BAB 2
LANDASAN TEORI

2.1 #KaburAjaDulu

Tagar #KaburAjaDulu muncul sebagai bentuk ekspresi kritik dan ketidakpuasan sebagian masyarakat, khususnya generasi muda, terhadap kondisi sosial, politik, ekonomi, serta kebijakan yang dirasa tidak memihak pada masyarakat luas di Indonesia [7]. Fenomena ini mulai kembali viral seiring dengan meningkatnya isu-isu seperti tingginya tingkat pengangguran, kesempatan kerja, ketidakstabilan politik, korupsi, serta persoalan hak asasi manusia [8]. Menurut survei yang dipublikasikan oleh CNN, sebanyak 41% Generasi Z Indonesia (kelahiran 1997–2009) mempertimbangkan untuk pindah ke luar negeri, baik untuk studi maupun berkarir, dalam rangka mencari peluang hidup yang lebih baik [8].

Meskipun keinginan generasi muda untuk mencari kehidupan yang lebih baik di luar negeri semakin tinggi, hal ini juga memunculkan kekhawatiran baru, terutama terkait maraknya praktik migrasi non-prosedural. Kementerian Luar Negeri Republik Indonesia, melalui Direktur Perlindungan WNI Judha Nugraha, memperingatkan bahwa tren seperti tagar #KaburAjaDulu dapat menyesatkan bila tidak diiringi dengan pemahaman terhadap prosedur legal yang benar. Data Kemlu menunjukkan bahwa sebanyak 67.000 WNI terlibat pelanggaran keimigrasian, banyak di antaranya menjadi korban penipuan daring, eksploitasi, hingga perdagangan orang akibat berangkat melalui jalur ilegal [9]. Oleh karena itu, pemerintah menekankan pentingnya edukasi serta kesadaran untuk selalu menempuh jalur resmi seperti pengurusan visa kerja dan kontrak yang sah sebelum bekerja di luar negeri.

2.2 Media Sosial X

Media sosial X (sebelumnya dikenal sebagai Twitter) telah berkembang menjadi *platform* utama untuk komunikasi cepat dan diseminasi informasi, baik di tingkat individu maupun institusi. Penelitian menunjukkan bahwa X dapat berperan sebagai saluran efektif dalam mitigasi bencana, di mana lembaga seperti BNPB memanfaatkan *platform* ini untuk menyebarkan peringatan dini dan panduan keselamatan secara *real-time*, meskipun risiko terhadap penyebaran informasi keliru tetap signifikan [10]. Di bidang akademik, penelitian oleh

Zedda *et al.* menunjukkan bahwa X masih berperan sebagai salah satu *platform* utama dalam strategi promosi dan kurasi konten jurnal ilmiah, meskipun terdapat peningkatan kekhawatiran terkait penyebaran misinformasi serta penggunaan bot dalam otomatisasi konten [11].

Selain perannya dalam komunikasi dan akademik, X juga menjadi wadah penting dalam dinamika politik dan sosial. Pada tahun 2025, Syahfiraputri *et al.* menemukan bukti bahwa penggunaan X sebagai media pemberitaan memiliki dampak signifikan terhadap tingkat partisipasi politik online serta memengaruhi aksi politik *offline*, khususnya di kalangan mahasiswa Jawa Barat [12]. Namun demikian, sejak *rebranding* dan berlangsungnya peninjauan ulang terhadap kebijakan moderasi konten serta akses API, sejumlah studi melaporkan peningkatan signifikan dalam ujaran kebencian sekitar 50% sementara aktivitas bot tetap tinggi, didampingi kekhawatiran mengenai kekaburan algoritma dan hambatan akses data bagi peneliti [13, 14, 15].

2.3 Natural Language Processing

Natural Language Processing (NLP) merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*, AI) yang berfokus pada pengembangan sistem yang mampu memproses, memahami, dan menghasilkan bahasa manusia dalam bentuk teks maupun ucapan. NLP mengintegrasikan berbagai disiplin ilmu, seperti linguistik komputasional, ilmu komputer, dan statistika, guna mengatasi kompleksitas dan ambiguitas yang melekat pada bahasa alami. Pendekatan dalam NLP telah berkembang dari berbasis aturan (*rule-based systems*) menuju metode berbasis statistik, dan saat ini didominasi oleh teknik *deep learning* yang memanfaatkan arsitektur *Transformer* serta model pra-pelatihan skala besar seperti BERT dan GPT [16, 17].

