BAB III

METODOLOGI PENELITIAN

3.1 Objek Penelitian

Obyek pada penelitian ini adalah *tweet* yang diperoleh dari Twitter resmi kedua pasangan calon presiden 2019 dengan rentang waktu pengambilan *tweet* antara 1 Januari 2019 sampai 17 April 2019 yang sesuai dengan peraturan jadwal kampanye milik KPU. Nama Twitter resmi masing-masing pasangan calon adalah sebagai berikut:

1. @jokowi (Joko Widodo)
2. @KHMarufAmin_ (K.H. Ma'ruf Amin)
3. @prabowo (Prabowo Subianto)
4. @sandiuno (Sandiaga Salahuddin Uno)

Tweet-Tweet tersebut nantinya dikelompokkan menggunakan algoritma Improved K-Means kemudian dicocokkan sesuai misi masing-masing pasangan calon pemilihan presiden tahun 2019. Misi masing-masing pasangan calon dapat dilihat pada Tabel 3.1. Setelah misi masing-masing pasangan calon cocok, kelompok misi tersebut kemudian dipisah-pisah sesuai waktu debat pilpres 2019. Waktu debat capres 2019 dapat dilihat pada Tabel 3.3.

Tabel 3.1 Misi Masing-Masing Pasangan calon

No.	Misi Pasangan calon 1	Misi Pasangan calon 2
1	Peningkatan kualitas manusia Indonesia.	Membangun perekonomian nasional yang adil, makmur, berkualitas, dan berwawasan lingkungan dengan mengutamakan kepentingan rakyat Indonesia melalui jalan politik-ekonomi sesuai Pasal 33 dan 34 UUD Negara Republik Indonesia Tahun 1945.
2	Struktur ekonomi yang produktif, mandiri, dan berdaya saing.	Membangun masyarakat Indonesia yang cerdas, sehat, berkualitas, produktif. dan berdaya saing dalam kehidupan yang aman, rukun, damai, dan bermartabat serta

No.	Misi Pasangan calon 1	Misi Pasangan calon 2
		terlindungi oleh jaminan sosial yang berkeadilan tanpa diskriminasi.
3	Pembangunan yang merata dan berkeadilan.	Membangun keadilan dibidang hukum yang tidak tebang pilih dan transparan, serta mewujudkan persatuan dan kesatuan bangsa Indonesia melalui jalan demokrasi yang berkualitas sesuai dengan Pancasila dan UUD Negara Republik Indonesia Tahun 1945.
4	Mencapai lingkungan hidup yang berkelanjutan.	Membangun kembali nilai-nilai luhur kepribadian bangsa untuk mewujudkan Indonesia yang adil, makmur, bermartabat, dan bersahabat, yang diberkati oleh Tuhan Yang Maha Esa.
5	Kemajuan budaya yang mencerminkan kepribadian bangsa.	Membangun sistem pertahanan dan keamanan nasional secara mandiri yang mampu menjaga keutuhan dan integritas wilayah Indonesia.
6	Penegakan sistem hukum yang bebas korupsi, bermartabat, dan terpercaya.	
7	Perlindungan bagi segenap bangsa dan memberikan rasa aman pada seluruh warga.	
8	Pengelolaan pemerintahan yang bersih, efektif, dan terpercaya.	
9	Sinergi pemerintah daerah dalam kerangka Negara Kesatuan.	

(Sumber: Iqbal, 2019; Setiawati, 2019)

