

BAB III

METODE PENELITIAN

Metode penelitian mengenai *clustering* berita *online* mengenai Covid-19 dengan menggunakan metode *k-means clustering* dan *hierarchical clustering* ini terdiri dari 4 (empat) tahapan. Tahapan penelitian terdiri dari tahap pengumpulan data, tahap implementasi, tahap analisa hasil dan tahap penyusunan laporan.

3.1. Tahap Pengumpulan Data

Pada tahap ini dilakukan pengambilan data yang akan diolah oleh peneliti. Data yang digunakan dalam penelitian ini berasal dari situs berita online *cnnindonesia.com*. Berita yang digunakan dalam penelitian ini adalah berita dengan tag Covid-19. Pengumpulan data akan dilaksanakan dengan metode *web scraping* menggunakan *beautifulsoup4*. Metode *web scraping* dipilih karena dengan metode tersebut, pengambilan data dapat dilakukan secara otomatis. Sehingga memungkinkan untuk mengambil data sebanyak 228 artikel dalam waktu yang singkat.

3.2. Tahap Clustering Data

Proses *clustering* terdiri dari beberapa proses yaitu *text preprocessing*, *term-weighting*, *feature selection*, dan *clustering*.

3.2.1. Text Preprocessing

Teks yang akan dikelompokkan menggunakan *clustering* akan diproses terlebih dahulu untuk meningkatkan akurasi hasil *clustering*. Pada tahap ini, terdapat 4 proses yaitu *case folding*, *tokenization*, *filtering*, dan *stemming*.

a. Case folding

Menghapus karakter yang bukan merupakan huruf, seperti tanda baca dan angka. Kemudian mengubah seluruh teks menjadi *lowercase*.

b. Tokenization

Memenggal kata pada dokumen berdasarkan spasi dan tanda penghubung (-). Pada tahap ini dokumen yang awalnya merupakan satu string panjang akan dipenggal pada setiap kata, sehingga menjadi banyak string.

c. *Filtering*

Filtering adalah proses menghilangkan kata yang tidak signifikan. Kata-kata yang dihilangkan dalam proses *filtering* yaitu *stopwords*. *Stopwords* adalah kata-kata yang non-deskriptif, seperti “yang”, “di”, “dan”, “dari”, dan seterusnya. Dengan menghilangkan *stopwords*, ukuran index dan waktu pemrosesan dapat dikurangi. Proses *filtering* akan dilakukan menggunakan *library sastrawi stopwords* yang dikembangkan oleh peneliti menggunakan algoritma Nazief dan Adriani.

d. *Stemming*

Stemming adalah menghilangkan imbuhan yang ada pada kata sehingga kata tersebut kembali menjadi bentuk kata dasarnya. *Stemming* pada Bahasa Indonesia berbeda dengan *stemming* pada Bahasa Inggris. Pada Bahasa Inggris, mengembalikan kata ke bentuk kata dasarnya dapat dilakukan dengan menghapus *suffix* dari kata tersebut. Pada Bahasa Indonesia, imbuhan yang ditambahkan ada kata lebih bervariasi, sehingga untuk mengembalikan kata ke kata dasar diperlukan juga menghilangkan *prefix*. Kemudian ada kata-kata yang berubah bentuk ketika diberikan imbuhan. *Stemming*, dalam penelitian ini, akan dilakukan dengan menggunakan *library sastrawi stemming* yang dikembangkan oleh peneliti menggunakan algoritma Nazief dan Adriani.

Setelah *text preprocessing* selesai, maka akan dihasilkan *bag-of words* untuk diproses pada tahap selanjutnya.

3.2.2. *Term Weighting*

Setelah memperoleh *bag-of-words*, proses selanjutnya adalah *term weighting*. Pada tahap ini, setiap kata dihitung nilai bobotnya menggunakan *tf-idf*. Faktor Tf dan Idf ini yang berperan untuk memperbaiki nilai *recall* dan *precision*. Setelah proses ini selesai, akan dihasilkan *term-weight-matrix*. *Term-weight-matrix* berisi bobot setiap kata yang ada pada teks.

