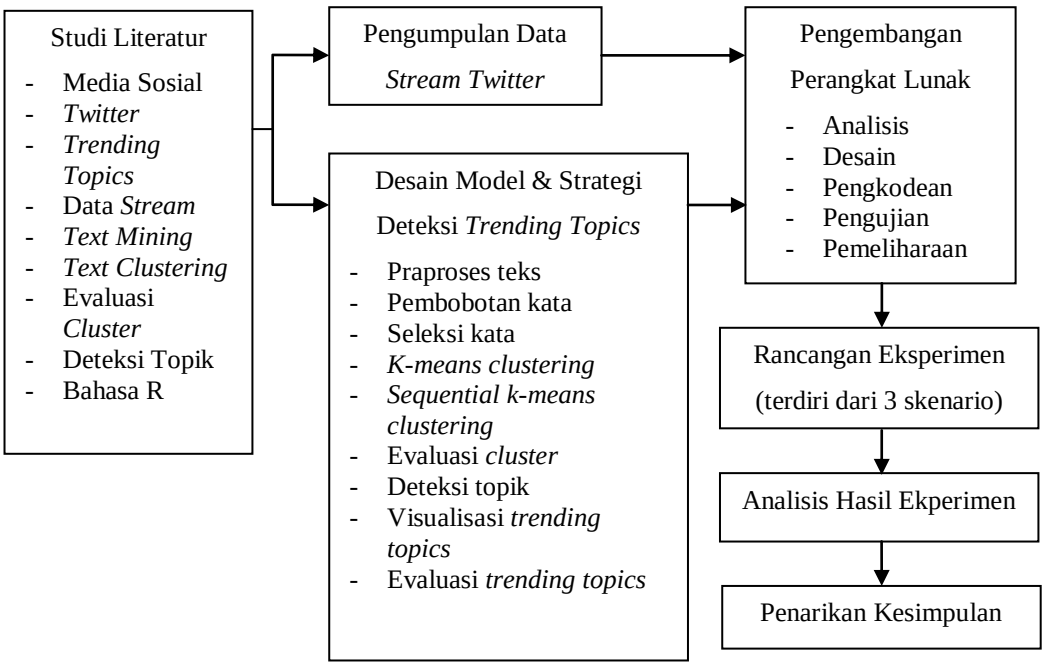
BAB III

METODOLOGI PENELITIAN

Pada bab 3 ini akan dijelaskan mengenai metodologi penelitian yang diusulkan untuk mendeteksi *trending topics* dari data *stream twitter* dengan pendekatan *sequential k-means*. Pemaparan dan penjelasan mengenai desain penelitian, pengembanangan perangkat lunak, serta alat dan bahan yang digunakan akan dijelaskan pada bab ini.

3.1 Desain Penelitian

Desain penelitian merupakan tahapan yang akan dilakukan penulis dalam melakukan penelitian. Tahapan penelitian yang dilakukan akan diilustrasikan pada gambar 3.1.



Gambar 3. 1 Desain Penelitian

Gambar 3.1 menunjukan desain penelitian yang telah dirancang oleh penulis dan terdiri dari tujuh tahap. Setiap tahap akan dijelaskan pada sub bab berikut secara berurutan.

3.1.1 Studi Literatur

Pada tahap ini, penulis melakukan studi literatur mengenai penelitian deteksi *trending topics twitter* untuk menemukan latar belakang dan rumusan masalah pada penelitian. Dengan melakukan studi literatur, penulis dapat menentukan tujuan dan metodologi dalam penelitian ini. Penulis melakukan studi literatur mengenai teori-teori yang dikaji yaitu media sosial *Twitter*, *Trending Topics*, Data *Stream*, *Text Mining*, *Document Clustering*, Evaluasi *Cluster*, Deteksi Topik serta Bahasa R. Teori-teori tersebut sangat penting untuk dipelajari karena semua teori tersebut menunjang keberhasilan penelitian ini.

3.1.2 Pengumpulan *Data Stream Twitter*

Data yang digunakan dalam penelitian ini diambil dari situs jejaring sosial *Twitter* yaitu *tweet*. *Tweet* yang digunakan merupakan *tweet* berbahasa Inggris yang berlokasi di kota New York. Lokasi ini dipilih karena menghasilkan volume *tweet* yang tinggi dengan konsisten. *Tweet* terbaru diambil secara *real-time* dari *Twitter Streaming* API, sehingga proses pengumpulan data *stream twitter* ini dilakukan secara *online* dan berkala. Teknik pengumpulan data dilakukan secara *streaming* karena penelitian ini membutuhkan data *real-time* untuk menghasilkan *trending topics* yang *real-time* pula. T*weet* diambil secara berkala dan ditentukan selama 60 detik sehingga akan menghasilkan satu blok data dengan jumlah data yang tidak terlalu banyak namun tidak sedikit. Waktu pengumpulan data dibatasi karena jumlah *tweet* yang tidak terbatas, dan proses analisis data dapat dilakukan dengan cepat.

