Preprocessing Text Mining Pada *Email Box* Berbahasa Indonesia

Gusti Ngurah Mega Nata[1)], Putu Pande Yudiastra[2)]
STMIK STIKOM Bali[1,2]
Jl. Raya Puputan No.86 Renon, telp: (0361) 244445/fax. (0361) 264773
e-mail: mega@stikom-bali.ac.id[1)], yudiastra87@gmail.com[2)]

Abstrak

Knowleage discovery atau menemukan pengetahuan dari client / customer perusahaan dapat kita dapatkan dari email yang pernah mereka kirim ke inbox email perusahaan. Informasi dari email box perusahaan dapat digunakan untuk email marketing. Teknik text mining sangat memungkinkan untuk mengekstrak teks email box. Namun, Text-Preprocessing pada document email sedikit berbeda dari dokumen berita, dimana dokument email memiliki kontak email pengirim, subjek, konten dan berkas (attacement). Selain itu, jumlah email dari satu orang pengirim bisa lebih dari satu dan saling berkaitan. Untuk mendapatkan informasi yang lengkap dari seorang pengirim maka semua email pengirim harus dijadikan satu dokumen dan kemudian diproses dengan teknik text mining. Pada paper ini dilakukan analisis text preprocessing pada email box berbahasa Indonesia untuk mendapatkan informasi yang berguna berdasarkan isi konten dari semua email yang pernah dikirm ke perusahaan kita. Tujuan dari panelitian ini adalah menemukan teknik text preprocessing pada email box berbahasa Indonesia dan sebagai dasar dari pengembangan knowleage discovery pada email box sebagai penunjang email marketing. Proses text preprocessing yang akan diajukan yaitu pertama semua email dari pengirim yang sama digabungkan menjadi satu dokumen kemudian dilakukan parsing / tokenizing, stopword removal dan stemming. Algoritma stemming yang digunakan yaitu porter stemming berbahasa Indonesia. Dari hasil analisis diketahui pengabungan semua email dari pengirim yang sama sebelum dilakukan text mining akan membuat informasi yang didapat dari seorang pengirim lebih banyak dan tidak parsial dalam banyak dokumen email.

Kata kunci: email box, *text mining*, porter *stemmer* Bahasa Indonesia, *text preprocessing.*

1. Pendahuluan

Electronok-mail merupakan media pengiriman pesan yang murah dan cepat menggunakan media internet. Pemanfaatan email pada era sekarang tidak lagi sebatas mengirim pesan tapi sudah menjadi teknik marketing dalam promosi produk. Email marketing biasanya dikirim oleh perusahaan dalam mempromosikan produk mereka ke setiap client atau member mereka. Pengiriman email yang berlebihan serta tidak sesuai dengan minat atau bisnis dari orang yang menerima akan mengabaikan atau bahkan dianggap *spam* (email sampah) oleh penerima. Akibat dari hal tersebut, hubungan bisnis antara perusahaan dengan client atau calon client menjadi tidak terjalin dengan baik. Maka, sebelum mengirim email promosi produk akan lebih baik jika bagian marketing mencari tahu minat dan beberapa informasi penting tentang penerima.

Salah satu sumber informasi penting untuk mengetahui minat dari penerima email adalah email-email yang pernah dikirim ke inbox perusahaan. Untuk mendapatkan informasi tersebut diperlukan teknik *text mining*[5][6]. Teknik *text mining* sudah sering digunakan untuk melakukan analisis dokumen teks, seperti dokumen berita[4]. Namun, document email sedikit berbeda dari dokumen berita, dimana dokument email memiliki kontak email pengirim, subjek, konten dan berkas (*attacement*). Konten email pada umumnya singkat dan saling berkaitan dengan isi attacement dan subjek email. Informasi lebih lengkap biasanya pada *attacement* namun file tersebut formatnya bisa bermacam-macam maka, sangat sulit untuk melakukan ekstrasi pada file tersebut. Jadi, bagian email yang sangat memungkinkan diekstrak adalah bagian subjek dan bagian konten yang berupa teks. Ditambah lagi email dari satu pengirim bisa lebih dari satu, itu artinya akan ada banyak dokumen email untuk satu pengirim.

