**Pembobotan *Vector Space Model* Korpus Twitter Tentang *Data Science Text Mining Text Retrieval* Menggunakan *Cosine Smiliarity***

**TUGAS KELOMPOK**

**Disusun Untuk Memenuhi Tugas Mata Kuliah Temu Kembali Informasi**

**Dosen Pengampu : Retnani Latifah, M.Kom**

**Disusun Oleh :**

|  |  |
| --- | --- |
| **MUHAMMAD REZA** | **2019470055** |
| **SELAMET SAPUTRA** | **2019470069** |
| **SYECHAN AHMAD ZIDAN** | **2019470110** |

PROGRAM STUDI TEKNIK INFORMATIKAFAKULTAS TEKNIKUNIVERSITAS MUHAMMADIYAH JAKARTA2022

# **DAFTAR ISI**

[**DAFTAR ISI** ii](#_Toc120082097)

[**DAFTAR GAMBAR** iii](#_Toc120082098)

[**BAB 1** 1](#_Toc120082099)

[**1.1.** **Latar belakang masalah** 1](#_Toc120082100)

[**1.2.** **Identifikasi Masalah** 2](#_Toc120082101)

[**1.3.** **Rumusan Masalah** 2](#_Toc120082102)

[**1.4.** **Batasan Masalah** 2](#_Toc120082103)

# **DAFTAR GAMBAR**

# **BAB 1**

## **Latar belakang masalah**

Saat ini sedang gencarnya *data science, text mining,* dan *text retrieval. Data science* menurut David M. Blei dan Padhraic Smyth adalah turunan atau perhitungan menggunakan statistika untuk melakukan prediksi (Blei & Smyth, 2017). *Text mining* adalah pengolahan dari kumpulan document yang dipecah menjadi teks untuk mengetahui informasi yang bermanfaat dengan menggunakan perhitungan kalkulasi matematika (Sabrani et al., 2020). I*nformation retrieval* atau pengambilan informasi atau text retrieval adalah tugas untuk mengambil informasi yang sesuai atau relevan dari kumpulan korpus yang mewakili permintaan (kueri) (Djenouri et al., 2021).

*Text Mining* adanya *preprocessing* dan ekstraksi fitur, tahap *preprocessing*  terdiri dari *case folding,stop word removal, stemming,word normalization* untuk mengatasi *overfitting* dari hasil *stemming* (Ma’rifah et al., 2020). ekstrasi fitur dalam text mining ada *term* *frequency* melihat setiap kata yang muncul didalam dokumen atau korpus, untuk menghitung *inverse document* diperlukan masing-masing kemunculan *term frequency* di setiap document atau korpus lalu dikalkulasi dengan rumus *Inverse Document Frequency* (IDF) .Maka dari hasil nilai ekstrasi fitur tf-idf ini digunakan untunk perhitung similaritas, dan untuk beberapa metode dalam pendekatan statistika (Setyawan et al., 2021).

*Vector Space Model* (VSM). Sebuah model yang digunakan untuk mengukur sebuah kueri antara suatu dokumen dengan suatu kata kunci atau *keyword* (Susanti et al., 2020). *Vector space* adalah geometri berdimensi besar, ruang yang batas-batasnya ditentukan oleh *vector*. Konsep dasar *vector space model* adalah menghitung jarak vector antara dokumen dengan kata kunci yang dimasukkan kemudian mengurutkan berdasarkan tingkat kedekatannya (Susanti et al., 2020). Salah satu model *Vector Space Model* adalah *Cosine Smiliarity* untuk memodelkan *document* text sebagai vector kata, dengan menggunakan kesamaan antara dua dokumen (Ma’rifah et al., 2020).

## **Identifikasi Masalah**

Berdasarkan permsalahan di latar belakang, permasalahan tersebut dilakukan identifikasi sebagai berikut :

1. Melakukan *preprocessing* *document* teks
2. Melakukan rkstraksi fitur teks menggunakan *Term Frequency Inverse Document Frequency*
3. Menghitung jarak teks dokumen menggunakan *Cosine Smiliarity*

## **Rumusan Masalah**

Berdasarkan permasalahan diatas, akan dilakukan perumusan atau kajian sebagai berikut :

1. Bagaimana melakukan preprocessing data teks *document*  ?
2. Bagaimana cara melakukan ekstraksi fitur text dari *document* menggunakan *Term Frequency Inverse Document Frequency* ?
3. Bagaimana cara menghitung jarak teks *document* twitter menggunakan *Cosine Smiliarity*  ?

## **Batasan Masalah**

1. Mengolah teks *document* dari twiiter tentang *data science text mining text retrieval*
2. Menggunakan data dari *twitter* berjumlah 83 dokument
3. Mengetahui hasil kedekatan jarak antara *document*