BAB 2 LANDASAN KEPUSTAKAAN

Landasan kepustakaan berisi uraian dan pembahasan tentang teori, metode, atau sistem dari literatur ilmiah, yang berkaitan dengan *sentiment analysis review* aplikasi *mobile* dengan menggunakan metode *Modified k nearest neighbour*. Dalam landasan kepustakaan terdapat landasan teori dari berbagai sumber pustaka yang terkait dengan teori dan metode yang digunakan dalam penelitian.

1.1. Kajian Pustaka

Kajian kepustakaan yang digunakan pada penelitian ini adalah penelitian yang dilakukan oleh Prima Afrianda Putri. Penelitian yang dilakuakan berjudul "*Implementasi Metode Improved K-Nearest Neighbor pada Analisis Sentimen Twitter Berbahasa Indonesia*" Penelitian yang telah dilakukan menggunakan metode *Improve* k *Nearest Neighbour* untuk analisis sentimen pada *tweet* dari user pada social *media twitter.* Sentimen yang dihasilkan adalah berupa sentimen positif atau negatif. Sebelum masuk ke metode, data terlebih dahulu melalui proses *preprocessing*. Proses *preprocesing* terdiri dari tahap pembersihan (*cleaning*), *tokenizing*, *filtering*, dan *stemming*. Data berupa *term* hasil dari proses *preprocessing* akan diberi bobot kata menggunakan *Term Frequency – Inverse Document Frequency* (TF-IDF). Hasil akurasi yang diperoleh dari klasifikasi dengan metode yang digunakan mencapai 87% untuk rata-rata *recall*, 82% untuk rata-rata *precission*, dan 84% untuk rata-rata *F-Measure*.

Penelitian terkait lainnya adalah mengacu pada penelitian yang dilakukan oleh Noviana Ayu Kumalasari pada tahun 2014. Penelitian yang berjudul *"Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Menentukan Tingkat Resiko Penyakit Lemak Darah (PROFIL LIPID)"*. Parameter yang digunakan dalam penelitian tersebut ada 4, yaitu kolesterol total, kolesterol HDL, kolesterol LDL, dan trigilserida. Data yang digunakan merupakan data yang diperoleh dari hasil tes darah pada Laboratorium Klinik Sejahtera, Probolinggo. Data yang digunakan merupakan data *continue*. Data akan mengalami beberapa proses yang ada pada proses klasifikasi MKNN. Diantaranya yaitu proses perhitungan nilai validitas, perhitungan jarak, perhitungan normalisasi, dan perhitungan nilai *weight voting*. Penelitian yang dilakukan menghasilkan akurasi maksimum sebesar 85,81% dengan 140 data latih untuk nilai k = 2.

Dengan mengacu pada beberapa penelitian tersebut, peneliti berasumsi bahwa dengan menggabungkan dua penelitian yang sudah ada yakni menganalisis sentimen dengan menggunakan metode *Modified K Nearest Neighbour*, akan mampu memperoleh hasil akurasi yang lebih baik dan bisa digunakan sebagai acuan untuk penelitian selanjutnya.

1.2. Review Aplikasi

Review dapat diartikan secara bahasa sebagai tinjauan. Tinjauan menurut Kamus Besar Bahasa Indonesia adalah suatu pemerikasaan yang dilakukan kepada suatu objek dengan teliti, dan hasil pemeriksaannya disajikan secara sistematis. Sedangkan aplikasi adalah program yang bisa digunakan untuk menyelesaikan kebutuhan pengguna dengan menjalankan fungsi-fungsi yang sudah dibuat. (Intalasari, 2016). Berdasarkan arti tersebut, dapat diambil kesimpulan bahwa arti dari *review* aplikasi adalah pemeriksaan terhadap suatu aplikasi. Hal ini diperlukan adalah untuk mengetahui apakah aplikasi yang di*review* sudah memenuhi kebutuhan pengguna atau belum. Apabila belum maka akan dilakukan perbaikan sehingga aplikasi akan menjadi lebih baik.

