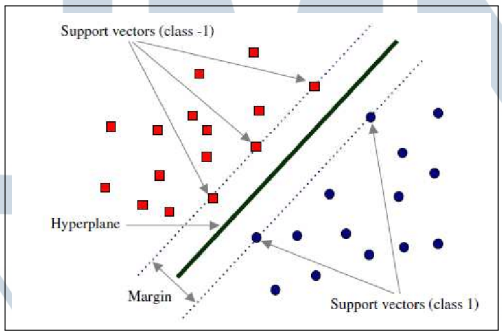
BAB 2
LANDASAN TEORI

2.1 Twitter

Twitter adalah media sosial berbentuk *microblogging* yang bertujuan agar penggunanya dapat mengunggah sesuatu berbasi teks yang disebut *"tweet"* dan menerima ataupun mengirim pesan. Twitter dirancang oleh Jack Dorsey dibulan Maret tahun 2006, lalu dipublikasikan dibulan Juli 2006. Ditahun 2022, Twitter diperkirakan memiliki kurang lebih 437 juta pengguna diseluruh dunia dan menetapkan Twitter sebagai salah satu dari 10 besar sosial media yang paling banyak digunakan [8].

2.2 Support Vector Machine

Support Vector Machine (SVM) adalah metode yang kuat untuk membangun pengklasifikasi. Hal ini bertujuan untuk membuat batas keputusan antara dua kelas yang memungkinkan prediksi label dari satu atau lebih vektor fitur. Batas keputusan ini, yang dikenal sebagai *hyperplane*, diorientasikan sedemikian rupa sehingga sejauh mungkin dari titik data terdekat dari masing-masing kelas. kelas. Titik-titik terdekat ini disebut vektor pendukung [9].



Gambar 2.1. Hyperplane SVM
sumber: IEEE Journal of Emerging and Selected Topics in Power Electronics

Dalam memperoleh garis hyperplane yang terbaik untuk memisahkan kelas disebuah data, maka diimplementasikan perhitungan margin hyperplane dan menemukan titik maksimal. Garis hyperlane didapatkan dengan memakai persamaan berikut.

$$(w.x_i) + b = 0 \tag{2.1}$$

Dalam data Xi , yang tergolong pada kelas -1 dapat dirumuskan seperti persamaan berikut.

$$(w.x_i + b) \leq 1, y_i = -1 \tag{2.2}$$

Sementara data Xi yang tergolong pada kelas +1 dapat dirumuskan seperti persamaan berikut.

$$(w.x_i + b) \geq 1, y_i = 1 \tag{2.3}$$

Keterangan :

- w : Barometer dari garis tegak lurus yang berada diantara garis *hyperplane* dan titik *support vector*) *hyperplane*.
- x : Titik data masukan Support Vector Machine
- y : Pemisahan antara *hyperplane* dan titik data terdekat untuk vektor bobot w dan bias b yang diberikan.
- b : Tolak ukur *hyperplane* yang dicari (nilai bias)

SVM juga merupakan algoritma yang menggunakan pemetaan nonlinier untuk untuk mentransformasikan data pelatihan asli ke dalam dimensi yang lebih tinggi. Dalam algoritma ini, setiap item data diplot sebagai titik dalam ruang ndimensional (dimana n adalah jumlah fitur) dengan nilai dari setiap fitur menjadi nilai dari suatu fitur tertentu. nilai dari setiap fitur menjadi nilai dari koordinat - koordinat tertentu. Kemudian, klasifikasi dilakukan dengan menemukan *hyperplane*, *hyperplane* yang membedakan dua kelas dengan sangat baik sehingga contoh-contoh dari kategori yang terpisah dibagi oleh celah yang jelas yang jelas yang selebar mungkin. Contoh-contoh baru kemudian dipetakan ke dalam ruang yang sama dan diprediksi termasuk dalam kategori-kategori berdasarkan sisi mana dari celah yang mereka jatuhi [10].