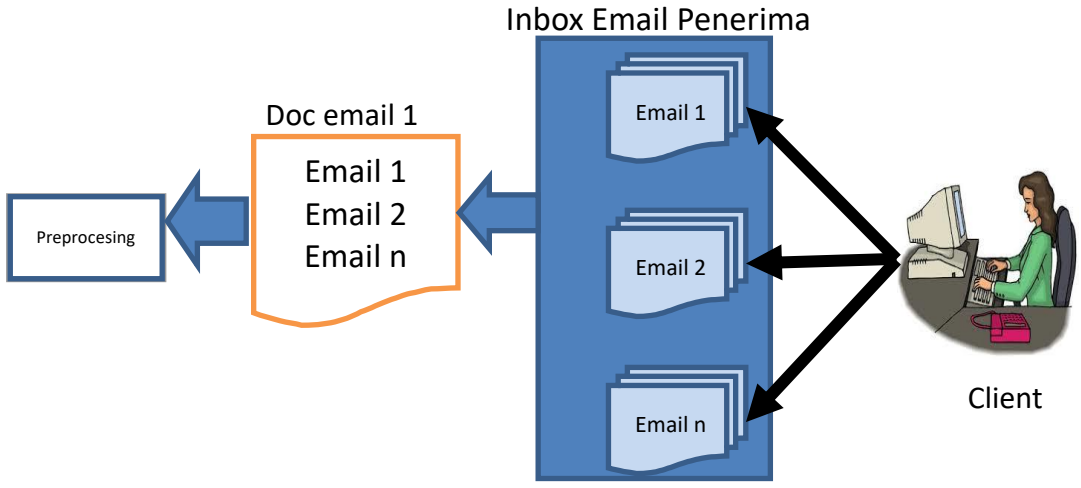
Dari penjelasan diatas, bahwa format dokumen email tidak sama dengan dokumen teks pada umunya, maka diperlukan *preprocessing* yang khusus untuk dokumen email. Pada paper ini dilakukan analisis *text preprocessing* pada email box dimana email dari sumber pengirim yang sama akan digabung

menjadi satu dokumen. Pengabubungan email-email dari satu sumber adalah untuk mendapatkan informasi yang lebih lengkap dan lebih jelas dibandingkan dengan tidak menggabungkan email dari sumber yang sama. Setelah email dari satu sumber digabung maka dilakukan proses *text preprocessing* untuk *text mining*. *Text-preprocessing* dilakukan karena email adalah data *unstructured* yang harus dirubah dahulu menjadi terstuktur sebelum dianalis lebih lanjut. Bentuk *unstructured* dari teks adalah kata – kata yang terurut (*sequence*) dan menyambung dalam sebuah dokumen. *Teks processing* diawali dengan memotong setiap kata yang ada dalam teks tersebut menjadi perkata. Process pemotongan teks menjadi kata – kata terpisah disebut *parsing / tokenizing* [6,7]. Setelah proses *tokenizing* setiap kata menjadi berdiri sendiri / tidak terikat dengan kata yang lain. Akibat dari pemisahlan kata tersebut, akan ada kata yang tidak memiliki arti yang relevan untuk menentukan ciri dari dokumen yang di *tokenizing* seperti "ini, itu, adalah, dan, atau" dan bayak lagi kata – kata sejenis [3]. Kata – kata yang tidak memiliki arti yang relevan tersebut disebut *stop word* [2,3]. Kumpulan dari *stop word* disebut *stop list* dan proses untuk menghapus *stop word* dalam dokumen disebut *stopword removal* [2]. Proses yang sulit dari *text-preprocessing* adalah proses *stemming*. Algoritma *Stemming* atau *tool stemmer* untuk Bahasa Indonesia sudah banyak dikembangkan diantaranya: Nazief dan Adriani dari Universitas Indonesia[2], Vega dari Univertas nasional singapura tahun 2001 [1], Arifin dan setiono dari Institut teknologi sepuluh November 2002 [1], *Porter Stemmer for Bahasa Indonesia* dikembangkan oleh Fadillah Z. Tala pada tahun 2003[3]. Namun pada penelitian ini Algoritma stemming yang digunakan yaitu porter *stemming* berbahasa Indonesia. Penggunakan algoritma porter stemmer karena proses menemukan kata dasar lebih cepat dan tidak memerlukan list kata dasar [1].

2. Metode Penelitian

A. Pengabungan Dokumen Email Pengirim

Dokument email memiliki kontak email pengirim, subjek, konten dan berkas (*attacement*). Konten satu email pada umumnya singkat sehingga sangat sedikit informasi yang didapat dari pengirim. Maka, Sebelum melakukan preprocessing semua dokumen email dari pengirim email yang sama dijadikan satu dokumen. Pengabungan semua dokumen email dari satu pengirim karena email – email tersebut memberikan satu informasi yang utuh untuk satu orang pengirim. Maka pada penelitian ini semua email yang pernah dikirim oleh satu alamat emai dijadikan satu dokumen teks email. Berikut adalah ilustrasi pengabungan email-email dari satu pengirim menjadi satu dokumen teks email.