1.3. Sentiment Analysis

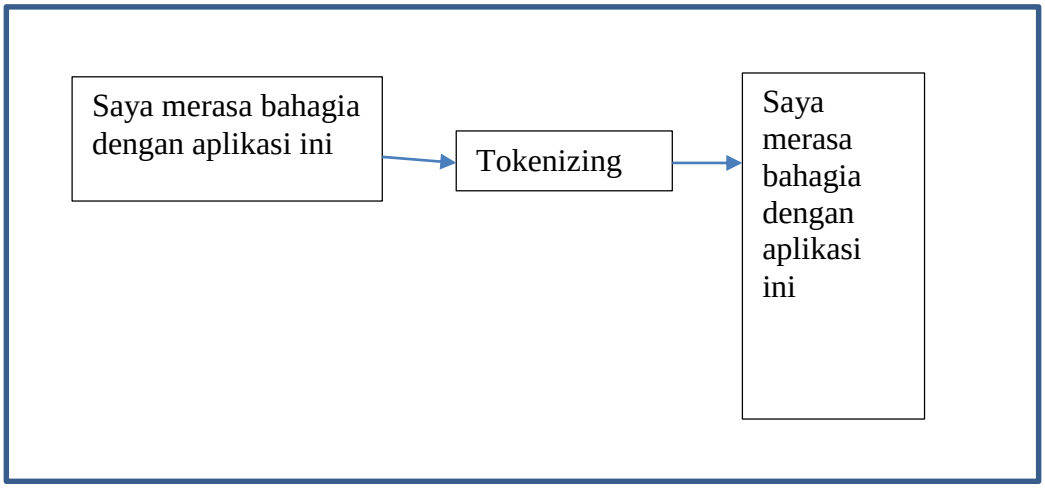
Sentiment Analysis merupakan teknik mengolah dan menggali data berupa teks untuk mengetahui informasi yang terkandung dalam data.(Putri, 2013). Dengan analisis sentimen, dapat diketahui dengan cepat dan tepat informasi yang terkandung dalam suatu data teks. Hal ini sangat bermanfaat untuk mengetahui bagaimana respon seorang pengguna terhadap suatu produk. Sehingga produk akan dikembangkan menjadi lebih baik dari sebelumnya. Selain itu, analisis sentimen bisa juga digunakan untuk mengetahui keadaan politik melalui tanggapan dari masyarakat terhadap kondisi politik yang saat ini.(Putri, 2013).

1.4. Preprocessing

Sebuah data teks atau dokumen memiliki struktur yang kurang baik sehingga perlu dilakukan perbaikan struktur data untuk mempercepat proses komputerisasi. Proses pengubahan tersebut dinamakan *teks prepropcessing*.(Marfian, 2015). Pada *preprocessing* terdapat beberapa tahap yang perlu dilakukan diantaranya adalah *Tokenizing*, *Filtering*, dan *Stemming*. Masing-masing proses tersebut saling berhubungan untuk mengubah data dokumen yang tidak terstruktur menjadi data yang terstruktur.

1.4.1. Tokenizing

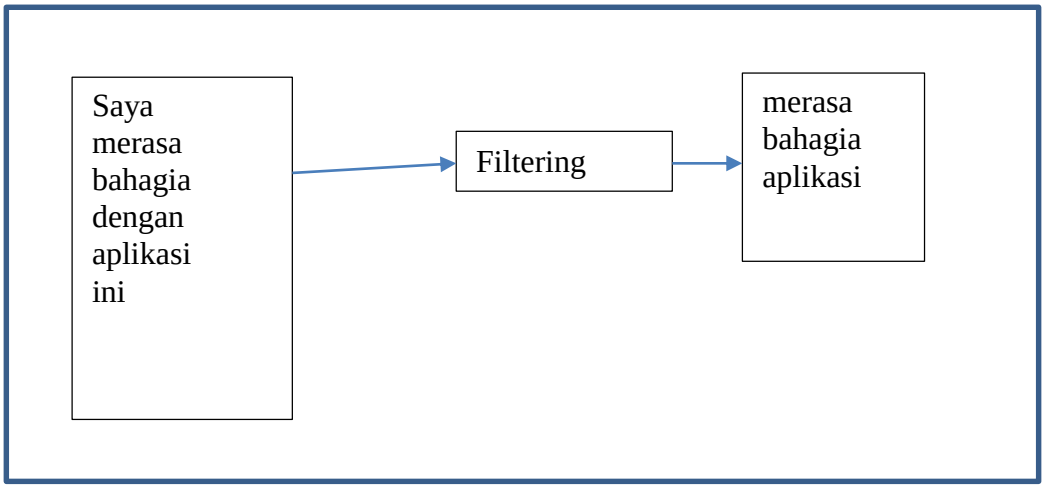
Tokenizing adalah proses untuk mengubah *input* kalimat menjadi sebuh barisan kata *(token)*. Setiap kata tersebut nantinya akan digunakan sebagai sebuah fitur tiap kelas.



