BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Pesatnya kemajuan teknologi memberi dampak pada seluruh aspek kehidupan secara langsung maupun tidak langsung, mulai dari aspek ekonomi, kesehatan, politik, tak terkecuali aspek pendidikan. Sebagai salah satu sarana penyelenggaraan pendidikan, perpustakaan merupakan institusi pengelola koleksi karya tulis, karya cetak, dan/atau karya rekam secara profesional dengan sistem yang baku guna memenuhi kebutuhan pendidikan, penelitian, pelestarian, informasi, dan rekreasi para pemustaka (UU No 43 tahun 2007 tentang Perpustakaan, 2007). Kemajuan teknologi informasi turut memberi perubahan pada layanan perpustakaan, dalam hal penyimpanan hingga pendayagunaan terhadap koleksinya (Hartono, 2017; Himmah & Azisi, 2019; Rahmawati, 2017). Hal itu dikarenakan koleksi seperti halnya laporan kegiatan hingga karya ilmiah yang dihasilkan masyarakat akademis telah bertransformasi menuju digital (Hartono, 2017).

Karya akhir merupakan karya ilmiah penelitian yang dihasilkan oleh mahasiswa sebagai syarat yang harus dipenuhi untuk memperoleh gelar akademik sesuai jenjangnya. Dalam proses penulisannya dibutuhkan banyak perujukan guna menguatkan basis teori pada penelitian yang akan dilakukan.

Kemutakhiran dan relevansi dari rujukan yang digunakan haruslah diperhatikan agar dapat menghasilkan karya ilmiah yang aktual (Rahma, 2017). Dengan menyimpan dokumen karya akhir dalam format digital pada layanan repositori institusional, dapat mempermudah pencarian rujukan oleh mahasiswa sebagai pemustaka yang nantinya akan digunakan dalam menulis karya ilmiah penelitiannya. Repositori institusional (*Institutional Repositories*) yang biasa disebut juga dengan repositori merupakan koleksi digital yang menangkap dan melestarikan hasil intelektual komunitas universitas berupa karya ilmiah (Crow, 2002). Penggunaan repositori merupakan sebagai tanggapan atas isu strategis yang dihadapi oleh institusi akademik. Isu yang dimaksud meliputi; (1) untuk melengkapi model penerbitan karya ilmiah yang ada dalam reformasi komunikasi, dan (2) berfungsi sebagai indikator nyata kualitas universitas sehingga dapat meningkatkan nilai publik, prestise, dan visibilitasnya. Repositori memiliki peran ganda berupa: (1) menangkap dan melestarikan *file* keluaran intelektual kolektif lembaga dan, (2) meningkatkan komunikasi ilmiah dengan menyediakan akses yang lebih besar pada keluaran tersebut (Richardson & Wolski, 2012). Penggunaan repositori oleh perpustakaan bertujuan untuk mempermudah dalam mengelola, menyimpan, dan mempublikasikan koleksi digital yang dimilikinya (Pramudyo & Hendrawan, 2018).

Selain sebagai penyimpan koleksi karya ilmiah, kemampuan repositori untuk menemukan informasi yang dicari oleh pemustaka juga sangat penting untuk diperhatikan. Oleh karena itu, repositori yang dapat mengambil informasi yang relevan dengan cepat dan efisien sangat dibutuhkan (Göker &

Davies, 2009). Semua kebutuhan tersebut akan terpenuhi apabila repositori dapat memberikan hasil pencarian yang relevan dengan kueri yang diketikkan oleh pemustaka, tanpa harus membaca sebagian bahkan seluruh dokumen sehingga mempercepat proses pencarian. Misal ketika pemustaka menggunakan kata "`sistem laravel`" sebagai kueri pencarian, maka repositori dapat menghasilkan keluaran dokumen karya akhir yang di dalamnya membahas topik seputaran "`sistem`" dan/atau "`laravel`". Seperti dokumen yang di dalamnya terdapat kata "`aplikasi`", "`software`" hingga "`informasi`" untuk kueri "`sistem`". Dan menghasilkan dokumen yang mengandung kata "`php`", "`website`" hingga "`framework`" untuk kueri "`laravel`", sehingga pemustaka tidak perlu lagi membuka dan membaca dokumen satu per satu.

