BAB II
TINJAUAN PUSTAKA

2.1 Kajian Pustaka Penelitian

Penelitian dengan tujuan membantu mengelompokkan data guna mendapatkan *topic modelling* terhadap suatu fenomena sudah banyak dilakukan dan sudah banyak diterapkan di berbagai bidang di Indonesia. Dalam penelitian sebelumnya, menunjukkan bahwa pemilihan metode yang sesuai di *topic modelling* akan sangat menentukan tingkat keberhasilan dalam penelitian.

Seperti penelitian [8] menerapkan *latent dirichlet allocation* (LDA) untuk mencari tren topik penelitian dosen pada jurnal JPTEI UNY. Korpus yang digunakan diambil dari Google Scholar. Model yang dibangun dapat menghasilkan informasi bahwa publikasi yang terdapat pada jurnal JPTEI UNY adalah seputar topik pendidikan vokasi, pengembangan sistem, media pembelajaran, dan sistem pembelajaran di SMK.

Penelitian [2] menggunakan korpus dokumen publikasi ilmiah berupa Tugas Akhir mahasiswa keperawatan. Penelitian tersebut juga menggunakan LDA. Penelitian ini bertujuan untuk menganalisis tren topik penelitian dengan pemodelan topik terhadap judul-judul penelitian di bidang keperawatan pada repositori jurnal SINTA. Representasi dokumen yang digunakan yaitu TF-IDF. Model yang dibangun menghasilkan topik tentang ibu hamil, fenomenologi, tingkat kecemasan, penderita hipertensi, tingkat stress, intensitas nyeri, penderita diabetes, kesehatan reproduksi, ibu nifas, kualitas tidur, asuhan keperawatan, tekanan darah, kinerja perawat, kualitas hidup, relaksasi otot dan lain-lain .

Penelitian [4] menggunakan korpus abstraksi skripsi mahasiswa yang berasal dari 584 judul. Penelitian ini bertujuan untuk menganalisis tren pada topik skripsi mahasiswa program studi sastra inggris di Universitas Islam Negeri Sunan Ampel Surabaya (UINSA). Representasi dokumen yang digunakan yaitu *Bag of*

Words atau jumlah kemunculan kata. Penentuan jumlah topik menggunakan metode percobaan, mulai dari dua sampai lima topik. Setiap percobaan menggunakan iterasi 100, 500, 1000, dan 5000 untuk mendapatkan model yang fit. Kesimpulannya, model yang fit yaitu jika topiknya adalah tiga. Setelah hasil tren topik diverifikasi oleh stakeholder (Program Studi Sastra Inggris UINSA) ternyata ada satu topik yang menghasilkan tren yang sesuai Program Studi Sastra Inggris UINSA.

Penelitian [10] menggunakan korpus sekumpulan laporan yang telah diberi label. Penelitian ini bertujuan untuk mengetahui topik tersembunyi pada dokumen pada laporan Tugas Akhir mahasiswa dengan mengimplementasi *gibbs sampling* dan mengetahui relevansi klasifikasi topik yang dihasilkan. Representasi dokumen yang digunakan yaitu TF-IDF. LDA yang digunakan sudah dimodifikasi dengan metode *collapsed gibbs sampling*. Penelitian ini menyimpulkan bahwa meotde probabilitas sebuah kata dipengaruhi oleh banyaknya jumlah topik dan dokumen. Penelitian [10] juga menyimpulkan bahwa LDA sensitif terhadap komposisi kata yang mengandung banyak kata umum sehingga mengurangi akurasi.

Penelitian [11] menggunakan korpus judul dan abstrak yang dibobotkan dengan TF-IDF dan dikelompokkan menggunakan K-Means Clustering. Penelitian tersebut juga menggunakan LDA-*gibbs sampling*. Penelitian ini bertujuan untuk membuat sistem rekomendasi topik tugas akhir berdasarkan kompetensi dalam transkrip akademik. Penelitian ini menyimpulkan bahwa semakin tinggi nilai kemiripan vektor kata inti dari sistem akan memberikan nilai probabilitas yang tinggi dalam topik terpilih.

Setelah melakukan studi literasi, maka dapat disimpulkan seperti pada Tabel 2.1. perbedaan antara penelitian terdahulu dengan penelitian yang diajukan.

