BAB 2
LANDASAN TEORI

2.1 Piala Dunia U-20

Piala dunia U-20 adalah kompetisi sepak bola internasional yang diselenggarakan oleh FIFA setiap dua tahun sekali. Negara-negara yang mengikuti turnamen tersebut dipilih melalui serangkaian proses seleksi dan Indonesia berkesempatan menjadi tuan rumah di piala dunia U-20 2023[14]. Bagi masyarakat Indonesia, ini tentu menjadi kebanggaan tersendiri karena Indonesia terakhir kali berpartisipasi pada ajang piala dunia U-20 pada tahun 1979 di Jepang[13]. Namun, karena adanya penolakan kepada timnas U-20 Israel untuk bermain di Indonesia dari pihak-pihak tertentu membuat Indonesia batal menjadi tuan rumah piala dunia U-20 2023.

2.2 Analisis Sentimen

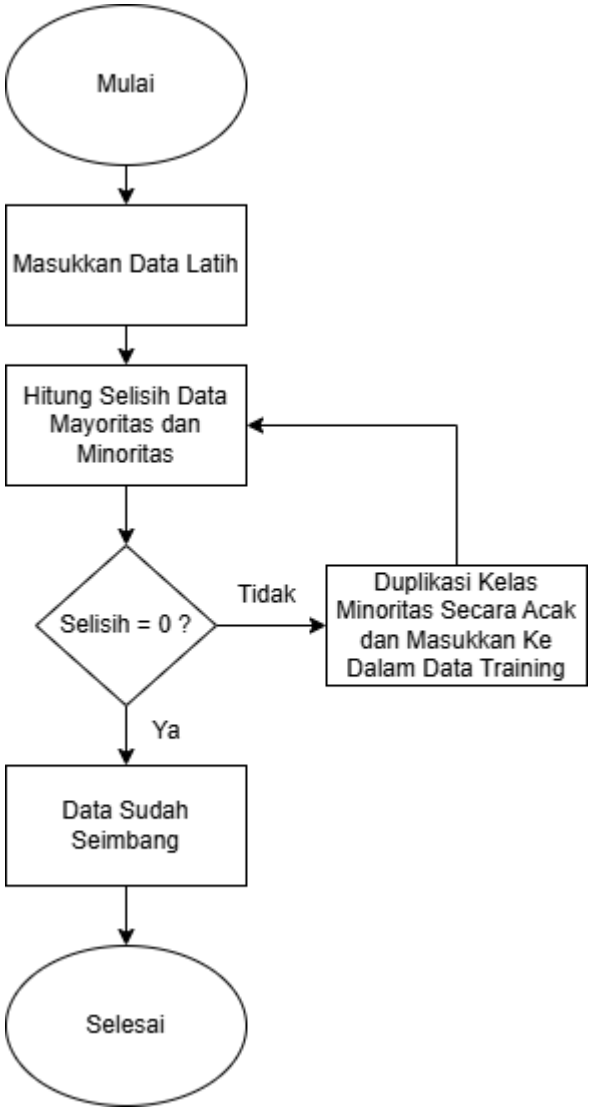
Analisis sentimen (sentiment analysis) adalah tahap untuk memperoleh informasi tentang suatu objek yang secara otomatis mengenali subjektivitas dari objek tersebut. Yang bertujuan untuk mengetahui apakah data teks yang dibuat oleh pengguna dapat digolongkan kepada opini positif, opini negatif atau opini yang cenderung netral [13]. Analisis sentimen ialah studi komputasi untuk menganalisis perasaan dan opini orang terhadap suatu entitas. Bidang analisis sentimen telah menjadi topik penelitian yang banyak dilakukan dalam beberapa dekade terakhir [13]. Dari paparan sebelumnya, dapat disimpulkan bahwa analisis sentimen adalah sebuah metode untuk memperoleh informasi mengenai pendapat seseorang terhadap suatu peristiwa. Analisis sentimen berguna untuk menentukan pandangan umum terhadap isu tertentu, peristiwa, kepuasan layanan, pergerakan harga saham, serta analisis persaingan berdasarkan data teks.

2.3 Analisis Sentimen Berbasis Aspek

Analisis sentimen berbasis aspek adalah teknik analisis yang memfokuskan pada aspek-aspek atau elemen-elemen spesifik dalam teks, seperti produk, layanan, atau kejadian tertentu, untuk mengetahui pendapat atau sentimen yang terkait dengan setiap aspek tersebut. Teknik ini melibatkan identifikasi, klasifikasi, dan evaluasi sentimen yang terkait dengan aspek-aspek tersebut dalam teks, baik itu dalam bentuk positif, negatif, atau netral. Hal ini membantu dalam memahami persepsi dan reaksi pengguna atau pelanggan terhadap berbagai aspek yang ditinjau[18].