3.2.3. *Feature Selection*

Untuk mengurangi dimensi tersebut maka perlu dilakukan *feature selection*. Metode yang digunakan adalah *document frequency feature selection* karena merupakan metode *feature selection* yang paling sederhana dengan waktu komputasi yang rendah. Pada proses ini akan dilakukan perhitungan jumlah dokumen yang mengandung kata tertentu. Kemudian akan

ditentukan *threshold* untuk menyeleksi kata yang signifikan. Hasil dari proses ini adalah *term-weight-matrix* yang memenuhi batas *threshold*.

3.2.4. Clustering

Proses *clustering* terdiri dari dua proses. Proses pertama adalah proses *hierarchical clustering*. Metode yang digunakan adalah *agglomerative hierarchical clustering* dimana sebelum dilakukan pengelompokan, setiap data yang ada akan dianggap sebagai *cluster*. Apabila terdapat jumlah data sebanyak n , dan k dianggap sebagai jumlah *cluster*, sehingga besarnya n adalah samadengan k ($n = k$). Selanjutnya, penelitian ini menggunakan *Euclidean Distance Space* untuk menghitung jarak antar *cluster* berdasarkan jarak rata-rata antar objek. Berdasarkan dari hasil perhitungan tersebut pilih jarak yang paling minimal kemudian gabungkan, maka besarnya n adalah $n-1$ ($n = n - 1$). Jarak *cluster* akan di-*update* ketika 2 (dua) *cluster* digabungkan.. Setelah proses ini selesai, maka akan dihasilkan sebuah dendogram, dendogram menggambarkan proses penggabungan *cluster* sehingga menjadi *cluster* yang lebih tinggi.

K-means clustering akan dijalankan dengan hasil dari *hierarchical clustering* sebagai titik awal. Metode *k-means* akan mengoptimalkan posisi *centroid* dengan melakukan hitungan berulang pada *centroid* dari tiap *cluster*. Penghitungan ini akan terus berlangsung hingga nilai *centroid* stabil atau batas iterasi tercapai. Setelah *k-means* mencapai *centroid* yang stabil, maka nilai *centroid* dianggap sudah akurat.

3.3. Tahap Analisa

Tahap ini terdiri dari 2 bagian, yaitu:

a. Intra Cluster Similarity dan Inter Cluster Similarity

Pada tahap ini dilakukan analisis hasil *clustering* dengan cara menghitung *average intra similarity* dan *average inter similarity* dari setiap *cluster*. Dengan mengamati nilai *similarity* tersebut, dapat dianalisa kualitas dari *cluster* yang terbentuk pada tahap sebelumnya. Nilai *average intra similarity* mencerminkan similaritas dari setiap dokumen yang ada pada satu *cluster*, bila nilai *average intra similarity* tinggi, maka *cluster* dianggap baik karena setiap dokumen yang ada dalam *cluster* memiliki *similarity* yang tinggi. Nilai *average inter similarity* mencerminkan jarak antara satu *cluster* dengan *cluster* lainnya, semakin kecil nilai *average inter similarity*, maka semakin baik pula *cluster* tersebut karena *cluster* tersebut terpisah dari *cluster* lainnya.

b. Shilouette Coefficient

Pada tahap ini dilakukan analisis hasil clustering menggunakan metode Shilouette Coefficient. Metode ini dapat menggambarkan mana dokumen mana yang berada di cluster yang benar dan mana dokumen yang berada di kluster yang salah. Dokumen yang dianggap berada di *cluster* yang benar akan memiliki nilai *shilouette coefficient* lebih dari 0, sedangkan dokumen yang dianggap tidak berada di cluster yang benar akan memiliki nilai *shilouette coefficient* kurang dari 0.

3.4. Tahap Pengambilan Kesimpulan

Tahap pengambilan kesimpulan dilakukan dengan mengambil kesimpulan dari hasil *clustering* data. Hasil analisa juga akan disimpulkan dalam tahap ini. Setelah kesimpulan diambil, akan dituliskan saran mengenai *clustering* berita *online*.