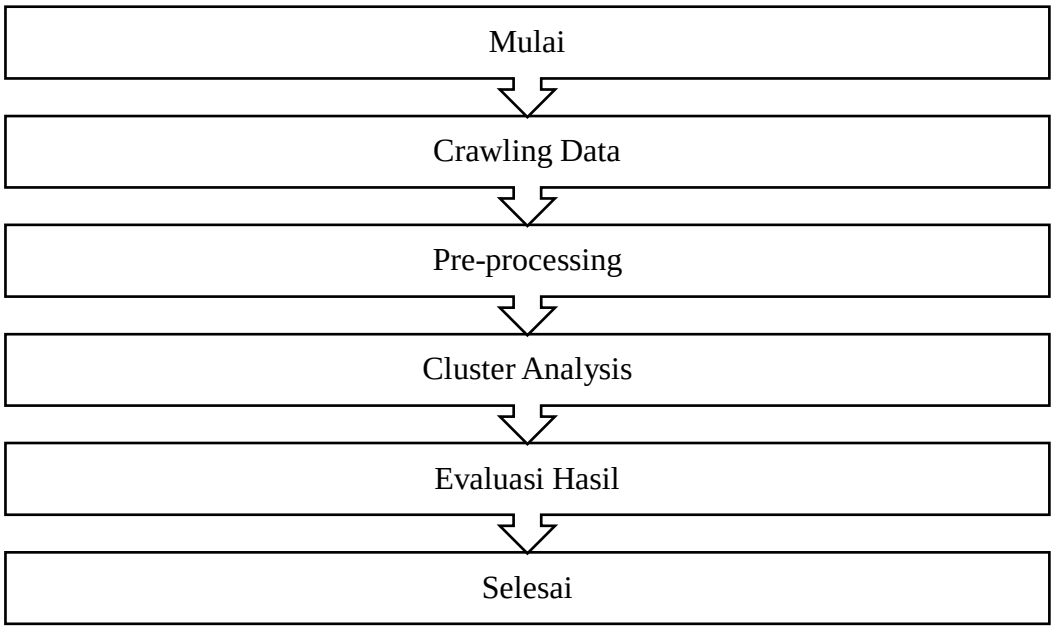
Tabel 3.2 Waktu Debat Calon Presiden dan Wakil Calon Presiden 2019

Urutan Debat	Tema	Waktu
Debat I	Hukum, HAM, Korupsi, dan Terorisme	17 Januari 2019
Debat II	Energi, pangan, infrastruktur, sumber daya alam, lingkungan hidup.	17 Februari 2019
Debat III	Pendidikan kesehatan, ketenagakerjaan, sosial dan budaya.	17 Maret 2019
Debat IV	Ideologi, pemerintahan keamanan serta hubungan internasional.	30 Maret 2019
Debat V	Ekonomi dan kesejahteraan sosial, keuangan, investasi, serta industri.	13 April 2019

(Sumber: Tim, 2019; Farisa, 2019)

3.2 Tahapan Penelitian

Tahapan pada penelitian ini secara garis besar terbagi menjadi 4 (empat) tahap yaitu *crawling data, pre-processing, cluster analysis,* dan evaluasi hasil. Tahapan penelitian ini dapat dilihat pada Gambar 3.1.



Gambar 3.1 Diagram Tahapan Penelitian

3.2.1 Crawling Data

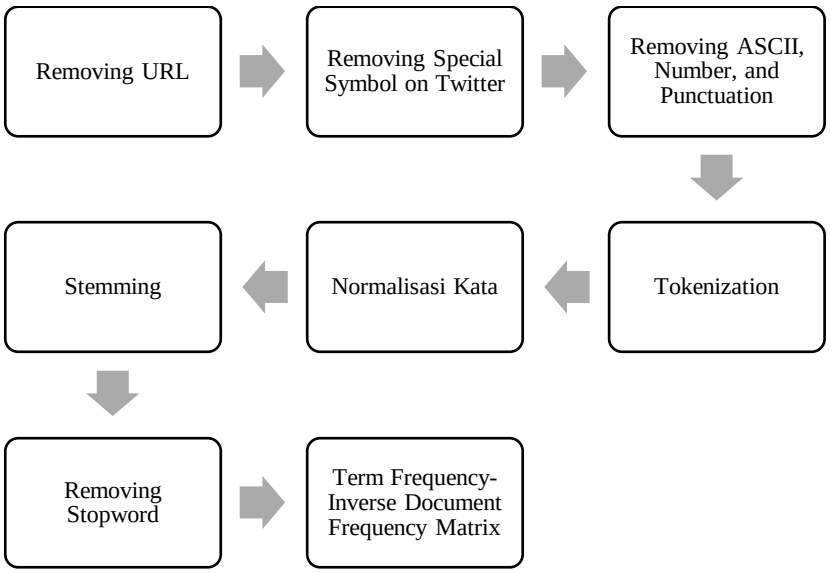
Melakukan pengambilan data tweet melalui aplikasi *Rapid Miner Studio*. Pengambilan data tweet dilakukan dengan menggunakan *Application Programming Interface(API)* Twitter Developer. Data hasil crawling tweet dapat dilihat pada Tabel 3.3.

