BAB III
METODOLOGI PENELITIAN

Pada bab ini menjelaskan langkah-langkah yang akan ditempuh dalam menyusun skripsi, yaitu perancangan, implementasi, dan pengujian sistem yang akan dibuat. Sistem ini dirancang untuk menghasilkan aplikasi yang mampu mengkategorikan pesan singkat berbahasa Indonesia pada jejaring sosial Twitter dengan metode klasifikasi *naive bayes*. Kesimpulan dan saran disertakan sebagai catatan atas aplikasi dan kemungkinan arah untuk pengembangan aplikasi selanjutnya. Urutan langkah-langkah yang akan dilakukan pada penelitian digambarkan pada blok diagram sebagai berikut :

3.1 Studi Literatur

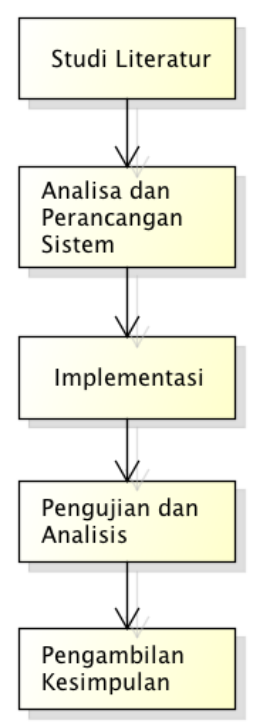
Langkah ini dilakukan untuk mendalami konsep dan teori-teori yang akan diterapkan pada "Pengkategorian Pesan Singkat Berbahasa Indonesia Pada Jejaring Sosial Twitter dengan Metode Klasifikasi Naive Bayes". Pada langkah ini merupakan proses awal yang sangat penting dalam implementasi sebuah penelitian karena merupakan pengetahuan dasar yang digunakan. Metode yang akan dipelajari dalam studi literatur yaitu :

1. Metode *Stemming*
2. Metode *Naive Bayes Classifier*

3.2 Analisis Kebutuhan Sistem

Pada bagian ini dilakukan analisa terhadap kebutuhan-kebutuhan yang dibutuhkan dalam merancang sistem. Kebutuhan sistem yang akan disusun pada bagian ini meliputi perangkat keras dan perangkat lunak. Secara umum, sistem memiliki fungsi untuk mengolah pesan singkat pada jejaring sosial Twitter berbahasa Indonesia dengan proses-proses terkait dengan pengolahan teks, kemudian dengan menggunakan metode *naive bayes* sistem akan melakukan pengkategorian terhadap

dokumen sehingga pesan singkat sudah dalam kondisi memiliki kategori tersendiri. Gambar 3.1 berikut akan menjelaskan skema aliran data dalam sistem.



Gambar 3.1 Diagram Blok Penelitian

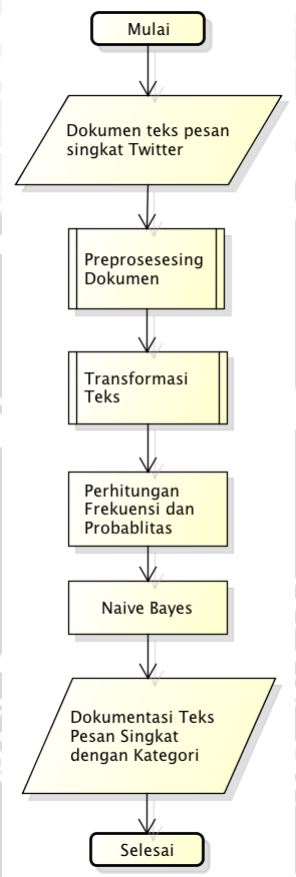
Aplikasi yang akan dibuat adalah aplikasi untuk mengkategorikan pesan singkat pada jejaring sosial Twitter menggunakan teknik klasifikasi *naive bayes.* Aplikasi yang dibuat akan terintegrasi secara langsung dengan jejaring sosial Twitter menggunakan fasilitas *Streaming* API (*Application Programming Interface*) yang disediakan oleh Twitter. Bahasa pemrograman yang digunakan adalah bahasa PHP yang terhubung dengan database yang berisi indeks dari data latih berupa dokumen RSS (*Really Simple Sindication*) beserta bobot dan kategorinya. Perangkat komputer standar dengan perangkat lunak Apache sebagai *web server* dan MySQL sebagai sistem manajemen basis data sebagai media untuk membuat aplikasi.

