Stemming pada Preprocessing Twit Berbahasa Indonesia dengan Mengimplementasikan Algoritma Fonetik Soundex untuk Proses Klasifikasi

Stemming in Indonesian Language Twit Preprocessing Implementing Phonetic Soundex Algorithm for Classification Process

Noviadrianti[1]

[1]Prodi S1 Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Noviadrianti@gmail.com

Abstrak

Twitter merupakan layanan jejaring sosial dan *microblogging* yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter, dengan kemajuan yang sangat pesat Twitter menjadi objek analisis yang sangat baik untuk berbagai kepentingan, salah satu penelitian yang diminati saat ini terhadap sosial media Twitter adalah analisa sentiment dan *opinion mining*. Untuk melakukan *opinion mining* terhadap Twitter menjadi kendala sendiri karena keterbatasan data Twitter dari twit pengguna yang hanya di batasi oleh 140 karakter, selain itu penelitian opinion *mining* biasanya hanya terfokus pada klasifikasi atau *clustering* data tetapi tidak banyak menjelaskan tahap *Preprocessing*, pada dasarnya *Preprocessing* yang baik akan menghasilkan proses *mining* yang baik juga, maka perlu berbagai cara untuk memaksimalkan proses *Preprocessing* pada Twitter salah satunya adalah dengan proses *stemming* dengan mengimplementasikan algoritma Soundex dimana algoritma ini diharapkan mampu memaksimalkan proses *stemming* pada *Preprocessing* untuk proses *mining* pada Twitter, selain itu metode ini akan di pasangkan dengan berbagai variasi algoritma pembobotan *Term Frequency (TF), Feature Term Presence (TP), Term Frequency-Inverse DocumentFrequency (TF-IDF)* untuk menemukan pasangan algoritma yang tepat untuk mendukung proses klasifikasi yang baik, klasifikasi dilakukan dengan metode *Naïve bayes* yang selanjutnya dapat di analisa bagaimana pengaruh algoritma soundex untuk *stemming* serta pengaruh algoritma pembobotan jika diterapkan pada proses klasifikasi, selain itu penelitian diharapkan mampu menghasilkan algoritma yang memberikan kontribusi yang baik untuk proses *stemming* data Twitter serta mempelajari bagaiman pengaruh algortitma pembobotan jika dipasangkan dengan algoritma soundex. Setalah dilakukan penelitian terhadap *stemming* dibandingkan hasil *stemming* algoritma soundex dengan porter maka didapatkan hasil untuk data uji sebanyak 300 twit bahwa soundex sedikit lebih unggul kemudian diklasifikasikan data hasil *stemming* dengan soundex dengan beberapa algoritma pembobotan didapatkan hasil nilai akurasi yang sama , berdasarkan analisis didapatkan bahwa algoritma pembobotan tidak berpengaruh kepada hasil klasifikasi.

Kata kunci : Twitter, Algoritma Soundex,Stemming, Preprocessing, Naïve bayes

1. Pendahuluan

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, 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 diharapkan mampu memberikan kontribusi baik dalam proses *Preprocessing* khususnya *stemming* untuk data yang akan di klasifikasi.

2. Dasar Teori /Material dan Metodologi/perancangan

2.1 Data Mining

Data mining merupakan disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data (Han and Kamber, 2006). Data mining sering juga disebut knowledge discovery in database (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan (Santosa, 2007).

2.2 Algoritma Fonetik Soundex

Algoritma soundex pertama kali dipatenkan oleh Margaret O'Dell and Robert C. Russell pada tahun 1918 (Donal, 1973) .Berdasarkan namanya, prosedur Soundex ini merupakan prosedur pencocokan string berdasarkan suara dari pelafalan kata. Soundex merepresentasikan kelas -kelas dari suara yang dapat diucapkan bersama – sama dari suatu tulisan pada bahasa tertentu. Berikut pembuatan kode soundex :

1. Sebuah string nama dimasukkan, huruf pertama akan diambil.
2. Berikan kode fonetis untuk masing-masing huruf
3. Jika terdapat dua atau lebih huruf berurutan dengan nomor (kode fonetis) yang sama, maka ambil salah satunya saja

2.3 Stemming

Menurut Talla (2013,p7), Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata. Dengan menghilangkan semua imbuhan (afixes) baik yang terdiri dari awalan (prefixes), sisispan(infixes), akhiran(suffixes), dan confixes (kombinasi dari awalan dan akhiran) pada kata turunan. Stemming digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar kata tersebut sesuai dengan struktur morfologi bahasa Indonesia yang baik dan benar.

