BAB II
TINJAUAN PUSTAKA

2.1. *Cluster Analysis on Journal*

Analisis klaster adalah teknik statistik canggih yang digunakan di berbagai bidang untuk mengidentifikasi *Cluster* atau klaster alami dalam suatu kumpulan data. Penerapannya mencakup berbagai disiplin ilmu seperti biologi, sosiologi, psikologi, persebaran, dan banyak lagi. Ketika membahas makalah atau jurnal yang membahas analisis klister[5].

Pentingnya analisis klaster sebagai alat serbaguna untuk pengenalan pola dan ekstraksi wawasan. Penerapannya yang luas dan metodologi yang berkembang terus meningkatkan signifikansinya di berbagai bidang. Dengan menekankan pentingnya memahami metodologi, menafsirkan hasil, dan mengatasi tantangan, makalah ini menetapkan landasan untuk kemajuan berkelanjutan dan penerapan analisis klaster dalam lanskap analisis data yang terus berkembang.

2.2. DASK

Dask adalah perpustakaan komputasi paralel fleksibel dengan Python yang dirancang untuk menskalakan tugas komputasi dari mesin tunggal ke cluster besar. Ini menyediakan alat untuk mengaktifkan paralelisme dan bekerja dengan kumpulan data besar yang tidak sesuai dengan memori[6]. Dask beroperasi dengan baik di lingkungan mulai dari laptop hingga cluster berskala besar.

Komponen Utama Dask:

1. *Dask Arrays*

Ini adalah versi array NumPy yang terdistribusi dan paralel. Mereka memecah array besar menjadi potongan-potongan kecil yang dapat dioperasikan secara paralel, memungkinkan komputasi yang tidak sesuai dengan memori.

2. *Dask DataFrames*

   Mirip dengan Pandas DataFrames, Dask DataFrames bekerja pada kumpulan data besar yang tidak muat dalam memori. Mereka memungkinkan operasi paralel dan di luar inti untuk manipulasi dan analisis data.

3. *Dask Bags*

   Tas Dask menyediakan cara untuk bekerja dengan data semi-terstruktur (seperti JSON atau file teks) secara paralel. Mereka dapat menangani koleksi objek Python yang berubah-ubah.

4. *Dask Dealyed*

   Fitur ini mengubah kode Python biasa menjadi alur kerja paralel dengan menunda eksekusi fungsi hingga secara eksplisit diminta untuk menghitung hasilnya. Hal ini sangat berguna untuk membuat algoritma paralel khusus atau menangani komputasi yang tidak teratur.

Dask memungkinkan komputasi paralel dengan memecah tugas menjadi komponen yang lebih kecil yang dapat dijalankan secara paralel di seluruh inti atau mesin. Ini mengoordinasikan tugas-tugas ini dan mengelola aliran data di antara tugas-tugas tersebut, mengoptimalkan grafik komputasi untuk eksekusi yang efisien[7].

Keuntungan Dask:

1. *Scallability*

   Dask berkembang dari mesin tunggal hingga cluster dengan ribuan node, menyediakan antarmuka yang konsisten untuk komputasi paralel.

2. *Out-of-Core Processing*

   Dask menangani kumpulan data yang lebih besar dari memori yang tersedia dengan membagi data menjadi beberapa bagian yang dapat dikelola, sehingga memungkinkan komputasi pada data yang berada di disk.

3. *Integration with Existing Libraries*

   Ini terintegrasi dengan baik dengan perpustakaan Python populer seperti NumPy, Pandas, dan Scikit-learn, memperluas kemampuannya untuk komputasi skala besar.

4. *Task Schedulling*

Penjadwal tugas Dask secara dinamis menjadwalkan tugas untuk dieksekusi, mengoptimalkan pemanfaatan sumber daya, dan meminimalkan waktu komputasi.

Penggunaan Dask:

1. *Big Data Processing*

   Menganalisis dan memproses kumpulan data besar yang tidak sesuai dengan memori.

2. *Machine Learning*

   Menskalakan algoritme pembelajaran mesin untuk menangani kumpulan data besar.

