BAB III
METODE PENELITIAN

3.1. Bahan Penelitian

Proses pengumpulan data diperoleh dari Buletin yang diterbikan oleh APTIKOM setiap bulannya. Bahan penelitian berupa kumpulan kata pada artikel.

3.2. Peralatan Penelitian

Penelitian ini tentunya membutuhkan peralatan yang dapat membantu proses penelitian yaitu *Hardware* dan *Software*. Spesifikasi perangkat keras terdiri dari Processor Core i5-6200, 8192 RAM, 1 TB HDD. Kemudian, perangkat lunak yang dimanfaatkan untuk pengolahan bahan penelitian adalah bahasa pemrograman R Studio. R Studio digunakan sebagai text editor dalam penelitiaan ini. Lalu, Google Chrome yang digunakan untuk browsing, mencari referensi dari artikel maupun jurnal. Selanjutnya, Microsoft Office 2016 berfungsi untuk membuat laporan dan proposal.

3.3. Lokasi Penelitian dan Waktu Penelitian

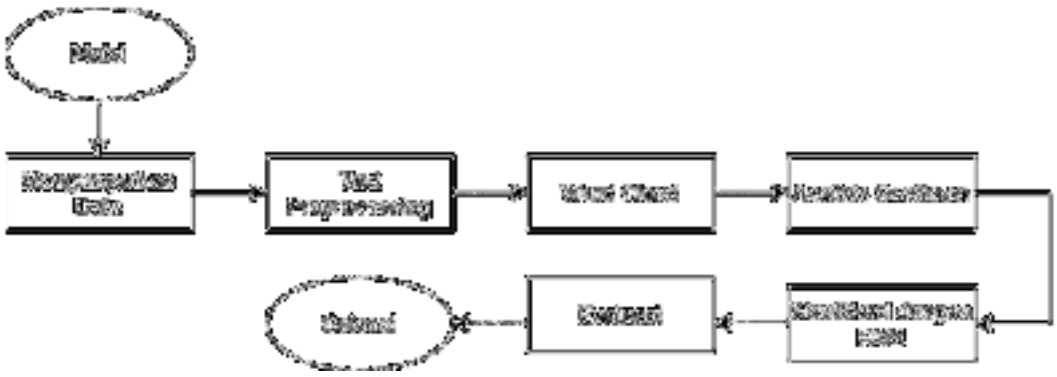
Penelitian ini dilakukan di laboratorium riset dan teknologi Gedung A lantai II Universitas Buana Perjuangan Karawang. Kemudian, pelaksanaan penelitian berlangsung sejak bulan Januari. Rincian pelaksaan penelitian ditunjukkan pada Tabel 3.1

Tabel 3. 1 Jadwal penelitian

	Januari				Februari				Maret				April				Mei				Juni				Juli			
Kegiatan	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Studi Literatur																												
Penulisan Proposal																												
Analisis Data																												
PreProcessing																												
Implementasi																												
Klasifikasi																												
Evaluasi																												
Pengujian																												
Penulisan Laporan																												
Tugas Akhir																												
Yudisium																												

3.4. Prosedur Penelitian

Adapun prosedur penelitian yang digambarkan dengan bentuk *Flowchart* dalam Gambar 3.1. Penelitian ini dimulai dengan menganalisis data yang diambil dari Buletin APTIKOM. Kemudian, data yang akan diolah dengan *Text Preprocessing* agar terstruktur dengan baik dan menghilangkan *noise*. Setelah itu data yang telah diolah akan dilakukan analisis sentimen untuk mendapatkan kelas positif, negatif atau netral dan dapat divisualisasikan pada klasifikasi kelas sentimen.



Gambar 3. 1 Prosedur penelitian

3.5. Analisis Data

Data yang dipakai merupakan sebuah data primer. Data tersebut diambil dari Buletin ATPIKOM yang diterbitkan sebulan satu kali. Data diambil secara acak sesuai tema yang ada pada Buletin APTIKOM dan data ini akan diolah atau diklasifikasi menggunakan algoritma *K-Nearest Neighbor*. Untuk mendapatkan hasil klasifikasi tersebut ada beberapa tahapan yaitu sebagai berikut:

- Mengumpulkan data

  Data yang diambil yaitu data dari Buletin yang diterbitkan oleh APTIKOM setiap bulannya. Pada penelitian ini menggunakan Buletin yang membahas *Artificial Intellegence*.