2.4 Random Over Sampling

Algoritma ROS adalah salah satu metode yang digunakan untuk menangani ketidakseimbangan kelas dalam dataset. Pada proses ROS, data kelas minoritas dipilih secara acak kemudian ditambahkan ke dalam data latih. Proses ini dilakukan secara berulang sampai jumlah data kelas minoritas sama dengan jumlah kelas mayoritas[10]. Langkah yang dilakukan di awal adalah dengan menghitung selisih antara jumlah data kelas mayoritas dan kelas minoritas dalam dataset. Setelah itu, dilakukan perulangan sebanyak hasil perhitungan selisih data sambil membaca data dari kelas minoritas secara acak, kemudian dimasukkan ke dalam data training[9]. Untuk lebih jelas dapat dilihat pada gambar 2.1.



Gambar 2. 1 Flowchart Random Over Sampling

2.5 TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah sebuah metode yang digunakan dalam pemrosesan teks untuk mengukur tingkat relevansi kata tertentu dalam sebuah dokumen atau koleksi dokumen (korpus). Perhitungan TF (Term Frequency) mengindikasikan jumlah kemunculan suatu kata dalam suatu dokumen atau teks tertentu, sementara DF (Document Frequency) merupakan jumlah dokumen atau teks yang mengandung kata atau term tersebut. Dengan kombinasi TF dan DF, TF-IDF dapat memberikan

bobot yang lebih tinggi pada kata-kata yang muncul secara unik dalam dokumen tertentu tetapi tidak muncul secara umum dalam korpus. Hal ini membantu mengidentifikasi kata-kata kunci yang memiliki nilai informasi yang lebih tinggi dalam analisis teks[16].

Persamaan TF[17] dapat dirumuskan dengan persamaan 2.1 sebagai berikut:

$$TF_{ij} = \frac{f_i(d_j)}{\sum_{i=1}^{k} f_i(d_j)} \qquad (2.1)$$

Dimana :

TF_{ij} : Term Frequency *i* pada dokumen *j*

$f_i(d_j)$: Frekuensi kemunculan term *i* pada dokumen *j*

$\sum_{i=1}^{k} f_i(d_j)$: Total term pada dokumen *j*

Persamaan IDF dapat dirumuskan dengan persamaan 2.2 sebagai berikut:

$$IDF = \log\left(\frac{1+D}{1+DF_j}\right) + 1 \qquad (2.2)$$

Dimana:

IDF : Inverse Document Frequency

D : Jumlah Dokumen

DF_j : Jumlah dokumen yang berisi term *j*.

Persamaan TF-IDF dapat dirumuskan dengan persamaan 2.3 berikut.

$$W(i.j) = tf_{(ij)} * IDF \qquad (2.3)$$

Dimana:

$W(i.j)$: bobot kata *j* terhadap dokumen *i*

$tf_{(ij)}$: banyaknya kemunculan kata *j* pada dokumen *i*

IDF : Inverse Document Frequency

2.6 K-Nearest Neighbors

Metode *k-Nearest Neighbor* (k-NN) merupakan salah satu metode klasifikasi klasik yang paling sederhana[8]. Metode ini bekerja dengan cara mencari k-tetangga terdekat dari data yang akan diprediksi berdasarkan jaraknya dalam ruang fitur. Jika suatu data memiliki tetangga terdekat yang mayoritasnya berasal dari suatu kelas, maka data tersebut akan diklasifikasikan ke dalam kelas tersebut. Metode k-NN tidak memerlukan proses pelatihan yang kompleks, namun perlu memperhatikan parameter k yang menentukan jumlah tetangga terdekat yang digunakan untuk melakukan prediksi[19]. Berikut langkah-langkah yang digunakan untuk melakukan klasifikasi algoritma K-NN[20]:

1. Menentukan parameter k, dengan minimal nilai K adalah 1 dan jumlah maksimalnya dari data latih yang ada.
2. Melakukan perhitungan jarak antara data latih dan data uji. Perhitungan yang paling sering digunakan untuk menghitung jarak pada perhitungan algoritma K-NN adalah Euclidean. Untuk menghitung jarak dengan rumus persamaan 2.4.

$$\boldsymbol{Euclidean} = \sqrt{\sum_{\boldsymbol{i}}^{\boldsymbol{n}} =} \ \mathbf{1}(\boldsymbol{pi} - \boldsymbol{qi})^{\mathbf{2}} \qquad (2.4)$$

Keterangan:

pi : data pelatihan

qi : pengujian data

i : variabel data

n : ukuran data

3. Selanjutnya mengurutkan jarak yang telah terbentuk dimulai dari yang terbesar ke yang terkecil
4. Tentukanlah jarak yang paling dekat dengan barisan K
5. Cocokkan dengan kelas yang seimbang
6. Menemukan besaran kelas dari tetangga terdekat dan menugaskan pada kelas yang tersebut sebagai data kelas yang akan dievaluasi dan diklasifikasi.