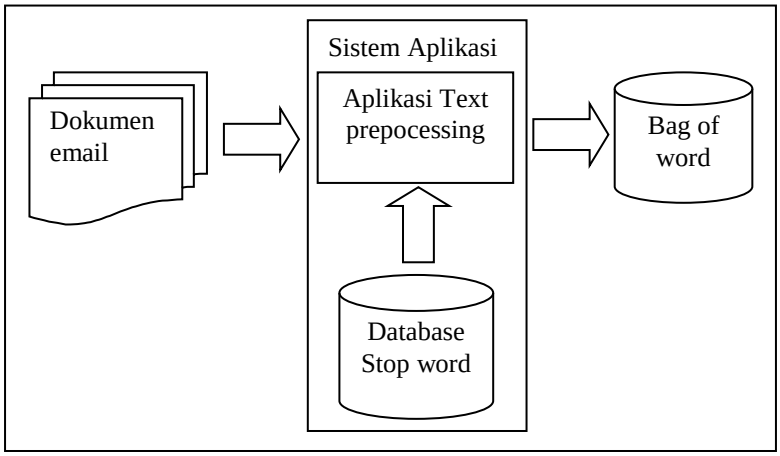


Gambar 1. Ilustrasi pengabungan dokumen email pengirim

B. Gambaran Umum Sistem

Gambaran umum sistem dari penelitian ini digambarkan dengan block diagram. Dalam gambaran umum sistem ini terdapat inputan, proses dan output dari sistem. Inputan sistem berupa dokumen email, proses yaitu sistem aplikasi text-*preprocessing* yang terdiri dari aplikasi dan database

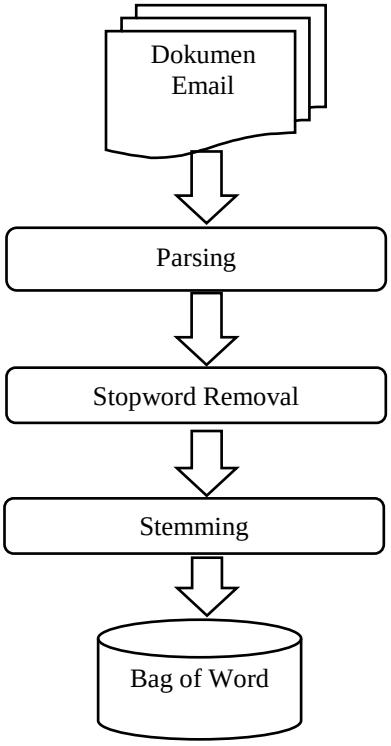
stop word, dan output yang dihasilkan berupa *bag of word*. *Bag of word* yang dihasilkan berupa kata dasar yang disimpan dalam tabel database relasional.



Gambar 2. Gambaran Umum Sistem *Text-Processing*

C. Alur penelitian

Metode penelitian ini fokus pada proses *text pre-processing* pada email box berbahasa Indonesia. *Text-Preprocessing* diawali dengan ditentukannya dokumen email yang akan diproses, kemudian proses *parsing/tokenizing* dilakukan hasil dari *parsing* berupa kumpulan kata yang telah terpisah dari kalimat. Kata-kata tersebut disimpan pada memori untuk selanjutnya dilakukan penghapusan kata – kata yang tidak memiliki makna seperti kata penghubung atau pelengkap kalimat yang disebut proses *stopword removal*. Langkah terahir dari *text-prprocessing* yaitu *stemming*. Proses *stemming* menggunakan algoritma *porter stemmer* Bahasa Indonesia. Hasil dari stemmer tersebut berupa kumpulan kata dasar yang terdaapat dalam dokumen email yang disebut *bag of word*.



Gambar 3. Proses text preprocessing

Penjelasan pada masing – masing tahapan dari proses *text preprocessing* diatas adalah sebagai berikut:

1. *Parsing / Tokenizing*

Teks adalah data unstructured yang harus dirubah dahulu menjadi terstuktur sebelum dianalis lebih lanjut. Teks email dimasukan kedalam aplikasi yang disimpan kedalam array 1 dimensi. Kata – kata dalam kalimat dibagi berdasarkan spasi.