Gambar 2.1 Proses Tokenizing

1.4.2. Filtering

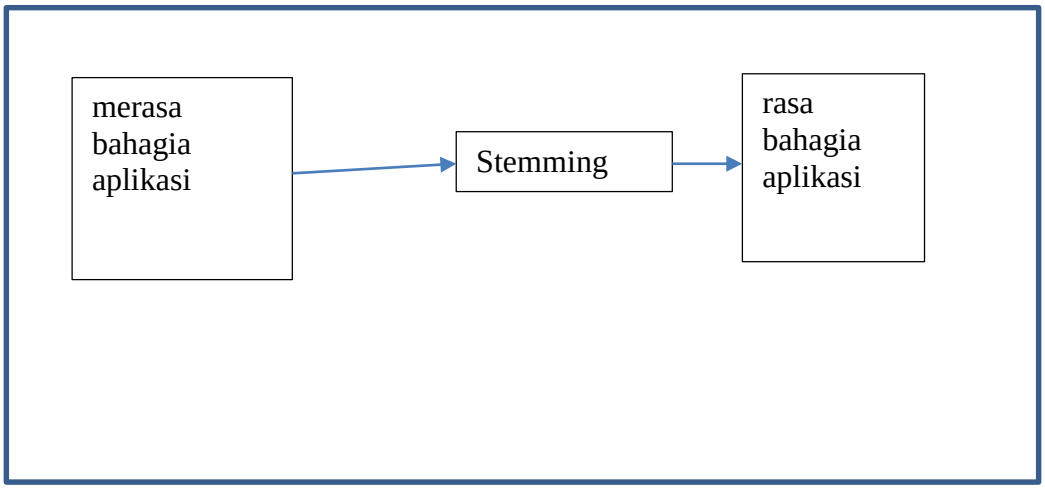
Filtering adalah proses untuk menghilangkan kata-kata yang tidak penting dari hasil *Tokenizing*. Cara menghilangkannya bisa menggunakan *stopword* atau *stoplist*. *Stoplist* berisi daftar kata-kata yang tidak penting yang bisa diinputkan oleh *developer* program atau *library* yang sudah dibuat.



Gambar 2.2 Proses Filtering

1.4.3. Stemming

Setelah dilakukan proses *Filtering*, kata-kata yang sudah di *filter* akan dilakukan proses *stemming* yaitu mengambil kata dasar dari setiap kata.



Gambar 2.3 Proses Stemming

1.5. Perhitungan Bobot Kata

Term Frequency (TF) merupakan jumlah kemunculan sebuah *term* pada sebuah dokumen. Jika sebuah kata (*term*) sering muncul pada sebuah dokumen, maka *query* yang mengandung kata tersebut harus mendapatkan dokumen tersebut. *Term Frequency* ini didasari pada aspek lokal pada TF-IDF. Lokal *weight*, atau yang biasa disebut dengan TF (*term frequency*) yang memiliki fungsi yaitu untuk menentukan bobot dari *term* pada dokumen tertentu, yang pada dasarnya menghasilkan estimasi berdasarkan frekuensi atau *relative frequency* dari kata pada suatu dokumen. (Kao, 2007).

Term Frequency yang digunakan adalah *Raw* TF. Pada Raw TF ini, nilai TF sebuah *term* dihitung berdasarkan kemunculan *term* tersebut dalam dokumen.

