BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Tinjauan pustaka yang dilakukan oleh peneliti mencari beberapa penelitian yang terkait, baik itu secara objek penelitian atau pun secara metode. Peneliti mengambil artikel dari beberapa sumber yang didapat. Perbedaan dengan penelitian yang akan dilakukan.

Penelitian pertama Nur Alamsyah dan Muhammad Rasyidan (2019) melakukan penelitian dengan judul "Deteksi Plagiarisme Tingkat Kemiripan Judul Skripsi Pada Fakultas Teknologi Informasi Menggunakan Algoritma Winnowing". Metode yang digunakan Menggunakan Algoritma Winnowing. Menghasilkan sistem yang dapat melakukan proses pengecekan otomatis dan menampilkan hasil tingkat plagiarisme yang digunakan oleh koordinator tugas akhir. Perbedaan dengan penelitian yang akan dilakukan yaitu metode cosine similarity disertai dengan pembobotan TF-IDF, selain itu akan dibuat sebuah tools sehingga bisa dicek secara langsung. Dalam penelitian ini tidak dilakukan proses uji sedangkan penelitian yang akan dilakukan akan dibandingkan antara menggunakan stemming atau tidak.

Penelitian kedua Ade Riyani, Muhammad Zidny Naf'an, Auliya Burhanuddin (2019) melakukan penelitian dengan judul "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen". Metode yang digunakan Cosine Similarity dan Pembobotan TF-IDF. Metode cosine similarity dan

pembobotan TF-IDF telah berhasil mendeteksi kemiripan suatu dokumen. Proses stemming pada preprocessing sangat berpengaruh terhadap nilai kemiripan hasil jika dilakukan stemming lebih tinggi Perbedaan dengan penelitian yang akan dilakukan akan dibuat sebuah aplikasi yang bisa membantu untuk proses pengujian yang dilakukan. Selain itu data yang dibutuhkan berupa luang lingkup dan judul skripsi sehingga lebih mencakup banyak aspek dalam kategori data.

Penelitian ketiga Eric Siswanto, Yo Ceng Giap (2020) melakukan penelitian dengan judul "Implementasi Algoritma Rabin-Karp Dan Cosine Similarity Untuk Pendeteksi Plagiarisme Pada Dokumen". Metode yang digunakan Cosine Similarity dan Rabin-Karp. Algoritma Rabin-karp dan metode cosine similarity dapat di terapkan di aplikasi sesuai dengan rule. Perbedaan dengan penelitian yang akan dilakukan pada penelitian ini data akan dibandingkan dengan sumber data tidak terbatas setiap perbandingan dengan 2 dokumen, selain itu dilakukan pembobotan terlebih dahulu dengan pembobotan TF-IDF.

Penelitian keempat yang dilakukan oleh Windy Pramudita, dkk (2021) melakukan penelitian dengan judul "Design of Undergaduated Thesis Plagiarism Detection System (Title and Anstract), Based on Matlab, Using Winnowing Algorithm". Metode yang digunakan Menggunakan Algoritma Winnowing. Hasil dari penelitian ini merupakan aplikasi yang dibangun dengan tools matlab yang mampu melakukan deteksi nilai kemiripan cukup baik dengan data uji sejumlah 400 judul. Perbedaan dengan penelitian yang akan dilakukan penggunaan metode yang

berbeda yaitu akan menggunakan cosine similarity dan target data yang digunakan menggunakan judul dan ruang lingkup, selain itu implementasi dalam bentuk web.

Penelitian kelima yang dilakukan oleh Bei Harira Irawan, Manase Sahat H Simarangkir dan Erlinna (2021) melakukan penelitian dengan judul "Deteksi Kemiripan Judul Skripsi Menggunakan Algoritma Levenshtein Distance Pada Kampus STMIK Mic Cikarang". Hasil penelitian berupa perhitungan dengan mampu melakukan deteksi kemiripan judul dengan kategori yang ada dengan nilai yang cukup rendah. Perbedaan dengan penelitian yang akan dilakukan objek data yang dilakukan tidak hanya judul akan tetapi berupa ruang lingkup, selain itu metode yang digunakan juga berbeda yaitu dengan terlebih dahulu dilakukan dengan pembobotan dan perbandingan dengan cosine similarity.