3. *Scientific Computing*

   Memparalelkan simulasi ilmiah, analisis data, dan tugas komputasi.

Dask menyederhanakan proses komputasi paralel dengan Python, memungkinkan ilmuwan dan insinyur data menangani komputasi skala besar secara efisien dan bekerja dengan kumpulan data yang melebihi memori yang tersedia. Fleksibilitas dan kemudahan integrasinya dengan pustaka Python yang ada menjadikannya alat yang berharga untuk berbagai aplikasi intensif data[7].

2.3. *Topic Modelling*

Topic Modelling adalah teknik yang digunakan dalam pemrosesan bahasa alami (NLP) dan pembelajaran mesin untuk menemukan topik abstrak dalam kumpulan dokumen. Ini sangat berguna ketika menangani data teks dalam jumlah besar dan ingin memahami tema atau topik mendasar yang ada. Salah satu

algoritma yang paling populer untuk *Topic Modelling* adalah Latent Dirichlet Allocation (LDA). LDA berasumsi bahwa setiap dokumen dalam koleksi dapat direpresentasikan sebagai campuran topik, dan setiap topik merupakan distribusi probabilitas atas kata-kata. Algoritme ini bekerja dengan mencoba secara berulang-ulang untuk menetapkan kata-kata pada topik dan dokumen pada distribusi topik dengan cara yang paling mewakili data yang diamati[2].

Output dari proses *Topic Modelling* biasanya meliputi:

1. *Topics,* adalah kumpulan kata-kata yang cenderung muncul bersamaan dalam dokumen. Kata-kata ini mewakili tema atau subjek yang ada dalam dataset.
2. *Topics Distribution,* dan untuk setiap dokumen, proporsi topik yang ada dalam dokumen tersebut.

Pemodelan topik dapat diterapkan di berbagai bidang, termasuk:

1. *Information Retrieval*: Mengorganisir dan mencari koleksi dokumen dalam jumlah besar.
2. *Content Recommendation*: Mengidentifikasi konten yang relevan bagi pengguna berdasarkan minat mereka.
3. *Content Analysis*: Memahami tren, tema, dan sentimen dalam data teks.
4. *Market Research*: Mengidentifikasi opini dan preferensi pelanggan dari ulasan, survei, dll
5. *Academic Research*: Menganalisis sejumlah besar makalah akademis untuk menemukan tema atau ide yang ada.

Namun, penting untuk diperhatikan bahwa meskipun pemodelan topik dapat mengungkap pola dalam data, penafsiran topik ini sering kali memerlukan pertimbangan dan konteks manusia. Para peneliti dan analis biasanya menyaring dan menafsirkan topik-topik ini untuk memperoleh wawasan yang bermakna dari hasil penelitian[2].

2.1. TF – IDF

TF-IDF, kependekan dari Term Frekuensi-Invers Dokumen Frekuensi, adalah statistik numerik yang digunakan dalam pemrosesan bahasa alami dan pengambilan informasi untuk mengukur signifikansi suatu istilah dalam dokumen atau kumpulan dokumen. Hal ini didasarkan pada gagasan bahwa pentingnya suatu istilah dalam sebuah dokumen meningkat secara proporsional dengan seberapa sering istilah tersebut muncul dalam dokumen, namun diimbangi oleh frekuensi istilah tersebut di seluruh korpus[8].

Alasan di balik penggunaan TF-IDF adalah untuk menyoroti istilah-istilah yang sering muncul dalam dokumen individual dan relatif jarang terjadi di seluruh korpus. Hal ini membantu dalam mengidentifikasi istilah-istilah yang khas pada dokumen tertentu dan oleh karena itu mungkin lebih menunjukkan isi atau topik dokumen.[9]

Korelasi antara TF-IDF dan K-means dalam analisis cluster pemodelan topik terletak pada peran mereka yang saling melengkapi dalam proses mengidentifikasi dan mengelompokkan dokumen serupa berdasarkan kontennya. TF-IDF adalah teknik yang digunakan untuk merepresentasikan dokumen secara numerik dengan memberi bobot pada pentingnya istilah di dalamnya, sedangkan

K-means adalah algoritma pengelompokan yang mempartisi sekumpulan dokumen ke dalam kelompok atau cluster berbeda berdasarkan representasi fiturnya[10].

