# BAB II
# TINJAUAN PUSTAKA

## 2.1 Studi Literatur

Penelitian yang digunakan sebagai bahan studi literatur dalam penelitian ini, antara lain:

Penelitian (Supriadi, et al., 2020) tentang "Sensitivitas Sistem Pencarian Artikel Bahasa Indonesia Menggunakan Metode *n-gram* Dan *Tanimoto Cosine*". Penelitian ini bertujuan membuat suatu sistem dengan mengimplementasikan *n-gram* dan *Tanimoto Cosine* untuk melakukan pencarian suatu dokumen menggunakan kata kunci di dalam sekumpulan dokumen. Penelitian ini membantu peneliti lebih memahami dan mendalami penggunaan metode *n-gram* dalam proses *information retrieval* dalam membuat sistem rekomendasi artikel teknologi informasi dan komunikasi (TIK) sebagai acuan studi literatur yang sesuai.

Penelitian (Sugianto, et al., 2013) tentang "pembuatan aplikasi *predictive text* menggunakan metode *n-gram-based"*. Penelitian ini bertujuan membuat aplikasi untuk melakukan prediksi kata dengan menggunakan *n-gram* sebagai metode dasar dalam melakukan proses prediksi kata. *Predictive text* merupakan sebuah fitur dalam pengetikan yang mempunyai tujuan mengurangi *keystroke* saat pengetikan dengan langkah memprediksi kata yang terlihat berdasarkan huruf yang diketikan. Penelitian ini membantu peneliti dalam memahami dan mendalami cara menggunakan metode *n-gram* untuk memprediksi kata sebagai salah satu dasar dalam membuat sistem rekomendasi artikel teknologi informasi dan komunikasi (TIK) sebagai acuan studi literatur yang sesuai.

Penelitian (Azizurahman, et al., 2011) tentang "analisis dan implementasi metode *n-gram* pada *information retrieval"*. Penelitian ini bertujuan menganalisis dan mengimplementasikan *n-gram* sebagai metode dasar melakukan proses *information retrieval*. *Information retrieval* merupakan pengambilan informasi dari dokumen-dokumen yang didasarkan pada isi dan konteks dari dokumen-dokumen tersebut. Penelitian ini membantu peneliti dalam memahami dan mendalami cara menggunakan metode *n-gram* dalam proses *information retrieval* sebagai salah satu

dasar dalam membuat sistem rekomendasi artikel teknologi informasi dan komunikasi (TIK) sebagai acuan studi literatur yang sesuai.

Penelitian (Indranandita, et al., 2011) tentang "sistem klasifikasi dan pencarian jurnal dengan menggunakan metode *naive bayes* dan *vector space model"*. Penelitian ini bertujuan membuat sistem klasifikasi dan pencarian jurnal dengan metode *Naive Bayes* dan *Vector Space Model* dengan pendekatan *Cosine* yang diharapkan membantu pengguna dalam penentuan topik/kategori dan menghasilkan daftar artikel berdasarkan urutan tingkat kemiripan. Penelitian ini membantu peneliti dalam memahami dan mendalami sistem pencarian artikel dan juga membandingkan metode *vector space model* dengan metode *n-gram* yang digunakan peneliti sebagai metode dasar dalam membuat sistem rekomendasi artikel teknologi informasi dan komunikasi (TIK) sebagai acuan studi literatur yang sesuai.

Perbedaan penelitian ini dengan penelitian sebelumnya adalah sistem yang dirancang dan dibangun merupakan sistem rekomendasi dan juga hanya merekomendasikan artikel teknologi informasi dan komunikasi (TIK). Sistem rekomendasi artikel teknologi informasi dan komunikasi (TIK) berbasis web dengan *n-gram* sebagai metode dasar penelitian

## 2.2 Landasan Teori

### 2.2.1 Sistem Rekomendasi

Sistem rekomendasi merupakan suatu sistem yang mempunyai tujuan untuk menolong pengguna dengan cara memberikan petunjuk, ide atau saran disaat pengguna dihadapkan pada banyak informasi. Menurut (Ricci, et al., 2011), rekomendasi yang ditawarkan diharapkan bisa menunjang pengguna dalam mengambil keputusan, seperti barang apa yang dapat dibeli, buku apa yang dapat dibaca, dan lagu apa yang dapat didengarkan.

Dalam penelitian ini yaitu sistem rekomendasi artikel mengharuskan *user* untuk memasukkan kata kunci atau *query* agar sistem dapat merekomendasikan artikel yang relevan kepada *user*. Menurut (Manning, et al., 2008) *query* merupakan

formula yang digunakan *user* untuk mencari informasi yang dibutuhkannya dengan bentuk kata yang paling sederhana berupa kata kunci atau *keyword*.