Tabel 3.3 Tweet Hasil Crawling

No	Text
1	Terima kasih kepada seluruh rakyat yang telah menggunakan hak suara dengan sebaik-baiknya dalam Pemilu 2019 hari ini. Apa pun hasilnya, kita tetap bersatu dan bersaudara. Semoga Allah SWT meridhai segenap ikhtiar kita untuk kemajuan bangsa Indonesia -- rumah kita bersama. https://t.co/h9L49CMUXX
2	Saya telah menggunakan hak pilih saya sebagai warga negara Indonesia. Bagaimana dengan Anda? https://t.co/E98u2byrpG
3	Selamat memilih, Indonesia ???? https://t.co/mviv7roNtS
4	Aku penuhi panggilanMu, ya Allah, aku penuhi panggilanMu. Tidak ada sekutu bagiMu, aku penuhi panggilanMu. Sesungguhnya segala puji, nikmat dan kerajaan bagiMu. Tidak ada sekutu bagiMu. https://t.co/KbC5tr8Xlv

3.2.2 Pre-processing

Setelah mendapatkan dokumen, perlu dilakukan *pre-processing* untuk mengolah dokumen agar dapat meningkatkan efisiensi dan mempersingkat waktu saat dokumen diproses pada tahap selanjutnya. Hasil pada tahap *pre-processing* akan berupa matriks TF-IDF. *Pre-processing* dibagi menjadi beberapa tahap sesuai dengan Gambar 3.2.



Gambar 3.2 Tahapan *Pre-processing*
(Sumber: Hidayatullah, 2017, telah diolah)

3.2.2.1 *Removing Uniform Resource Locator(URL)*

Pesan Twitter biasanya berisi *URL* seperti misalnya http://t.co/dXXdCPih23 pada tweet prabowo tanggal 11 april 2019. URL dihapus karena *preprocessing* fokus pada kata-kata yang ada dalam tweet (Hidayatullah, 2017). Contoh *input* dan *output* tahapan ini ditampilkan pada Gambar 3.3.

-	**Nilai**
Input	"Terima kasih kepada seluruh rakyat yang telah menggunakan hak suara dengan sebaik-baiknya dalam Pemilu 2019 hari ini. Apa pun hasilnya, kita tetap bersatu dan bersaudara. Semoga Allah SWT meridhai segenap ikhtiar kita untuk kemajuan bangsa Indonesia -- rumah kita bersama. https://t.co/h9L49CMUXX"
Output	"Terima kasih kepada seluruh rakyat yang telah menggunakan hak suara dengan sebaik-baiknya dalam Pemilu 2019 hari ini. Apa pun hasilnya, kita tetap bersatu dan bersaudara. Semoga Allah SWT meridhai segenap ikhtiar kita untuk kemajuan bangsa Indonesia -- rumah kita bersama.”

Gambar 3.3 *Input* dan *Output* Tahap *Removing URL*

3.2.2.2 *Removing Special Symbols on Twitter*

Twitter memiliki simbol khusus pada tweetnya seperti hashtag (#), nama pengguna (@username), dan retweet (RT). Karakter-karakter ini akan diberi tindakan pada tahap ini. Username (@) dan retweet (RT) akan dihapus karena tidak memuat topik pembicaraan. Namun, simbol hashtag tidak dihapus seluruhnya melainkan dihapus lambang(#)nya saja karena kata dalam hashtag masih mengandung makna pada tweet tersebut (Hidayatullah, 2017). Contoh *Input* dan *Output* tahapan ini ditampilkan pada Gambar 3.4.