Dari hasil penjabaran penelitian terkait dapat dilihat tabel perbandingan penelitian yang dapat dilihat dalam tabel 2.1

Tabel 2.1 Kajian Penelitian Terdahulu

Nama Peneliti	**Topik**	**Metode**	**Hasil**	**Perbandingan Penelitian**
Nur Alamsyah dan Muhammad Rasyidan (2019)	Deteksi Plagiarisme Tingkat Kemiripan	Algoritma Winnowing	Sistem dapat menampilkan hasil tingkat	Penelitian yang akan dilakukan dengan metode yang berbeda di

Nama Peneliti	Topik	Metode	Hasil	Perbandingan Penelitian
	Judul Skripsi		plagiarisme yang digunakan oleh koordinator tugas akhir.	dahului dengan metode TF-IDF selain itu juga dilakukan perbandingan akurasi dengan tahapan stemming dan tanpa stemming.
Ade Riyani,Muhammad Zidny Naf'an, Auliya Burhanuddin (2019)	Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen	Cosine Similarity dan Pembobotan TF-IDF	Metode cosine similarity dan pembobotan TF-IDF telah berhasil mendeteksi	Penelitian yang dilakukan perbandingan akurasi dengan tahapan stemming dan tidak selain itu dokumen yang di lakukab

Nama Peneliti	Topik	Metode	Hasil	Perbandingan Penelitian
			kemiripan suatu dokumen. Proses stemming pada tahapan processing sangat berpengaruh terhadap nilai kemiripan hasil jika dilakukan stemming lebih tinggi	perbandingan merupakan judul skripsi dan ruang lingkup.
Eric Siswanto, Yo	Algoritma	Cosine	Algoritma	Penelitian yang

Nama Peneliti	Topik	Metode	Hasil	Perbandingan Penelitian
Ceng Giap(2020)	Rabin-Karp Dan Cosine Similarity dalam Pendeteksi Plagiarisme Pada Dokumen	Similarity dan Rabin-Karp	Rabin-karp dan metode cosine similarity dapat di implementasikan di aplikasi sesuai dengan rule.	akan dilakukan dengan metode yang berbeda di dahului dengan metode TF-IDF selain itu juga dilakukan perbandingan akurasi dengan tahapan stemming dan tanpa stemming
Windy Pramudita, dkk (2021)	Deteksi pagiarisme judul dan abstract Skripsi.	Algoritma Winnowing	aplikasi yang dibangun dengan tools matlab yang mampu	Penelitian yang akan dilakukan dengan metode yang berbeda di dahului dengan metode TF-IDF

Nama Peneliti	Topik	Metode	Hasil	Perbandingan Penelitian
			mendeteksi nilai kemiripan cukup bak dengan data uji sejumlah 400 judul.	selain itu juga dilakukan dengan implementasi sistem berbasis web
Bei Harira Irawan, Manase Sahat H Simarangkir , Erlinna (2021)	Deteksi Kemiripan Judul Skripsi Menggunakan Algoritma Levenshtein Distance	Algoritma Levenshtein Distance	Mampu mendeteksi kemiripan judul dengan kategori yang ada dengan nilai yang cukup rendah.	Penelitian yang akan dilakukan dengan metode yang berbeda di dahului dengan metode TF-IDF selain itu juga dilakukan perbandingan akurasi dengan tahapan

Nama Peneliti	Topik	Metode	Hasil	Perbandingan Penelitian
				stemming dan tanpa stemming selain itu tidak hanya judu lakan tetapi juga dengan ruang lingkup.

Dari hasil kajian pustaka yang dilakukan peneliti akan membangun sebuah sistem deteksi kemiripan judul skripsi dengan metode cosine similarity dengan pembobotan dengan TF-IDF. Sistem yang dibangun mampu dapat melakukan pre processing text terlebih dahulu dengan tahapan stemming untuk merubah kata imbuhan ke kata dasar serta adanya filter *stopword* untuk menghilangkan kata hubung atau kata sambung. Hasil dari aplikasi berupa tingkat kemiripan yang ditampilkan dalam nilai presentase 0 sampai dengan 100%. Selain itu sistem akan dilakukan perbandingan dengan proses yang dilakukan dengan stemming maupun yang tidak menggunakan tahapan stemming. Sistem ini akan dibangun dengan

teknologi berbasis web dengan bahasa program PHP dan basis data MySQL dengan bantuan library sastrawi untuk preprocessing dokumen

2.2 Text Mining

Text mining merupakan penambangan teks atau secara luas di definisikan sebagai proses pengetahuan intensif dimana pengguna melakukan interaksi dengan koleksi dokumen dari waktu ke waktu dengan menggunakan seperangkat alat analisis (All Farizi, 2015). Text mining merupakan penambangan data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa hubungan antar dokumen.