` IDF merupakan *Inverse Document Frequency* yang berfungsi mengurangi bobot suatu *term* jika kemunculannya banyak tersebar di seluruh dokumen. Biasanya TF dan IDF adalah satu kesatuan yang baik digunakan untuk jumlah dokumen yang besar dan memiliki kata-kata yang banyak.

Berikut adalah rumus yang digunakan untuk mengitung nilai IDF:

$$IDF(t) = \log \frac{N}{df} \quad (2.1)$$

Keterangan :

IDF = Inverse Document Frequency

N = Jumlah dokumen

df = Banyak dokumen yang mengandung term

1.6. Modified K-Nearest Neighbour

Modified K-Nearest Neighbour (MK-NN) merupakan pengembangan dari metode *K-Nearest Neighbour* yang menggunakan algoritma *supervised* dengan hasil dari *query instance* yang dikelompokkan berdasarkan mayoritas dari kategori KNN. (putri, 2013). Pada metode MK-NN terdapat nilai *validitas* untuk menghitung jumlah titik pada setiap label yang sama dengan data tersebut. (kumalasari, 2014).

Pada MKNN terdapat Validitas data latih yang berfungsi untuk menghitung jumlah titik pada kelas yang sama dari data tersebut. berikut adalah formula untuk menghitung nilai validitas (Mutiara, 2004)

$$validitas\ (x) = \frac{1}{k}\sum_{i=1}^{k} S(label(x), (label(Ni(x))) \qquad (2.2)$$

Keterangan :

K : jumlah titik terdekat

label (x) : kelas x

label (Ni(x)) : label kelas titik terdekat x

Fungsi S berfungsi untuk menghitung kesamaan data uji dengan tetangga terdekat. Tetangga terdekat berdasarkan nilai IDF yang telah dirutkan dari yang terbesar. Nilai S bisa diperoleh dari persamaan S(*a*,*b*). Dimana *a* adalah kelas data uji pada data latih, sedangkan *b* adalah kelas dari tetangga terdekat yang telah terurut. S bernilai 1 jika nilai $a = b$. S bernilai 0 jika nilai $a \neq b$. Data uji yang dimaksud dalam perhitungan validitas adalah data latih yang diujikan, bukan data uji yang ada pada proses pengujian (*testing*).

Pada metode MKNN, digunakan rumus *Weight voting* untuk masing-masing tetangga dari data uji. Nilai dari Weight voting dapat diperoleh dari persamaan berikut. (Mutiara, 2004).

$$W(i) = \frac{1}{d+\alpha}\ , \qquad (2.3)$$

Keterangan :

d = jarak *Euclidian*

α = nilai *regulator smoothing*

pada penelitian ini digunakan α = 0,5. *Weight voting* ini kemudian dijumlahkan untuk setiap kelasnya, untuk menentukan kelas yang terpilih, diambil dari nilai *Weight voting* pang paling besar.

Dalam metode MKNN, masing-masing k tetangga terdekat dihitung dengan menggunakan persamaan (2.3). Kemudian, nilai validitas dari tiap data yang telah dihitung sebelumnya dikalikan dengan hasil *weight voting* berdasarkan jarak.

$$W(i) = validitas\ (i) * \frac{1}{d+\alpha}\ , \qquad (2.4)$$

Dimana :

W(i) : *Weight voting*

Validitas (i) : Nilai Validitas

d : Jarak Euclidean

berikut adalah langkah-langkah dalam proses *Modified K-Nearest Neighbour* (kumalasari, 2014):

1. Menentukan nilai *k* tetangga terdekat
2. Menghitung validitas data latih
3. Menghitung jarak
4. Menghitung *weighted voting*
5. Menentukan kelas data uji

1.7. Evaluasi

Proses evaluasi merupakan proses untuk mengukur bagaimana sistem mampu bekerja dengan benar sesuai dengan kebutuhan. Dalam penelitian ini menggunakan akurasi sebagai proses evaluasi nya. Nilai akurasi dapat dihitung menggunakan rumus sebagai berikut

$$Akurasi = \frac{\sum data\ uji\ benar}{\sum data\ uji} * 100\% \quad (2.5)$$