

Pada tahap ini peneliti akan melakukan implementasi pada sistem yang dibangun yaitu dengan melakukan tahap preprocessing dan text mining.

I.5.6. Pengujian

Pada tahap ini peneliti akan melakukan pengujian pada sistem yang dibangun. Tahap ini akan menguji fungsionalitas dan metode yang diterapkan pada sistem.

I.5.7. Penarikan Kesimpulan

Pada tahap peneliti akan menyimpulkan hasil dari sistem yang dibangun dan memberikan saran untuk pengembangan sistem kedepannya.

I.6. Sistematika Penulisan

Sistematika penulisan laporan ini dibagi dalam beberapa bab dengan pokok pembahasan secara umum sebagai berikut:

Bab I Pendahuluan

Berisi latar belakang masalah, Perumusan masalah, maksud dan tujuan, batasan masalah, metodologi penelitian beserta sistematika penulisan.

Bab II Landasan Teori

Berisi tentang teori pendukung, dan istilah -istilah yang digunakan.

Bab III Analisis Dan Perancangan Sistem

Berisi tentang analis dan perancangan sistem, Tahap Analisis Sistem mencakup Analisi masalah, Analisis data masukan, Analisis Kebutuhan Fungsional dan Analisis Kebutuhan NonFungsional. Tahap Perancangan Sistem mencakup Perancangan Aliran Data , Perancangan Data dan Perancangan Antarmuka.

Bab IV Implementasi Dan Pengujian Sistem

Berisi tentang implementasi dari hasil analisis dan perancangan yang telah dibuat, disertai juga dengan pengujian perangkat lunak yang dibangun.

Bab V Kesimpulan Dan Saran

Berisi tentang kesimpulan dan saran dari metode yang digunakan dan sistem yang dibuat.