2.4. *Principal Component Analysis*

Principal Component Analysis (PCA) dan analisis klaster, meskipun merupakan teknik yang berbeda, dapat saling melengkapi bila diterapkan secara bersamaan pada topik seperti pemodelan topik[3][11]. Bagaimana *Principal Component Analysis* dapat terlibat dalam analisis cluster yang berasal dari pemodelan topik[12]:

1. *Dimensionality Reduction*

   PCA terutama digunakan untuk reduksi dimensi. Saat menangani keluaran pemodelan topik, yang mungkin melibatkan sejumlah besar fitur (kata atau istilah), PCA dapat diterapkan untuk mereduksi dimensi. Setiap topik yang dihasilkan dari pemodelan topik dapat dianggap sebagai dimensi, dan PCA membantu mengidentifikasi topik paling penting (komponen utama) yang berkontribusi signifikan terhadap varians dalam kumpulan data.

2. *Identifying Key Topics*

   PCA dapat membantu dalam memahami topik mana yang paling berkontribusi terhadap variasi antar dokumen. Dengan mengurangi dimensi, memvisualisasikan dan menafsirkan hubungan antar topik menjadi lebih mudah. Hal ini membantu dalam mengidentifikasi topik atau tema paling signifikan yang ada dalam kumpulan data.

3. *Clustering Based on Reduced Dimensons*

   Setelah menggunakan PCA untuk mereduksi dimensi, komponen utama yang dihasilkan dapat digunakan dalam analisis cluster. Algoritme *Cluster* (seperti K-means atau *Cluster* hierarki) dapat beroperasi lebih efisien pada kumpulan dimensi yang lebih kecil yang diperoleh dari PCA, menjadikannya kurang intensif secara komputasi sambil mempertahankan esensi topik aslinya.

4. *Enhancing Interpretability*

   PCA dapat meningkatkan interpretasi cluster yang berasal dari pemodelan topik. Hal ini membantu dalam memvisualisasikan dan memahami hubungan antar topik, sehingga berpotensi mengungkap pola atau korelasi mendasar yang mungkin tidak langsung terlihat dari distribusi topik aslinya.

5. *Feature Selecton for Better Clustering*

   PCA juga membantu pemilihan fitur dengan menyoroti topik yang paling relevan. Hal ini dapat memberikan hasil *Cluster* yang lebih baik karena topik yang paling signifikan (komponen utama) memberikan kontribusi yang lebih berarti terhadap pembedaan klaster.

6. *Preprocessing for Improved Clustering*

Menerapkan PCA sebelum *Cluster* mungkin membantu dalam mengatasi masalah terkait dimensi tinggi, multikolinearitas antar topik, atau fitur gangguan yang berasal dari pemodelan topik.

2.5. *T-Distributed Stochastic Neighbor Embedding (t-SNE)*

T-Distributed Stochastic Neighbor Embedding (t-SNE) adalah teknik reduksi dimensi lain yang sering digunakan dalam konteks *Cluster* dan visualisasi, terutama ketika menangani data berdimensi tinggi yang berasal dari teknik seperti pemodelan topik[4], [13]. Berikut cara t-SNE dapat digunakan dalam analisis cluster yang berasal dari pemodelan topik:

1. *Dimensionality Reduction for Visualization*

   Mirip dengan PCA, t-SNE bertujuan untuk mereduksi data berdimensi tinggi ke ruang berdimensi lebih rendah. Namun, t-SNE lebih fokus pada pelestarian struktur lokal dan hubungan antar titik data. Ketika diterapkan pada keluaran pemodelan topik (seperti distribusi kata untuk topik dalam dokumen), t-SNE dapat membantu memvisualisasikan cluster dalam ruang berdimensi lebih rendah (seringkali 2D atau 3D), menjaga kesamaan lokal antara topik atau dokumen[14].