-	**Nilai**
Input	"@MRomahurmuziy @DPP_PPP @lukmansaifuddin Salam satu jempol dari saya dan @MRomahurmuziy ?? #01IndonesiaMaju"
Output	"Salam satu jempol dari saya dan 01IndonesiaMaju"

Gambar 3.4 *Input* dan *Output* tahapan *Removing Special Symbols*

3.2.2.3 *Removing Symbol ASCII, Numbers, and Punctuation*

Pesan Twitter biasanya berisi simbol, angka, dan tanda baca. Semua ini akan dihapus menggunakan operator *Process Documents* (Hidayatullah, 2017). Contoh *input* dan *output* tahapan ini ditampilkan pada Gambar 3.5.

-	Nilai
Input	"Terima kasih kepada seluruh rakyat yang telah menggunakan hak suara dengan sebaik-baiknya dalam Pemilu 2019 hari ini. Apa pun hasilnya, kita tetap bersatu dan bersaudara. Semoga Allah SWT meridhai segenap ikhtiar kita untuk kemajuan bangsa Indonesia -- rumah kita bersama.
Output	"Terima kasih kepada seluruh rakyat yang telah menggunakan hak suara dengan sebaik-baiknya dalam Pemilu 2019 hari ini. Apa pun hasilnya, kita tetap bersatu dan bersaudara. Semoga Allah SWT meridhai segenap ikhtiar kita untuk kemajuan bangsa Indonesia rumah kita bersama."

Gambar 3.5 *Input* dan *Output* tahapan *Removing Symbol ASCII*

3.2.2.4 *Tokenization*

Tokenization digunakan untuk memisahkan kata per kata yang ada pada dokumen tweet. Kata dalam dokumen perlu dipisah karena pemrosesan dokumen menggunakan perhitungan kemiripan berdasarkan kata yang dikandung. Pada tahap ini juga dilakukan pengubahan kata yang mengandung huruf besar menjadi huruf kecil. Hal ini dikarenakan bahasa pemrograman bersifat *case sensitive*. Contoh *tokenization* dapat dilihat pada Gambar 3.6.

-	Nilai
Input	Selamat memilih Indonesia
Output	Selamat | memilih | Indonesia

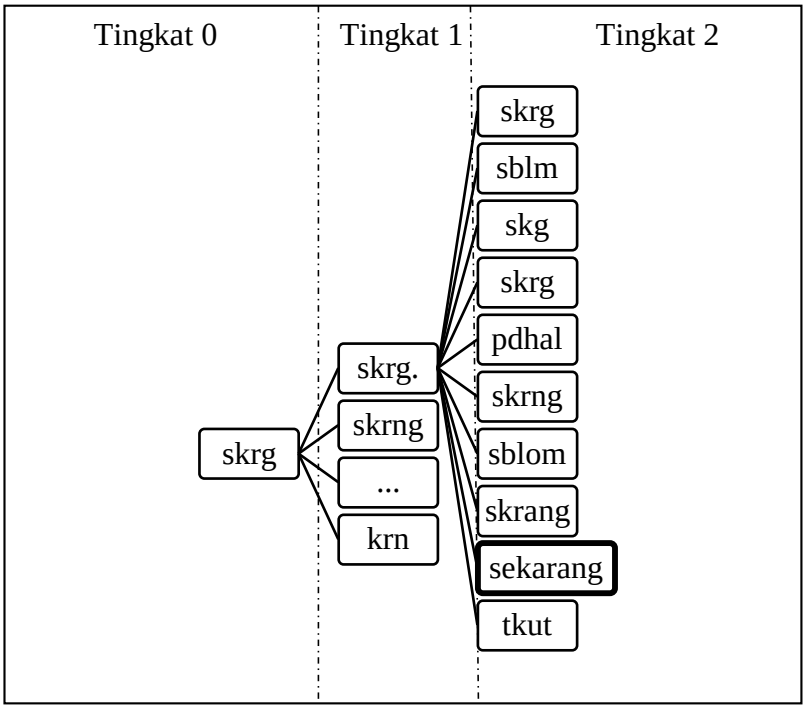
Gambar 3.6 *Input* dan *Output Tokenization*

