

BAB 3

METODE PENELITIAN

Pada penelitian ini, metode yang digunakan adalah metode *K-Means Clustering* untuk melakukan *clustering* data teks. Sebelum melakukan *clustering*, akan dilakukan pengumpulan data dengan menggunakan teknik *web scraping* melalui twitter pada topik *thrifting*. Setelah data terkumpul, data akan dibersihkan dan dilakukan *clustering*. Fokus penelitian ini terletak pada penentuan *cluster* berdasarkan data yang didapat dari *twitter* untuk selanjutnya dilihat perbandingan data antar *cluster*.

3.1 BAHAN DAN ALAT PENELITIAN

Bahan penelitian yang perlu disiapkan pada penelitian ini yaitu data serta informasi mengenai minat masyarakat terhadap pakaian bekas, informasi dari penelitian sebelumnya serta informasi melalui *hashtag* pada *Twitter* dengan menggunakan *hashtag* *#import* *#thrifting* *#pakaianbekas* *#pakaianimport*.

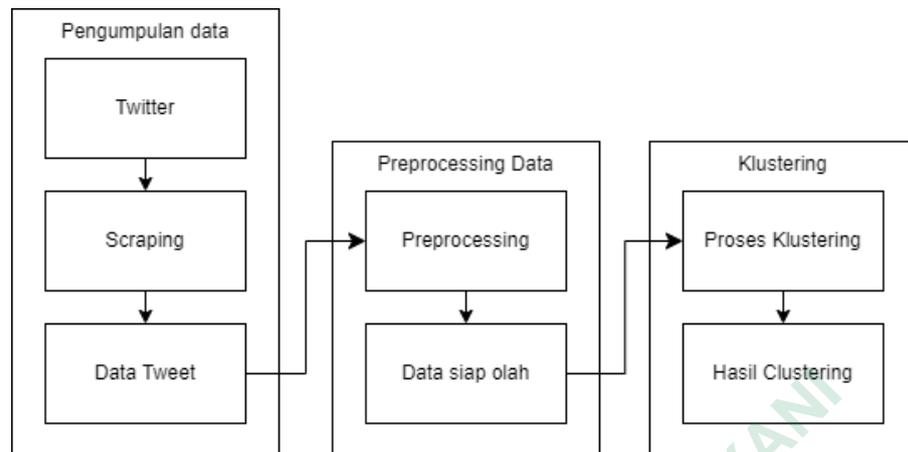
Data tersebut kemudian diolah dengan menggunakan beberapa alat yang digunakan dengan spesifikasi tertentu. Alat yang digunakan dalam penelitian ini adalah komputer dengan spesifikasi cukup untuk menjalankan sistem operasi dan perangkat lunak pengembangan serta koneksitas Internet.

Sistem Operasi dan program-program aplikasi yang dipergunakan dalam dalam pengembangan aplikasi ini adalah:

1. Sistem Operasi: *Windows 10* atau yang lebih baru.
2. Aplikasi *web Jupyter Notebook*
3. Bahasa pemrograman *Phyton*.

3.2 JALAN PENELITIAN

Dalam penelitian ini diperlukan diagram alur untuk membantu dalam jalannya penelitian. Jalan penelitian ditunjukkan pada gambar 1 yaitu diagram alur penelitian.



Gambar 3. 1 Diagram Alur Penelitian (Kharisma & Aesy, 2023)

Pada awal proses akan dilakukan pengumpulan data dengan menggunakan scraping pada twitter dengan menggunakan *selenium*. *Library selenium* merupakan *library* yang berfungsi untuk melakukan otomatis testing dari *web*. Penggunaan *library selenium* adalah karena *library* tersebut cukup mudah digunakan. *Scraping* akan dilakukan dengan menggunakan beberapa kata kunci yang berkaitan dengan *thrifting*. *Hashtag* yang digunakan adalah *#import #thrifting #pakaianbekas #pakaianimport* serta beberapa keyword lain diluar hashtag seperti “*thrifting pemerintah*”. Setelah data diambil kemudian data disimpan dalam bentuk *csv* untuk selanjutnya dilakukan *preprocessing*.

Tahap selanjutnya adalah tahap *preprocessing*. Tahap ini merupakan tahap memproses data mentah menjadi data siap olah. Dalam melakukan *preprocessing*, data akan dibersihkan dengan meliputi beberapa proses seperti menghapus data duplikat, *case folding*, *punctuation removal*, *stemming*, dan *TF-IDF*. Menghapus data duplikat dilakukan secara manual dengan menggunakan *Microsoft Excel*. Penggunaan *Microsoft Excel* untuk menghapus data duplikat adalah agar data dapat dibaca dan diteliti satu per satu. Beberapa data yang telah terkumpul terdapat data yang menggunakan bahasa inggris sehingga perlu diubah kedalam bentuk bahasa Indonesia. Selain itu data juga banyak yang

menggunakan singkatan dalam penulisan sehingga perlu diperbaiki agar dapat dilakukan *preprocessing* tahap selanjutnya.

Setelah dilakukan penghapusan data duplikat, selanjutnya data masuk ke fase *preprocessing* tahap selanjutnya yaitu *case folding*. *Case folding* dilakukan dengan menggunakan *library* “re” (*Regular Expression*). *Regular Expression* merupakan *library* untuk mengecek huruf huruf pada kalimat. Pada kasus ini, *case folding* akan dilakukan untuk menyetarakan huruf menjadi huruf kecil semua. Setelah dilakukan *case folding* selanjutnya dilakukan tahap *punctuation removal* yang merupakan tahap menghapus punctuation atau tanda baca dari kalimat. penghapusan dilakukan dengan menggunakan *library* yang sama yaitu *library* “re”.

Tahap selanjutnya pada fase *preprocessing* adalah *stemming*. *Stemming* dilakukan untuk mengubah kata menjadi kata dasar. *Stemming* akan dilakukan dengan menggunakan *library* *sastrawi*. *Sastrawi* merupakan *library* untuk *stemming* untuk data berbahasa indonesia. Selain dilakukan *stemming*, pada tahap ini juga dilakukan penghapusan *stopword* atau kata sambung seperti “yang”, “di”, “ke”, dan lainnya. Kemudian selanjutnya data dilakukan *TF-IDF*. *Preprocessing* dengan *TF-IDF* dilakukan dengan menggunakan *library* *sklearn* modul *tfidfvectorizer*. *TF-IDF* dilakukan untuk menghitung berapa kali kata tertentu muncul dalam satu kalimat. Sehingga *clustering* dapat dilihat berdasarkan keyword yang sering muncul pada data.

Setelah data dilakukan *preprocessing* selanjutnya data akan masuk ke fase *processing*. Pada fase ini data akan dilakukan *processing* dengan *K-Means* menggunakan *library* *sklearn* modul *cluster*. Tahap awal sebelum dilakukan *clustering* adalah pencarian *elbow* untuk menentukan nilai K optimum. Setelah didapatkan nilai K maka selanjutnya adalah melakukan *cluster* dengan *K-Means*. Hasil *clustering* akan ditampilkan dalam bentuk diagram batang. Kemudian berdasarkan diagram batang dari

hasil *clustering* tersebut. Maka akan terlihat opini masyarakat di *twitter* mengenai *thrifting*.

UNIVERSITAS JENDERAL ACHMAD YANI
PEPUSTAKAAN
YOGYAKARTA