BAB I
PENDAHULUAN

1.1. Latar Belakang

ArXiv dimungkinkan oleh format file TeX yang ringkas, yang memumngkinkan makalah ilmiah dengan mudah dikirim melalui internet dan direnderdi sisi klien. Artikel dimulai sebagai arsip fisika atau *scientific* yang disebut arsip pracetak LANL, tetapi segera diperluas untuk emmasukkan astronomi, matematika, ilmu komputer, biologi kuantitatif dan yang terbaru pada saat itu adalah ilmu statistik. Pada penelitian kali ini, penulis bertujuan untuk mengajukan analisis pengujian terhadap dataset artikel yang ada pada journal arXiv dengan menggunakan dataset yang berasal dari Kaggle.com yang berisikan data record terkait informasi artikel yang ada pada journal arXiv. Analisis pada penelitian kali ini akan menggunakan Metode *K-means* dikarenakan Metode ini adalah Metode yang cocok untuk digunakan pada dataset yang bersifat kategorikal data dan melihat referensi yang dijadikan acuan pada penelitian kali ini adalah analisis Topic Modelling yang terikat kuat dengan *Natural Language Processing* dengan mengimplementasikan Metode Cluster dan algoritma Dimensionality Reduction.

Metode K-means nampaknya sangat cocok digunakan dalam kasus analisis pemodelan topik jika dilihat berdasarkan referensi yang digunakan yang menyatakan bahwa K-means menggunakan titik tengah atau centroid dari setiap titik dalam cluster diperoleh untuk memberikan suatu nilai, untuk setiap poin. Cluster tersebut kemudian akan digunakan sebagai referensi untuk menentukan label topik yang akan diberikan. Clustering dengan metode K-means sangat bermanfaat untuk analisis data karena data yang diolah akan terbentuk secara terorganisir. Metode K-means menggunakan nilai centroid setiap

cluster, dan pada setiap cluster terdapat sekumpulan nilai fitur yang dapat mengidentifikasi setiap kelompok cluster yang akan dibuat. Dengan memberikan nilai bobot, centroid juga digunakan sebagai bentuk interpretasi kualitatif suatu kelompok yang diwakili oleh setiap cluster.

Beberapa referensi penelitian terdahulu menunjukkan bahwa algoritma reduksi dimensi (PCA) dan t-SNE merupakan algoritma yang ekuivalen untuk digunakan dengan metode K-means karena kedua algoritma ini mempunyai sifat yang mirip dan sifat yang berbeda yaitu linier dan non linier. Hasil visualisasi, pada kalimat lain dikatakan jika suatu benda mempunyai sifat yang sama atau dikatakan keduanya mempunyai sifat linier maka hasil yang diperoleh tidak sebaik menggunakan algoritma linier dan non linier.

Dengan mengetahui skala halaman sitasi jurnal arXiv.com, penelitian ini bertujuan untuk menganalisis dan mengelompokkan data serupa berdasarkan topik pemodelan serupa menggunakan dataset jurnal arXiv situs Kaggle.com di atas. Penelitian ini bertujuan untuk memberikan hasil visualisasi berbasis informasi dan topik terkait model dokumen yang terkandung melalui analisis cluster dengan melakukan beberapa prosedur pemrosesan bahasa alami.

Metode K-means pertama kali digunakan oleh James MacQueen pada tahun 1967, meskipun idenya datang dari Steinhaus Hugo pada tahun 1956[1]. K-means merupakan metode yang awalnya dirancang untuk dapat mencari nilai optimal suatu partisi dan menggunakan nilai k sebagai determinannya. nilai awal. centroid (titik tengah), hal ini menunjukkan dalam bentuk lain bahwa nilai k merupakan atribut yang sangat penting dalam analisis menggunakan metode K-means dan penentuan nilai k dapat dilakukan

dengan menggunakan nilai strain k yang diperoleh dengan metode elbow[1][2]. Nilai distorsi akan berkurang jika jumlah nilai cluster ditambah, namun hal ini juga dapat berdampak pada pembagian dimensi menjadi dua kali jumlah cluster yang valid.

Penelitian sebelumnya telah dilakukan oleh Maksim Ekin Eren pada tahun 2019, beliau bersama tim melakukan penelitian analisis kasus Covid-19 dengan topik Klaster Sastra Covid-19, penelitian ini menggunakan metode K-means[2]. Kumpulan data yang digunakan berupa review dan artikel terkait perkembangan epidemi Covid-19 dari CORD19-Dataset. Dalam penelitiannya, ia dan timnya mengatakan bahwa penggunaan algoritma pereduksi dimensi (PCA) meningkatkan dan memfasilitasi tingkat kinerja dan efisiensi metode K-means. Melihat nilai deformasi yang diperoleh, mereka menggunakan nilai tetap untuk k = 20 untuk mendapatkan label untuk setiap record data dalam dataset dan kemudian menyusun ulang hasil yang diperoleh dengan menggunakan algoritma t-SNE akan memberikan distribusi probabilitas yang dapat mewakili konektivitas semua data yang digunakan dan memberikan dimensi yang lebih kecil. Penelitian ini menggunakan 49.000 catatan dalam kumpulan datanya dan menemukan akurasi 79 menggunakan pengklasifikasi K-Neighbours[2].