3.2.2.5 *Normalisasi Kata*

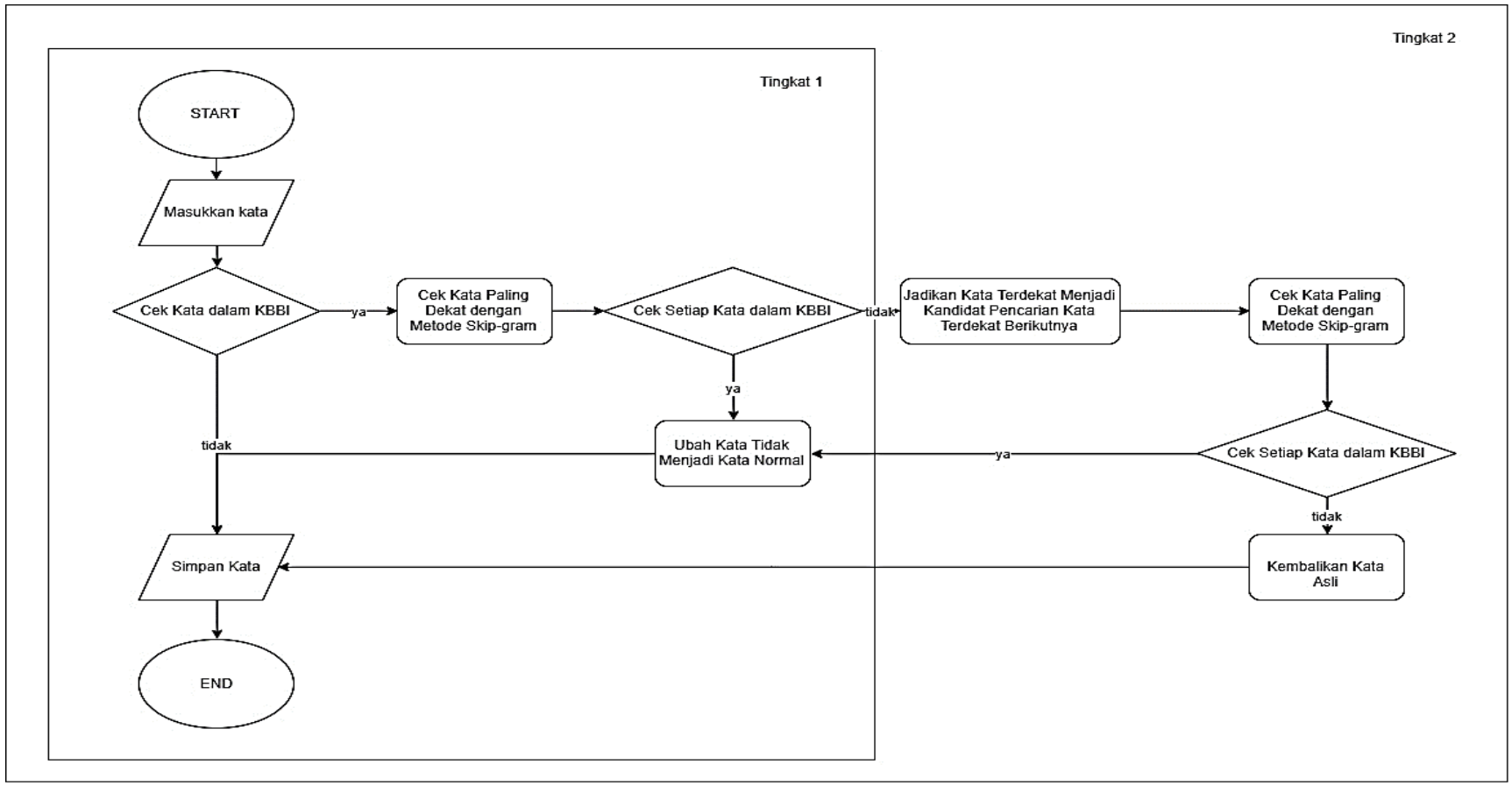
Teknik normalisasi digunakan guna menormalisasi bentuk kata yang tidak baku menjadi kata baku. Teknik normalisasi penelitian ini menggunakan metode Skip-Gram dengan *library fasttext* dan modul *gensim. FastText* menggunakan data *training* berasal dari https://commoncrawl.org/ dan Wikipedia.

