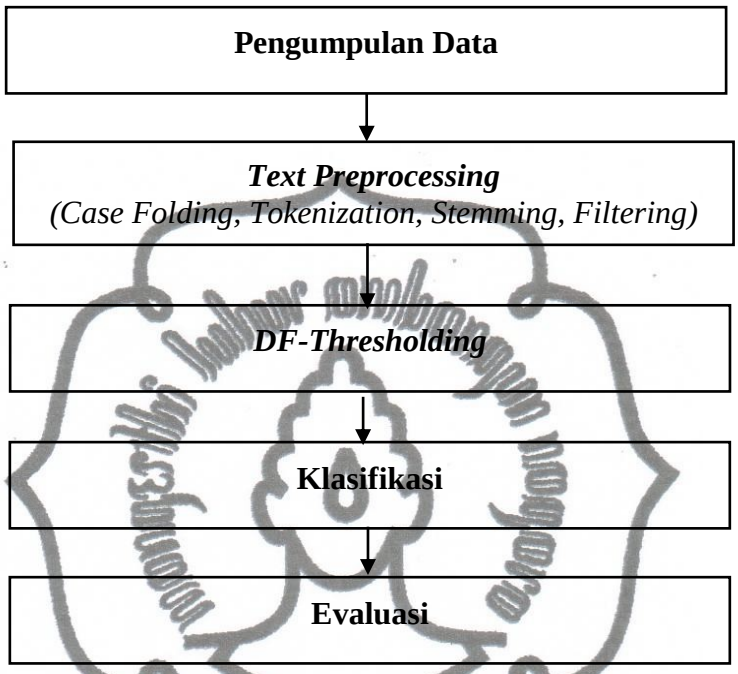


BAB III
METODOLOGI PENELITIAN

Tahap-tahap yang dilakukan dalam menyelesaikan penelitian ini dijelaskan pada gambar berikut ini.



Gambar 3.1 Tahapan Penelitian

3.1 Pengumpulan Data

Data yang dibutuhkan dalam penelitian ini merupakan data berita yang dimonitoring oleh Lembaga Pengolahan dan Penyedia Informasi (PPI), Dirjen Informasi dan Komunikasi Publik (IKP), Kementerian Komunikasi dan Informatika.

3.2 Text Preprocessing

Seluruh data akan melalui tahapan *text preprocessing* yang terdiri dari beberapa proses, yaitu :

1. *Case Folding*

   Pada tahapan ini, semua karakter selain huruf di dalam data berita dihilangkan dan semua huruf diubah menjadi huruf kecil (*lowercase)*.

2. *Tokenization*

   Data berita yang telah melalui proses *case folding* akan melalui proses *tokenization* yang bertujuan untuk mengubah bentuk *string* menjadi *token-token.*

3. *Stemming*

   Selanjutnya data akan melalui proses *stemming* untuk menghilangkan imbuhan pada kata sehingga semua kata menjadi kata dasar atau *root word*.

4. *Filtering*

   Proses terakhir dari *text preprocessing* adalah *filtering* yang berujuan untuk menghilangkan *stopwords.*

Hasil dari *text preprocessing* ini kemudian disimpan di dalam database kata.

3.3 DF-Thresholding

Sebelum masuk pada proses klasifikasi, dilakukan terlebih dahulu proses seleksi fitur dengan menggunakan *DF-Thresholding*. Tujuan dari proses ini adalah untuk mengurangi dimensi data. Pada proses ini, dilakukan perhitungan jumlah dokumen yang mengandung kata tertentu. Selanjutnya menentukan *threshold*, apabila jumlah data kurang dan lebih dari *threshold*, maka kata tersebut tidak digunakan pada proses klasifikasi. *Threshold* dipilih berdasarkan pada beberapa nilai yang dilakukan dengan kombinasi batas atas dan batas bawah.

3.4 Klasifikasi

Proses klasifikasi dibagi menjadi dua tahap, yaitu *training* dan *testing*. Pada proses *training*, masing-masing data berita diproses dan setiap kata dihitung jumlah kemunculannya. Data-data ini yang kemudian akan digunakan sebagai bahan pembelajaran pada proses *testing* untuk menentukan suatu data berita masuk pada kelas isu tertentu.

