**BAB 2**

# **LANDASAN TEORI**

## **Landasan Teori**

1. **Twitter**

Twitter merupakan salah satu sosial media yang digunakan untuk berinteraksi secara daring dengan pengguna lain dan atau bot. Pada twitter sendiri, terdapat istilah *tweet* yaitu teks status pengguna yang digunakan untuk memberikan informasi (Analisis Sentimen Data Twitter Tentang Pasangan Capres-Cawapres Pemilu 2019 Berbasis Metode Lexicon Dan Support Vector Machine).

1. **Text Mining**

Text mining menggunakan proses preprocessing dokumen, proses ini dapat menghasilkan sentimen kelas positif, negatif, dan netral (PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA). Text mining menurut Nurhuda dan Sihwi, proses menemukan informasi dari sekumpulan dokumen teks menggunakan metode analisis tertentu (Analisis Sentimen Data Twitter Tentang Pasangan Capres-Cawapres Pemilu 2019 Berbasis Metode Lexicon Dan Support Vector Machine). Umumnya, tahapan pada text mining adalah sebagai berikut:

1. *Tokenizing* merupakan tahapan pemotongan string input per kata penyusunnya ([Microsoft PowerPoint - Text Mining (pens.ac.id)](http://tessy.lecturer.pens.ac.id/kuliah/dm/6Text%20Mining.pdf)). Contohnya “Saya Ingin Makan” akan menjadi “Saya”, “Ingin”, “Makan”.
2. *Filtering* merupakan tahapan mengambil kata-kata penting dari hasil *Tokenizing*, salah satu algoritma yang digunakan adalah algoritma *stop list* atau *word list*.
3. *Stemming* merupakan tahapan mencari kata dasar dari tiap kata hasil *filtering*. Contohnya “Mengambil” menjadi “Ambil”
4. *Tagging* merupakan tahapan mencari bentuk awal dari tiap kata lampai atau kata hasil *stemming*.
5. *Analyzing* merupakan tahap penentuan keterhubungan kata-kata antar dokumen.
6. **Sentiment Analysis**

*Sentiment analysis* juga bisa disebut *Opinion Mining* merupakan sebuah riset yang meneliti opini, sentiment dan emosi secara tekstual lalu diklasifikasi menjadi kelompok sentiment negatif, netral dan positif (PENERAPAN SENTIMENT ANALYSIS PADA HASIL EVALUASI DOSEN DENGAN METODE SUPPORT VECTOR MACHINE).

1. **Pembobotan TF-IDF**

Bobot TF-IDF merupakan penilaian bobot setiap kata yang ada pada sebuah dokumen (PENERAPAN SENTIMENT ANALYSIS PADA HASIL EVALUASI DOSEN DENGAN METODE SUPPORT VECTOR MACHINE), TF-IDF merupakan salmetode pembobotan fitur yang kuat yang mengukur frekuensi dan distribusi kata untuk menentukan bobotnya. TF-IDF terdiri dari TF (*Term Frequency*) yaitu frekuensi kemunculan kata dalam dokumen dan IDF (*Inverse Document Frequency*) yaitu kebalikan dari frekuensi dokumen, nilai TF-IDF dapat diperoleh dengan menggunakan persamaan berikut:

1. Nilai *Term Frequency* (TF) diperoleh dengan nilai frekuensi kemunculan fitur *t* pada dokumen *d*
2. Nilai Inverse Document Frequency (IDF) diperoleh dari logaritma banyaknya dokumen *n* dibagi dokumen *df* yang mengandung fitur *t*
3. Nilai Term Frequency Inverse Document Frequency (TF-IDF) didapatkan dengan mengalikan nilai *TF* dan *IDF*

(PENERAPAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DENGAN TF-IDF N-GRAM UNTUK TEXT CLASSIFICATION).

