# **BAB 2**

# **LANDASAN TEORI**

## **Landasan Teori**

### **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 [5].

### ***Text Mining***

*Text mining* menggunakan proses *preprocessing* dokumen, proses ini dapat menghasilkan sentimen kelas positif, negatif, dan netral [6] *Text mining* menurut Nurhuda dan Sihwi, proses menemukan informasi dari sekumpulan dokumen teks menggunakan metode analisis tertentu [5]. Umumnya, tahapan pada *text mining* diawali dengan tahapan *preprocessing* sebagai berikut:

1. *Lowercasing* merupakan tahapan merubah semua huruf menjadi huruf kecil [7].
2. *Puctuation Removal* merupakan tahapan menghapus semua tanda baca karena umumnya tidak mempengaruhi makna sentimen [7]
3. *Tokenizing* merupakan tahapan pemotongan string input per kata menjadi token [7]. Contohnya “Saya Ingin Makan” akan menjadi “Saya”, “Ingin”, “Makan”.
4. *Slang Word Conversion* merupakan tahapan mengubah bahasa gaul atau singkatan menjadi bahasa baku [7]
5. *Stop Word Removal* merupakan tahapan menghapus kata penghubung atau *stopwords* yang tidak membawa makna sentimen apapun [7].
6. *Stemming* merupakan tahapan menghapus imbuhan dan mengambil kata dasar yang tersisa [7].
7. *Synonim* merupakan tahapan menggabungkan beberapa kata yang memiliki arti serupa atau sejenis [8].

### ***Sentiment Analysis***

*Sentiment analysis* juga bisa disebut *Opinion Mining* merupakan sebuah studi komputasi yang meneliti opini, pendapat, perilaku, dan emosi seseorang secara tekstual lalu diklasifikasi menjadi kelompok sentiment negatif, netral dan positif terhadap suatu individu, kejadian, atau topik [9]

### **Pembobotan TF-IDF**

Bobot TF-IDF merupakan penilaian bobot setiap kata yang ada pada sebuah dokumen, TF-IDF merupakan metode 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* [10]

(1)

1. Nilai Inverse Document Frequency (IDF) diperoleh dari logaritma banyaknya dokumen *n* dibagi dokumen *df* yang mengandung fitur *t* [10]

(2)

1. Nilai Term Frequency Inverse Document Frequency (TF-IDF) didapatkan dengan mengalikan nilai *TF* dan *IDF* [10]

(3)

### **Support Vector Machine (SVM)**

Support Vector Mahine diciptakan oleh Boser, Guyon, dan Vapnik [8] yang merupakan salah satu metode klasifikasi dengan tujuan menemukan *hyperplane* terbaik untuk memisahkan 2 kelas [11]. Algoritma ini akan menghasilkan *hyperplane* yang memisahkan sebuah plane menjadi 2 bagian pada masing-masing kelas yang dapat dilihat pada gambar 1.1

Gambar 2. 1 Contoh Hyperplane Support Vector Machine

Beberapa kernel yang dapat diterapkan pada SVM:

1. *Linear* Kernel merupakan fungsi kernel yang memisahkan 2 kelas dengan persamaan sebagai berikut:

(4)

Deskripsi :

= dataset *training*

= 1, 2, …, n

= label dari [8]

1. *Polynomial* Kernel merupakan fungsi kernel yang tidak terpisah secara *linear*, persamaan sebagai berikut:

(5)

Deskripsi :

= *training* data

= konstanta [8]

1. *Radial Basis Function* (RBF) Kernel merupakan fungsi untuk menganalisis data yang tidak terpisah secara linear dan memerlukan 2 jenis parameter (Gamma dan Cost).

(6)

Deksripsi :

= *Euclidean Distance*

= Parameter Independen menentukan tingkat pengurangan

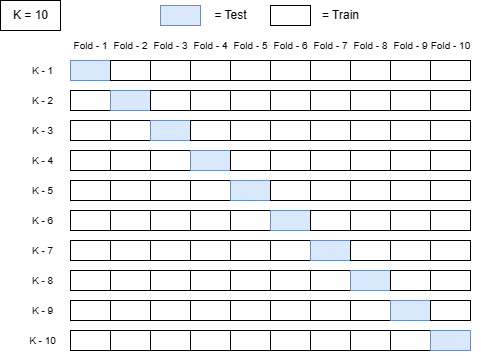
menuju 0 [8]

### **Python**

Python merupakan bahasa pemrograman digunakan oleh banyak platform oleh developer seperti website, mobile dan desktop [12]. Namun sejak berkembangnya *Machine Learning* python pun turut digunakan terutama dengan ketersediaan banyaknya *library* seperti scikit-learn, keras, tensorflow, openCV, dan lain-lain.