Untuk SVM, mungkin ada jumlah tak terbatas untuk memisahkan bidang-bidang *hyper*. Yang terbaik harus ditemukan yang akan memiliki kesalahan klasifikasi minimum pada tupel yang sebelumnya tidak terlihat. *Hyperplane* yang optimal adalah yang memiliki margin terbesar. Itulah sebabnya tujuan dari SVM

adalah untuk menemukan bidang hiper pemisah optimal yang memaksimalkan margin dari data pelatihan. *Hyperplane* pemisah optimal ini disebut *Maximum Marginal Hyperlane* (MMH) [10].

2.3 RUU PDP

RUU PDP adalah sebuah alat hukum yang perlu segera hadir di Indoneasia. Seperti yang tercantum dałam pasal 28 G ayat (1) di UUD 1945 bahwa, "Setiap orang berhak atas perlindungan diri pribadi, keluarga, kehormatan, martabat, dan harta benda yang di bawah kekuasaannya, serta berhak atas rasa aman dan perlindungan dari ancaman ketakutan untuk berbuat atau tidak berbuat sesuatu yang merupakan hak asasi" dan Pasal 28 H ayat (4) UUD tahun 1945 yang berbunyi, "setiap orang berhak mempunyai hak milik pribadi dan hak milik tersebut tidak boleh diambil alih secara sewenang-wenang oleh siapa pun". Oleh karena itu, RUU PDP ditunjukan untuk memenuhi hak masyarakat Inonesia atas perlindungan diri pribadi dan menumbuhkan kesadaran masyarakat serta menjamin pengakuan dan penghormatan atas pentingnya pelindungan data pribadi. Dalam RUU ini, data pribadi memiliki arti sebagai "Setiap data tentang seseorang baik yang teridentifikasi dan/atau dapat di identifikasi secara tersendiri atau dikombinasi dengan informasi lainnya baik secara langsung maupun tidak langsung melalui sistem elektronik dan/atau nonelektronik." Data pribadi terbagi menjadi dua macam yaitu, data umum seperti nama lengkap, agama, tenis kelamin, kewarganegaraan, dan lainya. Dan juga data yang sensitif seperti data biometric kesehatan, genetika, dan data yang bersifat konfidensial. [11].

2.4 Analisis Sentimen

Analisis sentimen adalah sebuah bidang yang lebih luas *natural language processing* yang memiliki tujuan untuk menganalisis sentiment, pendapat, skip, dan penilaian emosi dari individual tersebut tenting sesuatu kejadian, objek, atau individual lainya [12]. Analisis Sentimen dapat digunakan untuk mendapatkan informasi yang representatif seperti mendapatkan presentase sentimen positif dan negatif terhadap hal tertentu. Analisis Sentimen dapat dipecah menjadi tiga nilai yaitu positif, negatif, dan netral. Analisis Sentimen bertujuan untuk melakukan penilaian terhadap sikap yang berpendapat, emosi, kritik atau evualuasi yang disampaikan oleh seseorang terhadap produk ataupun tokoh masyarakat[13].

2.5 Feature Extraction

Proses ini bertujuan untuk mengubah kumpulan data yang besar menjadi fitur yang relevan. Dilakukan dengan memasukan data yang besar menjadi kelompok yang lebih kecil agar lebih mudah diproses. Proses ini dilakukan menggunakan TF-IDF *(term frequency-inverse document frequency)*, TF-IDF adalah ukuran statistik yang melakukan penilaian dari seberapa relevan suatu kata dengan dokumen dalam kumpulan dokumen.

$$tf(t,d) = \frac{n_i j}{\sum_k n_i, j} \tag{2.4}$$

$$idf = log \frac{N}{df_j} \tag{2.5}$$

$$w_{ij} = tf_{ij} \times idf \tag{2.6}$$

Keterangan:

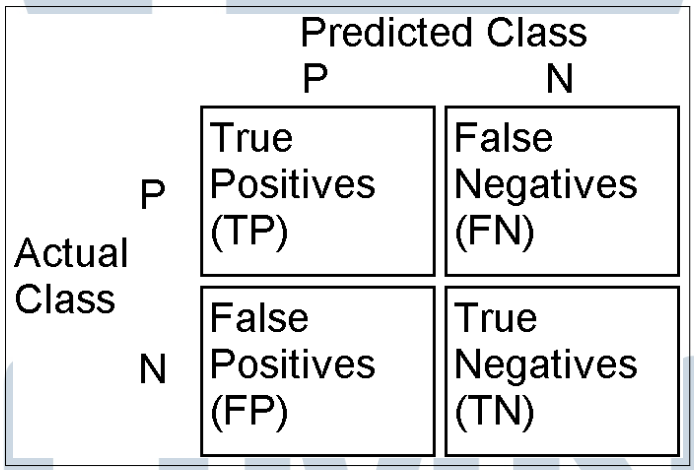
1. TF-lDF(w,d) : bobot suatu kata dalam keseluruhan dokumen
2. w : suatu kata (word)
3. d : suatu dokumen (document)
4. TF(w,d) : frekuensi kemunculan sebuah kata w dalam dokumen d
5. IDF(w) : inverse DF dari kata w
6. N : jumlah keseluruhan dokumen
7. DF(w) : jumlah dokumen yang mengandung kata w
8. $tf(t,d)$ = Frekuensi *term*
9. $n_i j$ = Total suatu *term* yang muncul pada suatu dokumen
10. $\sum_k n_i, j$ = Total seluruh kata dalam suatu dokumen

Cara kerjanya adalah mengalikan dua *matrix*, lalu berapa kali sebuah kata itu akan muncul dalam cuitan dan frekuensi kata tersebut di seluruh kumpulan data .

Bertujuan untuk menganalisis teks otomatis, dan menilai kata-kata dalam algoritme *machine learning* untuk *Natural Language Processing* (NLP).

2.6 Confusion Matrix

Confusion Matrix merupakan indikator penilaian performa untuk permasalahan klasifikasi pada *machine learning*, dimana terdapat dua kelas atau lebih. *Confusion Matrix* juga merupakan tabel dengan empat kombinasi yang terdiri dari nilai prediksi dan nilai aktual. Ada empat istilah yang meru- pakan representasi hasil proses klasifikasi pada confusion matrix yaitu (True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN)) seperti yang bisa dilihat di Gambar2.2. [14]



Gambar 2.2. TPFN
sumber: [15]

Keterangan :

1. TP (*True Positive*): jumlah data yang kelas aktual dan prediksinya merupakan kelas positif
2. FN (*False Negative*): total data yang kelas aktualnya merupakan kelas positif sedangkan kelas prediksinya merupakan kelas negatif.
3. FP (*False Positive*): banyaknya data yang kelas aktualnya merupakan kelas negatif sedangkan kelas prediksinya merupakan kelas positif.

4. TN (*True Negative*): banyaknya data yang kelas aktualnya merupakan kelas negatif sedangkan elas prediksinya merupakan kelas negatif

Berikut diketahui rumus untuk mencari *accuration, precision, recall,* dan *F1-Score* sebagai berikut.

$$Accuration = \frac{TP+TN}{TP+FN+FP+TN} x100 \tag{2.7}$$

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive} \tag{2.8}$$

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative} \tag{2.9}$$

$$F1-Score = 2X\frac{Recall*Precision}{Recall+Precision} \tag{2.10}$$

1. *Accuration* adalah rasio prediksi yang tepat dari seluruh data.
2. *Precision* adalah rasio prediksi positif yang benar dibandingkan dengan seluruh data prediksi positif.
3. *Recall* adalah perbandingan *True Positive* (TP) dengan jumlah data prediksi positif sebenarnya.
4. *F1-Score* adalah rata-rata dari *precision* dan *recall*. Seperti yang diketahui didalam *precision* dan *recall* terdapat *false positive* dan *false negative*sehingga juga perlu untuk mempertimbangkan keduanya.

UNIVERSITAS MULTIMEDIA NUSANTARA