3.3 Merancang Aplikasi

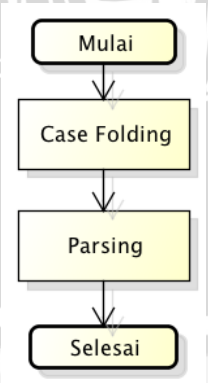
Pada langkah perancangan sistem aplikasi ini dilakukan setelah semua kebutuhan sistem didapatkan dari langkah analisa kebutuhan sistem yang telah dirancan sebelumnya. Rancangan yang sudah disusun akan dapat diterapkan pada perancangan aplikasi sampai bagaimana aplikasi berjalan. Adapun rancangan yang didapat adalah sebagai berikut :

1. Pengguna akan masuk pada aplikasi Twitter *client* yang terhubung secara langsung melalui API yang disediakan oleh Twitter dnegan *user interface* berupa *desktop*, pada tampilan awal pengguna akan melihat *timeline* yang berisi pesan singkat pada halaman beranda.
2. Sistem melakukan tahapan-tahapan *preprocessing* dimana pada proses ini dilakukan *parsing, case folding* dan *transformation* berupa *stopword removal* dan *stemming*. Proses akan dilanjutkan pada tahap berikutnya dengan menggunakan kata-kata yang dianggap penting saja.
3. Pada proses yang terpisah dari aplikasi Twitter *client* yang dijelaskan sebelumnya, data latih yang bersumber dari dokumen RSS diproses sama seperti tahap ke dua. Dokumen RSS yang telah memiliki label berupa kategori yang berasal dari sumber berita diproses *text preprocessing* untuk selanjutnya dibuat tabel frekuensi kemunculan kata disertai perhitungan probabilitas.
4. Aplikasi Twitter *client* yang menampilkan pesan singkat melakukan perhitungan berdasarkan pada data hasil perhitungan oleh data latih, sehingga pesan singkat yang ada pada Twitter akan memiliki kategori sesuai dengan perhitungan menggunakan metode *naive bayes.*

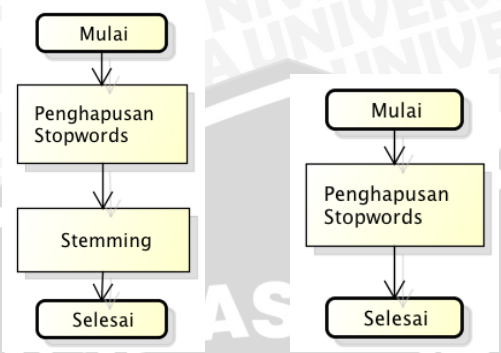
Untuk mempermudah alur jalan dari sistem, maka semua proses yang dijelaskan pada tahap perancangan ini digambarkan dalam diagram alir berikut :



Gambar 3.2 Diagram Alir Keseluruhan Sistem



Gambar 3.3 Diagram Alir *Preprocessing*



(a) (b)

Gambar 3.4 Diagram Alir *Transformation* (a) Menggunakan Proses *Stemming* dan (b) Tanpa *Stemming*

3.4 Implementasi Pembuatan Aplikasi

Pada tahap ini akan dilakukan implementasi pembuatan aplikasi yang mengacu pada tahap perancangan yang dilakukan sebelumnya. Bahasan implementasi pembuatan aplikasi dibagi menjadi dua, yaitu penjelasan mengenai lingkungan sistem dan pengimplementasian aplikasi.