### 2.2.2 Studi Literatur

Studi literatur dilakukan dengan mencari referensi penelitian yang relevan dengan sesuatu yang diteliti. Tujuan studi literatur dilakukan adalah agar peneliti mempunyai pemahaman yang lebih luas dan mendalam terhadap permasalahan serta sebagai dasar teori dalam melakukan penelitian tentang sistem rekomendasi artikel teknologi informasi dan komunikasi (TIK) sebagai acuan studi literatur menggunakan metode *n-gram*.

### 2.2.3 *Website*

Menurut (Maryono & Istiana, 2008), *Website* bisa diartikan menjadi kumpulan halaman yang digunakan untuk menampilkan informasi, gambar bergerak, suara atau gabungan dari semuanya itu baik yang bersifat statis maupun dinamis yang menciptakan suatu rangkaian halaman yang saling terkait dan dihubungkan menggunakan *link-link.*

### 2.2.4 Web *Scrapping*

*Webscraping* merupakan proses mengekstraksi informasi dan data di dalam *website* secara otomatis kemudian menyimpannya dengan format yang diinginkan seperti teks, *csv* atau *json*. Teknik *webscraping* yang digunakan pada tahap pengumpulan data pelatihan ini adalah HTML *parsing*. HTML *parsing* merupakan salah satu teknik yang sering digunakan dalam proses *parsing* atau menguraikan data, beberapa data yang diperoleh seperti teks, *link, screen* dan lain-lain. Teknik ini dilakukan menggunakan *javascript* dan menargetkan halaman HTML linear atau *nested*.

### 2.2.5 *Text Preprocessing*

#### 2.2.5.1 Pengertian *Text Processing*

*Text preprocessing* merupakan langkah penting dalam *Text Mining* dan *Natural Language Processing* (NLP) untuk mengubah teks yang tidak terstruktur menjadi teks yang lebih terstruktur karena data teks biasanya berisi format khusus

seperti angka, tanda baca dan kata yang tidak membantu dalam proses selanjutnya seperti kata sambung, dll. Menurut (Feldman & Sanger, 2007) *text preprocessing* merupakan tahapan awal proses yang dilakukan dalam mempersiapkan teks menjadi data yang dapat diolah lebih lanjut.

#### 2.2.5.2 Tahapan *Text Processing*

*Text preprocessing* ini terdiri dari beberapa tahap yaitu *case folding*, *tokenizing*, *filter stopword* dan *stemming*. Berikut ini penjelasannya:

a) *Case folding* adalah proses untuk menyeragamkan bentuk huruf menjadi huruf kecil (*lowercase*), dengan tujuan menjadikan *lowercase* sebagai bentuk standar pada teks.
b) *Tokenization* adalah proses untuk memecah teks dalam suatu artikel menjadi kumpulan kata atau disebut token. Pada proses ini juga menghilangkan karakter lain seperti angka dan tanda baca.
c) *Filter Stopword* adalah proses untuk menghapus kata-kata yang tidak relevan dan kata-kata yang tidak mempunyai makna tersendiri jika dipisahkan dengan kata lain.
d) *Stemming* adalah proses untuk mencari dan mengelompokkan kata-kata yang memiliki kata dasar dengan arti yang serupa namun memiliki bentuk yang berbeda karena mendapatkan imbuhan yang berbeda.

Pada tahap *filter stopword* dan *stemming* menggunakan *library* sastrawi yang diambil dari https://github.com/sastrawi/sastrawi untuk membantu agar proses tersebut dapat berjalan dan bekerja dengan baik.

### 2.2.6 *N-Gram*

*N-gram* merupakan salah satu metode yang biasanya digunakan dalam *Natural Language Processing* dan *Text Mining*. *N-gram* merupakan sekumpulan kata yang mempunyai ukuran panjang n kata. Misalnya, n-gram ukuran 1 disebut *uni-gram*, ukuran 2 menjadi *bi-gram*, ukuran 3 disebut *tri-gram*, dan seterusnya. (Sugianto, et al., 2013).

Dalam proses *text mining*, salah satu langkah awal yang dilakukan yaitu mengatur dan menyusun text dengan cara tertentu sehingga dapat dieksplorasi dan

dianalisis secara kuantitatif dan kualitatif. Dalam prosesnya, menggunakan teknologi *natural language processing* untuk menerapkan prinsip komputasional linguistik dengan metode *n-gram* untuk menguraikan dan menginterpretasikan data yang ada, untuk dapat melakukan hal-hal seperti mengelompokkan, mengategorikan, penandaan teks, meringkas data, menciptakan taksonomi, mengekstraksi informasi tentang frekuensi kata dan hubungan antar entitas data.