Proses klasifikasi pada penelitian ini dilakukan dengan cara meng-*update* data *training*. Misalnya, dalam penelitian ini digunakan sebanyak 395 data berita yang terbit terlebih dahulu sebagai data *training* awal, kemudian sebanyak 98 data digunakan sebagai data *testing*. Untuk proses selanjutnya, dilakukan *update* data *training* dengan menambahkan 98 data tersebut, sehingga data *training* yang digunakan untuk proses kedua adalah sebanyak 493 data. Dengan demikian, data *training* akan terus bertambah.

Proses klasifikasi dilakukan menggunakan *Multinomial Naive Bayes* dan *Multionomial Naive Bayes* dengan pembobotan *TFIDF*.

3.4.1 Klasifikasi *Multinomial Naive Bayes*

Proses klasifikasi dilakukan dengan menggunakan metode *Multinomial Naive Bayes.* Adapun langkah-langkahnya adalah sebagai berikut :

1. Menghitung data *prior* masing-masing kelas dengan menggunakan rumus 2.4.
2. Menghitung probabilitas kata ke-n data berita dengan menggunakan rumus 2.5.
3. Menghitung probabilitas suatu dokumen masuk ke dalam suatu kelas dengan rumus 2.3.
4. Menentukan kelas dokumen dengan memilih nilai probabilitas tertinggi.

3.4.2 Klasifikasi *Multinomial Naive Bayes* dengan Pembobotan TFIDF

Pada proses klasifikasi *Multinomial Naive Bayes* dengan pembobotan TFIDF, data *term* akan dihitung jumlahnya berdasarkan kemunculannya pada suatu dokumen (TF) dan seluruh dokumen (DF) terlebih dahulu. Dari nilai DF, selanjutnya akan dihitung nilai *inverse (IDF)* yang kemudian dikalikan dengan nilai TF. Hasil pembobotan TFIDF tersebut yang akan digunakan dalam proses klasifikasi. Adapun proses klasifikasi dilakukan dengan menggunakan rumus 2.6.

3.5 Evaluasi

Proses evaluasi pada penelitian ini menggunakan perhitungan akurasi, *precision* dan *recall* dari hasil klasifikasi yang disajikan dengan tabel *confusion matrix* pada Tabel 3.1.

Tabel 3.1 Confusion Matrix

Realita	Sistem				
	Kelas-1	Kelas-2	....	Kelas-n	Total
Kelas-1	True Positive	Error	....	Error	Total Kelas-1
Kelas-2	Error	True Positive	...	Error	Total Kelas-2
...	Error	Error	...	Error	...
Kelas-n	Error	Error	...	True Positive	Total Kelas-n
	Prediksi Kelas-1	Prediksi Kelas-2	...	Prediksi Kelas-n	

Adapun rumus perhitungannya adalah sebagai berikut (Power, 2011) :

$$Accuracy : \frac{\text{TP(Kelas}-1)+ \text{TP(Kelas}-2)+..+\text{TP(Kelas}-n)}{Total\ (Kelas-1)+Total(Kelas-2)+\cdots+Total(Kelas-n)} \quad (3.1)$$

$$Precision : \frac{\text{TP(Kelas}-i)}{Prediksi\ (Kelas-i)} \quad (3.2)$$

$$Recall : \frac{\text{TP(Kelas}-1)}{Total\ (Kelas-i)} \quad (3.3)$$

Keterangan :

TP (*True Positive*) = data yang diklasifikasikan sesuai dengan kelas sebenarnya.

Error = data yang diklasifikasikan tidak sesuai dengan kelas sebenarnya.

Accuracy = total jumlah seluruh data yang diklasifikan dengan benar dibagi dengan total jumlah data.

Precision = jumlah data yang diklasifikasikan dengan benar dibagi jumlah data terprediksi pada kelas tertentu.

Recall = jumlah data yang diklasifikasikan dengan benar dibagi jumlah data pada kelas sebenarnya.