2. *Visualizing Clusters*

   Setelah menerapkan t-SNE, cluster yang berasal dari pemodelan topik dapat divisualisasikan dalam ruang berdimensi lebih kecil. Visualisasi

ini dapat memberikan wawasan tentang hubungan antara topik atau dokumen, menunjukkan topik mana yang cenderung lebih sering muncul secara bersamaan atau bagaimana dokumen dikelompokkan berdasarkan kesamaan kontennya[14].

3. *Revealing Subtle Patterns*

   Setelah menerapkan t-SNE, cluster yang berasal dari pemodelan topik dapat divisualisasikan dalam ruang berdimensi lebih kecil. Visualisasi ini dapat memberikan wawasan tentang hubungan antara topik atau dokumen, menunjukkan topik mana yang cenderung lebih sering muncul secara bersamaan atau bagaimana dokumen dikelompokkan berdasarkan kesamaan kontennya[14].

4. *Interpretability of Clusters*

   Representasi visual yang dihasilkan oleh t-SNE dapat membantu dalam menafsirkan cluster yang berasal dari pemodelan topik. Ini membantu dalam mengidentifikasi kelompok topik yang memiliki kesamaan dan memahami kedekatan atau pemisahannya dalam ruang berdimensi tereduksi[14].

5. *Parameter Sensitivity*

   Performa t-SNE bisa jadi sensitif terhadap parameternya, seperti parameter kebingungan yang mengontrol jumlah tetangga terdekat yang

dipertimbangkan selama pengoptimalan. Penyetelan parameter yang cermat sangat penting untuk mendapatkan visualisasi yang bermakna[14].

6. *Complementing PCA*

   Meskipun PCA efektif dalam menangkap struktur global, t-SNE lebih fokus dalam menjaga hubungan lokal. Dalam beberapa kasus, penggunaan PCA dan t-SNE secara berurutan dapat memberikan pemahaman komprehensif tentang struktur dalam data[14].

Intinya, t-SNE berfungsi sebagai alat yang ampuh untuk memvisualisasikan dan menafsirkan hubungan dan cluster yang diperoleh dari pemodelan topik[15]. Kemampuannya untuk mengungkap struktur lokal dan hubungan antar titik data dapat memberikan wawasan berharga tentang pola inheren dalam distribusi topik atau cluster dokumen. Ketika digunakan bersama teknik lain seperti PCA, t-SNE berkontribusi pada analisis data berdimensi tinggi yang lebih komprehensif dari pemodelan topik[16], [17].

2.6. *K-means Cluster*

Cluster K-means adalah algoritme pembelajaran mesin tanpa pengawasan populer yang dapat diterapkan pada topik cluster yang berasal dari pemodelan topik[2], [18], [19]. Berikut ini cara K-means dapat digunakan dalam analisis klaster dari pemodelan topik:

1. *Clustering Topics*

Singkatnya, *Cluster* K-means dapat digunakan untuk mengatur dan mengelompokkan topik yang berasal dari pemodelan topik ke dalam kelompok yang koheren. Ini memfasilitasi interpretasi dan pemahaman kesamaan tematik antar topik, membantu dalam analisis dan eksplorasi data tekstual yang kompleks. Namun, pertimbangan yang cermat terhadap jumlah cluster dan evaluasi hasil sangat penting untuk interpretasi dan wawasan yang bermakna.

2. *Identifying Topic Groups*

   K-means mempartisi topik menjadi K cluster, dengan K adalah angka yang telah ditentukan sebelumnya dan dipilih oleh analis. Algoritme ini secara berulang menetapkan topik ke klaster berdasarkan kesamaannya, yang bertujuan untuk meminimalkan jumlah kuadrat jarak antara topik dan pusat massa klasternya masing-masing.

3. *Interpretaton of Topic Clusters*

   Cluster yang dihasilkan mewakili kelompok topik yang memiliki karakteristik serupa atau muncul bersamaan dalam dokumen. Analis dapat menafsirkan kelompok ini untuk membedakan tema, subjek, atau pola umum di seluruh korpus.

