BAB 1 PENDAHULUAN

1.1 Latar Belakang Masalah

Twitter merupakan layanan jejaring sosial dan *microblogging* yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter, seiring dengan banyaknya pengguna Twitter, Twitter menjadi salah satu media penyebar informasi yang sangat cepat, informasi yang beredar dapat menjadi bahan analisis untuk mengidentifikasi kecenderungan terhadap suatu objek, hal ini menjadi bahan pertimbangan baik perorangan atau organisasi untuk melakukan analisa data melalui Twitter, salah satu analisa data yang dapat dilakukan adalah proses *data mining.*

Data mining merupakan proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang bermanfaat yang tersimpan di dalam database besar (Turban et al, 2005). Salah satu cabang *data mining* adalah *Text mining*. Untuk melakukan proses *Text mining* data tidak langsung diproses karena teks yang akan dilakukan proses *Text mining* pada umumnya memiliki beberapa karakteristik diantaranya adalah memiliki dimensi yang tinggi, terdapat *noise* pada data, dan terdapat struktur teks yang tidak baik, maka ada beberapa tahapan proses sebelum *Text mining* salah satunya adalah *Preprocessing*, tahapan *Preprocessing* sangat penting dalam data mining karena *Preprocessing* membantu kinerja *data mining,* namun sebagian orang terkadang tidak terfokus pada proses ini dan lebih fokus terhadap algoritma yang diterapkan, semakin berkembangnya penelitian tentang *Social Network Analyzing* (SNA) seperti pada Twitter khususnya maka perlu adanya penelitian tentang beberapa faktor pendukung.

Dalam tugas akhir ini saya melakukan penelitian terhadap tahapan *Preprocessing* khususnya *stemming* dan *term weighting* pada data Twitter (twit pengguna) berbahasa Indonesia hingga data siap untuk dilakukan klasifikasi, hal ini di latar belakangi oleh karena belum adanya algoritma stemming yang memberikan tingkat akurasi hingga 100%, beberapa algoritma stemming yang sering dipakai saat ini seperti nazief adriani hanya mencapai akurasi 93% (Jelita Asian) dan untuk algoritma adaptasi yaitu algoritma porter hanya memeberikan akurasi sebesar 82%- 95% (Susastra Wiguna, 2013),sementara untuk tahapan penelitian term weighting hal ini guna untuk meningkatkan tahapan preprocessing agar hasil preprocessing data lebih maksimal, pada proses *Preprocessing* banyak tahapan yang harus dilakukan dan terdapat beberapa tools yang digunakan untuk memaksimalkan proses *stemming* tersebut salah satunya adalah dengan mengimplementasikan algoritma Fonetik dalam proses *stemming* pada *Preprocessing* Algoritma ini mampu memberikan kontribusi baik dalam proses *Preprocessing* khususnya *stemming* untuk data yang akan di klasifikasi.

1.2 Perumusan Masalah

Dalam Tugas Akhir ini, dirumuskan beberapa masalah sebagai berikut:

a. Bagaimana mengimplementasikan algoritma Fonetik Soundex dalam proses *Stemming* pada *Preprocessing* data twit berbahasa Indonesia?
b. Bagaimana kontribusi algoritma Fonetik soundex untuk proses *stemming* pada Twitter berbahasa Indonesia?
c. Algoritma apa yang cocok untuk pembobotan jika di pasangkan dengan algoritma Fonetik Soundex untuk *Text mining*?

1.3 Tujuan

Tujuan yang ingin dicapai dari pembuatan Tugas akhir adalah sebagai berikut:

a. Mengimplementasikan algoritma Fonetik Soundex pada proses *stemming* pada *Preprocessing* Twitter berbahasa Indonesia.
b. Menganalisis hasil *stemming* terhadap data yang telah dilakukan proses *stemming* dengan algoritma Soundex .
c. Melakukan pembobotan dengan beberapa metode *term weighting* untuk mengetahui pengaruh pembobotan pada proses klasifikasi serta mengetahui algoritma pembobotan yang tepat untuk *Preprocessing* Twitter.

1.4 Batasan Masalah

Batasan masalah dari Tugas akhir yang akan dibuat adalah:

a. Tugas akhir terfokus terhadap bagaimana implemenatasi soundex dalam *Preprocessing* khususnya *stemming* pada data Twitter berbahasa Indonesia dan sejauh mana kontribusi untuk memaksimalkan proses *stemming.*
b. Algoritma untuk proses klasifikasi menggunakan algoritma *Naïve bayes*
c. Algoritma pembobotan yang digunakan pada pengerjaan tugas akhir ini adalah *Term Frequency (TF), Feature Term Presence (TP), Term Frequency-Inverse DocumentFrequency (TF-IDF).*
d. Sebagai acuan pembanding digunakan algoritma *stemming* yang sering digunakan yaitu algoritma porter.
e. Pengujian hasil *stemming* dengan pengetahuan penulis serta sedikit bantuan kamus besar Bahasa Indonesia online karena data twit jarang menggunakan Bahasa baku
f. Data yang dibobotkan hanya data yang memiliki *value* di *lexicon*

1.5 Metodologi Penyelesaian

a. Studi Literatur

Langkah ini bertujuan untuk memahami dasar teori yang berhubungan dengan *Preprocessing* untuk membantu mendukung penyelasaian Tugas Akhir.