Kemajuan teknologi dalam NLP memungkinkan berbagai aplikasi cerdas, mulai dari penerjemahan otomatis, analisis sentimen, hingga pengenalan entitas dan dialog interaktif. Namun, tantangan utama tetap ada, antara lain dalam hal bias data, keterbatasan untuk bahasa dengan sumber daya rendah (*low-resource languages*), serta kebutuhan komputasi yang tinggi. Selain itu, upaya integrasi teori *fuzzy* dengan NLP menjadi salah satu arah penelitian yang menjanjikan, karena dapat membantu dalam menangani ketidakpastian dan ambiguitas bahasa alami secara lebih efektif [18]. Dengan perkembangan yang pesat, NLP terus menjadi bidang riset yang sangat dinamis dan relevan dalam mendukung interaksi manusia

dan mesin di berbagai domain.

2.4 Analisis Sentimen

Analisis sentimen merupakan cabang dari pemrosesan bahasa alami (*Natural Language Processing*) yang bertujuan untuk mengidentifikasi, mengekstrak, dan mengklasifikasikan opini atau emosi dalam teks, seperti positif, negatif, atau netral. Teknik ini memungkinkan analisis data dalam skala besar secara otomatis, menjadikannya alat yang penting dalam memahami respons publik terhadap isu-isu tertentu melalui media sosial (Medhat et al., 2020) [19].

Studi oleh Yao dan Gillen (2023) menunjukkan bahwa analisis sentimen dapat digunakan untuk memetakan opini publik terhadap proyek infrastruktur seperti kereta cepat HS2 di Inggris. Melalui data dari media sosial Twitter, penelitian tersebut berhasil mengungkap dinamika persepsi masyarakat secara *real time*, memperlihatkan potensi metode ini sebagai alat evaluasi kebijakan publik yang berbasis data [20].

Lebih lanjut, Khan et al. (2024) menyatakan bahwa analisis sentimen secara umum efektif untuk mengukur opini publik terhadap fenomena sosial dan peristiwa aktual [6]. Dengan data yang berasal langsung dari masyarakat, analisis ini mampu menangkap emosi kolektif tanpa perlu melalui survei konvensional yang cenderung terbatas dalam cakupan dan waktu.

2.5 Term Frequency - Inverse Document Frequency (TF-IDF)

Term Frequency – Inverse Document Frequency (TF-IDF) merupakan salah satu metode representasi teks ke dalam bentuk numerik yang banyak digunakan dalam pemrosesan bahasa alami (*Natural Language Processing*), khususnya dalam tahap ekstraksi fitur [21]. Teknik ini bertujuan untuk menilai seberapa penting sebuah kata dalam suatu dokumen relatif terhadap kumpulan dokumen (*corpus*).

TF-IDF terdiri dari dua komponen utama, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) [22]. *Term Frequency* menunjukkan seberapa sering suatu kata muncul dalam satu dokumen. Semakin tinggi frekuensi kemunculan suatu kata, semakin besar nilai TF-nya. Sementara itu, *Inverse Document Frequency* berfungsi untuk mengurangi bobot kata-kata umum yang muncul di hampir semua dokumen, dengan memberi nilai yang lebih kecil terhadap kata-kata yang sering muncul dalam banyak dokumen.

Secara matematis, TF-IDF dari suatu kata t dalam dokumen d terhadap kumpulan dokumen D dihitung menggunakan rumus:

$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) \times \text{IDF}(t,D) \tag{2.1}$$

dengan:

- $\text{TF}(t,d) = \dfrac{f_{t,d}}{\sum_k f_{k,d}}$, yaitu rasio frekuensi kata t terhadap jumlah seluruh kata dalam dokumen d,
- $\text{IDF}(t,D) = \log\left(\dfrac{N}{n_t}\right)$, di mana:
  - N adalah jumlah total dokumen dalam korpus,
  - n_t adalah jumlah dokumen yang mengandung kata t.

Nilai TF-IDF yang tinggi menunjukkan bahwa suatu kata sering muncul dalam satu dokumen namun jarang muncul di dokumen lain, sehingga kata tersebut dianggap penting dan relevan untuk membedakan dokumen tersebut dari yang lain. Sebaliknya, kata yang sering muncul di hampir semua dokumen akan memiliki nilai IDF rendah, sehingga bobot TF-IDF-nya juga rendah.