4. *Determining Optimal Number of Clusters*

Penentuan jumlah cluster yang optimal (nilai K) sangatlah penting. Berbagai teknik seperti metode siku atau skor siluet dapat membantu mengidentifikasi jumlah kelompok yang paling sesuai untuk menangkap variasi topik yang bermakna.

5. *Visualizing Clustered Topic*

   Teknik visualisasi dapat diterapkan untuk menampilkan hubungan antar topik dalam cluster. Misalnya, memvisualisasikan distribusi topik dalam kelompok atau menggunakan teknik seperti penskalaan multidimensi dapat memberikan wawasan tentang kedekatan dan kesamaan topik.

6. *Refinement and Evaluation*

   Penyempurnaan klaster mungkin diperlukan dengan bereksperimen dengan berbagai teknik pra-pemrosesan, parameter model, atau bahkan mempertimbangkan metode *Cluster* hierarki untuk menangkap hubungan kompleks antar topik.

Singkatnya, *Cluster* K-means dapat digunakan untuk mengatur dan mengelompokkan topik yang berasal dari pemodelan topik ke dalam kelompok yang koheren. Ini memfasilitasi interpretasi dan pemahaman kesamaan tematik antar topik, membantu dalam analisis dan eksplorasi data tekstual yang kompleks. Namun, pertimbangan yang cermat terhadap jumlah cluster dan evaluasi hasil sangat penting untuk interpretasi dan wawasan yang bermakna[18], [20].

2.7. *Uniform Manifold Approximation and Projection (UMAP)*

UMAP (Uniform Manifold Approximation and Projection) adalah teknik reduksi dimensi lainnya, mirip dengan t-SNE, yang biasa digunakan untuk memvisualisasikan data berdimensi tinggi di ruang berdimensi lebih rendah sambil mempertahankan struktur global[21]. Dalam hal visualisasi dengan label t-SNE menggunakan UMAP, hal ini melibatkan penyematan kluster atau label turunan t-SNE ke dalam plot UMAP. Berikut cara penerapannya:

1. *Dimensionality Reductioin with t-SNE*

   t-SNE diterapkan untuk mereduksi data berdimensi tinggi (misalnya, distribusi topik dari dokumen) ke dalam ruang berdimensi lebih rendah, biasanya 2D atau 3D. Langkah ini menangkap hubungan lokal antar titik data, yang bertujuan untuk merepresentasikan data berdimensi tinggi dalam bentuk yang lebih dapat dipahami secara visual[22].

2. *Clusering or Labeling with t-SNE*

   t-SNE sering membantu dalam mengelompokkan atau memberi label pada titik data berdasarkan representasi dimensinya yang dikurangi. Misalnya, dokumen dengan distribusi topik serupa dapat dikelompokkan atau diberi label sesuai dengan cluster yang berasal dari t-SNE[23].

3. *Further Redcution and Visualization with UMAP*

   UMAP kemudian diterapkan pada data berlabel t-SNE. UMAP berupaya menciptakan representasi data yang lebih global dan

berpotensi lebih dapat diinterpretasikan dengan tetap mempertahankan struktur lingkungan yang dibangun oleh t-SNE[24].

4. *Embedding of Visualized Data*

   Setelah visualisasi UMAP dihasilkan, cluster atau label yang diturunkan dari t-SNE dapat ditumpangkan atau diwarnai ke plot UMAP. Hal ini memungkinkan visualisasi cluster t-SNE di ruang yang telah diubah oleh UMAP[24].

5. *Interpretation of Visualized Data*

   Plot yang dihasilkan, menggabungkan keunggulan t-SNE dan UMAP, memungkinkan inspeksi visual terhadap cluster atau grup berlabel yang berasal dari t-SNE dalam konteks global yang berpotensi lebih dapat ditafsirkan yang disediakan oleh UMAP[24].

Singkatnya, kombinasi UMAP dengan label t-SNE memungkinkan pendekatan multi-langkah untuk memvisualisasikan data berdimensi tinggi, yang berpotensi memberikan wawasan lebih dalam tentang struktur dan hubungan cluster atau grup berlabel dalam data[21].