3.4.1 Lingkungan Sistem

Lingkungan sistem yang akan dijelaskan adalah perangkat keras dan perangkat lunak yang digunakan dalam pembuatan sistem aplikasi.

Perangkat keras yang digunakan dalam pembuatan sistem aplikasi pengkategorian pesan singkat pada jejaring sosial Twitter berbahasa Indonesia adalah sebagai berikut :

- Prosesor Intel Core 2 Duo 2.26 GHz
- Memori RAM 2 GB 1067 MHz DDR3
- Hardisk 160.4 GB
- Monitor 13"
- Keyboard
- Mouse

Perangkat lunak yang digunakan dalam pembuatan aplikasi adalah sebagai berikut :

- Apache Web Server
- PHP
- MySQL
- FileZilla
- Aptana Studio
- Squel Pro
- Sistem Operasi Mac OSX Snow Leopard
- Framework Code Igniter

3.4.2 Implementasi Aplikasi

Pada bagian ini menjelaskan implementasi dari hasil perancangan dan analisis kebutuhan yang dijelaskan sebelumnya. Teks yang akan diproses dalam sistem baik data latih yang bersumber dari teks RSS yang memiliki kategori dan data uji berupa teks pesan singkat berbahasa Indonesia dari jejaring sosial Twitter secara umum akan diproses dengan cara yang sama. Proses-proses yang akan dilakukan oleh sistem secara terperinci akan dijelaskan pada bagian ini.

Proses *text mining* yang dilewati dalam proses pembentukan pengetahuan dan proses klasifikasi memiliki tiga tahap yaitu *text preprocessing, text transformation,* perhitungan frekuensi kata, dan *pattern discovery.*

3.4.2.1 Perancangan *text preprocessing*

Pada tahap ini akan dilakukan beberapa proses pada teks baik data latih yaitu teks RSS maupun data uji berupa teks pesan singkat pada jejaring sosial Twitter. Langkah-langkah yang dilakukan adalah proses *case folding* dan *parsing. Case folding* adalah mengubah semua huruf dalam teks menjadi huruf kecil. Sedangkan proses *parsing* sederhana yang dilakukan adalah memecah sebuah teks menjadi kumpulan kata-kata tanpa memperhatikan keterkaitan antar kata dan peran atau

kedudukannya dalam kalimat. Karakter yang diterima dalam pembentukan kata adalah karakter huruf saja. Dengan demikian, seperti kata ulang yang ada dalam kaidah bahasa Indonesia akan diurai menjadi dua kata bukan satu kesatuan kata.

3.4.2.2 Perancangan *text transformation*

Tahap *text transformation* akan dilakukan proses penghilangan *stopword* dan proses *stemming.*

a. Penghapusan *Stopword*

Penghapusan *stopword* dilakukan dengan menghapus kata-kata yang dianggap tidak mempengaruhi maksud dari isi atau makna dari keseluruhan kalimat. Kata-kata yang setelah di *parsing* untuk setiap kata akan melakukan pengecekan ke dalam daftar kata *stopword.* Jika kata merupakan *stopword* maka kata tersebut akan dibuang. Jika bukan maka kata tersebut akan melalui tahap *stemming.* Kriteria penghapusan dilakukan pada kata-kata atau karakter sebagai berikut :

- Kata-kata dalam bahasa Indonesia berjenis kata depan, kata sambung, kata hubung, dan lain-lain
- Angka dan karakter lain selain karakter huruf
- URL atau *link* pada halaman web lain
- *Whitespace* atau area kosong yang tidak memiliki teks

b. *Stemming* bahasa Indonesia

Dalam bahasa Indonesia imbuhan terdiri dari sufiks (akhiran), infiks(sisipan), dan prefiks (awalan). *Stemming* merupakan proses untuk mendapatkan kata dasar dari setiap kata yang ada pada dokumen. Proses ini bertujuan untuk mendapatkan kata penting pada kalimat. Proses *stemming* melakukan proses perubahan dari kata-kata yang ada menjadi kata dasar yang nantinya digunakan dalam pembobotan kalimat. Banyak metode yang bisa diterapkan dalam melakukan proses *stemming* bahasa Indonesia, namun dalam penelitian ini peneliti mengimplementasikan *stemming* dengan metode yang digagas oleh Arifin dan Setiono dimana konsep detail sudah dijelaskan pada bab 2.

