BAB I
PENDAHULUAN

2.1 Latar Belakang

Seiring dengan perkembangan teknologi, jumlah data teks pun semakin banyak. Dengan didukung keberadaan internet menjadikan data teks mudah menyebar luas sampai dikenal dengan istilah *Big Data*. Walaupun jumlah data teks sangat banyak, tetapi data teks yang diolah untuk menjadi sebuah informasi yang memiliki pengetahuan masih sedikit [1].

Text mining adalah proses ekstraksi pola berupa pengetahuan dari sebagian besar jumlah data teks, data teks dapat berupa *paper*, berita, Al-quran, dan Hadist. [1]. *Text mining* merupakan teknik yang digunakan untuk menangani masalah *klasifikasi*, *clustering*, *information extraction*, dan *information retrieval* [1]. *Text mining* menerapkan konsep dan teknik *data mining* untuk mencari pola dalam teks, yaitu proses penganalisaan teks dengan mencari informasi yang bermanfaat untuk tujuan tertentu. Berdasarkan ketidakteraturan struktur data teks, maka proses *text mining* memerlukan beberapa tahap awal untuk mempersiapkan teks menjadi lebih terstruktur.

Untuk memperoleh informasi yang bermanfaat dalam proses text mining, harus dilakukan beberapa penelitian dan percobaan. Oleh karena itu, *text mining* telah banyak dikembangkan oleh para ilmuwan dalam bidang komputasi. Diantaranya adalah Charu .C Aggarwal, Philip S, dan Yuchen Zhao yang berjudul "*On Text Clustering with Side Information"*, dilakukan percobaan 50.080 jurnal, data menggunakan *informasi sampingan* dengan menggunakan algoritma *COATES* (*Content and Auxiliary Attribute Based on Text Clustering*) dan menghasilkan algoritma *clustering* yang dianggap effisien untuk proses *clustering* dengan memanfaatkan *informasi sampingan* [12]. Pada Jurnal *"An Effective Clustering Approach for Mining Text Data Using Side Information"* yang ditulis oleh Monica. M, dan Ganesh.J, mengenalkan Algoritma *COATES* untuk *clustering* dan *COLT* untuk *klasifikasi* [8].

Pada Jurnal yang ditulis oleh Shilpa S. Raut dan V. Maral yang berjudul "*Text Clustering and Classification on The Use of Side Information*",

Pengenalan Algoritma *COATES* yang dianggap sangat efektif dalam penggunaan *informasi sampingan* [9].

Jurnal yang ditulis oleh Neha Tiwari dan Gaima Singh yang berjudul "*A Framework For Mining Of Text Data With The Application Of Side Information*" ditambahkan fungsi *JACARD* untuk menghitung jarak minimumnya dalam algoritma *COATES* [6].

Mrunal V. Uspani, dan Rucha C. Samant menulis jurnal yang berjudul "*Meta Information Based On Text Clustering and Classification with the Use of COATES and COLT Algorithm*", dan " *Clustering and Classification based on Meta Information using COATES and COLT Algorithm*", dikembangkan *informasi sampingan* yang berupa meta informasi yang digunakan untuk membantu proses *clustering* dan klasifikasi [11, 7].

Nikhil Patankar dan Sailee Salkar didalam jurnalnya yang berjudul "*On the use of Side Information Based Improved K-means Algorithm for Text Clustering*" telah dikembangkan penggunaan *informasi sampingan* berdasarkan pada algoritma *COATES* dengan menggunakan algoritma *k-means* dan menggunakan data sebagai objek percobaannya [10].

Dijelaskan oleh Shraddha S. Bhanuse, Shailes D. Kamble, Sandeep M. Kakde didalam jurnalnya yang berjudul "*Text mining using Metadata for Generation of Side Information*", dijelaskan bahwa *metadata* merupakan bagian dari *informasi sampingan* dan merupakan meta informasi dari data yang bersifat informatif yang dapat membantu proses *text mining* dengan *clustering*. Oleh karena itu dalam paper tersebut diusulkan untuk mengaplikasikan teknik *clustering* dengan menggunakan algoritma *COATES* pada *metadata* dalam paper sebagai *informasi sampingan* data. Dijelaskan bahwa *metadata* dapat berupa judul, abstrak, *publisher*, *keyword* dari sebuah paper [1].

Alasan utama untuk merancang algoritma *clustering* dalam *text mining* yang efektif adalah dengan meningkatnya jumlah data tekstual. Dalam penambangan teks, banyak masalah yang diangkat karena beberapa domain aplikasi seperti informasi web, data digital, dan jaringan yang berbeda dalam domain ini, sejumlah besar *informasi sampingan* dikaitkan dengan dokumen. Cukup sulit untuk menghitung pentingnya *informasi sampingan*, karena

penggabungan *informasi sampingan* dapat mempengaruhi kualitas proses penambangan. Untuk itu adanya ruang lingkup perbaikan yaitu mengambil *metadata* sebagai *informasi sampingan*, dimana *metadata* ini mencakup sebagian besar informasi dari sebuah data.

Metadata adalah informasi terstruktur yang mendeskripsikan, menjelaskan, menemukan, atau setidaknya menjadikan suatu informasi mudah untuk ditemukan kembali, digunakan, atau dikelola. Dengan pengertian sederhana, *metadata* adalah informasi yang ditanam pada sebuah *file* yang isinya berupa penjelasan tentang *file* tersebut [1, 2]. *Metadata* dalam hadist dapat berupa sanad dari sebuah hadist. *Metadata* dalam sebuah hadist bersifat informatif karena dapat merepresentasikan informasi matan hadist yang memiliki sanad yang berbeda.