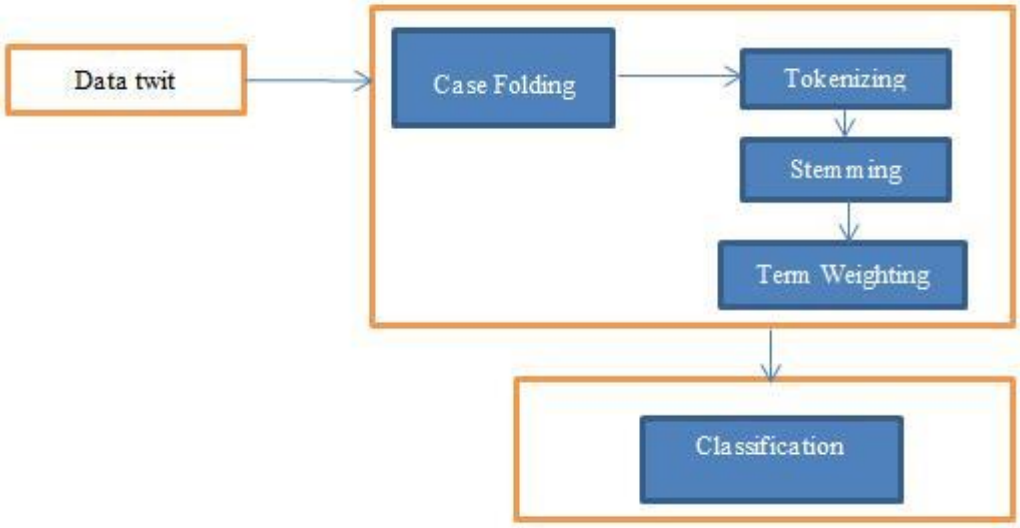
b. Pengumpulan data

Mengumpulkan data yang berasal dari pengguna Twitter

c. Analisis kebutuhan perancangan perangkat lunak

Analisis terhadap kebutuhan perangkat lunak agar memiliki gambaran umum seperti apa perangkat lunak yang akan di bangun, yang

selanjutnya merancang perangkat lunak yang sesuai dengan kebutuhan. Gambar 1.1 menjelaskan gambaran umum analisis perangkat lunak



Gambar 1.1 Perancangan Perangkat Lunak

d. *Case Folding*

Proses *Case Folding* adalah proses penghilangan tanda baca, karakter apapun selain huruf kemudian mengubah menjadi bentuk *lower case*, tujuan dari proses ini adalah untuk mempermudah dalam proses selanjutya.

e. *Tokenizing*

Proses *tokenization* adalah proses pemotongan string *input* berdasarkan tiap kata yang menyusunnya. Pemecahan kalimat menjadi kata - kata tunggal dilakukan dengan men-scan kalimat dengan pemisah seperti delimeter, *white space (spasi,tab dan newline)* . selanjutnya masuk ke tahap *stopword removal* (penghilangan kata stop*word*). Stopword removal merupakan teknik atau tahapan untuk menghapus kata-kata yang sering muncul dalam suatu dokumen teks tapi tidak memberikan arti yang penting pada teks tersebut selanjutnya data di berikan *value* dengan menggunakan database lexicon.

f. Proses *stemming*

Proses *stemming* adalah proses pengubahan setiap kata menjadi kata akarnya seperti “mengubah” menjadi “ubah”.Dalam tahap *stemming* ini di implementasikan algoritma Fonetik Soundex

g. Proses *Term weighting*

Proses *Term weighting* adalah proses pembobotan setiap kata dalam *Preprocessing.*

h. *Classification*

Dalam proses ini data yang telah dilakukan proses *Preprocessing* dilakukan klasifikasi untuk dilakukan analisis perbandingan performansi serta kecocokan pasangan algoritma Fonetik soundex dengan algoritma *term weighting.*

i. Implementasi

Mengimplementasikan analisis yang telah dilakukan dengan pembuatan sistem yang sesuai dengan kebutuhan dan perancangan yang telah dilakukan.

j. Pembuatan Laporan Tugas Akhir