### ***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 total *fold*. Selanjutnya, salah satu kelompok K tersebut akan dijadikan sebagai data *testing* dan sisa kelompok akan dijadikan sebagai data *training* [13].

Contoh ilustrasi cara kerja K-Fold Cross Validation ditampilkan pada gambar 1.2

Gambar 2. 2 Ilustrasi Cara Kerja 10-Fold Cross Validation

### ***Confusion Matrix***

*Confusion Matrix* merupakan sebuah metode berupa matriks yang berfungsi untuk menilai kinerja akurasi klasifikasi berdasarkan dataset dan label sebenarnya.

Penerapan confusion matrix terlihat pada tabel berikut:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Prediksi Positif** | **Prediksi Netral** | **Prediksi Negatif** |
| **Label Positif** | True Positif | False Netral | False Negatif |
| **Label Netral** | False Positif | True Netral | False Negatif |
| **Label Negatif** | False Positif | False Netral | True Negatif |

Tabel 2. 1 Contoh Tabel Confusion Matrix

Dengan keterangan True Negatif adalah hasil prediksi dan label sebenarnya negatif. False Negatif adalah hasil prediksi adalah negatif sedangkan label sebenarnya bukan negatif. True Netral adalah hasil prediksi dan label sebenarnya netral. False Netral adalah hasil prediksi netral sedangkan label sebenarnya bukan netral. True Positif adalah hasil prediksi dan label sebenarnya positif. False Positif adalah hasil prediksi positif sedangkan label sebenarnya bukan positif [6].

### **Pearson Correlation**

Pearson Correlation merupakan metode untuk mencari hubungan linear antara 2 variable atau lebih. Hasil dari Pearson Correlation merupakan koefisien korelasi yang berkisar antara angka 0 sampai 1 yang berlaku untuk angka negatif [8]. Dalam menguji validitasnya menggunakan Pearson Product Moment Correlation dengan persamaan sebagai berikut:

(7)

Dimana:

= Jumlah Data

= Hasil Jumlah antara X dan Y setiap pasang

= Hasil Jumlah Variable X

= Hasil Jumlah Variable Y

= Hasil Jumlah Variable X Kuadrat

= Hasil Jumlah Variable Y Kuadrat

## **Penelitian Terdahulu**

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

Tabel 2. 2 Tabel Penelitian Terdahulu

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Penulis** | **Tahun** | **Judul** | **Kesimpulan** |
| 1 | Dionisia Bhisetya Rarasati, Josef Cristian Adi Putra | 2021 | Correlation Between Twitter Sentiment Analysis with Three  Kernels Using Algorithm Support Vector Machine (SVM)  Governor Candidate Electability Level | Pada algoritma Support Vector Machine dengan studi kasus pemilihan gubernur DKI Jakarta, kernel dengan akurasi terbaik adalah Gaussian RBF 90.58%, diikuti dengan Linear 85.87%, dan Polynomial 78.5 |
| 2 | Hendrik Setiawan, Ema Utami, Sudarmawan | 2021 | Analisis Sentimen Twitter Kuliah Online Pasca Covid-19 Menggunakan  Algoritma Support Vector Machine dan Naive Bayes | Penelitian analisis sentiment pengguna twitter terhadap kuliah online pasca covid-19 memiliki akurasi sebesar 85% dengan algoritma SVM, sedangkan akurasi 81.2% menggunakan *Naïve Bayes*. |
| 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 | Elly Indrayuni | 2018 | KOMPARASI ALGORITMA NAIVE BAYES DAN SUPPORT VECTOR  MACHINE UNTUK ANALISA SENTIMEN REVIEW FILM | Penelitian analisis sentiment pengguna twitter terhadap review film menggunakan *confussion matrix* menunjukan bahwa akurasi algoritma SVM sebesar 90% dan untuk algoritma *Naïve Bayes* sebesar 84.5% |
| 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%, precision 90%, recall 88%, dan f1-score 89% |

Sehingga berdasarkan penelitian terdahulu para penulis menggunakan algoritma SVM dan mendapatkan akurasi terbaik dibandingkan algoritma lain dalam melakukan analisis sentimen dengan data dari twitter. Dan untuk kernel yang dipakai akan menggunakan kernel *Gaussian Radial Basis Function* dikarenakan kernel tersebut memiliki tingkat akurasi terbaik berdasarkan jurna [8]. Dikarenakan pemilihan presiden diikuti oleh masyarakat luas, dan pemilihan presiden akan sangat berkaitan dengan pandangan masyarakat terhadap masing-masing calon, oleh sebab itu maka penulis melakukan penelitian *sentiment analysis* dengan algoritma *Support Vector Machine* dalam melakukan analisis sentimen terhadap pemilihan calon presiden tahun 2024.

## **Kerangka Berpikir**

Gambar 2. 3 Gambar 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, *precision*, *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.