Text mining memiliki tugas-tugas yang berhubungan dengan analisa teks dengan jumlah yang besar, penemuan pola serta penggalian informasi yang mungkin berguna dari suatu teks dapat dilakukan. Dalam text mining untuk mendapatkan pola dari suatu teks, sumber- sumber data yang akan diolah adalah dari koleksi dokumen. Pola-pola menarik tersebut tidak ditemukan diantara catatan database yang sudah normal melainkan dalam data teks yang tidak terstruktur di dalam koleksi dokumen-dokumen tersebut. Secara garis besar dalam melakukan implementasi text mining terdiri dari dua tahap besar yaitu pre processing dan processing.

2.2.1 Text Prepocessing

Text pre processing adalah tahapan untuk mempersiapkan teks menjadi data yang akan mengalami pengolahan pada tahapan berikutnya. Inputan awal pada proses ini adalah berupa dokumen utuh (Mustaqhfiri, 2012). Text *pre processing* pada penelitian ini terdiri dari beberapa tahapan, yaitu: proses case folding, proses pemecahan kalimat menjadi kata (*tokenizing*), proses filtering kata dengan menghilangkan kata stopword, dan proses stemming.

a. *Case Folding*

Case folding adalah tahapan proses mengubah semua huruf dalam teks dokumen menjadi huruf kecil semua, serta menghilangkan karakter selain a-z dan dianggap sebagai delimiter.

b. Pemecahan Kalimat (*Tokenizing*)

Pemecahan kalimat yaitu proses memecah string teks dokumen yang panjang menjadi kumpulan kalimat-kalimat. Dalam memecah dokumen menjadi kalimat-kalimat menggunakan fungsi explode(), dengan tanda spasi “ ” sebagai delimiter untuk memotong string dokumen. Dengan menghilangkan tanda-tanda tersebut dokumen akan terpotong menjadi kata.

c. *Filtering* kata / *stopword*

Filtering merupakan proses menghilangkan *stopword*. *Stopword* adalah kata-kata yang sering kali muncul dalam dokumen namun artinya tidak deskriptif dan tidak memiliki keterkaitan dengan tema tertentu. Di dalam bahasa Indonesia *stopword* adalah kata penghubung dan dapat disebut sebagai kata tidak penting, misalnya “di”, ”oleh”, “pada”, ”sebuah”, ”karena” dan lain sebagainya.

d. *Stemming*

Stemming adalah proses mencari akar (root) kata dari tiap token kata yaitu dengan pengembalian suatu kata yang mempunyai imbuhan ke bentuk dasarnya (*stem*). Seperti contoh kata "memakan" akan ke bentuk dasar menjadi kata "makan" dengan menghilangkan imbuhan awal "me".

2.2.2 Processing

Tahap processing adalah tahap terpenting dari seluruh proses text mining. Tahap ini berusaha menemukan pola atau pengetahuan dari keseluruhan teks. Teknik yang di gunakan pada tahap ini adalah dengan melakukan pembobotan (*weighting*) terhadap term dari hasil tahap *pre pocessing*. Setiap term di berikan bobot sesuai dengan skema pembobotan yang di pilih, baik itu pembobotan lokal, global atau kombinasi keduanya. Banyak aplikasi menerapkan pembobotan kombinasi berupa hasil kali bobot lokal *term frequency* dan *global inverse document frequency* yang ditulis dengan TF-IDF.

2.3 TF-IDF

TF-IDF (*Term Frequency Inverse Document Frequency*) merupakan metode yang digunakan untuk menentukan nilai frekuensi sebuah kata di dalam sebuah dokumen atau artikel dan juga frekuensi di dalam banyak dokumen. Perhitungan ini menentukan seberapa relevan sebuah kata di dalam sebuah TF-IDF (Evan, 2014). TFIDF adalah sebuah algoritma yang umumnya digunakan untuk pengolahan data besar. Algoritma TF-IDF melakukan pemberian bobot pada setiap kata kunci di setiap kategori untuk mencari kemiripan kata kunci dengan kategori yang tersedia. Sebelum melakukan pembobotan maka akan dilakukan lima tahap pencarian text *pre processing* yaitu pemecahan kalimat, *case folding, tokenizing, filtering,* dan *stemming*, lalu selanjutnya dilakukan proses menghitung bobot TF-IDF, bobot *query relevance* dan bobot similarity (Marlinda dan Rianto, 2013).