- *Text Preprocessing*

  Sebelum dilakukannya proses klasifikasi, teks dokumen atau data harus disiapkan terlebih dahulu, dan proses tersebut dinamakan *Text* Preprocessing. Tahapan *Text Preprocessing* ini berguna agar data teks yang masih terdapat banyak *noise* atau biasa disebut tidak terstuktur menjadi lebih terstuktur.

  Terdapat beberapa tahapan pada *Text Preprocessing* yaitu *Case Folding*, *Tokenizing*, menghapus angka, *filtering*, dan TF IDF.

  a. *Case Folding*

     Case Folding adalah sebuah proses untuk merubah huruf kapital menjadi huruf standar. Proses ini dilakukan guna mempermudah pencarian dikarenakan tidak semua dokumen teks konsisten dengan huruf kapital.

b. *Tokenizing*

Tokenizing ini adalah sebuah proses memecah kalimat menjadi kata-kata yang dilakukan untuk menjadi sebuah kalimat menjadi lebih bermakna. Ada beberapa tahapan pada proses ini yang pertama itu adalah melakukan normalisasi kata dengan mengubah semua karakter huruf menjadi huruf kecil atau *Case Folding*. Proses *tokenizing* diawali dengan menghilangkan delimiter-delimiter yaitu simbol dan tanda baca yang ada pada teks tersebut seperti @, $, &, tanda koma (,), tanda titik (.), tanda tanya (?), tanda seru (!). Tahap selanjutnya yaitu melakukan proses pengurairan teks yang semula berupa kalimat-kalimat yang berisi kata-kata. Umumnya setiap kata akan terpisahkan dengan karakter spasi, proses tokenisasi mengandalkan karakter spasi pada dokumen teks untuk melakukan pemisahan. Hasil dari proses ini yaitu kumpulan kata saja.

c. Menghapus Angka

Tahapan ini memulai memasuki tahapan proses *filtering* yang berguna membersihkan kata sebelum di proses salah satunya dalam bentuk angka. Untuk dilakukan proses analisis sentimen, data yang bisa diolah hanya berbentung teks sehingga semua yang berhubungan dengan angka atau nomor akan dibersihkan pada proses *Text Preprocessing*.

d. *Filtering*

Filtering atau tahap filterisasi adalah tahapan mengambil kata-kata penting dari hasil token. Pada tahapan ini biasanya menggunakan Algoritma *stoplist* untuk membuat kata yang kurang penting dan *wordlist* untuk menyimpan kata yang penting. *Stopword* adalah kata-kata yang tidak dekriptif dan bukan merupakan kata penting dari suatu dokumen sehingga dapat dibuang. Beberapa contoh *Stopword* yaitu “yang”, “dan”, “di” dan masih banyak lagi. Dalam tahap *filtering* ini menggunakan *stoplist/stopword* supaya kalimat yang sering muncul

pada suatu dokumen dapat dihilangkan sehingga menyisakan kalimat yang penting dan mempunyai arti yang sudah bisa di proses ke tahapan selanjutnya.

e. TF IDF

Term Frequency-Invers Document Frequency (TF-IDF) merupakan tahapan terakhir pada proses *text processing* untuk melakukan pembobotan kata. TF IDF adalah sebuah statistik numerik yang dapat menunjukkan relevansi kata kunci dengan dokumen tertentu. Selain itu, TF IDF juga dapat mengetahui kata apa yang sering muncul pada suatu dokumen.

- *Wordcloud*

  Wordcloud merupakan sebuah represintasi data yang berbentuk teks. *Wordcloud* juga dapat berguna dalam berbagai konteks untuk memvisualisasikan data teks secara menarik. *Wordcloud* biasanya digunakan untuk menampilkan data teks atau kata yang sering muncul.

- Analisis Sentimen

  Analisis Sentimen merupakan suatu proses yang berfungsi untuk mendapatkan sentimen penulisan pada teks apakah tergolong positif, negatif, netral. Analisis sentimen adalah bagian dari penelitian domain *text mining* yang sudah banyak digunakan di tahun 2013. Analisis sentimen dapat membantu dalam berbagai kemungkinan domain, kecenderungan penelitian yang membahas analisis sentimen ini berfokus pada opini yang menghasilkan sentimen positif dan negatif. Dengan analisis sentimen biasanya dapat mengelompokkan sebuah sentimen positif atau negatif.

