Nama :Gian Rofiqri Andra

NIM :24917036

BAGIAN 6

1. Perbedaan Text Classification dan Text Clustering

Jawab

Perbedaan utama antara text classification (klasifikasi teks) dan text clustering (pengelompokan teks) terletak pada ketersediaan label data dan tujuannya:

- Text Classification (Supervised Learning):
    - Tujuan: Untuk memprediksi atau menetapkan label/kategori yang sudah ditentukan sebelumnya ke sebuah dokumen teks baru.
    - Proses: Membutuhkan data latih (training data) yang sudah memiliki label. Model "belajar" dari data berlabel ini untuk mengenali pola yang terkait dengan setiap kategori.
    - Contoh: Klasifikasi email sebagai "Spam" atau "Bukan Spam". Kategori "Spam" dan "Bukan Spam" sudah didefinisikan dari awal.
- Text Clustering (Unsupervised Learning):
    - Tujuan: Untuk menemukan struktur atau pengelompokan alami (disebut cluster) di dalam kumpulan data teks, tanpa ada label atau kategori yang ditentukan sebelumnya.
    - Proses: Algoritma mengelompokkan dokumen berdasarkan kesamaan (similaritas) konten di antara mereka. Dokumen dalam satu cluster akan mirip satu sama lain dan berbeda dengan dokumen di cluster lain.
    - Contoh: Mengambil 10.000 ulasan produk dan mengelompokkannya secara otomatis untuk menemukan topik-topik utama yang dibicarakan (misalnya, cluster ulasan tentang "harga", cluster ulasan tentang "pengiriman", cluster ulasan tentang "kualitas baterai").

Secara singkat, klasifikasi memberi label berdasarkan kategori yang ada (terawasi), sedangkan clustering menemukan kelompok berdasarkan kesamaan (tidak terawasi).

Referensi Ilmiah (Soal 1):

- Makalah: Defiyanti, S. (2017). Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining. ResearchGate. Publikasi No: 314266899.
- Posisi: Halaman 1, Paragraf 3 (di bawah "Public Full-text 1").
- Kutipan Pendukung: "Klasifikasi adalah pemprosesan untuk menemukan sebuah model atau fungsi yang menjelaskan dan mencirikan konsep atau kelas data, untuk kepentingan tertentu. ... Clustering digunakan untuk pengelompokkan data berdasarkan kemiripan pada objek data dan sebaliknya meminimalkan kemiripan terhadap kluster yang lain."

2. Kapan Text Clustering Digunakan?

Jawab

Text clustering dapat dilakukan (dan sangat bermanfaat) pada situasi di mana kita memiliki data teks dalam volume besar yang tidak terstruktur dan tidak memiliki label (unlabeled data). Teknik ini digunakan ketika kita belum mengetahui kategori-kategori yang ada di dalam data tersebut dan ingin menemukannya secara otomatis.

Situasi dan Kondisi Bermanfaat:

- Analisis Eksploratif: Saat pertama kali berhadapan dengan kumpulan data teks yang besar (misalnya, arsip berita, tumpukan log server, atau data media sosial) dan ingin mendapatkan gambaran umum tentang topik-topik utama yang terkandung di dalamnya.
- Tidak Ada Kategori Pasti: Ketika tidak ada kategori yang jelas atau telah ditentukan sebelumnya untuk mengorganisir dokumen.
- Topic Modeling: Untuk mengidentifikasi tema-tema atau topik-topik tersembunyi yang muncul berulang kali dalam koleksi dokumen.
- Segmentasi: Untuk mengelompokkan pengguna atau pelanggan berdasarkan teks yang mereka hasilkan (misalnya, mengelompokkan pelanggan berdasarkan isi ulasan atau keluhan mereka).

Contoh Kasus Penggunaan:

Sebuah perusahaan ingin menganalisis ribuan umpan balik (feedback) pelanggan yang masuk melalui formulir "hubungi kami" di website mereka. Umpan balik ini berbentuk teks bebas dan tidak memiliki kategori. Perusahaan tidak mungkin membaca satu per satu secara manual.

Dengan menggunakan text clustering, perusahaan dapat secara otomatis mengelompokkan ribuan umpan balik ini ke dalam beberapa cluster. Setelah proses clustering selesai, mereka mungkin menemukan cluster seperti:

- Cluster 1: Berisi keluhan tentang "waktu pengiriman yang lama".
- Cluster 2: Berisi pujian tentang "layanan pelanggan yang ramah".
- Cluster 3: Berisi pertanyaan teknis tentang "cara reset password".
- Cluster 4: Berisi saran untuk "fitur baru pada aplikasi".