2.7 Preprocessing

Preprocessing adalah tahap persiapan data sebelum dilakukan analisis lebih lanjut yang bertujuan untuk memastikan data siap digunakan dengan efektif dan lebih akurat dalam proses analisis[22]. Adapun tahapan-tahapan yang akan dilakukan dalam tahap preprocessing sebagai berikut :

2.7.1 Cleaning

Tahap cleaning merupakan bagian penting dari proses preprocessing di mana data diperiksa, dibersihkan, dan disesuaikan agar sesuai dengan kebutuhan analisis atau pemodelan yang akan dilakukan [14].

2.7.2 Case Folding

Case folding adalah proses dalam pemrosesan teks di mana semua huruf dalam teks diubah menjadi huruf kecil atau huruf besar. Tujuannya

adalah untuk menciptakan konsistensi dalam teks, sehingga memungkinkan pengolahan lebih lanjut seperti pencarian dan analisis teks dilakukan dengan lebih efisien. Misalnya “Kita” akan diubah menjadi “kita” [5].

2.7.3 Tokenizing

Tahapan ini memisahkan deretan kata atau tweet menjadi token. Selain itu juga menghilangkan karakter-karakter tertentu (seperti tanda baca, karakter, angka, tag HTML, link URL, username, dan lain sebagainya) dan mengubah semua token ke bentuk huruf kecil (lower case) [27].

2.7.4 Normalization

Pada tahap ini, dilakukan proses normalisasi untuk mengubah semua kata non-baku menjadi kata baku sesuai dengan Kamus Besar Bahasa Indonesia (KBBI)[5].

2.7.5 Stemming

Stemming adalah proses mengembalikan kata-kata yang berimbuhan ke bentuk kata dasarnya. Tujuan dari stemming yaitu mengubah kata-kata menjadi kata dasar, termasuk kata benda, kata sifat, kata kerja, dan lain-lain. Secara garis besar algoritma stemming dapat diklasifikasikan menjadi tiga kategori yaitu truncating methods, statistical methods, dan mixed methods[3]

2.7.6 Stopword Removal

Stopword adalah kata-kata yang sering diabaikan atau dihapus saat melakukan pemrosesan teks atau analisis teks karena dianggap tidak memiliki makna atau kontribusi yang signifikan terhadap pemahaman konten teks. Biasanya, stopword terdiri dari kata-kata umum seperti

"yang", "dan", "di", "dari", "ke", "sebuah", dan sejenisnya dalam suatu bahasa tertentu[11].

2.7.7 Convert Negasi

Konversi negasi adalah tahap untuk memproses kata-kata negatif. Kata negasi dapat mengubah makna sentimen suatu dokumen, sehingga kata negasi akan digabungkan dengan kata yang mengikutinya. Contoh kata negasi termasuk "tidak," "bukan," "jangan," dan lain-lain[21].

2.8 Confussion Matrix

Confusion matrix merupakan metode yang umum digunakan untuk menghitung akurasi. Dalam evaluasi keakuratan hasil pencarian, nilai recall, precision, accuracy, dan error rate akan dihitung[24]. Confusion matrix digunakan untuk menilai akurasi dari suatu proses klasifikasi yang telah dilakukan. Akurasi ini menunjukkan persentase prediksi yang tepat.[24].

$$Total = TPosPos + PFNon + PFNeg + NonFP + TNonNon + NonFNeg + NegFP + NegFnet + TNegNeg$$

