2.LANDASAN TEORI

2.1 Tinjauan Pustaka

Tinjauan Pustaka akan menjelaskan beberapa landasan teori yang berkaitan dengan proses pembuatan skripsi.

2.1.1 TF-IDF (Term Frequency - Inverse Document Frequency)

TF-IDF adalah sebuah metode yang digunakan dalam pemrosesan bahasa untuk mengetahui seberapa banyak sebuah kata muncul dalam suatu dokumen. *Term Frequency* bekerja dengan cara mencari frekuensi banyaknya kata yang ingin dicari dengan dokumen yang ada (Simha,2021). Sedangkan *Inverse Document Frequency* melihat seberapa banyak sebuah kata muncul dalam seluruh koleksi dokumen dengan cara membagi total jumlah dokumen dengan total jumlah dokumen yang mengandung kata yang diinginkan (Simha, 2021). Untuk mendapatkan nilai TF-IDF, harus dilakukan perkalian antara *Term Frequency* dan *Inverse Document Frequency* sehingga dapat menunjukkan seberapa penting sebuah kata dalam dokumen dibandingkan dengan seluruh koleksi dokumen. Semakin banyak sebuah kata yang terdapat pada koleksi dokumen akan menurunkan nilai TF-IDF. Formula dari *term frequency inverse document frequency* dapat dilihat sebagai berikut.

$$TF = \frac{Jumlah\ kemunculan\ kata\ pada\ dokumen}{Jumlah\ total\ kata\ pada\ dokumen} \quad (2.1)$$

$$IDF = \frac{Total\ jumlah\ dokumen\ pada\ koleksi}{Jumlah\ dokumen\ yang\ mengandung\ kata + 1} \quad (2.2)$$

$$TFIDF = TF \times IDF \quad (2.3)$$

2.1.2 Phrase Recognition

Phrase Recognition adalah sebuah teknik untuk mengidentifikasi sebuah frasa atau *multi* kata. Suatu *multi* kata dapat dikatakan sebagai konstituen yang dimana memiliki jabatan yang lebih tinggi dari kata individual, karena ia dapat menyimpulkan dengan arti lebih ketika diartikan secara berkesinambungan yang dimana apabila diartikan secara individual akan memiliki kesimpulan yang berbeda (Ortmann, 2021). *Phrase Recognition* dapat membantu beberapa

tugas pemrosesan bahasa antara lain analisis teks, klasifikasi teks, dan pemahaman konten. Dengan menggunakan *Phrase Recognition* kata kata *multi* kata dapat diidentifikasi sehingga pemrosesan bahasa dapat mendapatkan hasil yang maksimal dan relevan. Sebagai contoh "Apa kabar dunia" akan dapat dibagi seperti berikut

- Unigram : "apa", "kabar", "dunia"
- Bigram : "apa kabar", "kabar dunia"
- Trigram : "apa kabar dunia"

2.1.3 Cosine Similarity

Cosine Similarity digunakan untuk mengukur kemiripan dari 2 vektor yang didapat dari kata kata pada suatu dimensi. Pengukuran dilakukan dengan menghitung seberapa kecil perbedaan sudut kosinus antara 2 vektor (Park et al, 2020). Nilai yang dihasilkan berada pada nilai 0 sampai 1, dimana nilai semakin mendekati 1 akan dinilai semakin mirip karena menunjukkan bahwa kedua vektor memiliki arah yang mirip atau sejajar. *Cosine Similarity* sendiri banyak digunakan untuk membandingkan dokumen dokumen dalam bidang *text mining* (Lindang et al, 2022). Formula dari *cosine similarity* dapat dilihat sebagai berikut.

$$Cosine = \frac{A \cdot B}{||A|| \times ||B||} \quad (2.4)$$

Keterangan :

$A \cdot B$: perkalian *dot product* vektor A dan B

$||A||$: panjang vektor A

$||B||$: panjang vektor B

2.2 Tinjauan Studi

Pada skripsi ini akan menggunakan beberapa penelitian lain yang berkaitan dan telah dilakukan sebelumnya sebagai tinjauan studi, berikut adalah beberapa penelitian yang sudah dilakukan dalam topik Sistem Pengecekan Kemiripan Antar Teks:

2.2.1 Assessing Short Answers in Indonesia Using Semantic Text Similarity Method and Dynamic Corpus (Hasanah et al, 2020)

Masalah yang diangkat di penelitian ini adalah kesulitan dalam penilaian otomatis jawaban singkat para pelajar dimana sistem penilaian otomatis yang sudah ada kurang dapat

bekerja dengan baik dengan bahasa indonesia. Hal ini disebabkan karena kurangnya data mengenai bahasa indonesia di *website* seperti Thesaurus dan WordNet.

Metode yang digunakan di penelitian ini adalah dimana peneliti menggunakan *Dynamic Corpus* dengan mengambil data jawaban dari 5 pelajar dengan nilai tertinggi atau menyiapkan 5 data jawaban untuk setiap pertanyaan. Hasil dari *Dynamic Corpus* juga dibantu dengan menggunakan modul Gensim. Setelah itu akan diproses menggunakan metode *Semantic Text Similarity* untuk mengetahui tingkat kemiripan jawaban.

Hasil dari penelitian ini adalah dihasilkannya sebuah sistem penilaian otomatis untuk jawaban singkat bahasa indonesia para pelajar yang dapat memiliki keakuratan mendekati penilaian manusia.

Perbedaan penelitian yang dilakukan dengan skripsi ini adalah penelitian ini berfokus pada kemiripan jawaban singkat bahasa indonesia, sedangkan skripsi ini berfokus pada kemiripan topik skripsi mahasiswa berdasarkan judul, kata kunci, abstrak, latar belakang dan kesimpulan.

2.2.2 Sentence Similarity Measurement for Bengali Abstractive Text Summarization (Masum, 2019)

Masalah yang diangkat pada penelitian ini adalah untuk mendapatkan keakuratan dalam ringkasan yang dibuat oleh mesin dengan manusia. Sehingga membutuhkan sebuah sistem yang dapat memberikan nilai keakuratan ringkasan antara mesin dan manusia agar peneliti dapat meningkatkan keefektifitasan sistem ringkasannya.

Metode yang digunakan dalam penelitian ini adalah data akan diubah menjadi bentuk vektor menggunakan *Word2Vec* yang nantinya akan dihitung jarak kemiripan suatu kata menggunakan *Word Mover's Distance*. Peneliti juga menggunakan *Jaccard Similarity* untuk mendukung metode yang digunakan.

Hasil dari penelitian ini adalah dihasilkannya nilai keakuratan kemiripan jawaban antara mesin dan manusia sehingga peneliti dapat meningkatkan keefektifitasan sistem ringkasannya.

Perbedaan penelitian yang dilakukan dengan skripsi ini adalah tujuan yang ingin dicapai oleh peneliti, dimana penelitian ini bertujuan untuk mendapatkan data nilai keakuratan kemiripan sistem ringkasannya dengan manusia. Sedangkan skripsi ini bertujuan untuk mendapatkan dan menginformasikan kepada mahasiswa mengenai apakah topik skripsi yang diajukan sudah pernah dibuat sebelumnya.