

BAB 2

LANDASAN TEORI

2.1 Pemilu

Pemilu atau pemilihan umum merupakan sarana masyarakat untuk memilih anggota legislatif (DPR, DPD, dan DPRD) dan eksekutif (presiden dan wakil presiden) secara langsung, umum, bebas, rahasia, jujur, dan adil berdasarkan Pancasila dan Undang-Undang Dasar Negara Republik Indonesia Tahun 1945 [13]. Tujuan pemilu adalah untuk mendirikan struktur pemerintahan, meneruskan upaya untuk memaknai kemerdekaan, dan menjaga integritas dari Negara Kesatuan Republik Indonesia.

2.2 Debat Pilpres

Debat merupakan cara untuk menyampaikan ide melalui argumen yang dilakukan secara logis disertai bukti untuk mendukung kasus yang dibahas [14]. Debat pilpres merupakan salah satu bentuk sosialisasi yang memberikan gambaran nyata kepada calon pemilih. Debat pilpres menjadi sarana masyarakat untuk melihat keahlian paslon dalam melakukan komunikasi. Debat pilpres di Indonesia diselenggarakan oleh Komisi Pemilihan Umum (KPU) dengan aturan dan format tertentu yang telah ditetapkan sebelumnya.

2.3 Natural Language Processing

Natural language processing atau NLP merupakan cabang *artificial intelligence* atau AI yang digunakan untuk memahami, memproses, dan menghasilkan bahasa manusia [15]. NLP diimplementasikan pada banyak hal seperti mesin pencari, layanan terjemahan mesin, asisten suara, *chatbot*, dan analisis sentimen.

Tujuan utama dari NLP adalah membuat mesin yang dapat mengerti dan memahami bahasa manusia dengan cara memprosesnya terlebih dahulu untuk menghasilkan respons yang sesuai [16].

2.4 Analisis Sentimen

Analisis sentimen merupakan salah satu bentuk pengaplikasian NLP. Analisis sentimen adalah proses yang dilakukan untuk mendapat sentimen dalam sebuah teks [17]. Sentimen tersebut dibagi menjadi 3 yaitu positif, netral, dan negatif.

Analisis sentimen merupakan proses klasifikasi yang dibagi menjadi 3 level klasifikasi yang meliputi [18]:

1. *Document level*: merupakan analisis sentimen yang mengklasifikasi opini positif atau negatif pada suatu dokumen.
2. *Sentence level*: merupakan analisis sentimen yang mengklasifikasi opini positif dan negatif pada setiap kalimat.
3. *Aspect level*: merupakan analisis sentimen untuk mengklasifikasi suatu kalimat yang dapat memiliki opini yang berbeda untuk aspek yang berbeda dari entitas yang sama.

2.5 Data Preprocess

Data *preprocess* merupakan tahapan untuk mempersiapkan data mentah. Data *preprocess* bertujuan untuk memproses data mentah yang tidak berstruktur menjadi data yang lebih terstruktur [19]. Terdapat beberapa proses dalam data *preprocess* yaitu:

1. *Text Cleaning*: suatu proses pembersihan teks untuk mengurangi *noise*.
2. *Case Folding*: suatu proses untuk menyamakan *case* dalam suatu teks. Dalam hal ini mengubah seluruh karakter menjadi huruf kecil.
3. *Normalize*: suatu proses untuk mengubah kata yang tidak baku menjadi kata baku.
4. *Stopword Removal*: suatu proses untuk menghilangkan kata yang tidak penting atau tidak memiliki makna dalam suatu kalimat.
5. *Tokenizing*: suatu proses untuk memenggal teks menjadi bagian-bagian yang disebut dengan *token*.

6. *Stemming*: suatu proses untuk menguraikan suatu kata berimbuhan ke bentuk dasarnya.
7. *Data Labelling*: suatu proses untuk memberi label sentimen pada data. Sentimen dibagi menjadi sentimen positif, netral, dan negatif.

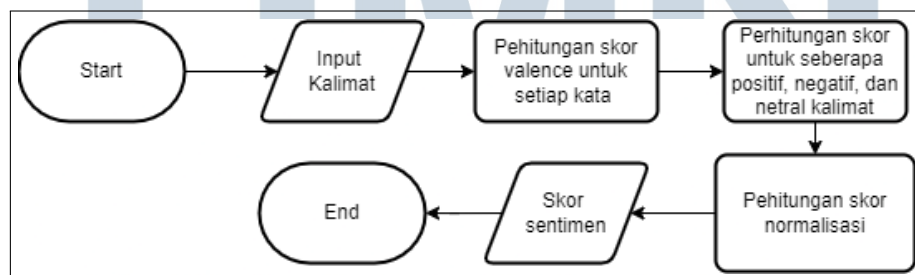
2.6 Valence Aware Dictionary and Sentiment Reasoner