$$Accuracy = TPP + TNetTNet + TNegTNeg \qquad (2.5)$$

$$Presisi = \frac{TPP}{TPP + NetFP + NegFP} \quad (2.6)$$

$$Recall = \frac{TPP}{TPP + PFNet + PFNeg} \quad (2.7)$$

$$F1Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (2.8)$$

2.9 Studi Literatur

Pada penelitian[1] membahas penggunaan algoritma resampling (ROS dan RUS) dan algoritma klasifikasi KNN untuk mengatasi ketidakseimbangan kelas dalam data. Hasil pengujian menunjukkan bahwa ROS lebih efektif daripada RUS dalam menyeimbangkan data, dengan peningkatan akurasi mencapai 15.79% untuk G-Mean dan 2.08% untuk F-Measure. Pada penelitian[4] membahas penggunaan teknik random over sampling (ROS) dan particle swarm optimization (PSO) untuk mengatasi ketidakseimbangan data dalam klasifikasi risiko kesehatan ibu hamil. Dengan hasil pengujian yang menunjukkan peningkatan akurasi dan kinerja model klasifikasi, teknik ini dapat membantu tenaga medis dalam mengidentifikasi faktor risiko kehamilan secara lebih akurat. Pada penelitian[9] membahas penggunaan model LightGBM dan teknik Random Oversampling untuk mengatasi ketidakseimbangan data dalam klasifikasi website phishing. Hasil eksperimen menunjukkan bahwa model LightGBM memberikan performa terbaik dengan akurasi, recall, F1-score, dan ROC yang baik, mencapai 96,9% akurasi dan 99,7% ROC. Penggunaan random oversampling juga terbukti efektif dalam meningkatkan deteksi website phishing.

Pada penelitian[12] menggunakan metode CFS untuk seleksi atribut dan Resample (Random Over Sampling) untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan algoritma J48. Dengan penambahan Adaboost, hasil penelitian menunjukkan peningkatan signifikan dalam akurasi hingga mencapai 92,3%, yang lebih optimal dibandingkan dengan penelitian sebelumnya. Pada penelitian[13] membahas analisis sentimen terhadap pembatalan Indonesia sebagai tuan rumah Piala Dunia FIFA U-20 menggunakan metode Naïve Bayes dan pendekatan Lexicon Based. Data dari Twitter diolah melalui tahap preprocessing sebelum dianalisis sentimennya. Hasil menunjukkan mayoritas tweet mengekspresikan kekecewaan terhadap keputusan FIFA, dengan akurasi analisis mencapai 84% dan nilai precision, recall, dan f-measure yang tinggi.

Tabel 2. 1 Penelitian Terdahulu

Reverensi	Masalah	Metode	Hasil
1	Masalah yang dibahas dalam jurnal tersebut adalah ketidakseimbangan data dalam klasifikasi, yang dapat mempengaruhi kinerja algoritma klasifikasi terutama pada dataset dengan Imbalance Ratio yang berbeda . Hal ini merupakan fokus masalah dalam bidang machine learning dan data mining, yang dapat mengakibatkan performa yang buruk pada algoritma klasifikasi jika tidak ditangani dengan baik .	ROS dan KNN	Dari hasil semua pengujian terlihat Ros+KNN dapat meningkatkan Accuracy Sebesar 0.09 atau 15.79% untuk performa dari G-Mean dan 0,01 untuk F-Measure 2.08%. Untuk mengetahui apakah nilai G-Mean dan FMeasure pada k-NN berbeda secara signifikan dengan performa k-NN+SMOTE, maka pengujian dilakukan dengan metode Wilcoxon Sing Rank Test dengan taraf Hasil pengujian
3	Masalah yang diangkat dalam jurnal tersebut adalah klasifikasi emosi dari komentar konsumen terhadap produk Natasha Skin Care yang diperoleh dari Twitter. Penelitian ini bertujuan	ROS dan Naïve Bayes	Algoritma resampling ROS dapat meningkatkan nilai recall, precision, dan F1-measure. Setelah menggunakan algoritma resampling ROS, rata-rata nilai precision, recall, dan F1-Measure tertinggi

	untuk memahami emosi pelanggan agar dapat meningkatkan kualitas pelayanan dan membantu calon konsumen dalam pengambilan keputusan. Selain itu, terdapat tantangan dalam klasifikasi emosi, seperti ketidakseimbangan kelas dan beberapa kelas emosi yang tidak dapat diklasifikasikan dengan baik.		terdapat pada kelas no emotions. Yaitu precision bernilai 96,64%, recall bernilai 76,36%, dan F1-Measure bernilai 85,93%.
25	ketidakseimbangan kelas dalam klasifikasi data, khususnya dalam konteks klasifikasi status kesejahteraan rumah tangga. Ketidakseimbangan ini terjadi ketika jumlah data dalam kelas minoritas jauh lebih sedikit dibandingkan dengan kelas mayoritas. Hal ini dapat menyebabkan bias dalam hasil klasifikasi, sehingga akurasi dan	ROS dan KNN	Hasil dari penelitian menunjukkan bahwa penggunaan teknik oversampling, khususnya Random Oversampling (RO), memberikan hasil klasifikasi yang lebih baik dibandingkan dengan data tanpa oversampling. Klasifikasi menggunakan metode K-Nearest Neighbor (KNN) dengan menghasilkan nilai sensitivitas 0,643, spesifisitas 0,805, G-mean 0,719, dan akurasi 78,873%

	kinerja algoritma klasifikasi, seperti K-Nearest Neighbor (KNN), menjadi tidak optimal.		pada data yang telah dioversampling.