BAB I
PENDAHULUAN

1.1 Latar Belakang

Berkembangnya media cetak dalam memberikan informasi dari masa ke masa terasa sangat pesat seiiring perkembangan teknologi. Informasi tentunya tidak hanya didapatkan melalui media cetak, melainkan menggunakan teknologi internet. Seiring dengan meningkatnya pengguna internet di era digital ini, banyak pihak yang secara bebas dapat memberikan informasi sehingga bertumpuk diinternet.

Informasi merupakan fungsi penting yang dapat mempengaruhi atau menambah pengetahuan seseorang. Oleh karena itu, kemudahan dalam menemukan informasi sangatlah diperlukan, agar seseorang dapat menemukan informasi yang relevan.

Pencarian informasi dalam Ilmu Komputer, dilakukan dalam bentuk teknologi sistem temu kembali informasi, atau *Information Retrieval System* (IRS) [1]. Salah satu implementasi dari IRS adalah mesin pencari informasi atau *search engine*. Mesin ini memudahkan pengguna untuk mendapatkan informasi, sehingga pengguna hanya perlu mengetikkan kata kunci, kemudian mesin memberikan sekumpulan informasi yang relevan dengan kata kunci tersebut. Sebelum itu, perlu dilakukan pelatihan data atau dasar dari suatu sistem sehingga dapat membentuk suatu *search engine*.

Solusi dalam membentuk sistem dasar untuk mendapatkan informasi yang relevan sudah banyak dikembangkan melalui cabang ilmu *Data Mining* (Penggalian Data) [2]. *Data Mining* merupakan inti ilmu dalam pengambilan informasi. Terdapat anak cabang dari *Data Mining* yang secara khusus membahas mengenai data teks, yaitu *Text Mining* (Penambangan Teks) [3].

Pengambilan informasi dari data teks yang relevan (*Text Mining*), diantaranya dapat menggunakan Algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF) [4]. Pada metode ini, kata kunci ditentukan oleh kemunculan kata didalam suatu dokumen, sehingga timbul masalah hubungan antar kata dalam dokumen seperti sinonim dan polisemi. Kemudian *Latent Semantic Indexing* (LSI), LSI telah mampu mengatasi masakah sinonim dan polisemi [5]. Meski demikian tingkat pekerjaan LSI hanya sampai pada level kata dan dokumen. Pada

kenyataannya, jumlah dokumen dalam tumpukan dokumen yang disebut korpus sangatlah besar, sehingga hubungan antar dokumen di dalam korpus perlu dianalisis juga. Akhir-akhir ini banyak peneliti menggunakan *Latent Dirichlet Allocation* (LDA) untuk mendapatkan informasi yang relevan dari suatu dataset yang besar [6]–[9]. LDA dapat dikatakan hampir sempurna dalam melakukan tugas sebagai metode penyajian informasi, karena mampu mengatasi masalah internal dokumen yaitu masalah polisemi dan sinonim. Tidak hanya itu, LDA juga dapat mengatasi masalah intra dokumen karena dapat bekerja sampai dengan tingkat korpus [7]. Metode yang dipolulerkan oleh David Blei pada tahun 2003 ini merupakan metode berbasis topik yang menggunakan penalaran untuk menentukan topik dokumen [9].

Metode LDA banyak diterpakan pada data teks artikel berita berbahasa Indonesia dalam menentukan topik, salah satunya telah dipaparkan oleh R.Kusumaningrum, M.Ihsan Aji, dan lainnya pada jurnal "*Classification of Indonesian News Articles based on Latent Dirichlet Allocation*" pada tahun 2016 [10]. Pada jurnal tersebut menerapkan metode LDA untuk klasifikasi artikel berita dengan parameter alpha, betta, dan jumlah topik yang beragam sehingga mendapatkan nilai akurasi sebesar 70%. LDA juga dapat diaplikasikan kepada data dari sosial media seperti yang dilakukan oleh Muh. Fajriyanto pada jurnalnya yang berjudul "Penerapan Metode Bayesian dalam Model Latent Dirichlet Allocation di Media Sosial" pada tahun 2018 [11]. Data yang digunakan merupakan *tweet* yang berasal dari *platform twitter*. Penelitian tersebut bertujuan untuk mendapatkan probabilitas ketakutan terror bom yang terjadi pada tahun 2018 di Surabaya dengan probabilitas sebesar 0,10057. Selain itu, terkait dengan penelitian yang penulis lakukan, terdapat pada penelitian rekan kuliah penulis yaitu Nanda Priatna dengan judul "Analisis perbandingan Algoritma *Clustering K-Means* & *DBScan* untuk Data Teks Terjemah Hadits Menggunakan Ekstraksi Ciri *Hybrid*" pada tahun 2019 [12]. Pada penelitian tersebut, data terjemah hadits dikelompokkan menggunakan *K-Means* dan *DBScan.* Kemudian dibandingkan hasilnya untuk menemukan mana metode pengelompokkan yang lebih baik.