2.4 *Term weighting*

Setiap dokumen mengandung beberapa kata yang berbeda-beda. Hal yang perlu diperhatikan dalam pencarian informasi dari koleksi dokumen yang heterogen adalah pembobotan kata, karena setiap kata memiliki tingkat kepentingan yang berbeda dalam dokumen. Oleh karena itu diberikan sebuah bobot *term* (Manning, et al. 2009). Ada beberapa metode pembobotan yang akan digunakan dalam tugas akhir diantaranya yaitu:

1. *Term Frequency* (TF)
   metode ini menghitung berapa kali jumlah kemunculan fitur fi dalam sebuah dokumen , misal sebuah kata muncul dua kali maka bobotnya adalah 2.
2. *Feature Term Presence* (TP)
   metode ini menunujukan ada atau tidaknya fitur fi di dalam dokumen jika ada maka nilai nya 1 jika tidak ada maka nilainya 0.
3. *Term Frequency-Inverse DocumentFrequency* (TF-IDF).
   Algoritma TF-IDF adalah algoritma yang berdasarkan nilai statistik kemunculan kata – kata dalam dokumen. TF (*Term Frequency*) menyatakan banyaknya suatu kata muncul dalam dokumen maupun kalimat. DF (*DocumentFrequency*) adalah banyaknya dokumen yang mengandung satu kata dalam segmen publikasi. TF IDF adalah nilai bobot dari hasil nilai kata TF dan nilai inverse IDF yang didefinisikan (Feldman, et al. 2007) :

$$IDF(w) = \log\left(\frac{N}{DF(W)}\right)$$

$$\text{TFIIDF}(w,d) = \text{TF}(w,d) \text{x IDF}(w)$$

Keterangan:

TF-IDF(w,d) : Bobot suatu kata dalam keseluruhan dokumen
w : Suatu kata (*word*)
d : Suatu dokumen (document)
TF(w,d) : Frekuensi kemunculan sebuah kata w dalam dokumen d
IDF(w) : *Inverse* DF dari kata w

2.5 Preprocessing

Preprocessing melibatkan melakukan beberapa tugas dengan tujuan menciptakan data yang akan digunakan sebagai masukan dalam fase *data mining* (Romer, et al. 2006). *Preprocessing* pada tugas akhir ini akan melibatkan beberapa langkah agar data dapat diproses selanjutnya oleh algoritma *Text mining*, proses ini dapat melibatkan beberapa tahapan diantaranya *Case Folding*, *Tokenizing*, Proses *stemming*, Proses *Term weighting*.

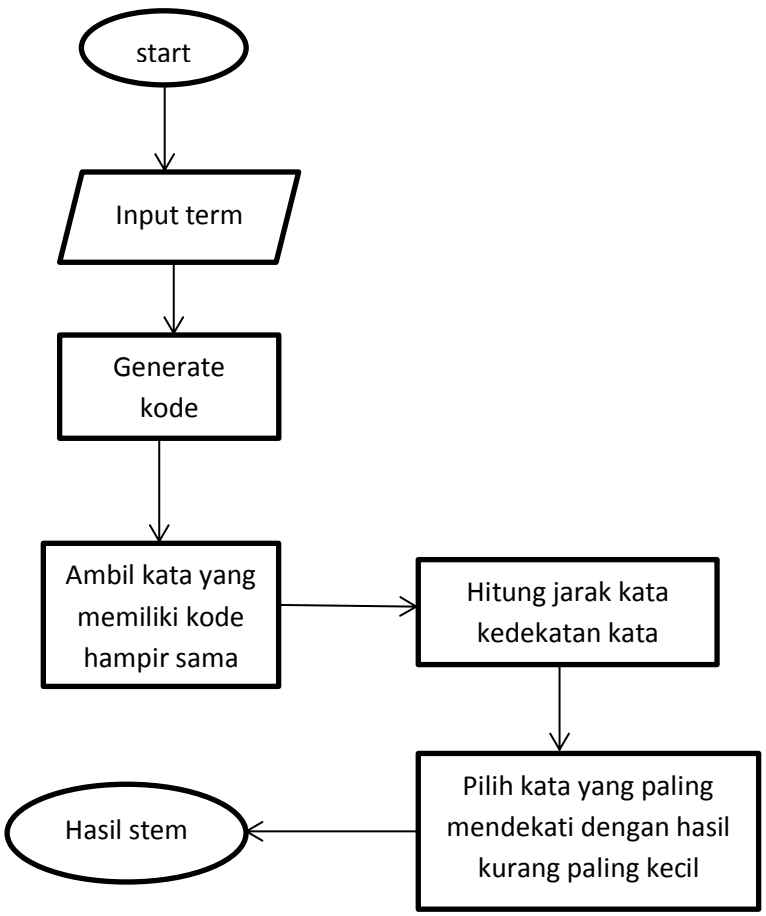
3.Pembahasan

3.1 Dataset

Dataset menggunakan dataset yang berasal dari jejaring sosial Twitter terdiri dari 300 twit pengguna dengan topik yang mengandung komentar terhadap suatu kota. Untuk mendapatkan data tersebut melalui proses pengambilan secara offline kemudian di kumpulkan dalam dokumen excel selanjutnya *file* yang sudah berisi kumpulan twit di *upload* kedalam sistem yang selanjutnya akan melakukan proses *Preprocessing*