Training data menggunakan metode CBOW dengan dimensi 300 dan kharakter n-gram berukuran 5. Hasil *Training* CBOW dapat diunduh di https://fasttext.cc/docs/en/crawl-vectors.html. Hasil *training* tersebut digunakan untuk memprediksi normalisasi kata dari kata tidak baku dengan metode Skip-gram sesuai penelitian thomas mikolov (Tomas Mikolov, 2013). Penelitian Edouard Grave telah melakukan penelitian tentang penggantian kata berdasarkan kelasnya menggunakan *library fasttext* dan modul *gensim* (Grave, Piotr, Guphta, joulin, & Mikolov, 2018).

Penelitian ini menormalisasi kata yang tidak normal menjadi kata-kata KBBI edisi ke-V dengan cara pengecekan secara online. Proses pencarian normalisasi kata dapat dilihat di gambar 3.8. Contoh *input* dan *output* tahapan normalisasi kata ditampilkan pada gambar 3.7.



Gambar 3.7 Ilustrasi Pencarian Kata Normalisasi



Gambar 3.8 *Flowchart* normalisasi

3.2.2.6 *Stemming Modifikasi Enhanced Confix Stripping Stemmer*

Teknik ini dilakukan untuk menghilangkan imbuhan pada suatu kata sehingga didapatkan kata dasar (*root word*) dari kata tersebut. *Stemming* dalam penelitian ini menggunakan *Stemming Nazief-Adriani* berbahasa Indonesia. Contoh *input* dan *output* tahapan *stemming* ditampilkan pada Gambar 3.9.

-	Nilai					
Input	hari-hari	yang	menyenangkan	bersama	jan	ethes
Output	hari	yang	senang	bersama	jan	ethes

Gambar 3.9 *Input* dan *Output* Tahap *Stemming*

3.2.2.7 *Removing Stopwords*

Setelah mendapatkan kata dalam dokumen, diperlukan suatu metode untuk menghilangkan kata-kata yang tidak memiliki makna atau kata yang tidak penting yang terdapat di dokumen (Stopword). Kata-kata tersebut perlu dihilangkan untuk bisa mendapatkan hasil yang lebih maksimal. Contoh *Input* dan *Output* Tahapan *Removing Stopwords* ditampilkan pada Gambar 3.10.

-	Nilai					
Input	hari	yang	senang	bersama	Jan	ethes
Output	hari		senang		Jan	ethes

Gambar 3.10 *Input* dan *Output* Tahap *Removing Stopwords*

3.2.2.8 *Term Frequency-Inverse Document Frequency Matrix*

Setelah dokumen tweet selesai dilakukan dibersihkan maka dokumen tersebut akan dilakukan pembobotan kata. Pembobotan diberikan pada setiap kata berdasarkan frekuensi kemunculan kata pada dokumennya. Pembobotan perlu dilakukan agar dapat diketahui kata yang mempunyai kepentingan yang tinggi dan rendah.

Hal ini perlu dilakukan karena pada tahap selanjutnya dokumen tweet akan dibutuhkan untuk mencari kelompok tweet. Sehingga tweet dengan topik yang sama dikelompokkan. Contoh Tabel TF-IDF ditampilkan pada Tabel 3.4.