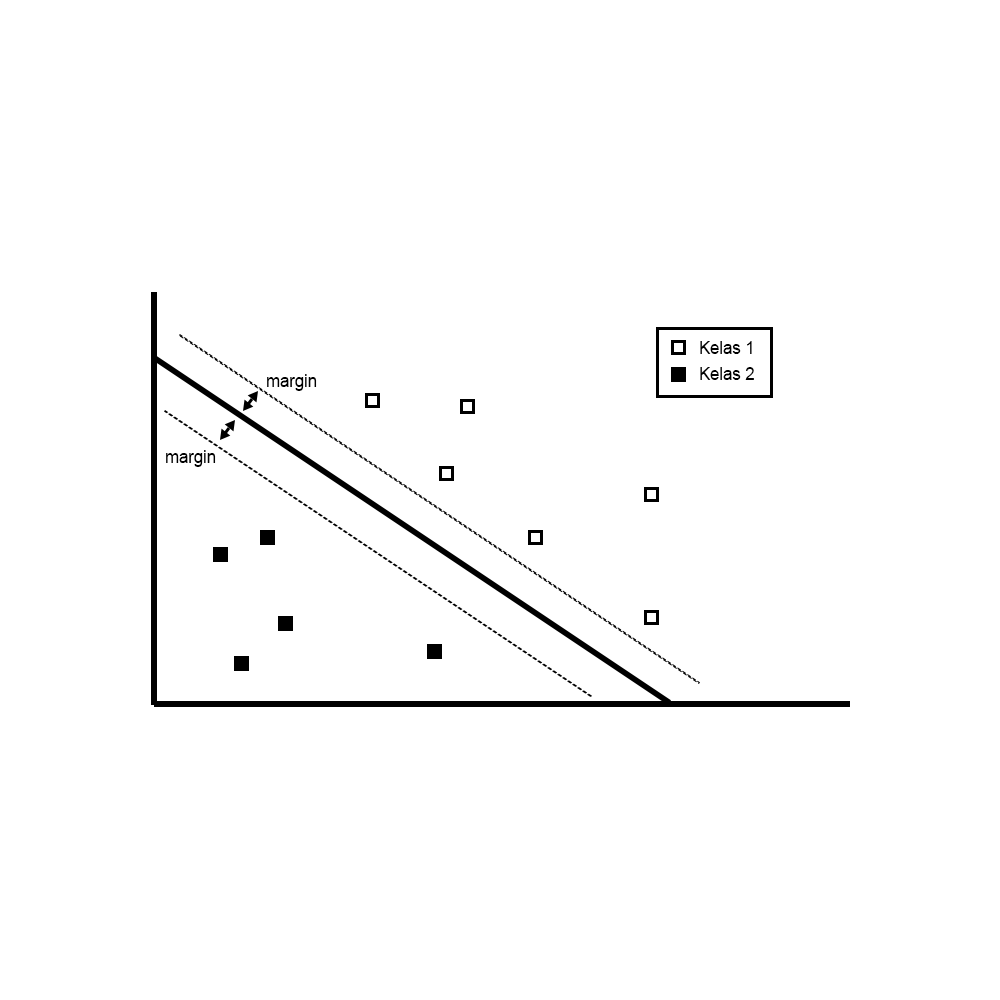
1. **Support Vector Machine (SVM)**

Support Vector Mahine diciptakan oleh Vladimir Vapnik yang merupakan salah satu metode klasifikasi dengan metode machine learning (Supervised Learning) yang akan melakukan prediksi kelas berdasarkan hasil *training* (PENERAPAN SENTIMENT ANALYSIS PADA HASIL EVALUASI DOSEN DENGAN METODE SUPPORT VECTOR MACHINE). Algoritma ini akan menghasilkan *hyperplane* yang memisahkan sebuah plane menjadi 2 bagian pada masing-masing kelas yang dapat dilihat pada gambar 1. Berikut adalah rumus perhitungan *hyperplane* pada algoritma SVM:

Dengan keterangan:

W = parameter *hyperlane* yang dicari (garis tegak lurus antara garis *hyperlane* dengan titik support vector)

x = titik data masukan SVM

b = parameter hyperlane yang dicari (nilai bias)

Beberapa kernel yang dapat diterapkan pada SVM:

1. *Linear* Kernel merupakan fungsi kernel yang memisahkan 2 kelas dengan persamaan sebagai berikut:
2. *Polynomial* Kernel merupakan fungsi kernel yang tidak terpisah secara *linear*, persamaan sebagai berikut:
3. *Radial Basis Function* (RBF) Kernel merupakan fungsi untuk menganalisis data yang tidak terpisah secara linear dan memerlukan 2 jenis parameter (Gamma dan Cost).

(Analisis Sentimen Data Twitter Tentang Pasangan Capres-Cawapres Pemilu 2019 Berbasis Metode Lexicon Dan Support Vector Machine)

1. **Python**

Python merupakan bahasa pemrograman digunakan oleh banyak platform oleh developer seperti website, mobile dan desktop (PEMBELAJARAN PEMROGRAMAN PYTHON DENGAN PENDEKATAN LOGIKA ALGORITMA). Namun sejak berkembangnya *Machine Learning* python pun turut digunakan terutama dengan ketersediaan banyaknya library seperti scikit-learn, keras, tensorflow, openCV, dan lain-lain.