*N-gram* juga biasanya digunakan dalam model bahasa statistik untuk mengatasi masalah ketersebaran kata yang menyebabkan urutan kata yang paling mungkin tidak terlihat di saat dalam pelatihan, dengan solusi membuat asumsi bahwa probabilitas urutan kata hanya bergantung n kata sebelumnya. Pemodelan bahasa dengan *n-gram* biasanya digunakan dalam pengenalan suara, penandaan bagian ucapan, pengenalan karakter optik, pengenalan tulisan tangan, mesin terjemahan, penguraian, pengambilan informasi.

### 2.2.7 *TF-IDF*

Menurut Robertson dalam (Maarif, 2015) *TF-IDF* atau *Term Frequency Inverse Documen Frequency* adalah pendekatan yang paling banyak digunakan untuk menghitung bobot setiap kata dalam *informasion retrieval*. Pendekatan ini juga dikenal efisien, sederhana, dan akurat.

*TF-IDF* digunakan untuk menentukan bobot setiap term atau kata, untuk mengetahui seberapa penting suatu kata terhadap sebuah dokumen. Nilai *TF* atau *Term Frequency* suatu kata menunjukkan betapa pentingnya kata tersebut dalam dokumen itu. Sedangkan Nilai *IDF* atau *Inverse Documen Frequency* suatu kata mencerminkan betapa pentingnya kata tersebut dalam seluruh dokumen itu.

#### 2.2.7.1. *Term Frequency* (*TF)*

*Term frequency* berfungsi untuk memberikan bobot pada suatu *term* dalam sebuah dokumen berdasarkan frekuensi kemunculannya pada dokumen tersebut. Formula *TF* yang digunakan dalam penelitian ini adalah *Raw TF*, yang artinya nilai *TF* diberikan berdasarkan jumlah kemunculan suatu kata dalam sebuah dokumen. Semakin tinggi frekuensi kemunculan suatu *term* dalam sebuah dokumen, maka semakin besar juga nilai *TF*-nya.

#### 2.2.7.2. *Inverse Documen Frequency (IDF)*

*Inverse Documen Frequency* berfungsi untuk memberikan bobot pada sebuah *term* dalam suatu dokumen berdasarkan frekuensi dokumen yang mengandung *term* tersebut. IDF juga berfungsi untuk mengurangi dominasi nilai *term* yang sering muncul di banyak dokumen. Term yang sering muncul dibanyak dokumen biasanya merupakan *term* umum sehingga nilainya tidak penting. Formula *Inverse Documen Frequency (IDF)* yang digunakan dalam penelitian ini adalah *Smooth IDF,* yang dapat dihitung menggunakan persamaan sebagai berikut:

$$IDF = ln\frac{(N+1)}{(DF+1)} + 1 \quad (2.1)$$

Berikut merupakan rumus atau persamaan untuk menghitung bobot dengan metode *TF-IDF*:

$$TF.IDF = TF \times IDF \quad (2.2)$$

### 2.2.8 *Cosine Similarity*

Menurut (Ariantini, et al., 2016), *cosine similarity* merupakan ukuran kemiripan dua vektor yang di mana hasilnya minimal 0 atau tidak mirip dan maksimal 1 atau sangat mirip. *Cosine similarity* digunakan dalam *Information Retrieval* atau Sistem Temu Balik Informasi untuk mengukur kemiripan antara dokumen dalam *database* dan *query* masukan *user*. Berikut merupakan persamaan *cosine similarity*:

$$cosine\ similarity = \cos(\theta) = \frac{A \cdot B}{|A|\,|B|} = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (2.3)$$

Keterangan pada persamaan di atas:

$A$ = vektor A

$B$ = vektor B

$A_i$ = bobot *term* ke- i dalam vektor A

$B_i$ = bobot *term* ke- i dalam vektor B

$i$ = jumlah term dalam kalimat

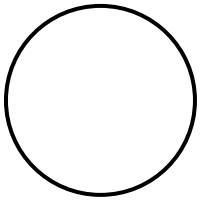
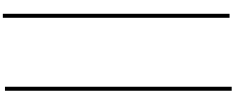
$n$ = jumlah vektor

Dalam penelitian ini persamaan di atas dapat artikan menjadi ukuran kemiripan antara artikel dalam *database* dan *query* masukan *user*, maka A adalah bobot setiap *term query* masukan *user*, dan B adalah bobot setiap *term* artikel yang sedang dibandingkan. Setelah mendapatkan nilai *cosine similarity* (CS) setiap artikel, dilanjutkan dengan proses pengurutan peringkat di mana semakin besar nilai CS-nya maka peringkatnya akan semakin tinggi yang berarti semakin relevan artikel tersebut terhadap *query* yang dimasukkan *user*.