Valence Aware Dictionary and sentiment Reasoner atau VADER merupakan *tools* untuk melakukan analisis sentimen berbasis leksikon dan *rule-based*. Setiap kata dalam leksikon dinilai seberapa positif atau negatif [20]. VADER bekerja dengan menganalisis suatu kalimat untuk melihat kata yang terdapat pada leksikon. VADER menghasilkan 4 metrik penilaian dalam suatu kalimat yaitu positif, netral, negatif, dan *compound* atau nilai normalisasi [21]. VADER digunakan untuk melakukan proses *labelling* pada *dataset* secara otomatis.

Untuk mengetahui apakah kalimat memiliki sentimen positif, netral, atau negatif, berikut merupakan standardisasi berdasarkan nilai *compound*:

1. positif: *compound* lebih besar sama dengan 0,05
2. netral: *compound* lebih kecil 0,05 dan lebih besar -0,05
3. negatif: *compound* lebih kecil sama dengan -0,05

Cara kerja VADER dalam melakukan analisis sentimen ditunjukkan pada Gambar 2.1 [22].



Gambar 2.1. Flowchart VADER

Gambar 2.1 menunjukkan bagaimana VADER melakukan analisis sentimen. VADER melakukan analisis berdasarkan skor *valence* atau nilai emosional dari setiap kata. Skor tersebut digunakan untuk menghitung seberapa positif, negatif, dan netral suatu kalimat. Setelah itu, dilakukan normalisasi nilai berdasarkan skor positif, negatif, dan netral.

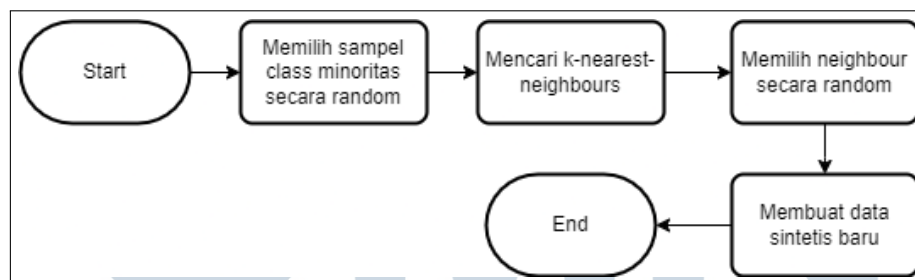
2.7 Bag of Words

Bag of words atau BOW merupakan vektor *input* berukuran tetap untuk mengubah teks dengan panjang yang berbeda. BOW melakukan normalisasi pada vektor untuk menunjukkan frekuensi dari setiap kata atau *token* [23]. BOW menghasilkan akurasi yang lebih baik pada algoritma Naive Bayes dibandingkan *feature extractor* lainnya seperti *Term Frequency-Inverse Document Frequency* atau TF-IDF [24].

2.8 Synthetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique atau SMOTE merupakan teknik untuk menyeimbangkan *class* pada *dataset*. SMOTE melakukan duplikasi data pada *class* minoritas agar memiliki jumlah yang sama dengan *class* mayoritas [25]. *Dataset* yang tidak seimbang dapat menyebabkan model salah dalam melakukan klasifikasi di mana model cenderung untuk mengklasifikasi data sebagai *class* mayoritas [26]. Untuk itu, diperlukan penyeimbangan *dataset* dengan SMOTE dengan membuat data sintetis untuk *class* minoritas.

Cara kerja SMOTE dalam melakukan *oversampling* ditunjukkan pada Gambar 2.2 [27].



Gambar 2.2. Flowchart SMOTE

Gambar 2.2 menunjukkan bagaimana SMOTE melakukan *oversampling* pada *class* minoritas. SMOTE memilih sampel dengan *class* minoritas secara *random*. Setelah itu, SMOTE mencari *neighbours* terdekat menggunakan algoritma KNN. Setelah didapatkan *neighbours* terdekat, SMOTE memilih *neighbour* secara *random*. Setelah itu, SMOTE membentuk data sintetis berdasarkan sampel dan *neighbour*.

2.9 Naive Bayes

Naive Bayes merupakan metode klasifikasi yang menggunakan model probabilitas dan statistik sederhana. Model Naive Bayes berdasar pada teorema Bayes dan mengasumsikan bahwa semua fitur data independen satu sama lain [28].