1. **K-Fold Cross Validation**

K-Fold Cross Validation merupakan salah satu bentuk pengujian yang berfungsi untuk menilai kinerja algoritma dengan cara kerja membagi sampel data secara acak lalu membagi data tersebut sebanyak nilai K dari k-fold. Selanjutnya, salah satu kelompok K tersebut akan dijadikan sebagai data *testing* dan sisa kelompok akan dijadikan sebagai data *training* (Penerapan Algoritme Modified K-Nearest Neighbour Pada Pengklasifikasian Penyakit Kejiwaan Skizofrenia).

Untuk perhitungan akurasi akan dihitung dengan rumusan berikut:

## **Penelitian Terdahulu**

Berikut ini merupakan penelitian yang menjadi referensi penulis dalam melakukan penelitian ini:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Penulis | Tahun | Judul | Kesimpulan |
| 1 | Valonia Inge Santoso, Gloria Virginia, Yuan Lukito | 2017 | PENERAPAN SENTIMENT ANALYSIS PADA HASIL EVALUASI DOSEN DENGAN METODE SUPPORT VECTOR MACHINE | Sistem berhasil melakukan *sentiment analysis* dengan metode SVM terhadap evaluasi dosen FTI UKDW Prodi Teknik Informatika tahun ajaran 2014/2015 menggunakan validasi *K-Fold Cross Validation* dengan akurasi tertinggi pada k = 6 sebesar 67.83%. |
| 2 | Fadholi Fat Haranto, Bety Wulan Sari | 2019 | IMPLEMENTASI *SUPPORT VECTOR MACHINE* UNTUK ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP PELAYANAN TELKOM DAN BIZNET | Penelitian sentiment pengguna twitter terhadap biznet dan telkom menggunakan algoritma SVM mendapat hasil rata-rata akurasi sebesar 79.6%, precision 76.5%, recall 72.8%, dan f1-score sebesar 74.6% untuk data Telkom. Sedangkan rata-rata akurasi 83.2%, precision 78.8%, recall 71.6%, dan f1-score 75% untuk Biznet. |
| 3 | Sri Diantika, Windu Gata, Hiya Nalatissifa, Mareanus Lase | 2021 | Komparasi Algoritma SVM Dan Naive Bayes Untuk Klasifikasi  Kestabilan Jaringan Listrik | Akurasi algoritma *Support Vector Machine* pada data kestabilan jaringan mendapatkan akurasi 98.8% semestara jika menggunakan *Naïve Bayes* sebesar 97.64%. |
| 4 | Fatihah Rahmadayana, Yuliant Sibaroni | 2021 | *Sentiment Analysis of Work from Home Activity using SVM with Randomized Search Optimization* | Penelitian ini menunjukan bahwa penggunakan metode *acronym extension* terhadap topik *Work From Home* mendapatkan hasil F1-score tertinggi sebesar 83.362%. Sedangkan pada penerapan SVM dengan *Randomized Search Optimization* menunjukan hasil sebesar 62.35% *tweet* positif, dan 37.65% *tweet* negatif. |
| 5 | Dedi Darwis, Eka Shintya Pratiwi, A. Ferico Octaviansyah Pasaribu | 2020 | PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA  DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK  INDONESIA | Hasil klasifikasi menggunakan metode SVM dibagi menjadi 3 kelas, yaitu 8% positif, 15% netral, dan 77% negatif. Lalu hasil pengujian nilai akurasi, precession, recall, dan F1-score dapat disimpulkan bahwa sentimen masyarakat terhadap kinerja KPK sangat kurang baik dengan presentase negatif sebesar 77% serta pengujian hasil akurasi sebesar 82%, precission 90%, recall 88%, dan f1-score 89% |

## **Kerangka Berpikir**

Kerangka alur penelitian yang dilakukan dalam membuat sistem *sentiment analysis* terhadap topik pemilu 2024. Dimulai dari mendapatkan dataset *tweet* dari kata kunci yang relevan seperti “Ganjar Pranowo”, “Prabowo Subianto”, dan “Anies Baswedan”, lalu melakukan *labeling* *tweet* secara manual oleh penulis, dilanjutkan dengan melakukan *pre-processing* pada dataset seperti melakukan *Cleansing*, *Tokenizing*, *Filtering*, *Stemming*, *Tagging*, dan *Analyzing*. Selanjutnya baru setiap kata diberi bobot dengan metode *TF-IDF* dan dilanjutkan dengan membagi data kedalam data *training* dan data *testing* menurut metode *10-Fold Cross Validation* untuk dimasukan ke dalam algoritma *Support Vector Machine*. Lalu dari setiap fold akan dilihat untuk nilai K terbaik dari segi akurasi, *precission*, *recall*, dan *f1-score*. Lalu hasil prediksi setiap capres akan dibandingkan secara manual oleh survey yang diterbitkan oleh portal berita untuk melakukan perbandingan hasil prediksi dengan survey yang disebarkan.