3.4.2.3 Perancangan Perhitungan Frekuensi Kata

Proses perhitungan frekuensi kata dilakukan setelah melalui proses *parsing* dan *stemming*. Pada proses ini akan menghitung jumlah kemunculan kata secara komulatif, sehingga kata yang mempunyai frekuensi tinggi pada sebuah dokumen akan dianggap sebagai kata penting pada dokumen tersebut. Hasil dari proses ini nantinya akan diproses untuk melakukan perhitungan peluang pada metode *naive bayes*.

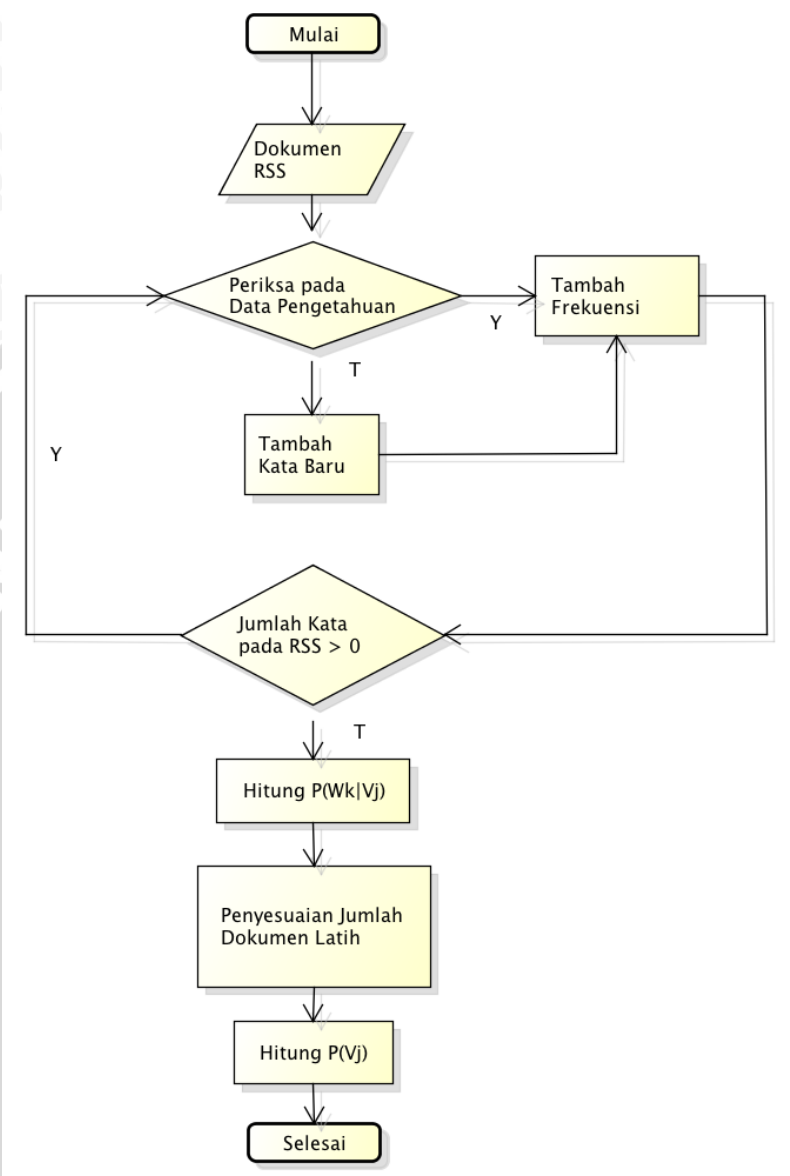
3.4.2.4 Perancangan *Pattern Discovery*

Pada penelitian ini metode *pattern discovery* yang digunakan adalah metode *naive bayes classifier*. Tahap ini terdapat dua bagian *learn naive bayes* dan *classifiy naive bayes*. *Learn naive bayes* berfungsi untuk membentuk pengetahuan berupa probabilitas dari perhitungan frekuensi kata sebelumnya, sedangkan pada *classifiy naive bayes* berfungsi untuk mengembalikan nilai perkiraan dari proses perhitungan dari target dokumen tersebut lebih cenderung ke sebuah kategori.

a. *Learn naive bayes*

Pada tahap *learn naive bayes* sistem akan melakukan proses pembentukan pengetahuan yang berasal dari data latih. Pengetahuan yang nantinya dihasilkan dari proses akan digunakan pada proses *classify