Tabel 2.1. Tabel Kajian Pustaka

No.	Judul dan Penulis	Permasalahan	Metode yang digunakan	Dataset	Hasil
1	Topic Modeling Penelitian Dosen JPTEI UNY pada Google Scholar Menggunakan Latent Dirichlet Allocation (Nurlayli, Akhsin. Nasichuddin, Moch Ari)	mencari tren topik penelitian dosen pada jurnal JPTEI UNY	LDA	Dataset diambil dari akun masing-masing dosen Jurusan Pendidikan Teknik Elektronika dan Informatika Universitas Negeri Yogyakarta (JPTEI UNY) pada Google Scholar	informasi bahwa publikasi yang terdapat pada jurnal JPTEI UNY adalah seputar topik pendidikan vokasi, pengembangan sistem, media pembelajaran, dan sistem pembelajaran di SMK
2	Pemodelan Topik Penelitian Bidang Keperawatan Indonesia Pada Repository Jurnal Sinta Menggunakan Metode Topic Modelling Lda (Latent Dirichlet Allocation) (Sahria, Yoga. Isnaini Febriarini, Nurul. Dwi Oktavianti, Pamulatsih)	menganalisis tren topik penelitian dengan pemodelan topik terhadap judul-judul penelitian di bidang keperawatan pada repositori jurnal SINTA	LDA dengan representasi TF-IDF	dokumen publikasi ilmiah berupa Tugas Akhir mahasiswa keperawatan dari repositori jurnal SINTA.	topik tentang ibu hamil, fenomenologi, tingkat kecemasan, penderita hipertensi, tingkat stress, intensitas nyeri, penderita diabetes, kesehatan reproduksi, ibu nifas, kualitas tidur, asuhan keperawatan, tekanan darah, kinerja perawat, kualitas hidup, relaksasi otot dan lain-lain
3	Topic Modelling Skripsi Menggunakan Metode Latent Diriclhet	menganalisis tren pada topik skripsi mahasiswa	LDA dengan representasi *Bag of Words*. Verifikasi	Lima ratus delapan puluh empat (584) abstraksi skripsi mahasiswa program studi sastra	hasil tren topik diverifikasi oleh stakeholder (Program

No.	Judul dan Penulis	Permasalahan	Metode yang digunakan	Dataset	Hasil
	Allocation (Iffan Alfanzar, Alif. Sudanawati Rozas, Indri)	program studi sastra inggris di Universitas Islam Negeri Sunan Ampel Surabaya (UINSA)	hasil topik dilakukan oleh pihak program studi sastra inggris UINSA	inggris UINSA.	Studi Sastra Inggris UINSA) ternyata ada satu topik yang menghasilkan tren yang sesuai Program Studi Sastra Inggris UINSA
4	Analisis Metoda Latent Dirichlet Allocation untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik (Setijohatmo, Urip T. Rachmat, Setiadi. Susilawati, Tati. Rahman, Yuda)	mengetahui topik tersembunyi pada dokumen pada laporan Tugas Akhir mahasiswa dengan mengimplementasi *gibbs sampling* dan mengetahui relevansi klasifikasi topik yang dihasilkan	LDA dengan estimasi *gibbs sampling* dan representasi dokumen yaitu TF-IDF	sekumpulan laporan Tugas Akhir yang telah diberi label	meotde probabilitas sebuah kata dipengaruhi oleh banyaknya jumlah topik dan dokumen. LDA sensitif terhadap komposisi kata yang mengandung banyak kata umum sehingga mengurangi akurasi
5	Pemodelan Topik dengan LDA untuk Temu Kembali Informasi dalam Rekomendasi Tugas Akhir (Purwitasari, Diana. Aida Muflichah.	membuat sistem rekomendasi topik tugas akhir berdasarkan kompetensi dalam transkrip akademik	Pembobotan kata menggunakan TF-IDF dan pengelompokkan kata menggunakan K-Means serta menggunakan metode LDA-	judul dan abstrak yang dibobotkan dengan TF-IDF dan dikelompokkan menggunakan K-Means Clustering	semakin tinggi nilai kemiripan vektor kata inti dari sistem akan memberikan nilai probabilitas yang tinggi dalam topik terpilih

No.	Judul dan Penulis	Permasalahan	Metode yang digunakan	Dataset	Hasil
	Novrindah Alvi Hasanah. Agus Zainal Arifin)		*gibbs sampling* sebagai topik modelingnya.		