Tabel 3.4 Hasil Perhitungan TF-IDF

ID	Aamiin	Abad	…	zuhri
1	0	0	…	0
2	0	0.12505	…	0
3	⋮	⋮	⋱	⋮
4	0	0	…	0

3.2.3 Cluster Analysis

Setelah tweet dihitung bobotnya menggunakan TF-IDF. Tweet akan dianalisis dengan jenis metode cluster analysis. Cluster analysis merupakan tahap pencarian kelompok tweet berdasarkan misi yang dikandung dalam setiap tweet. Algoritma yang digunakan adalah improved K-Means. Improved K-Means memiliki perbedaan di tahap inisiasi *centroid*. Diagram algoritma Improved K-Means secara garis besar dapat dilihat pada Gambar 3.11.

Gambar 3.11 Tahapan *Improved K-Means*

3.2.3.1 *Low Density Removal*

Improved K-Means berbeda dengan K-Means++ dan Tradisional K-Means. Improved K-Means menghitung kepadatan(*density*) setiap tweet dengan tweet lain untuk menentukan tweet tersebut cocok dikelompokkan atau tweet tersebut merupakan tweet yang sendirian(*isolated*). Penentuan

isolated tweet dibantu dengan ahli karena tweet terisolasi tersebut masih memungkinkan mengandung misi presiden. Berikut adalah tahapan penghapusan tweet:

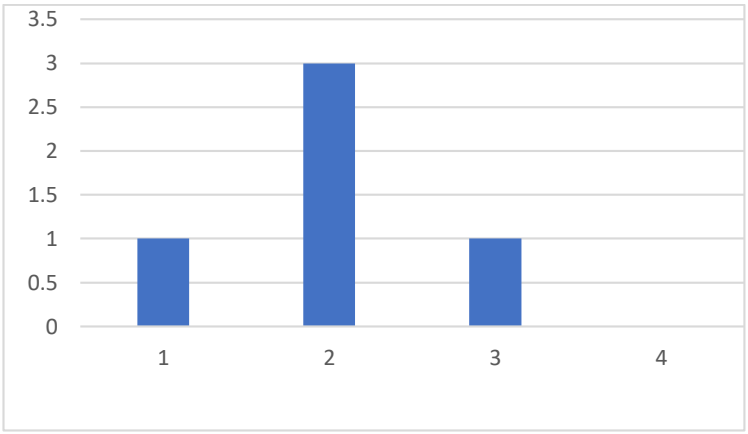
1. Hitung rata-rata jarak antara dua tweet dalam dataset D. Perhitungan ini digunakan untuk dasar perhitungan kepadatan/*distance* tiap tweet. Perhitungan nilai *euclidean* tweet pertama dan kedua yaitu $\sqrt{(0-0)^2+\cdots+(0.140941-0)^2+\cdots+(0-0)^2}$ sehingga didapatkan nilai 1.3129604373814387. Tabel perhitungan *Cosine Similarity* ditampilkan pada Tabel 3.16. Kemudian $MeanDist$ dicari dengan cara $\frac{1.3129604373814387+1.3944223235415487+\cdots+1.41421356237}{C_4^2}$ dan menghasilkan nilai 1.344721402.

   Tabel 3.5 Hasil Perhitungan Jarak *Cosine Similarity*

Pasangan		*Distance*
1	2	1.3129604373814387
1	3	1.3944223235415487
1	4	1.4057709128868021
2	3	1.1267476111595798
2	4	1.4142135623730947
3	4	1.414213562373095

2. Menghitung kepadatan/*density* dataset. Perhitungan kepadatan menggunakan rumus 2.8. Setelah itu, hitung *MeanDens* dengan cara $\frac{1+3+1}{4}$ yaitu 1.
3. Pindahkan dataset yang memiliki kepadatan tinggi ke koleksi A dan hapus *dataset* yang memiliki kepadatan rendah sesuai asumsi pakar dengan rumus $Dens(x_i) < \alpha * MeanDens(D)$ dengan parameter $0 < \alpha < 1$ (Xiong, 2016). Ahli akan menentukan α berdasarkan analisa tweet

dan menentukan tweet yang tidak memiliki korelasi terhadap misi presiden.



Gambar 3.12 Visualisasi *Density*

Pada contoh perhitungan ini, data yang dihapus adalah *dataset* ke-4 karena merupakan tweet terisolasi. Koleksi A dan koleksi B dapat dilihat pada Tabel 3.6.