- Ulasan Positif

  Jenis data pada sentimen positif merupakan hasil dari pelabelan secara otomatis dan telah dilakukan dengan baik. Proses ekstraksi pada sentimen positif dilakukan dengan cara berulang-ulang sampai menghaslkan kalimat positif pada buletin Aptikom. Sentimen positif tersebut dikenali dari frekuensi kata yang ada pada ulasan.

- Ulasan Negatif

  Untuk mendapatkan kalimat negatif pada Buletin APTIKOM maka dilakukan proses ekstraksi pada sentimen negatif secara berulang-ulang. Hasil ekstraksi sentimen negatif di identifikasi dari frekuensi kata yang ada pada ulasan, dan berlandaskan dengan relevansi kata pada topik yang telah merujuk sentimen negatif.

- Klasifikasi

  Klasifikasi dapat mempelajari beberapa kumpulan data sampai menemukan hasil aturan yang dapat memproses data yang masih baru. Maka dari itu, tahapan ini menjadi tahapan yang sangat penting dalam *text mining*.

- Evaluasi

  Pada proses terakhir yaitu evaluasi, penelitian ini menggunakan *confusion matrix* untuk melakukan tahapan pengujian atau evaluasi. Proses evaluasi yang menggunakan *confusion matrix* ini memang banyak digunakan pada penelitian yang menggunakan algoritma K-NN. Proses evaluasi ini berguna untuk menghitung tingkat akurasi algoritma. Selain tingkat akurasi, dengan *confusion matrix* juga dapat menghitung nilai *precision*, dan *recall*.

3.6. Implementasi

Penelitian ini diimplementasikan menggunakan R Studio dan diuji dengan *Confusion Matrix*. Sebelum mendapatkan hasil, ada beberapa tahapan terlebih dahulu seperti *text preprocessing* yang didalamnya ada beberapa proses seperti *Spelling normalization, case folding, tokenizing,* dan *filtering.* Setelah itu ada tahapan TF-IDF kemudian diklasifikasi menggunakan metode *K-Nearest Neighbor* (KNN) dan dilakukan pengujian dengan *Confusion Matrix.*

Term Frequency-Invers Document Frequency (TF-IDF) merupakan sebuah statistik numerik yang berguna untuk menunjukkan relevansi kata kunci dengan dokumen tertentu. TF-IDF ini memiliki 2 perhitungan yaitu *Term Frequency*(TF) dan *Invers Document Frequency*(IDF). Perhitungan TF-IDF dapat dilihat pada persamaan (3) dan (4).

$$TF = \begin{cases} 1 + log_{10}(tf_{t,d}), if\ tf_{t,d} > 0 \\ 0, if tf_{t,d} = 0 \end{cases} \quad (1)$$

$$IDF = \log\left(\frac{N}{dft}\right) \quad (2)$$

Keterangan:

$tf_{t,d}$: jumlah kemunculan *term*(t) pada dokumen(d), jika tidak ada termm atau t=0, maka TF menjadi 0

N : jumlah dokumen teks.

df_t : jumlah dokumen yang mengandung *term*(t)

Perkalian dari perhitungan TF dan IDF bisa mendapatkan bobot kata yang disebut TF-IDF, ditunjukkan pada persamaan (4):

$$W_{t,d} = TF\ X\ IDF \quad (3)$$

3.7. Pengujian

Pengujian pada penelitian ini digunakan untuk mengevaluasi performa algoritma (yang digunakan). Proses evaluasi dimulai dengan membuat matrik konfusi untuk menilai akurasi algoritma. Kemudian dilakukan pengujian performa algoritma menggunakan metode *Confusion Matrix*. Metode ini cukup membantu untuk melakukan proses analisis kualitas *classifier*.

Pengujian dilakukan dengan menghitung *accuracy, recall, precision, f-measure* dan ditampilkan dalam bentuk persentase.

A. *Accuracy*

   Akurasi merupakan hasil pada skala prediksi yang akurat

B. *Precision*

   Precision ialah skala pada jumlah teks yang signifikan dikalangan dokumen teks yang sudah di sortir menggunakan sistem

C. *Recall*

   Recall ialah skala pada jumlah teks yang relevan dikalangan dokumen teks pada koleksi.