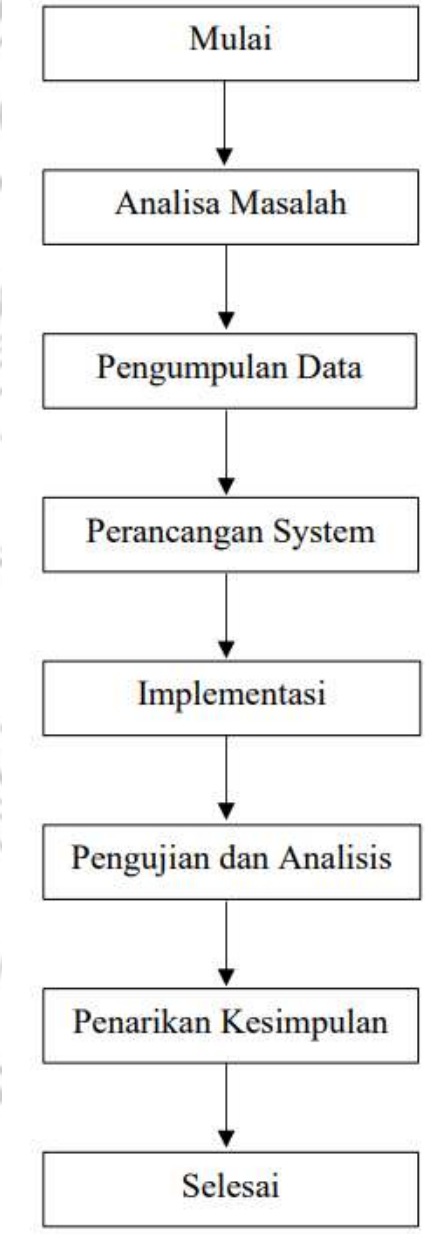
BAB III

METODE PENELITIAN

Pada proses penelitian ini terdapat beberapa tahapan-tahapan yang dilakukan, dapat dilihat pada Gambar 1 dibawah ini:



Gambar 1. Tahapan Penelitian

3.1. Analisa Masalah

Hasil Analisa pada latar belakang dengan rumusan masalah pada penelitian kali ini dengan menggunakan teknik clustering menggunakan metode *K-means,* didapati beberapa tahapan-tahapan pada pemrosesan dataset yang dilakukan, yaitu sebagai berikut:

1. Memuat dataset yang digunakan
2. Parsing teks dari fitur atau attribute [*abstract*] setiap dokumen dengan menggunakan teknik *Natural Language Processing.*
3. Transformasi setiap value pada attribute abstrak menjadi bentuk fitur vector dengan menggunakan Term Frequency-Inverse Document Frequency (TF-IDF)
4. Gunakan analisis algoritma *Principal Component Analysis (PCA)* untuk mengurangi besar dimensi dari data vector sebelumnya
5. Implementasikan visualisasi scatter plot dan pengurangan dimensi dengan menggunakan *t-Distributed Stochastic Neighbor Embedding (t-SNE)*
6. Gunakan t-SNE Embeddings sebagai input untu UMAP Embeddings
7. Implementasikan visualisasi dengan label k-means pada UMAP Embeddings

3.2. Pengumpulan Data

Pada tahap ini pengumpulan data yang akan digunakan pada penelitian kali ini adalah dengan menggunakan dataset Journal arXiv yang didapatkan dari situs Kaggle. Dataset ini berisikan data record terkait paper dan journal yang ada pada situs arXiv yang didalamnya terdapat beberapa atribut yaitu id, submitter, authors, title, comments, journal-ref, doi, report-no, category, license, abstract, vesion, update_date, dan author_parsed.

3.3. Variable Penelitian

Berdasarkan hasil dari pemaparan analisis masalah dan pengumpulan data pada tahapan sebelumnya, maka dapat ditetapkan untuk beberapa variable penentu yang menjadi kata kunci pada penelitian Topik Modelling kali ini. Variable atau atribut utama yang akan digunakan dan dijadikan sebagai variable penentu adalah kolom abstract, dan beberapa atribut utama lainnya yang digunakan adalah id, authors, title, doi, category, dan terakhir abstract. Contoh dataset yang sudah melalui tahap pre-processing awal dapat dilihat pada gambar 2.1 dibawah ini.

	id	authors	title	doi	category	abstract
0	0704.0033	Maxim A. Yurkin, Valeri P. Maltsev, Alfons G. ...	Convergence of the discrete dipole approximati...	10.1364/JOSAA.23.002578 10.1364/JOSAA.32.002407	[physics.optics, physics.comp-ph]	We performed a rigorous theoretical converge...
1	0704.0038	Maxim A. Yurkin, Alfons G. Hoekstra	The discrete dipole approximation: an overview...	10.1016/j.jqsrt.2007.01.034 10.1016/j.jqsrt.20...	[physics.optics, physics.comp-ph]	We present a review of the discrete dipole a...
2	0704.0479	T.Geisser	The affine part of the Picard scheme	None	[math.AG, math.KT]	We describe the maximal torus and maximal un...
3	0704.1445	Yasha Gindikin and Vladimir A. Sablikov	Deformed Wigner crystal in a one-dimensional q...	10.1103/PhysRevB.76.045122	[cond-mat.str-el, cond-mat.mes-hall]	The spatial Fourier spectrum of the electron...
4	0704.1476	Chris Austin	TeV-scale gravity in Horava-Witten theory on a...	None	[hep-th]	The field equations and boundary conditions ...