### 2.2.9 *Data Flow Diagram (DFD)*

Menurut (Kristanto, 2008) DFD adalah suatu model logika atau proses yang dibuat untuk menjelaskan dari mana data berasal, proses apa yang menghasilkan data tersebut, di mana data disimpan, interaksi antara data yang disimpan dan proses yang diterapkan, dan ke mana data keluar. Berikut merupakan notasi-notasi yang ada pada DFD.

**Tabel 2.1** Notasi-Notasi *Data Flow Diagram* (DFD)

| Notasi | Keterangan |
|---|---|
|  | Entitas Luar; digunakan untuk menggambarkan entitas di luar sistem yang memberikan masukan (*input*) atau keluaran (*output*) kepada sistem. |
|  | Aliran Data; digambarkan seperti anak panah untuk menunjukkan aliran data yang terjadi di antara entitas luar, proses dan penyimpanan data. |
|  | Proses; merupakan kegiatan/proses yang dilakukan oleh orang atau sistem di mana akan menggambarkan proses transformasi data dari aliran data masuk ke aliran data keluar. |
|  | Penyimpanan Data; merupakan simbol atau notasi untuk menandakan bahwa aliran data telah masuk dan disimpan ke suatu tabel dari *database*. |

Dalam Tabel 2.1 di atas dapat diperhatikan simbol dan keterangan notasi-notasi yang digunakan dalam merancang DFD. (Kristanto, 2008) menjelaskan bahwa perancangan *Data Flow Diagram* (DFD) dibagi menjadi 2 tingkatan yaitu:

#### 2.2.8.1 Diagram Konteks (*Context Diagram*)

Diagram konteks merupakan sebuah diagram yang sederhana untuk menggambarkan hubungan antara entitas luar yang memberikan masukan atau menerima keluaran dari sistem. Diagram konteks menunjukkan semua proses yang ada dalam lingkaran tunggal atau satu proses tunggal untuk mewakili keseluruhan sistem.

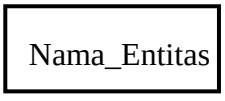
#### 2.2.8.2 DFD *Leveled*

DFD *leveled* ini menggambarkan semua proses yang ada dalam sistem. DFD *leveled* akan melakukan penurunan level untuk mempresentasikan proses yang sudah ada menjadi proses yang lebih spesifik. DFD *leveled* dimulai dari DFD level 0 atau diagram konteks, kemudian turun ke DFD level 1 dan seterusnya.

### 2.2.10 *Entity Relaptionship Diagram (ERD)*

Menurut (Firman, et al., 2016), *Entity Relaptionship Diagram* adalah model untuk menggambarkan hubungan antar data dalam *database* berdasarkan objek data dasar yang memiliki hubungan antar relasi. *Entity Relaptionship Diagram* untuk memodelkan struktur data, hubungan antara data untuk menggambarkannya, dan beberapa notasi dan simbol yang merupakan bagian dari *Entity Relaptionship Diagram* digunakan. Dalam notasi chen, simbol-simbol yang digunakan dalam ERD sebagai berikut:

**Tabel 2.2** Notasi-Notasi *Entity Relaptionship Diagram* (ERD)

| Simbol | Keterangan |
|---|---|
|  | Entitas; merupakan sebuah objek nyata yang unik dan dapat dibedakan dari objek lain. |
|  | Atribut; merupakan deskripsi dari karakteristik yang dimiliki oleh Entitas. |

| | |
|---|---|
|  | Atribut Kunci Primer; merupakan atribut yang memiliki data/karakteristik yang paling penting dan unik. Biasanya dalam bentuk angka. |
|  | Atribut Multinilai/*Multivalue*; merupakan atribut yang memiliki lebih dari satu nilai. |
|  | Asosiasi; merupakan garis penghubung antara entitas, atribut dan relasi. |
|  | Relasi; merupakan hubungan yang dimiliki antar entitas dari himpunan yang berbeda. |

Pada Tabel 2.2 terdapat simbol notasi-notasi yang digunakan dalam perancangan ERD, dan terdapat pula keterangan dari masing-masing notasi tersebut. Simbol atau notasi tersebut akan dihubungkan satu dan yang lain sesuai dengan fungsi masing-masing notasi dan sesuai dengan relasi yang ada.