3.1.3 Desain Model dan Strategi Deteksi *Trending Topics*

Pada penelitian ini akan dirancang sebuah model dalam mendeteksi *trending topics twitter* yang terdiri dari beberapa tahap dimana setiap tahap disertai dengan strategi yang akan membedakan penelitian ini dengan penelitian lain yang serupa. Model dan strategi penentuan *trending topics twitter* akan dijelaskan secara rinci pada bab 4. Setelah data terkumpul, proses yang akan dilakukan selanjutnya secara garis besar adalah sebagai berikut:

a) Praproses Teks

Praproses teks merupakan tahap awal pengolahan data teks sebelum membangun sebuah model. Pada proses ini dilakukan penghapusan kata dan hal-hal yang tidak diperlukan. Praproses teks pada penelitian ini disesuaikan dengan data masukan yaitu *tweet* berbahasa Inggris.

b) Pembobotan Kata

Pada tahap ini setiap dokumen teks hasil praproses teks akan dikonversi menjadi bentuk vektor. Setiap kata pada dokumen akan diberi bobot dengan menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Hasil dari tahap ini adalah matriks data yang berisi nilai TF-IDF dengan atribut berupa kata-kata penyusun dokumen.

c) Seleksi Kata

Pada tahap ini akan dilakukan seleksi kata yang merupakan atribut pada matriks data. Kata yang dipilih pada penelitian ini adalah kata benda yang akan menjadi kandidat nama topik. Hal ini dipertimbangkan karena pada umumnya topik adalah bentuk kata benda.

d) *K-Means Clustering*

Jika data masukan sudah terstruktur, tahap selanjutnya adalah analisis data dengan menggunakan teknik *clustering*. Pada penelitian ini, pengambilan data dilakukan secara berkala maka analisis data pun dilakukan secara berkala. Untuk *clustering* data blok pertama atau pada iterasi pertama, metode *clustering* yang digunakan adalah *k-means* klasik dimana hasil *clustering* akan digunakan untuk inisialisasi *cluster* pada iterasi selanjutnya.

e) *Sequential K-Means Clustering*

Untuk *clustering* blok data baru atau pada iterasi selanjutnya, metode *clustering* yang digunakan adalah *k-means* versi *online* atau disebut dengan *Sequential K-Means* dimana inputan datang secara inkremental dan *cluster centers* dapat diperbaharui. *Sequential k-means*

yang digunakan pada penelitian ini adalah variasi *sequential k-means* yang sudah dijelaskan pada bab 2.

f) Evaluasi *Cluster*

Evaluasi terhadap *cluster* yang terbentuk dilakukan untuk menguji kualitas *cluster*. Metode yang digunakan yaitu *Dunn Index*. Nilai evaluasi antara 0 sampai ∞, evaluasi *cluster* dikatakan baik jika nilai dari *dunn index* yang didapat tinggi.

g) Deteksi Topik

Setelah *cluster* terbentuk, langkah selanjutnya yaitu menentukan kata kunci atau label untuk setiap *cluster* dengan cara mencari bobot kata atau atribut terbesar pada setiap *cluster center*. Label ini akan menjadi perwakilan topik untuk setiap *cluster*. Setelah itu, setiap topik diurutkan berdasarkan jumlah anggota terbanyak dan dipilih lima topik teratas yang akan menjadi *trending topics*.

h) Visualisasi *Trending Topic*

Trending topic yang telah diperoleh kemudian divisualisasikan ke dalam bentuk histogram, sehingga *trending topics* dapat dilihat dan dipahami dengan mudah.

i) Evaluasi *Trending Topics*

Pada penelitian ini, *trending topics* adalah lima topik teratas berdasarkan jumlah anggota terbanyak. Topik yang akan menjadi tren adalah topik yang terus dibicarakan oleh pengguna *twitter* sehingga kata yang menjadi perwakilan topik akan terus muncul pada aliran data terbaru. Maka dari itu, topik yang sudah menjadi salah satu dalam jajaran *trending topics* perlu dievaluasi ulang untuk mengetahui apakah topik tersebut masih dibicarakan atau tidak. Cara untuk evaluasi *trending topic* yaitu dengan cara mengecek popularitas *trending topics* setiap 3 jam sekali dengan cara menghitung selisih jumlah *tweet* dalam 3 jam terakhir untuk setiap *cluster*. Jika ada penambahan jumlah minimal 100 *tweets,* maka topik atau *cluster* akan dipertahankan, jika tidak maka *cluster* akan dihapus.

3.1.4 Pengembangan Perangkat Lunak

Setelah model dan strategi deteksi *trending topics* selesai dirancang, langkah selanjutnya yaitu mengimplementasikan model kedalam kode program dan setiap proses pada model dijadikan fungsi pada perangkat lunak. Bahasa pemograman yang digunakan adalah bahasa R karena tersedianya *library* atau *package* yang mendukung proses-proses pada model deteksi *trending topics twitter*. Perangkat lunak dibangun dengan menggunakan model *Waterfall* yang terdiri dari analisis, desain, pengkodean, pengujian, dan pemeliharaan.