naive bayes*. Pada penelitian pengetahuan dibagai menjadi dua bagian, yaitu pengetahuan kata dan pengetahuan dokumen. Pengetahuan kata berisi semua jenis kata pada seluruh data latih, frekuensi kemunculan setiap kata untuk setiap kategori, dan nilai probabilitas $P(w_k \mid v_j)$. Sedangkan pengetahuan dokumen berisi jumlah dokumen data latih pada setiap kategori dan nilai probabilitas $P(V_j)$. Adapun proses pada tahapan *learn naive bayes* adalah sebagai berikut :

1. Setiap jenis kata yang muncul pada data latih dicocokan terlebih dahulu dengan data kata pada pengetahuan yang sudah ada.
    - Jika ada maka tambahkan angka jumlah kemunculan kata tersebut pada pengetahuan kata untuk kategori yang bersesuaian.
    - Jika tidak ada maka tambahkanlah kata baru dan juga jumlah kemunculan kata tersebut pada pengetahuan kata untuk kategori yang bersesuaian.
2. Setelah semua kata dan frekuensi kemunculannya ditambahkan pada pengetahuan kata langkah selanjutnya adalah penghitungan ulang probabilitas pada pengetahuan sesuai dengan rumus $P(w_k \mid v_j)$ pada persamaan 2.8.
3. Tambahkan jumlah dokumen yang bersesuaian pada pengetahuan dokumen.
4. Melakukan perhitungan ulang $P(v_j)$ sesuai dengan rumus persamaan 2.7.