Tabel 3.6 Koleksi A dan B dari Dataset *D*

Index *Object* Koleksi A	**Index *Object* Koleksi B**
1	
2	
3	

3.2.3.2 Inisiasi *Centroid*

Inisiasi *Centroid* merupakan proses pencarian titik *centroid awal* dalam proses iterasi K-Means. Terdapat 3 (tiga) tahap dalam pencarian titik inisiasi *centroid*:

1. Pindahkan tweet yang memiliki kepadatan tertinggi ke koleksi B. Koleksi B ini merupakan kumpulan *centroid-centroid* yang akan nantinya akan digunakan untuk iterasi K-Means. Koleksi A dan B pada tahap ini dapat dilihat pada Tabel 3.7.

Tabel 3.7 Koleksi A dan B inisiasi *centroid* pertama

Index *Object* Koleksi A	**Index *Object* Koleksi B**
1	2
3	

2. Pindah koleksi A yang memiliki jarak terjauh dengan koleksi B ke koleksi B.
3. Ulangi tahap 5 sampai terdapat k tweet. Setelah didapatkan k data dalam koleksi B, koleksi B akan dijadikan centroid awal untuk $KMeans$.

Tabel 3.8 *Centroid* awal untuk k = 2

No	***Index* Objek *Centroid* Awal**
1	2
2	1

3.2.3.3 *K-Means*

Improved K-Means menggunakan K-Means tradisional untuk Iterasinya. K-Means ini digunakan untuk mencari *centroid* optimal. K-Means memerlukan banyak klaster sebagai target kelas yang akan dibuat. Pada penelitian ini menggunakan teknik pencarian berdasarkan nilai validasi optimal yang serupa dengan penelitian Chunhui Yuan (Yuan & Yang , 2019). Dua validasi yang dipilih *adalah Sum of Square distance to centroid* dan *Silhouette Coefficient* dikarenakan metode *Sum of Square distance to centroid* memiliki tingkat kompleksitas algoritma yang sederhana dan *Silhouette Coefficient* memiliki batas ukuran validasi yang jelas. metode *Sum of Square distance to centroid* hanya menghitung jarak rata-rata setiap *dataset* ke *centroid*-nya sedangkan *Silhouette Coefficient* memerlukan beberapa tahap perhitungan. Berikut ini adalah langkah-langkah perhitungan *Silhouette Coefficient*:

1. Menghitung rata-rata jarak i ke titik dalam satu cluster untuk mencari $b(i)$
2. Menghitung minimal rata-rata jarak i ke titik dalam cluster lain untuk mencari $a(i)$.
3. Menghitung $s(i)$ dengan mengguaan persamaan kemudian merata-rata $s(i)$.

Tabel 3.9 Hasil Perhitungan S*ilhouette Coefficient*

i	***a(i)***	***b(i)***	***s(i)***
1	1.394422	1.31296	-0.05842
2	1.126748	1.31296	0.141827
3	1.823959	0	0
***Mean* s(i)**			0.027802

Setelah nilai *k* diketahui, langkah berikutnya adalah melakukan iterasi K-Means. Iterasi K-Means bertujuan untuk mencari pusat klaster yang seharusnya. Terdapat 4 (empat) iterasi K-Means:

1. Menghitung jarak *Cosine Similarity* dengan rumus 2.5 antara setiap dataset dan setiap centroid. Kemudian memasukkan dataset ke dalam suatu cluster berdasarkan jarak terdekat dataset.
2. Menghitung rata-rata nilai term dari objek pada suatu cluster kemudian rata-rata nilai term tersebut dijadikan *centroid* baru.
3. Mengubah *centroid* lama menjadi rata-rata nilai term dari objek suatu cluster secara keseluruhan term jika terdapat perbedaan nilai term *centroid* baru dengan nilai *centroid lam*a.
4. Ulangi dari langkah pertama sampai tidak ada perubahan *centroid* (*convergence*).