Gambar 2.1 Variable atau Atribut Utama pada Dataset yang digunakan

3.4. Perancangan System

Perancangan System pada analisis penelitian kali ini adalah dengan menggunakan platform google colab, dengan menggunakan Bahasa pemrograman python dalam perancangannya serta implementasi pada tahapan-tahapan yang sudah disebutkan pada sub bab sebelumnya.

3.5. *PCA (Principal Component Analysis)*

Penjelasan Deskriptif ini secara harfiah memberi bukti bahwa komponen utama yang dikalkulasikan pada teknik *Principal Component Analysis* setara dengan *indicator* pengelompokkan cluster dengan metode *K-means*[25]. Dengan pemahaman lain artinya *Principal Component Analysis* dapat memproses *Cluster* sesuai pada fungsi objektif pada metode *K-means* clustering. Hal Ini menjadi bukti untuk mengimplementasikan metoode *K-means* clustering dengan menggunakan algoritma pereduksi dimensi *Principal Component Analysis* sangatlah cocok[25].

K-means Clustering memiliki karakteristik pada nilai *centroids* disetiap individu cluster-nya. Definisi pada metode *K-means* memiliki fungsi sebagai berikut:

$$J_K = \sum_{k=1}^{K} \sum_{i \in C_k} (X_i - m_k)^2$$

Fungsi diatas memiliki nilai x1,x2,x…, xn yang berarti nilai matriks pada suatu Data, dan *mk* memiliki arti sebagai pusat massa cluster *k*[2], yang dapat didefinisikan sebagai:

$$m_k = \sum_{i \in C_k} X_i / n_k$$

nk menjadi petunjuk sebagai banyaknya jumlah titik pada cluster *k*[2]. dengan hal ini fungsi *Principal Component Analysis* adalah sebagai berikut:

asumsikan matriks X mempunyai data sebagai $X = (x1, x2, \dots, xn)$

$Y = (y1, y2, \dots, yn)$ dan

$y_i = X_i - \bar{X}$, lalu

$$\frac{\sum_i (X_i - \bar{X})\,(X_i - \bar{X})^T}{n = YY^T}$$

Ak adalah nilai arah utama, nilai *Eigen* didefinisikan pada kondisi nilai *Bk* untuk *Principal Component Analysis* dapat kalkulasikan dengan fungsi berikut[2]:

$$YY^T A_k = \lambda_k A_k, Y^T Y B_k = \lambda_k B_k, B_{k,} B_k = Y^T A_k / \lambda_k^{1/2}$$

3.6. *t-SNE (t-Distributed Stochastic Neighbor Embedding)*

Pada dimensi yang memiliki nlai yang tinggi, *t-Distributed Stochastic Neighbor Embedding (t-SNE)* sangat selektif dalam hal memilih nilai ukuran kesamaan pada setiap data disetap satu titik cluster untuk dua tingkat high-dimensionality dan satu untuk penyisipan dua dimensi (2D). Lalu melakukan percobaan untuk mebentuk suatu embedding dua dimensi yang dapat mengurangi nilai *divergensi*[4].

Algoritma ini memvisualisasikan titik-titik pada satu ruang ke ruang yang lain dalam usaha untuk menekankan bentuk struktur information yang tetap sama.

Visualisasi pada konteks literatur artikel ilmiah juga sudah pernah digunakan diseluruh domain yang memiliki dimensi yang lebih luas. Contoh penelitian paper sebelumnya telah merepresentasikan bentuk visual 1,7 juta artikel penelitian ilmiah. Naskah ini disusun berdasarkan referensi atau acuan satu sama lain. Pada penelitian kali ini, penulisa menggunakan teks yang berisikan setiap artikel dengan abstract untuk menentukan tingkan kesamaan topik dengan lebih baik[15]. Contoh gambaran lainnya adalah dengan menggunakan dekomposisi tensor *Tucker* yang digunakan untuk mengelompokkan sebuah permain *Shakespeare* dalam usaha guna memberikan gambaran visual tentang bagaimana cara menerapkan tensor dekomposisi supaya bisa mendukung analisis cluster.