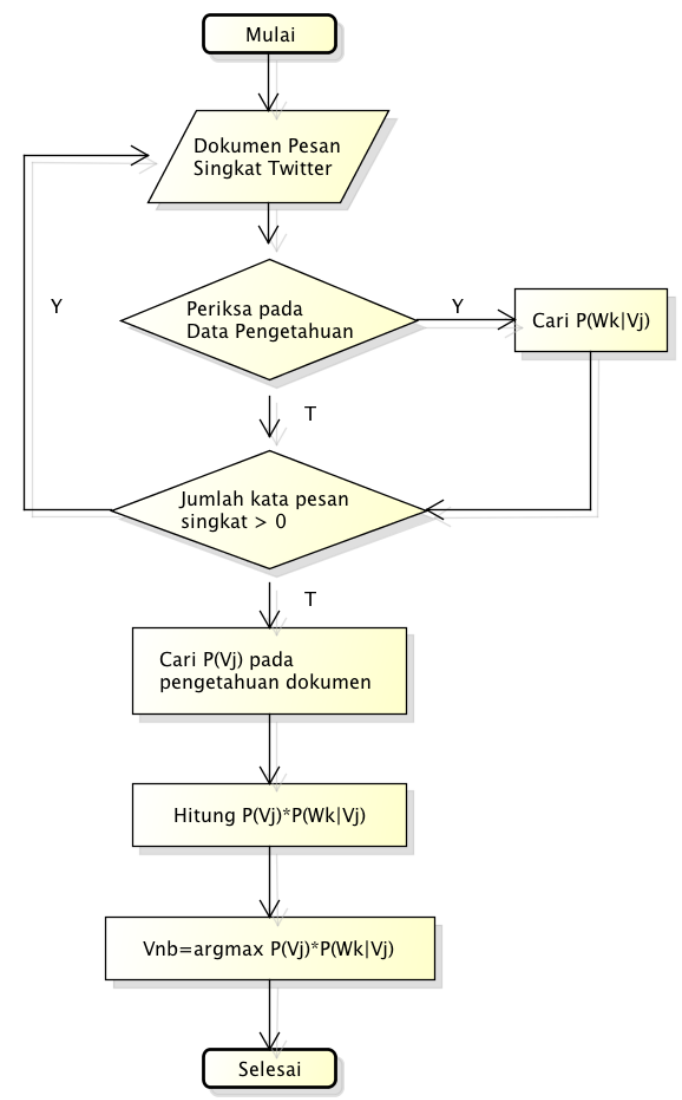


Gambar 3.5 Diagram Alir proses *learn naive bayes*

b. *Classify naive bayes*

Pada tahap ini dilakukan proses perhitungan data uji berdasarkan pengetahuan yang akan menghasilkan perkiraan atau estimasi kecenderungan kategori yang dimiliki oleh data uji. *Classify naive bayes* berusaha mencari nilai probabilitas tertinggi untuk mengklasifikasikan data uji pada kategori yang paling tepat. Tahapan yang dilakukan pada *classify naive bayes* adalah :

1. Setiap kata yang muncul dilakukan pencocokan ke dalam pengetahuan kata yang sebelumnya sudah ada
2. Dilakukan perhitungan rumus $P(v_j) \prod_k P(w_k \mid v_j)$ pada setiap kategori.
3. Setelah perhitungan sudah dilakukan pada seluruh kategori, maka akan dibandingkan nilai seluruh kategori dan untuk nilai yang terbesar maka dokumen akan dianggap dalam kategori tersebut.



Gambar 3.6 Diagram Alir proses *classify naive bayes*

3.5 Pengujian dan Analisis

Pada tahap ini dilakukan pengujian hasil kerja sistem yang telah dibuat dan melakukan evaluasi terhadap sistem sehingga mengetahui hasil dari sistem yang nantinya dijadikan sebagai kesimpulan untuk hasil dari pembuatan aplikasi pengkategorian pesan singkat berbahasa Indonesia pada jejaring sosial Twitter. Pengujian ini akan menguji 20 pesan singkat pada jejaring sosial Twitter yang akan membandingkan antara pengkategorian berasal dari sistem dan secara manual.

Proses untuk mempelajari pengaruh jumlah data latih terhadap efektifitas sistem klasifikasi maka dilakukan sepuluh kali uji coba dengan jumlah data latih yang berbeda, dengan proporsi yang berbeda dari keseluruhan dari data latih.

Efektifitas sistem klasifikasi akan dievaluasi menggunakan standar ukuran evaluasi *recall, precission* dan F_1 *measure* yang telah dijelaskan pada subbab 2.7. Hasil dari perhitungan *precission* dan *recall* akan membantu dalam penarikan kesimpulan dari hasil kerja sistem.

3.6 Pengambilan Kesimpulan

Tahap pengambilan kesimpulan dilakukan setelah melakukan pengujian dan analisis terhadap penelitian. Kesimpulan yang diambil berdasarkan pada hasil pengujian dan analisis, dengan adanya kesimpulan maka akan didapatkan inti dari hasil keseluruhan proses penelitian. Selain itu kesimpulan dapat memberi saran untuk pengembangan sistem selanjutnya.