2.2 Dasar Teori

2.2.1 Clustering

Clustering atau klasterisasi adalah metode pengelompokan data. Menurut Tan, 2006 *clustering* adalah sebuah proses mengelompokkan data kedalam kelompok / *cluster* yang memiliki tingkat kemiripan yang maksimum dan data antara satu *cluster* dengan *cluster* lainnya memiliki kemiripan yang minimum [12].

Hasil *clustering* yang baik akan menghasilkan tingkat kesamaan yang tinggi dalam satu kelas dan tingkat kesamaan yang rendah antar kelas. Kesamaan yang dimaksud merupakan pengukuran secara numerik terhadap dua buah objek. Semakin angkanya tinggi, maka objek memiliki tingkat kemiripan yang tinggi. Sebaliknya, jika angkanya rendah maka objek memiliki tingkat kemiripan yang rendah [12].

2.2.2 Pemodelan Topik

Pemodelan topik adalah salah satu teknik paling ampuh dalam penambangan teks untuk penambangan data, laten penemuan data, dan menemukan hubungan antara data dan dokumen teks. Pemodelan topik atau *topic modelling* merupakan salah satu pendekatan pada *text mining* yang cukup handal dalam menemukan data teks yang tersembunyi dan dapat menemukan hubungan antara teks yang satu dengan teks lainnya dari suatu *corpus* [13] . Menurut Verma & Gahier, 2015 konsep *topic modelling* terdiri dari entitas-entitas yaitu “kata”, “dokumen”, dan *“corpora”*. “kata” dianggap sebagai satuan dasar dari data diskrit dalam dokumen. “dokumen” adalah susunan N kata-kata. *“corpus”* adalah kumpulan M dokumen dan *“corpora”* merupakan bentuk jamak dari *corpus.* Sementara *“topic”* adalah distribusi dari beberapa kosakata yang bersifat tetap. Secara sederhana, setiap dokumen dalam *corpus* mengandung proporsi tersendiri dari topik-topik yang dibahas sesuai kata-kata yang terkandung di dalamnya [14].

2.2.3 Text Preprocessing

Text preprocessing adalah tahapan mengolah data teks mulai dari pembersihan data sampai dengan menjadi vektor sehingga siap dimodelkan. Banyak tahap yang menjadi bagian dari pembersihan data seperti *lower case, remove punctuation, stopword removal,* dan penanganan *slang words.* Setelah data bersih kemudian diubah menjadi vektor. Proses vektorisasi didahului dengan *tokenizing* dan *stemming. Tokenizing dan stemming* tidak selalu digunakan dalam setiap tahapan pemrosesan teks tetapi tergantung kepada tujuan pemrosesan teks. Berikut adalah penjelasan proses-proses yang ada dalam *text preprocessing.*

1. Remove Punctuation

   Remove punctuation adalah tahap dimana menghilangkan tanda baca yang tidak dipakai dalam data yang nantinya dipakai dalam penelitian.

2. Stopword Removal

   Stopword removal adalah tahap dimana menghilangkan kata-kata yang tidak dipakai dalam data yang nantinya dipakai dalam penelitian.

3. Lower Case

   Lower case adalah tahap dimana untuk membuat huruf kapital menjadi huruf kecil.

4. Penanganan Slang word

   Slang word adalah kata-kata tidak baku yang sering digunakan dalam kehidupan sehari-hari. Menghilangkan *slang word* bertujuan untuk menghilangkan kata tidak baku yang tidak dibutuhkan dalam penelitian.

5. Tokenizing

   Tokenizing adalah tahap dimana memisahkan sebuah kalimat menjadi kata per kata. *Tokenizing* dapat mempermudah dalam membaca kata yang sudah melalui proses *text processing.* Contoh *tokenizing* dari kalimat “saya makan jagung” adalah “saya”, “makan”, “jagung”. “saya”, “makan”, “jagung” merupakan kata yang sudah di *tokenizing.*

6. Stemming

   Stemming adalah tahap dimana merubah kata menjadi kata dasarnya. Proses

stemming berguna untuk mendapatkan bentuk kata aslinya karena kata dasar yang sudah diberi imbuhan bisa berubah maknanya. Contoh *stemming* dari kata "kedudukan" adalah "duduk". "kedudukan" dan "duduk" mempunyai makna yang berbeda, sehingga *stemming* akan berpengaruh kepada model *Neuro Linguistic Programming* (NLP).