Nilai TF-IDF meningkat secara proporsional berdasarkan jumlah atau banyaknya kata yang muncul pada dokumen, tetapi diimbangi dengan frekuensi kata dalam korpus. Variasi dari skema pembobotan TF-IDF sering digunakan oleh mesin pencari sebagai alat utama dalam mencetak nilai (*scoring*) dan peringkat (*ranking*) sebuah relevansi dokumen yang diberikan user.

TF-IDF pada dasarnya merupakan hasil dari perhitungan antara TF (Term Frequency) dan IDF (Inverse Document Frequency). Banyak cara untuk menentukan nilai yang tepat dari kedua statistik yang ada. Dalam kasus term frequency tf (t, d), cara yang paling sederhana adalah dengan menggunakan raw frequency di dalam

dokumen, yaitu berapa kali term t muncul di dokumen d. Jika menyatakan raw frequency t sebagai f (t,d), maka skema tf yang sederhana adalah tf (t, d) = f (t,d). Nilai idf sebuah term (kata) dapat dihitung menggunakan persamaan 2.1

$$IDF = log10\ (\frac{D}{dfi}) \quad (2.1)$$

D adalah jumlah dokumen yang berisi term (t) dan dfi adalah jumlah munculnya (frekuensi) kata terhadap D.

Adapun algoritma yang digunakan untuk menghitung bobot (W) masing - masing dokumen terhadap kata kunci (query) dapat dilihat dalam persamaan 2.2

$$Wd,t\ =\ tf\ d,t\ *\ IDft \ (2.2)$$

Keterangan :

d = dokumen ke–d

t = kata ke–t dari kata kunci

W = bobot dokumen ke–d terhadap kata ke–t

tf = term frekuensi/frekuensi kata

Setelah bobot (W) masing-masing dokumen diketahui, maka dilakukan proses urutan (*sorting*) dengan semakin besar nilai W, semakin besar tingkat kesamaan (*similarity*) dokumen tersebut terhadap kata yang dicari, demikian pula sebaliknya.

2.4 Cosine Similarity

Menurut (Firdaus, 2019) metode *cosine similarity* merupakan metode perhitungan jarak antara vector A dan B yang menghasilkan sudut cosine x diantara kedua vector tersebut. Nilai sudut kosinus antara dua vector menentukan kesamaan dua buah objek yang dibandingkan. dimana nilai terkecil adalah 0 dan nilai terbesar adalah 1. Rumus perhitungan similarity dapat dilihat dalam persamaan 2.3

$$A \cdot B = A_1 B_1 + \cdots + A_n B_n \quad (2.3)$$

Dengan $A \cdot B$ merupakan dot product. Dot product merupakan nilai yang mengekspresikan sudut antara dua vektor. Dot product merupakan scalar nilai hasil dari operasi dua vektor yang memiliki jumlah komponen yang sama. Jika vektor A dan B memiliki komponen sebanyak n maka dot product dapat dihitung dengan rumus berikut:

Dot product dapat dihitung dengan menjumlahkan product dari masing-masing komponen pada kedua vektor. Jika vektor A dan vektor B merupakan vector 3 dimensi, maka perhitungan dot product dapat dilihat dalam persamaan 2.4

$$A \cdot B = A_x B_x + A_y B_y + A_z B_z \quad (2.4)$$

Sedangkan |A| merupakan panjang vektor. Panjang vector dapat dihitung dengan rumus yang dapat dilihat dalam persamaanm 2.5

$$|A| = \sqrt{x_1^2 + x_2^2 + x_1^2 + x_3^2} \quad (2.5)$$

Perhitungan untuk menentukan nilai persentase kemiripan antar dokumen, maka persentase kemiripan didapat dengan melakukan perkalian nilai cosine similarity terhadap 100. Berikut rumus untuk menentukan nilai persentase kemiripan yang dapat dilihat dalam persamaan 2.6

$$\text{Kemiripan } (\%) = \text{Sim}(\text{A}, \text{B}) \quad (2.6)$$