3.7. *K-means*

K-means clustering adalah salah satu teknik Machine Learning yang biasa dipakai dalam mengolah data dengan skala yang besar. Teknik K-means dapat mengelompokkan data atau

menyatukan sekumpulan data berdasarkan bentuk kategorikal data tersebut dengan menggunakan dasar nilai mean terdekat. Pada tahapan pengimplementasian metode sebelumnya dapat dikatakan beberapa poin yang penting antara label dalam mengelompokkan dan cluster yang dapat dilihat pada hasil plot yang dihasilkan oleh *t-Distributed Stochastic Neighbor Embedding (t-SNE)* yang nantinya dan K-means dan di implementasikan secara mandiri untuk setiap X1 dan X2[15]. Dilain hal tersebut, *t-Distributed Stochastic Neighbor Embedding (t-SNE)* dapat mereduksi data menjadi dua dimensi, K-means dapat mengelompokkan representasi data dengan dimensi yang lebih tinggi, oleh karena itu keterikatan pada ekuivalensi antara metode K-means dan algoritma *t-Distributed Stochastic Neighbor Embedding (t-SNE)* dapat dikatakan satu ekuivalensi.[18]

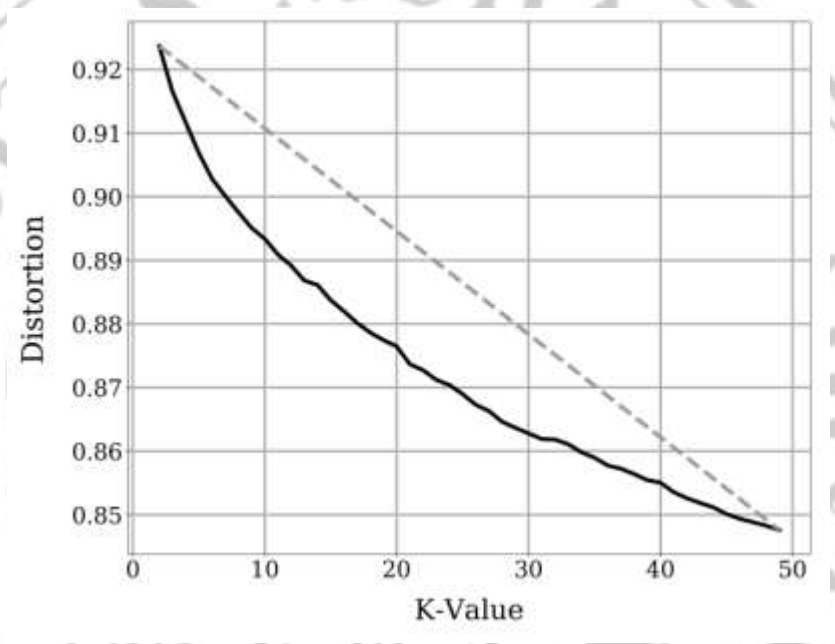
Pada penerapan metode sebelumnya, dapat diperhatikan poin penting antara label grup dan label cluster, yang nantinya dapat dilihat pada hasil plot *t-Distributed Stochastic Neighbor Embedding (t-SNE)* dan K-mean, dan diterapkan secara independen untuk setiap x_1 dan x_2. Sementara *t-Distributed Stochastic Neighbor Embedding (t-SNE)* mereduksi data menjadi dua dimensi, K-means mensintesis representasi data berdimensi lebih tinggi.

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \left\| x^i - \mu_k \right\|^2$$

Pada fungsi diatas memberikan arti bawah $w_{ik} = 1$ pada setiap titik data x^i apabila termasuk kedalam cluster *k,* apabila tidak terrmasuk maka, $w_{ik} = 0$. Dan artinya *k* dapat dikatakan sebagai pusat centroid dari cluster x^i[16].

Dengan menentukan nilai *k* dengan menggunakan metode *Elbow*, nilai *k* yang ditentukan dengan menggunakan metode *Elbow* ini akan lebih optimal, sedangkan apabila meningkatkan nilai *k* atau menggunakan nilai yang lebih tinggi dapat menyebabkan hasil distorsi yang lebih

rendah dan mengakibatkan cluster yang berdimensi lebih kecil. Faktor lain yang mengatakan bahwasanya metode *K-means* diterapkan pada ruang diemnsi yang lebih tinggi, adalah dari segi structural yang tidak dapat dikatakan secara akurat pada dua dimensional plotnya. Seperti pada gambar 3.2 distorsi terjadi dua kali pada angka antara 10 s/d 20, namun nilai distorsi terjadi pada titik akhir di angka 20 sehingga ini menjadi opsi pilihan yang lebih optimal dan variatif dalam menentukan nilai optimal untuk *k* value.



Gambar 3.2 Contoh Grafik Distorsi Metode Elbow