Pada penelitian ini, LDA akan diterapkan pada data teks terjemah hadits berbahasa inggris, yaitu terjemah Hadits Shahih Bukhari, Sahih Muslim, Abu-Daud, dan Malik's Muwatta sebanyak 903 hadits. Parameter yang digunakan untuk

mengevaluasi metode LDA yaitu *Precision, Recall, Accuracy*, dan *F-Measure*. Sehingga penulis tertarik melakukan penelitian yang berjudul **"IMPLEMENTASI ALGORITMA LATENT DIRICHLET ALLOCATION PADA DATA TEKS TERJEMAH HADITS BAHASA INGGRIS"**

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, maka dalam skripsi ini dibuat rumusan masalah sebagai berikut:

1. Bagaimana cara mendapatkan informasi menggunakan *Latent Dirichlet Allocation* (LDA)?
2. Bagaimana hasil evaluasi *Latent Dirichlet Allocation* (LDA) pada dataset terjemah hadits bahasa inggris?

1.3 Batasan Masalah

Adapun batasan masalah dalam skripsi ini sebagai berikut:

1. *Dataset* yang digunakan yaitu berupa terjemahan Hadits Shahih Bukhari, Shahih Muslim, Abu-Daud, dan Malik's Muwatta sebanyak 903 hadits dalam bahasa inggris. Terdapat lima kategori hadits yaitu: *Adzan, Wudlu, Zakat Knowledge,* dan *Tawheed*.
2. Metode yang digunakan untuk mendapatkan informasi yaitu *Latent Dirichlet Allocation* (LDA).
3. Metode evaluasi yang digunakan yaitu *Precision, Recall, Accuracy,* dan *F-Measure*.
4. Bahasa pemograman yang digunakan dalam penelitian yaitu *Python 3.6*.

1.4 Tujuan Penelitian

Tujuan dari skripsi ini adalah untuk:

1. Menentukan cara mendapatkan informasi menggunakan *Latent Dirichlet Allocation* (LDA).
2. Mendapatkan hasil evaluasi *Precision, Recall, Accuracy,* dan *F-Measure* pada dataset terjemah hadits bahasa inggris.

Adapun manfaat dari skripsi ini adalah sebagai berikut:

1. Pembaca dapat mengetahui cara mendapatkan informasi menggunakan metode *Latent Dirichlet Allocation* (LDA) yang dijalankan pada bahasa pemograman *Python 3.6.*
2. Pembaca dapat mengetahui hasil evaluasi *Latent Dirichlet Allocation* (LDA) pada dataset terjemah hadits bahasa inggris.

1.5 Metode Penelitian

Metode penelitian pada skripsi ini menggunakan pendekatan studi literatur. Pengkajian dilakukan dengan mencari referensi literatur berupa buku, jurnal, karya ilmiah, dan artikel yang berkaitan dengan metode-metode yang digunakan: Terjemah Hadits, algoritma *Latent Dirichlet Allocation* (LDA) dan evaluasi data teks (*Precision, Recall, Accuracy,* dan *F-Measure*).

1.6 Sistematika Penulisan

Berdasarkan sistematika penulisan, skripsi ini terdiri atas lima bab serta daftar pustaka di mana dalam setiap bab terdapat beberapa subbab:

BAB I PENDAHULUAN

Bab ini berisi tentang latar belakang penelitian, rumusan masalah dari pelitian Skripsi, batasan masalah yang membatasi penelitian ini, tujuan penelitian, metode penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bagian ini mengemukakan landasan teori yang menunjang dalam penulisan skripsi, seperti *information retrieval, feature extraction, text preprocessing, latent dirichlet allocation,* dan *gibbs sampling*.

BAB III IMPLEMENTASI ALGORITMA LATENT DIRICHLET ALLOCATION UNTUK DATA TEKS TERJEMAH HADITS DALAM BAHASA INGGRIS

Bab ini berisi pembahasan utama dari penelitian ini, yang diawali dari pengumpulan data, pra pengolahan teks, dan pengolahan teks menggunakan *Latent Dirichlet Allocation* (LDA).

BAB IV EVALUASI HASIL EKSTRAKSI TOPIK

Pada bab ini akan di paparkan mengenai analisis hasil pengolahan teks dan evaluasi yang sudah dilakukan di bab III. Klasifikasi dan evaluasi menggunakan *Precision, Recall, Accuracy,* dan *F-Measure*, kemudian menganalisis hasil evaluasi dan analisis parameter alpha dan beta.

BAB V PENUTUP

Bab ini berisi kesimpulan dari pembahasan yang telah dikaji. Selain itu peneliti memberikan saran untuk pengembangan lebih lanjut dari penelitian tersebut.

DAFTAR PUSTAKA