TF-IDF banyak digunakan dalam berbagai aplikasi *text mining* dan *machine learning* karena mampu merepresentasikan teks menjadi fitur numerik yang dapat diolah lebih lanjut oleh algoritma klasifikasi.

2.6 Support Vector Machine

Support Vector Machine (SVM) merupakan salah satu metode klasifikasi yang termasuk dalam *supervised learning*, di mana model dilatih menggunakan data berlabel untuk mempelajari pemisahan antar kelas. SVM bekerja dengan mencari *hyperplane* optimal yang memaksimalkan *margin*, yaitu jarak antara *hyperplane* dan data dari masing-masing kelas [23].

Konsep utama SVM adalah untuk menemukan pemisahan yang paling kuat antara kelas positif dan negatif, bahkan ketika data tidak sepenuhnya dapat dipisahkan secara linear. Secara umum, SVM berupaya mencari *hyperplane* yang memberikan margin terbesar terhadap *support vectors*, yaitu data yang paling dekat dengan *hyperplane*.

Secara matematis, *hyperplane* dinyatakan sebagai:

$$\vec{w} \cdot \vec{x} + b = 0 \tag{2.2}$$

di mana:

- $\vec{w}$ adalah vektor bobot,
- $\vec{x}$ adalah vektor fitur dari data,
- b adalah bias.

Adapun persamaan tersebut dikembangkan berdasarkan perbedaan *output* kelas y_i. Pertidaksamaan berikut menunjukkan kondisi pemisahan linear berdasarkan label kelas. Untuk sampel dengan kelas positif ($y_i = +1$), maka:

$$\vec{w} \cdot \vec{x}_i + b \geq 1, \quad y_i = +1 \tag{2.3}$$

Sedangkan untuk sampel dengan kelas negatif ($y_i = -1$), berlaku:

$$\vec{w} \cdot \vec{x}_i + b \leq -1, \quad y_i = -1 \tag{2.4}$$

Kedua pertidaksamaan ini digunakan dengan asumsi bahwa kedua kelas data dapat dipisahkan secara linear. Namun, dalam banyak kasus, data di dunia nyata sering kali tidak dapat dipisahkan secara linear, sehingga diperlukan pendekatan tambahan seperti *soft margin* ataupun *kernel trick* untuk menangani kondisi tersebut [24].

Tujuan utama SVM adalah untuk memaksimalkan margin atau secara ekuivalen meminimalkan fungsi berikut:

$$\min_{\vec{w},b} \frac{1}{2} \|\vec{w}\|^2 \tag{2.5}$$

Pada kondisi data yang tidak sepenuhnya terpisah secara linear, digunakan pendekatan *soft margin* dengan menambahkan variabel *slack* ξ_i, sehingga permasalahan optimisasi menjadi:

$$\min_{\vec{w},b,\xi} \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n} \xi_i \tag{2.6}$$

dengan batasan:

$$y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{2.7}$$

di mana:

- C adalah parameter *trade-off* antara kesalahan klasifikasi dan ukuran margin,
- ξ_i adalah variabel *slack* untuk mengizinkan pelanggaran margin pada sampel ke-i.

Untuk data yang tidak dapat dipisahkan secara linear, SVM memanfaatkan teknik *kernel trick*, yaitu mentransformasikan data ke bentuk baru agar lebih mudah dipisahkan. Proses ini dilakukan dengan memetakan data ke dalam ruang fitur berdimensi lebih tinggi, sehingga pemisahan antar kelas menjadi lebih jelas [23, 24].

Fungsi *kernel* yang umum digunakan antara lain:

- Linear: $K(\vec{x}_i,\vec{x}_j) = \vec{x}_i\cdot\vec{x}_j$
- *Polynomial*: $K(\vec{x}_i,\vec{x}_j) = (\vec{x}_i\cdot\vec{x}_j + 1)^p$
- *Radial Basis Functio*n (RBF): $K(\vec{x}_i,\vec{x}_j) = \exp\left(-\frac{\|\vec{x}_i-\vec{x}_j\|^2}{2\sigma^2}\right)$
- *Sigmoid*: $K(\vec{x}_i,\vec{x}_j) = \tanh(\alpha\vec{x}_i\cdot\vec{x}_j + \beta)$

Dengan pendekatan ini, SVM terbukti sangat andal dalam menangani berbagai masalah klasifikasi, termasuk di bidang *natural language processing* dan *sentiment analysis*.