7. Vektorisasi

Vektorisasi adalah proses merubah kata menjadi angka dalam bentuk *matrix. Matrix* nantinya akan berisi nilai atau bobot dari sebuah kata. Metode vektorisasi antara lain *Bag of Word, Term Frequency, Term Frequency-Inverse Document Frequency* (TF-IDF), *Word2vec, wang2vec, Fastext,* dan lain-lain.

2.2.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) adalah model probabilistik generatif dari koleksi data diskrit seperti korpus teks. LDA merupakan model *Bayesian* hirarki tiga tingkat, dimana setiap *item* koleksi dimodelkan sebagai campuran terbatas atas serangkaian set topik. Setiap topik dimodelkan sebagai campuran tak terbatas melalui set yang mendasari probabilitas topik. Probabilitas topik memberikan representasi eksplisit dari sebuah dokumen [15]. Ide dasar LDA yaitu dokumen terdiri dari beberapa topik. LDA adalah model statistik dari kumpulan dokumen yang berusaha untuk merepresentasikan ide tersebut. Proses LDA bersifat generatif melalui *imaginary random process* pada model yang di asumsikan bahwa dokumen berasal dari topik tertentu yang berisi kata-kata [15].

Cara kerja LDA adalah menghitung *joint probability distribution* dengan cara melakukan sampling satu persatu terhadap setiap variabel lainnya alias *full conditional probability.* Dalam pengelompokkan topik memiliki dua bentuk distribusi probabilitas yang harus dicari yaitu, distribusi probabilitas dokumen pada suatu dokumen dan distribusi probabilitas kata pada suatu topik [16]. Probabilitas topik pada dokumen merupakan nilai probabilitas tiap topik pada suatu dokumen. Misal pada dokumen I mempunyai probabilitas topik A senilai

X, mempunyai probabilitas topik B senilai Y dan seterusnya sesuai dengan jumlah topik [16]. Nilai tersebut dapat kita temui di persamaan 2.1.

$$\rho(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(Z_{dn}|\theta_d) p(w_{dn|} Z_{dn}, \beta) \right) d\theta_d \quad (2.1)$$

Dimana persamaan 2.1 dan Gambar 3.3. menjelaskan bahwa:

M adalah jumlah dokumen;

N adalah jumlah kata dalam dokumen tertentu;

D adalah *term;*

α adalah *dirichlet prior* pada sebaran topik per dokumen;

β adalah *dirichlet prior* pada sebaran kata per topik;

θ_i adalah sebaran topik untuk dokumen i;

Langkah-langkah penerapan LDA:

1. Menentukan jumlah topik, jumlah iterasi, parameter alpha dan beta.
2. Melakukan iterasi sesuai dengan jumlah yang ditentukan. Pada tiap iterasi kita menghitung distribusi probabilitas topik pada suatu dokumen dan distribusi probabilitas pada suatu topik.
3. Setelah memiliki dua nilai distribusi, kita bisa memperbarui topik pada tiap kata dengan cara melakukan *sample* dengan bobot. Dimana bobot didapatkan dari perkalian nilai distribusi probabilitas topik pada dokumen dengan distribusi probabilitas kata pada topik.

2.2.5 Term Frequency-Inverse Document Frequency

Metode TF-IDF merupakan suatu cara untuk memberikan bobot hubungan suatu kata (*term*) terhadap suatu dokumen. Metode ini menggabungkan dua konsep untuk menghitung bobot, yaitu frekuensi

munculnya sebuah kata di dalam sebuah dokumen dan *inverse* frekuensi dokumen yang mengandung kata tersebut. Semakin banyak kata yang muncul di suatu dokumen, menunjukkan seberapa penting kata itu di dalam dokumen tersebut. Semakin banyak dokumen yang mengandung kata tersebut, menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antar kata dan dokumen akan tinggi apabila frekuensi kata tersebut tinggi didalam dokumen dan frekuensi keseluruhan dokumen yang mengandung *term* tersebut yang rendah pada sekumpulan dokumen. Rumus TF-IDF ada beberapa macam. Salah satu rumus TF-IDF dihitung dengan menggunakan persamaan (2.2).

$$W_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right) \quad (2.2)$$

Dengan masing-masing keterangan yaitu:

$W_{i,j}$ adalah jumlah kata-i pada dokumen ke-j.

N adalah jumlah total dokumen.

df_i adalah banyaknya dokumen yang mengandung kata ke-i.