Ini memungkinkan perusahaan untuk dengan cepat memahami isu-isu utama tanpa harus mendefinisikan kategori dari awal.

Referensi Ilmiah (Soal 2):

- Makalah: M. A. F. G. R. (2018). Penerapan Metode Clustering Text Mining Untuk Pengelompokan Berita Pada Unstructured Textual Data. Jurnal Sisfokom (Sistem Informasi dan Komputer), 7(1).
- Posisi: Halaman 2, Bagian "Pendahuluan" (Paragraf 4 di bagian itu).
- Kutipan Pendukung: "Clustering dipakai ketika tidak diketahuinya bagaimana data harus dikelompokkan. Clustering dapat digunakan untuk membantu menganalisis berita dengan mengelompokkan secara otomatis berita yang memiliki kesamaan atau kemiripan."

3. Menentukan Jumlah Klaster Optimal (K) dalam K-Means

Jawab

Menentukan jumlah klaster (k) yang optimal adalah langkah krusial dalam K-Means, karena algoritma ini mengharuskan kita untuk menentukan nilai k di awal. Dua metode yang paling umum digunakan adalah Elbow Method dan Silhouette Method.

a. Elbow Method (Metode Siku)

Metode ini bekerja dengan mengukur seberapa padat dan terpisah klaster yang dihasilkan, yang dihitung menggunakan Sum of Squared Errors (SSE). SSE adalah jumlah dari kuadrat jarak antara setiap titik data ke centroid (pusat klaster) terdekatnya.

- Cara Kerja:
    1. Jalankan algoritma K-Means untuk berbagai nilai k (misalnya, k=1 hingga k=10).
    2. Untuk setiap nilai k, hitung nilai SSE total.
    3. Buat grafik (plot) yang menunjukkan nilai k pada sumbu-X dan nilai SSE pada sumbu-Y.
    4. Grafik ini biasanya akan menurun. Kita mencari titik "siku" (elbow)—yaitu, titik di mana penurunan nilai SSE mulai melambat secara drastis. Titik k tepat sebelum pelambatan inilah yang dianggap sebagai k optimal.
- Logika: Menambahkan lebih banyak klaster akan selalu mengurangi SSE, tetapi setelah titik siku, penambahan klaster baru tidak memberikan pengurangan SSE yang signifikan (peningkatannya minimal).

b. Silhouette Method (Metode Siluet)

Metode ini mengukur seberapa baik setiap titik data telah dikelompokkan. Metode ini mengukur dua hal:

1. Kohesi (a): Seberapa mirip sebuah titik data dengan titik lain di dalam klaster yang sama (jarak rata-rata internal).
2. Separasi (b): Seberapa berbeda sebuah titik data dengan titik-titik di klaster terdekat lainnya (jarak rata-rata eksternal).

- Cara Kerja:
    1. Hitung Silhouette Score untuk setiap titik data. Skor ini berkisar dari -1 hingga +1.
        - +1: Titik data sangat padat di dalam klasternya dan jauh dari klaster lain (ideal).
        - 0: Titik data berada sangat dekat dengan batas antar klaster.
        - -1: Titik data mungkin salah diklasifikasikan ke klaster yang salah.
    2. Jalankan K-Means untuk berbagai nilai k (misalnya, k=2 hingga k=10).
    3. Untuk setiap k, hitung rata-rata Silhouette Score dari semua titik data.
    4. Buat grafik (plot) yang menunjukkan k pada sumbu-X dan rata-rata Silhouette Score pada sumbu-Y.
- Logika: Nilai k yang menghasilkan rata-rata Silhouette Score tertinggi dianggap sebagai jumlah klaster yang optimal, karena ini menunjukkan klaster yang paling padat (kohesif) dan paling terpisah (separasi baik).

Referensi Ilmiah (Soal 3):

- Makalah: Nisa, K., & S, A. (2023). Cacah Klaster pada Klasterisasi dengan Algoritma K-Means Menggunakan Silhouette Coeficient dan Elbow Method. JuTI (Jurnal Ilmiah Teknologi Informasi), 4(2).
- Posisi: Halaman 1, Abstrak. (Metode Elbow juga dijelaskan di Bagian 2.3.1 dan Silhouette di Bagian 2.3.3).
- Kutipan Pendukung: "Untuk memperoleh cacah klaster yang optimal digunakan Silhouette Coefficient methode dan Elbow methode."

LINK Artike: https://www.antaranews.com/berita/4375475/bpdpks-menilai-riset-jadi-kunci-hadapi-kampanye-hitam-sawit-ri