Tugas perbaikan

Bab 7

Nama : muhammad makhlufi makbullah

Kelas : TK 45 01

NIM : 1103210171

Bab 7: Working with Text Data

Bab ini membahas bagaimana data teks dapat diolah dan direpresentasikan untuk digunakan dalam pembelajaran mesin. Teks, sebagai data tidak terstruktur, harus diubah menjadi bentuk numerik untuk diproses oleh algoritma. Berikut adalah poin-poin utama:

1. **Jenis Data Teks**: Mencakup berbagai format seperti dokumen panjang, teks pendek (tweet, ulasan), dan log sistem. Setiap format memerlukan pendekatan khusus untuk analisis.
2. **Studi Kasus Analisis Sentimen**: Contoh kasus menggunakan ulasan film untuk menentukan apakah sentimen dalam teks positif atau negatif. Studi ini mengilustrasikan langkah-langkah utama dalam pemrosesan teks.
3. **Representasi Data Teks**:
    - **Bag-of-Words (BoW)**: Metode dasar yang mengubah teks menjadi vektor berdasarkan frekuensi kata, tanpa memperhatikan urutan.
    - **TF-IDF (Term Frequency-Inverse Document Frequency)**: Menghitung bobot kata dengan memperhitungkan pentingnya kata dalam dokumen tertentu dibandingkan korpus secara keseluruhan.
4. **Preprocessing Data**: Meliputi tokenisasi, penghapusan kata umum (stopwords), stemming, dan lemmatization untuk membersihkan teks dan mengurangi dimensi fitur.
5. **n-Grams**: Mempertimbangkan urutan kata dengan menganalisis kelompok kata berurutan (misalnya, bigram atau trigram) untuk menangkap konteks yang lebih baik.
6. **Penerapan pada Dataset Ulasan Film**: Contoh praktis penerapan BoW dan TF-IDF pada dataset besar untuk membangun model klasifikasi teks, seperti Logistic Regression atau Naive Bayes.
7. **Koefisien Model**: Meninjau bobot fitur dari model untuk mengidentifikasi kata-kata yang paling berpengaruh terhadap prediksi.
8. **Model Topik dan Pengelompokan Dokumen**:

- **Latent Dirichlet Allocation (LDA)**: Teknik unsupervised learning untuk menemukan topik tersembunyi dalam dokumen. Membantu menganalisis tema utama tanpa label eksplisit.

9. **Ringkasan dan Pandangan ke Depan**:
    - Mencakup penggunaan BoW dan TF-IDF sebagai fondasi dasar.
    - Menyarankan eksplorasi teknik lanjutan seperti word embeddings (Word2Vec, GloVe) atau model berbasis deep learning (seperti BERT atau GPT) untuk representasi teks yang lebih kompleks dan akurat.