Berikut persamaan dari teorema Bayes.

$$P(C | X) = \frac{P(X | C) \cdot P(C)}{P(X)} \quad (2.1)$$

Di mana:

1. X: data dengan *class undefined*
2. C: hipotesis data X adalah *class* spesifik
3. $P(C|X)$: *posterior probability* atau probabilitas hipotesis C
4. $P(X|C)$: *likelihood* atau probabilitas X berdasarkan kondisi hipotesis C
5. $P(X)$: *predictor prior probability* atau probabilitas X

2.10 Confusion Matrix

Confusion matrix adalah tabel yang terdiri atas 4 kombinasi dari nilai aktual dan nilai prediksi [5]. *Confusion matrix* digunakan untuk mengukur performa suatu model dalam melakukan klasifikasi.

Tabel 2.1. Tabel confusion matrix

N	Aktual Positif	Aktual Negatif
Prediksi Positif	TP	FP
Prediksi Negatif	FN	TN

Tabel 2.1 merupakan bentuk dari *confusion matrix* di mana:

1. N: jumlah data
2. TP / *True Positive*: prediksi positif dan benar positif
3. FP / *False Positive*: prediksi positif namun aktual negatif
4. FN / *False Negative*: prediksi negatif namun aktual positif

5. TN / *True Negative*: prediksi negatif dan benar negatif.

Berdasarkan nilai *confusion matrix*, dapat dilakukan pengukuran performa dengan menghitung *accuracy*, *precision*, *recall*, dan *F1 score* [29].

2.10.1 Accuracy

Akurasi menggambarkan akurasi model dalam melakukan klasifikasi dengan benar. Akurasi dirumuskan dengan Rumus 2.2.

$$\frac{\text{Jumlah prediksi benar}}{\text{total prediksi}} \quad (2.2)$$

2.10.2 Precision

Precision menggambarkan akurasi model antara data yang diminta dengan hasil prediksi model. *Precision* dirumuskan dengan Rumus 2.3.

$$\frac{TP}{TP + FP} \quad (2.3)$$

2.10.3 Recall

Recall menggambarkan kemampuan model dalam menemukan kembali informasi. *Recall* dirumuskan dengan Rumus 2.4.

$$\frac{TP}{TP + FN} \quad (2.4)$$

2.10.4 F1 Score

F1 score menggambarkan rata-rata *precision* dan *recall* terbobot. *F1 score* dirumuskan dengan Rumus 2.5.

$$\frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.5)$$

2.11 Pearson Correlation Coefficient

Koefisien korelasi Pearson merupakan metode statistik untuk mengukur kuat lemahnya hubungan antara dua variabel [30]. Koefisien korelasi Pearson mengukur

arah hubungan dan kekuatan linear antara dua variabel. Nilai korelasi Pearson didapatkan melalui Rumus 2.6.

$$r_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (2.6)$$

Di mana:

1. x: variabel pertama
2. y: variabel kedua
3. n: banyaknya pengamatan

Nilai korelasi Pearson berada di antara -1 dan 1 di mana jika bernilai positif maka menunjukkan hubungan searah atau jika satu variabel naik, maka variabel lainnya ikut naik dan jika negatif maka menunjukkan hubungan yang berkebalikan atau jika satu variabel naik, maka variabel lainnya turun. Tabel 2.2 menunjukkan pedoman derajat hubungan [6].

Tabel 2.2. Pedoman derajat korelasi Pearson

Interval Koefisien	Tingkat Hubungan
0,00-0,199	Sangat Rendah
0,20-0,399	Rendah
0,40-0,599	Sedang
0,60-0,799	Kuat
0,80-1	Sangat Kuat

Tabel 2.2 merupakan pedoman derajat korelasi Pearson.

2.12 Uji t

Uji t merupakan pengujian statistika untuk membandingkan rata-rata sampel dua kelompok. Jika berbeda secara signifikan, maka dapat disimpulkan bahwa rata-rata populasi berbeda dan menolak hipotesis nol atau mendukung hipotesis penelitian [31]. T hitung didapatkan dengan Rumus 2.7.

$$t_{hitung} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (2.7)$$

Di mana:

1. r: koefisien korelasi
2. n: jumlah sampel

Terdapat 3 jenis uji t yaitu:

1. Uji t sampel independen: digunakan pada dua kelompok sampel yang tidak berhubungan. Uji t independen dapat berupa *one tailed* atau memiliki arah tertentu dalam hipotesis dan *two tailed* atau tidak memiliki arah tertentu dalam hipotesis atau hanya melihat apa ada perbedaan secara signifikan.
2. Uji t sampel berpasangan: digunakan pada dua kelompok yang sama dengan perlakuan yang berbeda.
3. Uji t satu sampel: digunakan ketika ingin melihat apakah rata-rata sampel secara signifikan berbeda dari nilai tertentu.

