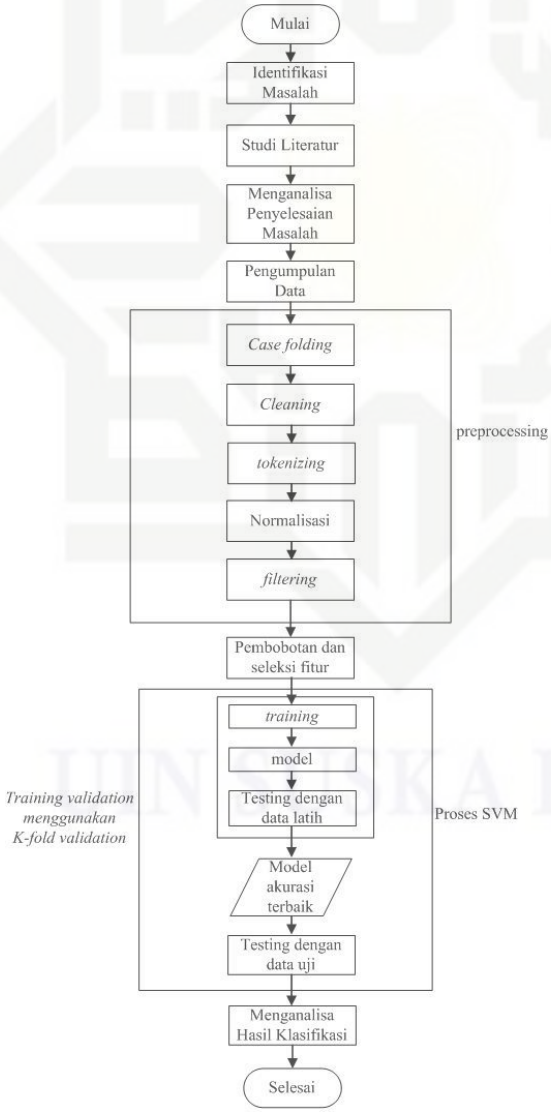
BAB III
METODOLOGI PENELITIAN

3.1 Tahapan Penelitian

Pada tugas akhir ini kasus yang diuraikan yaitu bagaimana menerapkan metode SVM pada Klasifikasi Bahasa yang Mirip (Bahasa Indonesia dan Bahasa Malaysia) pada Twitter. Adapun tahap-tahap yang dilakukan yaitu, sebagai berikut :



Gambar 3.1 Tahapan Penelitian

3.2 Identifikasi Masalah

Pada tahap pertama yang dilakukan untuk pengklasifikasian *tweet* yaitu harus diketahui dahulu apa saja kata yang paling sering digunakan di dalam bahasa Malaysia dan di dalam bahasa Indonesia. Kata tersebut akan menjadi acuan dalam pengklasifikasian antara bahasa Indonesia dan bahasa Malaysia.

3.3 Studi Literatur

Berdasarkan rumusan masalah, maka dilakukan pengkajian yang berhubungan dengan twitter, bahasa, *text mining*, *preprocessing* dan *Support Vector Machine* (SVM) yang akan digunakan dalam penelitian ini. Dalam mempelajari penelitian terkait klasifikasi teks, dilakukan pembelajaran dengan memperoleh sumber dari penelitian sebelumnya yang telah dilakukan seperti jurnal, tugas akhir terkait klasifikasi teks, dan artikel-artikel.

3.4 Menganalisa Penyelesaian Masalah

Pada tahap ini akan dilakukan analisa untuk menyelesaikan masalah pada penelitian yaitu, sebagai berikut :

3.4.1 Pengumpulan data

Dalam hal ini, dilakukan pengumpulan dataset dengan menggunakan *Application Programming Interface* (API). Dataset yang akan digunakan berasal twitter, dengan mengumpulkan data *tweet* secara *random*, yaitu dengan mencari *keyword* dari masing-masing kelas, dengan total data sebanyak 1000 data *tweet* yang terbagi menjadi 500 *tweet* bahasa Indonesia dan 500 *tweet* bahasa Malaysia.

3.4.2 Pelabelan

Pada tahap ini akan dilakukan pelabelan menjadi data latih dan data uji. Serta, akan dilabel kelas 1 untuk kelas bahasa Indonesia dan kelas 2 untuk kelas bahasa Malaysia. Adapun 1000 data *tweet* yang telah diperoleh, akan dibagi menjadi 800 data latih, yang terdiri dari 400 *tweet* bahasa Indonesia dan 400 *tweet* bahasa Malaysia dan 200 data uji, terdiri dari 100 *tweet* bahasa Indonesia dan 100 *tweet* bahasa Malaysia.

3.5 *Text Preprocessing*

Text preprocessing merupakan tahapan dalam membersihkan data yang akan digunakan dalam penelitian. T*ext preprocessing* bertujuan untuk menghilangkan *noise* dan menghilangkan kata yang tidak berpengaruh terhadap klasifikasi. Pada penelitian ini, tahapan *text preprocessing* yang akan digunakan yaitu *case folding, cleaning, tokenizing,* normalisasi, dan *filtering.*

3.6 Pembobotan dan Seleksi Fitur

Pembobotan dan seleksi fitur dilakukan dengan tujuan untuk dapat mengurangi *noise* dengan cara menghilangkan *feature* yang tidak tepat, sehingga akan dapat mempengaruhi pada tingkat akurasi yang lebih dalam klasifikasi. Pada penelitian ini, pembobotan kata dilakukan dengan menggunakan *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). Setelah masing-masing dari kata telah dibobotkan, maka tahap selanjutnya yaitu mengambil kata yang akan dijadikan fitur untuk klasifikasi. Adapun dalam mengambil kata yang akan dijadikan fitur dalam klasifikasi yaitu menggunakan *threshold* dengan nilai yang ditentukan. *Threshold* yang digunakan yaitu *threshold* dengan nilai 40.

3.7 Proses Support Vector Machine (SVM)

Pada proses SVM yang akan dilakukan adalah proses *training* dan proses *testing.* Pada proses *training* dilakukan dengan metode *grid search* dengan menggunakan K-*fold cross validation* untuk mendapatkan model terbaik dengan nilai parameter C dan γ, menentukan parameter C dan γ dilakukan untuk menghasilkan akurasi tertinggi.

3.8 Analisa Klasifikasi

Analisa klasifikasi bertujuan untuk proses dalam menganalisa kembali hasil klasifikasi dengan berdasarkan tahap-tahap yang sebelumnya telah dilakukan. Pada tahap analisa klasifikasi juga terdapat saran-saran yang bertujuan dalam membangun penelitian ini sehingga menjadikan penelitian selanjutnya agar lebih baik.