Universitas Pendidikan Ganesha, yang disingkat Undiksha sebagai perguruan tinggi negeri terbesar di Bali Utara nampaknya telah menggunakan repositori sebagai layanan penunjangnya. Namun sayangnya layanan tersebut tidak dapat mengakomodasi kebutuhan tersebut diatas dan juga tidak mendukung pencarian berbasis teks lengkap (*full text*), yang mana pencarian hanya didasarkan pada metadata dengan menggunakan *simple* dan *advanced search*. Jenis pencarian yang tidak berbasis *full text* ini mengakibatkan (1) kecenderungan menghasilkan dokumen dengan topik yang kurang relevan atau mengecewakan dengan pencarian pemustaka (Tramboo dkk., 2012). (2) Begitu juga waktu yang diperlukan untuk menemukan dokumen dengan topik yang relevan relatif lama karena pemustaka harus membaca sebagian bahkan seluruh isi dokumen. (3) Seiring waktu pertumbuhan data yang terjadi pada

repositori dipastikan terus mengalami peningkatan. Sehingga menjadi kebutuhan yang sangat mendesak untuk bagaimana menemukan informasi yang berguna secara efisien dari data yang melimpah (Shao & Qin, 2014). Kelimpahan data semestinya dapat diolah untuk ditransisi menjadi informasi agar memberi pemahaman hubungan data yang selanjutnya dapat memberikan *knowledge*/pengetahuan berupa pemahaman terhadap pola hingga dapat memengaruhi *wisdom*/kebijaksanaan untuk memahami prinsip dalam pengambilan keputusan (Cooper, 2014; Grataridarga & Hum, 2019).

Guna mengatasi permasalahan tersebut di atas, diperlukan solusi untuk dapat (1) memberikan hasil pencarian berupa dokumen dengan topik yang relevan terkait kueri masukan pemustaka, (2) menemukan dokumen yang sesuai dengan lebih cepat, dan juga (3) melayani permintaan akan kebutuhan informasi yang tinggi serta dapat memberi *insight* (wawasan) dengan melakukan pengolahan tumpukan data yang tersedia, alih-alih menjadi sampah data yang tidak berguna. Salah satu pendekatan yang dapat digunakan untuk mencapai solusi tersebut di atas yaitu, dengan menggunakan teknik dalam *text mining* yang dinamakan *topic modeling* (pemodelan topik).

Pemodelan topik merupakan proses identifikasi berbagai tema atau subjek laten/tersembunyi yang dibahas dalam koleksi dokumen berupa data tekstual, tema yang dimaksud adalah topik itu sendiri (Sawant & Kanawade, 2014). Pemodelan topik berguna untuk menganalisis informasi tingkat semantik serta menemukan, dan menjelaskan struktur tematik yang bertujuan untuk menemukan topik dari kumpulan dokumen secara otomatis (Kherwa & Bansal, 2018; Wu dkk., 2010). Pemodelan topik dipilih karena mampu dan

menjanjikan mengolah, meringkas, dan memahami data *unstructured* (tidak terstruktur) dalam dokumen yang berupa data tekstual yang kian waktu terus berkembang (Blei dkk., 2010). Karena komputer tidak dapat mengerti data tekstual secara utuh, maka diperlukan beberapa proses seperti tokenisasi, *stopword* hingga *stemming*. Proses itu berguna untuk mengubah bentuk data tidak terstruktur menjadi terstruktur yang direpresentasikan dalam *term document matrix* (TDM), berupa matriks vektor kata yang membentuk baris × vektor dokumen yang membentuk kolom (Guha, 2020). Teknik yang termasuk dalam *framework unsupervised learning* (pembelajaran tanpa pengawasan) ini dipilih karena data yang akan diteliti merupakan data tidak berlabel atau tidak memiliki *class* sehingga tidak dapat dilakukan pelatihan dengan data tersebut (Ray dkk., 2019).