Penelitian sebelumnya lainnya dilakukan oleh Nikhil Jangamreddy pada tahun 2019, dalam penelitiannya ia melakukan analisis yang melibatkan penerapan teknik reduksi dimensi yang digunakan untuk menghilangkan data yang berisik (data kotor) yang ada pada data sebelumnya dan sebelum menerapkan metode cluster dengan K-mean.[3] Dalam penelitiannya, ia melaporkan bahwa teknik reduksi dimensi menggunakan algoritma PCA ditemukan setara dengan nilai autoencoder linier pada batas tertentu. Hasil penelitian ini memberikan gambaran yang lebih jelas antara kombinasi teknik reduksi

dimensi dan kombinasinya dengan autoencoder yang diterapkan pada dataset IRIS untuk memvisualisasikan hasil buah cluster[3].

Penelitian sebelumnya lainnya dilakukan oleh Khalid Kahloot dengan melakukan analisis penelitian mengenai improvabilitas visualisasi menggunakan metode hybrid khususnya K-means dan t-SNE dengan judul Meningkatkan Visualisasi clustering dan kemampuan visualisasi t-SNE, Penelitian ini bertujuan untuk meningkatkan atau mengimprovisasi kinerja dalam, memperoleh cluster[4]. Kemudian hasil dan hasil visualisasi terbaik dibandingkan dengan metode gabungan K-means dan PCA menggunakan teknik reduksi dimensi yang sama. Dalam penelitiannya beliau membandingkan metode clustering lainnya seperti metode HDBSCAN, Spectral Clustering dan Gaussian Mixture dengan metode K-means, dan pada bagian 6 dijelaskan bahwa nilai eksak dan nilai mean dari standar deviasi yang diperoleh metode K-means dengan pada kumpulan data MNIST nilai akurasinya 80% dan nilai simpangan bakunya 80%, kemudian nilai akurasinya 56% dan simpangan bakunya 56% pada kumpulan datanya[4].

1.2. Rumusan Masalah

Berdasarkan pemaparan pada latar belakang diatas, dapat dirumuskan dan membentuk suatu rumusan masalah. Rumusan masalah yang didapatkan adalah seperti berikut:

1. Bagaimana cara meng-interpretasikan data yang digunakan dengan menggunakan metode *K-means* dan mengkombinasikannya dengan algoritma pereduksi dimensi *PCA* dan *t-SNE.*

2. Melakukan improvisasi pada hasil visual dengan mengimplementasikan Analisis *Topic Modelling* menggunakan Metode Kmeans dan menggabungkannya dengan algoritma reduction yaitu PCA dan membuat modelnya dengan visual *t-SNE* Model.
3. Bagaimana hasil atau tingkat efisiensi pada metode dan algoritma pada poin diatas kedalam penelitian dengan analisa Topic Modelling dengan menggunakan dataset arXiv dataset yang digunakan pada penelitian kali ini.

1.3. Tujuan Penelitian

Penelitian ini memiliki tujuan dan memiliki suatu pencapaian, hal ini dibuat agar penelitian ini tidak keluar dari konteks yang sudah dibuat sebelumnya, dan untuk memberikan improvisasi pada hasil dalam merepresentasikan persebaran data dan visualisasi data dengan menggunakan kombinasi antara metode *K-means* dan algoritma *dimensionality reduction PCA* dan *t-SNE*, dan memberikan hasil visualisasi yang informatif dan tidak hanya sekadar merepresentasikan persebaran data saja.

1.4. Batasan Masalah

Berdasarkan pemaparan rumusan masalah dan tujuan penelitian diatas, dibentuk lah sebuah Batasan masalah berikut, guna agar penelitian kali ini tidak keluar dari pokok permasalahan yang ada, berikut adalah Batasan masalah yang sudah disesuaikan pada rumusan masalah dan tujuan penelitian sebelumnya:

1. Penelitian ini memiliki kriteria yang digunakan pada dataset yang dipakai yaitu Journal Article arXiv dataset yang didapatkan dari situ Kaggle.

2. Metode yang dijadikan acuan analisis penelitian kali ini adalah metode *K-means* dan mengkombinasikannya dengan algoritma *dimensionality reduction* yaitu *PCA* dan *t-SNE*.
3. Penelitian ini bertujuan untuk melakukan analisis terkait Topik Modelling pada Journal Article arXiv dataset dan membuat hasil visualisasi yang informatif.
4. Menggunakan Bahasa pemrograman python dan menggunakan platform google collab dalam proses penelitiannya.