3.2 Implementasi

Bagaimana mengimplementasikan algortima fonetik soundex untuk proses stemming pada data twitter yang dilakukan preprocessing, serta bagaiman menganalisa pengaruh algoritma pembobotan pada hasil stemming dalam hal ini penulis mencoba mengimplementasikan bagaimana melakukan proses stemming berikut flowchart dalam melakukan stemming twit berbahasa Indonesia dengan mengimplementasikan algoritma fonetik soundex.



Gambar 3.0.1 Alur proses stemming

Pada Gambar 3. 1 dijelaskan tahapan *stemming* dengan menggunakan algoritma fonetik soundex, hingga *term* hasil t*okenizing* melalui dapat di proses menjadi akar kata yang sesuai , pada tahap ke empat dimana pengambilan kata yang memiliki kode fonetik yang sama menggunakan database kata dasar berbahasa Indonesia. Berikut ini contoh stemming yang dilakukan hasil pengujian sistem:

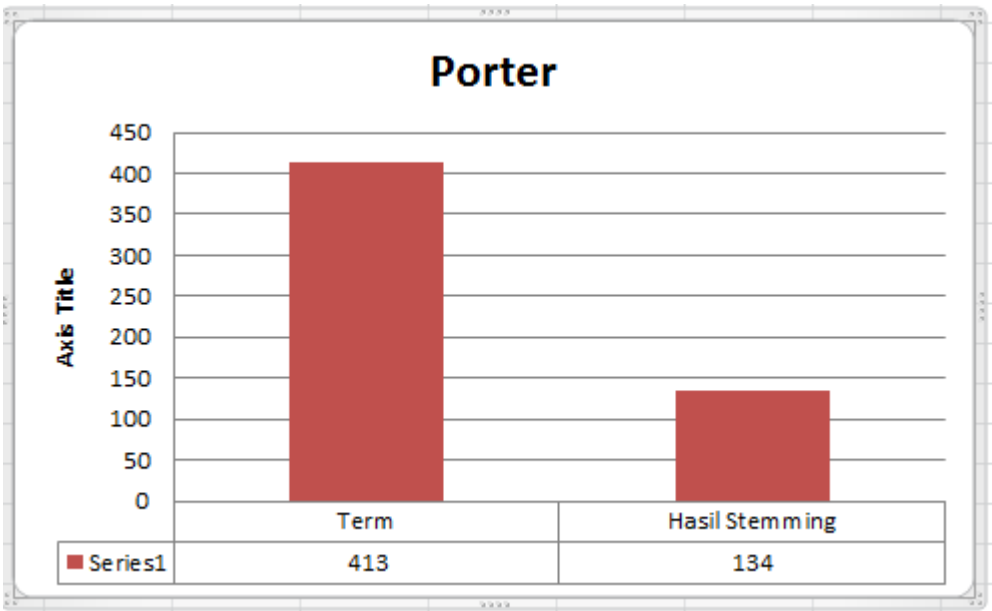
Tabel 3.0.1 Contoh perhitungan Stemming

Mengubah (8)	jarak
bah	5
ubah	4
uba	5

Maka kata stem yang diambil dari kata "mengubah" hasil dari perhitungan adalah "ubah", selanjutnya data hasil stemming dengan soundex diuji dengan berbagai algoritma pembobotan untuk melihat pengaruh algoritma pembobotan.