3.1.5 Rancangan Eksperimen

Rancangan eksperimen dilakukan agar eksperimen yang akan dilakukan lebih tearah dan terencana. Eksperimen akan dilakukan dalam tiga skenario. Pada setiap skenario, pengambilan data dilakukan di waktu yang berbeda sehingga data yang digunakan berbeda. Kemudian nilai parameter pada fungsi dibedakan pula untuk setiap skenario.

3.1.6 Analisis Hasil

Analisis hasil dilakukan setelah eksperimen pada tiga skenario selesai dilakukan. Informasi yang didapat dari tiga skenario akan dianalisis dan dicari kondisi apa saja yang terjadi dan apa saja perbedaan antara skenario satu dengan yang lainnya beserta dengan kelebihan dan kekurangannya. Hasil dari analisis akan dijabarkan pada bab 4 secara lengkap dan terstruktur.

3.1.7 Penarikan Kesimpulan

Setelah melaksanakan seluruh rangkaian kegiatan dalam penelitian, penulis perlu untuk menyimpulkan hasil yang didapatkan, juga menyampaikan keunggulan dan kelemahan penelitian. Kesimpulan yang disampaikan harus sejalan dengan tujuan dari penelitian dan menjawab rumusan masalah yang telah disampaikan pada bab pendahuluan. Selain itu, penulis juga perlu memberikan saran bagi peneliti selanjutnya yang akan membahas masalah yang berhubungan dengan penelitian ini agar penelitian yang dilakukan kedepannya dapat dilaksanakan dengan lebih baik.

3.2 Metode Pengembangan Perangkat Lunak

Metode pengembangan perangkat lunak yang digunakan pada penelitian ini adalah model *Sequential Linear* atau sering disebut juga sebagai *Waterfall* (Pressman, 2010). Model ini pertama kali yang diperkenalkan oleh Winston Royce sekitar tahun 1970, yaitu model yang bersifat sistematis, dimana setiap tahapan sistem akan dikerjakan secara berurutan mulai dari analisis, desain, implementasi, pengujian dan perawatan. Adapun proses-proses tersebut secara detail adalah sebagai berikut:

1. Analisis

   Pada tahap ini dilakukan analisis untuk menentukan kebutuhan, batasan, dan tujuan (*goal*) dari perangkat lunak sesuai yang diinginkan. Untuk memahami sifat program yang akan dibangun, pembuat perangkat lunak (analis) harus memahami domain informasi untuk perangkat lunak, fungsi yang diperlukan, perilaku dan kinerja perangkat lunak.

2. Desain

   Tahap ini difokuskan pada proses perancangan perangkat keras maupun perangkat lunak yang dilibatkan untuk menunjang sistem yang akan dibangun. Proses perancangan melibatkan identifikasi dan menggambarkan dasar sistem serta hubungan satu sama lain.

3. Implementasi

   Pada tahap ini dilakukan penerjemahan hasil desain ke dalam kode program yang bisa dibaca mesin komputer. Jika desain dilakukan secara rinci, pembuatan kode dapat dilakukan secara mekanis.

4. Pengujian

   Setelah kode program dihasilkan, tahap berikutnya adalah pengujian. Pengujian ini difokuskan pada internal perangkat lunak untuk memastikan bahwa setiap kode program diuji, dan pengujian pada fungsionalitas perangkat lunak untuk memastikan tidak terjadi eror, serta memastikan keluaran yang dihasilkan perangkat lunak sesuai dengan target keluaran yang diharapkan.

5. Perawatan

Pemeliharaan dilakukan untuk melakukan perbaikan pada perangkat lunak itu sendiri, maupun dengan tujuan untuk melakukan adaptasi terhadap perubahan kebutuhan pengguna.

3.3 Alat dan Bahan Penelitian

Alat yang digunakan pada penelitian ini adalah seperangkat komputer yang dilengkapi oleh perangkat keras dan perangkat lunak pendukung yang terhubung dengan internet. Sedangkan bahan yang digunakan adalah teks *tweet* yang diambil dari *Twitter Streaming API*.

3.3.1 Alat Penelitian

Dalam penelitian ini perangkat keras yang digunakan memiliki spesifikasi sebagai berikut:

1. Notebook dengan spesifikasi:
   - Processor Intel Atom N260
   - Memori 2 GB RAM
2. Modem

Sedangkan perangkat lunak yang digunakan yaitu:

1. Windows 7 Professional
2. RStudio
3. R Console versi 3.3.2

3.3.2 Bahan Penelitian

Bahan penelitian yang digunakan adalah kumpulan *tweets real-time* yang diperoleh dari *Twitter Streaming* API. *Tweet* yang digunakan merupakan *tweet* berbahasa Inggris yang berlokasi di kota New York. Data *tweet* yang didapat dari *Twitter Streaming* API adalah file dengan format *JavaScript Object Notation* (JSON). *Feature* atau atribut *tweet* yang digunakan dalam penelitian ini hanyalah atribut "*text*". Selain itu, peneliti juga menggunakan berbagai bahan sebagai penunjang dalam penelitian yaitu jurnal, buku, karya ilmiah, e-book, serta tulisan lain yang membantu peneliti dalam memahami aplikasi yang dibuat.