Terdapat beberapa algoritma dalam teknik pemodelan topik, antara lain *Latent Semantic Analysis* (LSA), *Non-Negative Matrix Factorization* (NNMF), *Probabilistic Latent Semantic Analysis* (PLSA), dan *Latent Dirichlet Allocation* (LDA) (Kherwa & Bansal, 2018; Wu dkk., 2010). LDA diusulkan oleh Blei pada tahun 2003 (Blei dkk., 2003) sebagai penyempurnaan dari algoritma pendahulunya, PLSA, dengan menggunakan pendekatan probabilistik generatif berdasarkan pada model topik statistik Bayesian (Padmaja dkk., 2018). Keunggulan utama LDA yang tidak dimiliki oleh LSA dan PLSA adalah kemampuan dalam reduksi dimensi (Padmaja dkk., 2018). Dengan penggunaan pendekatan probabilitas generatif, LDA mampu bekerja di level dokumen-topik saat LSA dan PLSA hanya dapat bekerja pada salah satu level saja (Kherwa & Bansal, 2018; Liu, 2013). Dalam

pengujian yang dilakukan, LDA terbukti efektif dapat menangkap dokumen yang relevan; mampu menangani *overfitting* yang dihasilkan oleh PLSA, yaitu ketidakkonsistenan tingkat kecocokan yang tinggi antara data latih dengan data uji akibat terlalu banyak parameter yang digunakan (Setijohatmo dkk., 2020). Penggunaan LDA akan menghasilkan distribusi topik pada tiap dokumen berdasarkan distribusi kata sebagai keluarannya, yang dapat diterapkan pada temu kembali informasi (*Information Retrieval*).

Hasil dari proses LDA sebagai keluaran pemodelan topik, berupa daftar topik yang terdapat di dalam dokumen repositori. Daftar topik tersebut direpresentasikan oleh distribusi kata atau *term*, yang diwakili oleh distribusi tertinggi pada masing-masing topik. Distribusi kata itulah yang akan dimanfaatkan ke dalam temu kembali informasi, yang mana tiap *term* kueri akan dihitung kecocokannya dengan model topik yang dihasilkan oleh LDA. Sehingga dengan menghitung kemiripan antara kueri dan model LDA diharapkan dapat memberikan hasil dokumen relevan, yang memuat informasi sesuai kebutuhan pemustaka. Dengan penggunaan pemodelan topik untuk pencarian informasi pada dokumen besar, dapat memberikan peningkatan yang signifikan dalam presisi ketika digunakan dalam bentuk yang sesuai (Park & Ramamohanarao, 2009). Pemodelan topik statistik seperti LDA menyediakan sarana untuk secara otomatis mengindeks, mencari, mengelompokkan, dan menyusun dokumen tidak terstruktur dan juga tidak berlabel (Blei dkk., 2003). Tugas tersebut di atas mampu diselesaikan dengan metode *topic modeling* (pemodelan topik) dengan

menemukan sekumpulan topik dalam dokumen, di mana topik merupakan kumpulan kata yang muncul bersamaan (Thomas, 2011).

Berdasarkan pemaparan tersebut di atas, peneliti tertarik untuk melakukan penelitian pemodelan topik dengan menggunakan algoritma LDA dengan studi kasus data abstrak yang terdapat pada repositori Undiksha. Bagian dokumen yang digunakan adalah bagian abstrak, abstrak dipilih karena merupakan sintesis keseluruhan dari isi dokumen karya ilmiah. Sebab mengandung konsep, pernyataan masalah, pendekatan, dan kesimpulan yang dirangkai sedemikian rupa sehingga saling berkaitan dan memiliki makna utuh yang kemudian dapat menggambarkan keseluruhan isi tulisan (Nasution, 2017). Bagian dokumen tersebut akan diterapkan ke dalam Sistem Temu Kembali (IR) untuk memudahkan pencarian dokumen karya ilmiah. Adapun judul penelitian ini adalah "Sistem Pemodelan Topik Pada Pencarian Dokumen Karya Akhir Menggunakan Metode *Latent Dirichlet Allocation* (LDA)".