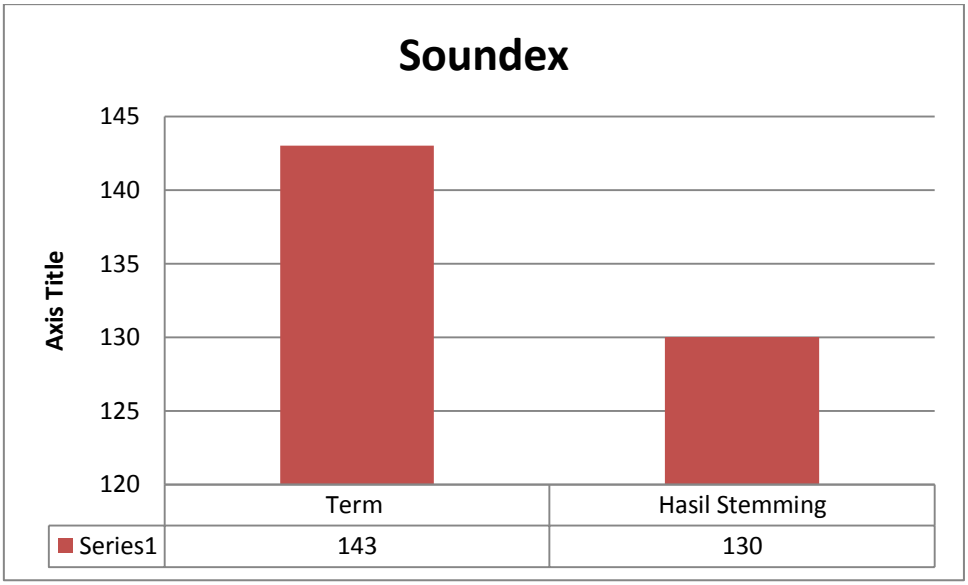
3.3 Pengujian

Pada pengujian ini digunakan data yang telah dilakukan proses *stemming* dengan implementasi algoritma soundex serta algoritma porter sebagai algoritma acuan (porter) pembanding, selain itu pengujian dilakukan perhitungan secara manual untuk mengetahui akurasi *stemming* yang dilakukan oleh kedua algoritma tersebut, yaitu algoritma soundex dan algoritma pembanding, berikut hasil Pengujian pada proses *stemming*. Berikut term yang dihasilkan masing – masing algoritma dijelaskan oleh Gambar 3.3 dan 3.4.



Gambar 3.2 Hasil stemming Porter

Data yang berasal dari 300 twit pengguna menghasilkan jumlah keseluruhan 413 *term* namun setelah melalui proses *stemming* data berkurang menjadi 134 untuk algoritma porter dan 130 untuk algoritma soundex nilai ini berbeda karena keterbatasan algoritma dalam melakukan *stemming*.



Gambar 3.3 Hasil stemming Soundex

Pengujian terhadap pengaruh data dengan berbagai macam algoritma pembobotan dengan melihat berbagai nilai yang berpengaruh terhadap klasifikasi seperti *precision, recall* dan akurasi untuk melihat pengaruh algoritma pembobotan. Untuk setiap jenis dokumen yang menggunakan algoritma pembobotan yang berbeda – beda, dari hasil klasifikasi ketiga dokumen didapatkan nilai yang sama, meskipun pembobotan berbeda setiap dokumen namun hasil menunjukan angka yang hampir sama baik itu *precission, recall* dan akurasi seperti dijelaskan Tabel 3.2 yang menjelaskan hasil klasifikasi twit positif sebanyak 146 dengan berbagai algoritma pembobotan.

Tabel 3.2 Hasil uji pengaruh Algoritma Pembobotan

	Recall	***Precision***	***Accuracy***
TF	0,774	0,926	113
TP	0,781	0,912	114
TFIDF	0,774	0,926	113
Rata-Rata	0,77633	0,9213	113.3333

4. Kesimpulan dan Saran

4.1 Kesimpulan

Dari hasil Pengujian sistem serta analisis hasil uji Maka dapat diambil kesimpulan bahwa :

1. Algoritma Soundex bisa menjadi alternative untuk melakukan *stemming Preprocessing* dengan tingkat akurasi pembobotan yang cukup tinggi
2. Algoritma pembobotan tidak terlalu berpengaruh dalam proses klasifikasi data mining sehingga untuk data Twitter pembobotan bisa dilakukan dengan menggunakan algoritma pembobotan yang diinginkan.

4.2 Saran

Berdasarkan hasil pengujian dan analisa terhadap hasil uji maka Penulis memberikan saran sebagai berikut

1. Untuk menyempurnakan *stemming* algoritma soundek ada baiknya algoritma di kombinasikan dengan algoritma penghilangan imbuhan karna akan jarak antar jumlah *term* semakin sedikit sehingga membantu mengurangi kesalahan yang mungkin terjadi
2. Algoritma soundex sangat bergantung pada database katadasar Bahasa Indonesia sehingga disarankan database kata dasar sudah diakui KBBI untuk meminimalisasi kesalahan pengambilan kata

3. Diharapkan ada penelitian – penelitian selanjutnya mengenai algoritma *stemming* untuk Bahasa Indonesia untuk mendapatkan tingkat akurasi *stemming* mencapai serratus persen.

Daftar Pusataka

Donal, Knuth. 1973. *The Art Of Computer Programming*, vol. 3: Sorting And Searching. Addison Wesley.

Feldman, R. & Sanger, J. 2007. *The Text mining Handbook*. New York: Cambridge University Press.

Han, J. dan Kamber, M. (2006), Data mining: Concepts and techniques (2nd ed,), Elsevier Inc.

Talla, Fadillah Z.A *Study of Stemming Effects on Information*,2003

Manning, Christoper D., Raghavan, Prabhakar, & Schutze Hinrich. 2009. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.