Berdasarkan uraian tersebut disusun suatu laporan tugas akhir yang berjudul "*Clustering* Hadist dengan Menggunakan Algoritma *COATES*" dengan menggunakan pendekatan yang memastikan pengelompokan karakteristik dari *informasi sampingan* dengan isi teks, hal ini akan memperbesar efek pengelompokan keduanya. Jenis data inti dari pendekatan ini adalah untuk menentukan pengelompokan dimana atribut dan *informasi sampingan* teks memberikan petunjuk yang sama tentang sifat pengelompokan yang mendasarinya, dan mengabaikan aspek-aspek di dalamnya.

Untuk mencapai tujuan tersebut, akan digunakan algoritma *Content and Auxiliary Attribute Based on Text Clustering* (*COATES*) dibantu dengan metode partisi klasik yang dikombinasikan dengan model probabilistik untuk mengelompokan data teks berdasarkan *cluster*. Pada *informasi sampingan*, proses evaluasi probabilistik mempartisi informasi untuk mengevaluasi berbagai lampiran dalam dokumen. Tujuannya adalah untuk memanfaatkan *informasi sampingan* yang berupa *metadata* dalam proses *text mining* dengan menggunakan algoritma yang efisien untuk masalah pengelompokan data berdasarkan *cluster* dan mengetahui algoritma yang efisien untuk digunakan sebagai inisialisasi *cluster*.

1.2 Rumusan Masalah

Adapun rumusan masalah dalam Tugas Akhir ini adalah:

1. Apa yang dimaksud dengan *metadata* dalam sebuah hadist?
2. Bagaimana memanfaatkan *metadata* dalam sebuah hadist dalam proses *clustering* ?
3. Bagaimana hasil *cluster* hadist yang dibentuk dengan menggunakan algoritma *Content and Auxiliary Attribute Based on Text Clustering*?

1.3 Batasan Masalah

Agar penelitian tetap fokus, maka dibatasi masalah sebagai berikut:

1. *Dataset* yang digunakan merupakan hadist Al-Muwwatta berbahasa inggris.
2. Metode *clustering* yang digunakan untuk inisialisasi *cluster* adalah *k-means*.
3. Metode untuk penentuan *centroid* yang digunakan adalah *neighbors link*.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dari Tugas Akhir ini adalah:

1. Dapat memanfaatkan *"informasi sampingan"* berupa *metadata* dari hadist Al-Muwwatta dalam proses *text mining*.
2. Dapat meng*clustering* data hadist Al-Muwwatta dengan menggunakan algoritma *COATES*.
3. Dapat mengetahui pengaruh hasil *cluster* yang dibentuk dengan menggunakan algoritma *COATES* dengan menggunakan inisialisasi *cluster* berdasarkan *k-means*.
4. Dapat mengetahui pengaruh *informasi sampingan* berupa *metadata* dari hadist Al-Muwwatta dalam proses *clustering*.

1.5 Metode Penelitian

Metode penelitian yang digunakan dalam penelitian Tugas Akhir ini adalah sebagai berikut:

a. Studi Literatur

Pengumpulan bahan-bahan referensi yang mendukung pengerjaan penelitian, mulai dari *text mining*, *clustering*, algoritma *COATES*, *text pre-processing*, dan algoritma inisialisasi *cluster* menggunakan *k-means*.

b. Analisis

Proses analisis ini menjadi salah satu metode utama yang dilakukan selama penelitian berlangsung. Diawali dengan analisis dari kondisi data secara real atau data sebenarnya, kemudian menganalisis setiap proses yang dilewati data dari mulai pengambilan data sampai proses *clustering* data.

c. Pembuatan *pseudocode* dengan python

Pada tahap ini akan dilakukan proses implementasi pembuatan pseudocode program dalam aplikasi komputer menggunakan bahasa pemograman yang telah ditentukan yaitu python. Pseudocode ini terdiri dari dua bagian yaitu:

1) *Pseudocode* untuk membaca data hadist Al-Muwwatta
2) *Pseudocode* untuk meng*cluster* hadist Al-Muwwatta

1.6 Sistematika Penulisan

Sistematika penulisan Tugas Akhir ini hanya memuat 5 bab. Dengan rincian sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini akan dipaparkan tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian serta sistematika penelitian dari masalah yang akan di kaji.

BAB II LANDASAN TEORI

Pada bab ini penulis akan memaparkan dari landasan teori yang dijadikan ukuran untuk membahas yang menjadi dasar teori pada masalah yang akan dibahas diantaranya *text mining*, *metadata*, *clustering*, algoritma *COATES*, *k-means*, *pre-processing*, *pembobotan* , dan pyhton.

BAB III PROSES *CLUSTERING* HADIST MENGGUNAKAN ALGORITMA *COATES*

Pada bab ini akan dipaparkan proses *clustering* hadist dari mulai tahap pengumpulan data, *text pre-processing*, pembobotan kata, inisialisasi *cluster* dengan algoritma *k-means*, sampai proses *clustering* dengan menggunakan algoritma *COATES*.

BAB IV ANALISIS HASIL *CLUSTERING* HADIST MENGGUNAKAN ALGORITMA *COATES*

Pada bab ini akan dipaparkan mengenai analisis hasil dari proses *clustering* yang meliputi analisis *data set* yang digunakan, tahap *text pre-processing*, pembobotan, hasil inisialisasi *cluster*, dan hasil *clustering*.

BAB V PENUTUP

Pada bab ini akan dipaparkan kesimpulan sebagai jawaban dari rumusan masalah yang diajukan serta saran untuk pengembangan tulisan yang berbeda dalam penulisan selanjutnya yang akan melanjutkan analisis untuk masalah yang telah dipaparkan.