1.2 IDENTIFIKASI MASALAH

Berdasarkan pada latar belakang yang telah diuraikan di atas, maka dapat diidentifikasi masalah penelitian sebagai berikut.

1. Pencarian cenderung menghasilkan dokumen yang kurang relevan dengan kueri yang dimasukkan pemustaka.
2. Pemustaka membutuhkan waktu yang relatif lama untuk menemukan dokumen yang relevan dengan membaca seluruh bagian dokumen.
3. Diperlukannya melakukan pengolahan dokumen dengan pemodelan topik sebagai antisipasi pertumbuhan data yang semakin meningkat.

1.3 RUMUSAN MASALAH

Berdasarkan latar belakang dan identifikasi masalah di atas, maka rumusan masalah penelitian sebagai berikut.

1. Bagaimana pengembangan Sistem Pemodelan Topik Pada Pencarian Dokumen Karya Akhir Menggunakan Metode *Latent Dirichlet Allocation* (LDA)?
2. Bagaimana kinerja dari Metode *Latent Dirichlet Allocation* (LDA) dalam melakukan Pemodelan Topik Pada Pencarian Dokumen Karya Akhir?

1.4 TUJUAN PENELITIAN

Berdasarkan rumusan masalah yang telah dikemukakan di atas, maka penelitian ini dilaksanakan dengan tujuan sebagai berikut.

1. Mengetahui pengembangan Sistem Pemodelan Topik Pada Pencarian Dokumen Karya Akhir Menggunakan Metode *Latent Dirichlet Allocation* (LDA).
2. Mengetahui kinerja Sistem Pemodelan Topik Pada Pencarian Dokumen Karya Akhir Menggunakan Metode *Latent Dirichlet Allocation* (LDA).

1.5 BATASAN MASALAH

Adapun batasan masalah dari penelitian yang berjudul "Sistem Pemodelan Topik Pada Pencarian Dokumen Karya Akhir Menggunakan Metode *Latent Dirichlet Allocation* (LDA)" sebagai berikut.

1. Hanya menggunakan data berupa abstrak karya ilmiah karena merupakan sintesa keseluruhan dari isi dokumen karya ilmiah.

2. Menggunakan abstrak dokumen karya akhir yang berbahasa Indonesia.
3. Sumber data berasal dari situs OAI (*Open Archives Initiative*) repositori Undiksha (https://repo.undiksha.ac.id/cgi/oai2).
4. Pengujian kinerja yang dilakukan pada sistem ini yaitu berupa uji relevansi hasil pencarian, yang di dalamnya mencakup *precision*, *recall*, dan *f-measure*.
5. Keluaran atau *output* dari penelitian ini adalah berupa modul sistem, yang berfungsi dalam melakukan pencarian dokumen karya akhir. Sehingga tidak ditujukan sebagai pengganti sistem repositori yang ada saat ini / sistem *existing*.

1.6 MANFAAT PENELITIAN

Sistem Pemodelan Topik Pada Pencarian Dokumen Karya Akhir Menggunakan Metode *Latent Dirichlet Allocation* (LDA) diharapkan memberikan beberapa manfaat sebagai berikut:

A. Manfaat Teoritis

Bagi peneliti, penelitian ini diharapkan akan mampu menambah wawasan serta membuat lebih mengerti, memahami dan mampu menerapkan materi pembelajaran yang telah diperoleh selama proses perkuliahan.

B. Manfaat Praktis

1. Bagi Masyarakat
   a. Dapat memberikan hasil pencarian yang relevan dengan kueri masukannya.
   b. Dapat mempermudah pemustaka dalam penemuan topik dalam dokumen.

c. Dapat memberikan *insight* tren topik penelitian.

2. Bagi Peneliti

   Hasil penelitian ini dapat memberikan pengetahuan dan wawasan yang lebih dalam tentang pengolahan dokumen teks dalam penemuan topik yang terkandung di dalamnya.