2. *Stopword removal*

Setelah proses *tokenizing* setiap kata menjadi berdiri sendiri / tidak terikat dengan kata yang lain. Akibat dari pemisahlan kata tersebut, akan ada kata yang tidak memiliki arti yang relevan untuk menentukan ciri dari dokumen yang di *tokenizing* seperti "*ini, itu, adalah, dan, atau*" dan bayak lagi kata – kata sejenis. Kata – kata yang tidak memiliki arti yang relevan tersebut disebut *stop word.* Kumpulan dari stop word disebut stop list dan proses untuk menghapus *stop word* dalam dokumen disebut *stopword removal.*

3. *Stemming*

Stemming adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya. Pada penelitian ini algoritma *stemmer* yang digunakan yaitu algoritma *porter stemmer* Bahasa Indonesia. Algoritma *porter stemmer* Bahasa Indonesia merupakan algoritma stemmer yang tidak menggunakan kamus kata dasar. Teknik algoritma porter Bahasa Indonesia dalam mencari kata dasar yaitu dengan melihat dan mengapus *Affixes* dari Bahasa Indonesia yang terdiri dari *Sufiks, prefix* dan konfiks [3]. Sehingga algoritma porter stemmer memiliki kecepatan yang lebih dibandingkan algoritma stemmer yang menggunakan kamus kata dasar seperti yang dikembangkan oleh Nazief dan Adriani dari Universitas Indonesia pada tahun 1996 dan 2007.

Dalam dokumen Bahasa Indonesia proses *stemming* sangat diperlukan sebelum proses *text mining* karena Bahasa Indonesia memiliki *prefixes, suffixes, infexes* dan *confixes* yang membuat suatu kata dasar dapat berubah manjadi banyak bentuk dan akibatnya membuat pencarian kata dasar menjadi sulit [1]. Berikut adalah arti dan contoh dari imbuhan dalam Bahasa Indonesia [3]:

a. *Sufiks* (Akhiran) adalah afiks yang ditambahkan pada bagian belakang kata dasar, misal "*-an, -kan,"* dan "*–i"*;

b. Prefiks (Awalan) adalah imbuhan yang ditambahkan pada bagian awal sebuah kata dasar atau bentuk dasar; awalan: *"per-"* adalah yang paling *produktif dalam bahasa Indonesia*

c. Konfiks (sifiks dan prefiks)afiks tunggal yang terjadi dari dua unsur yang terpisah (misal "*ke-...-an"* dalam kata "*kemerdekaan"*)

3. Hasil dan Pembahasan

Analisis yang dilakukan hanya pada hasil proses *text-preprocessing* dokumen email menjadi data tertruktur. Data terstruktur yang dihasilkan berupa tabel kata dasar dari dokumen email yang dianalisis. Hasil dari pengabungan semua email dari satu pengirim membuat proses pencarian informasi pengirim menjadi lebih singkat dan lebih mudah untuk dipahami. Jumlah dokumen yang diproses menjadi lebih sedikit karena pengabungan dokumen email dari sumber yang sama dimana sebelum digabung jumlah email yang dimiliki terdapat 85 email dari 11 pengirim kemudian email-email tersebut digabung berdasarkan pengirimnya dan menjadi 11 dokumen email. Walaupun, gabungan dari email tersebut berisi semua komunikasi dari pengirim Namun, informasi yang didapat adalah informasi sepihak dimana informasi penerima berada pada folder outbox yang tidak digabungan dalam satu file dikumen. Dari hasil *text-preprocessing* seperti *parsing, stop word removal* dan *stemming*, tidak ada perbedaan yang signifikan dengan proses *text-preprocessing* seperti pada *text-preprocessing* dokumen teks pada umumnya. Perbedaan yang ada terdapat pada *list of word*, *list of word* dari dokumen email perlu ditambahkan seperti

kata-kata salam sambutan, kepada, dan kata-kata salam perpisahan karena kata-kata tersebut kurang memberikan informasi dari seorang pengirim email.

Dari tahap *text-preprocessing*, proses stemming adalah yang paling penting dalam menemukan kata dasar. karena bagian dokumen email yang dilakukan proses stemming hanya pada bagian subjek dan bagian konten sehingga informasi yang didapat tidak lengkap. Namun pengabungan konten dari pengirim membuat informasi yang didapat sedikit lebih banyak dan tidak parsial dalam banyak dokumen email. *Bag of word* dari dokumen email pengirim tersebut dapat digunakan sebagai dasar analisis *data mining*, dokumen *classtering*, atau proses pencarian dokumen email.

4. Simpulan

1. Hasil analisis diketahui pengabungan semua email dari pengirim yang sama sebelum dilakukan *text mining* akan membuat informasi yang didapat dari seorang pengirim lebih banyak dan tidak *parsial* dalam banyak dokumen email.
2. Bagian dokumen email yang sangat memungkinkan dilakukan *text mining* adalah bagian subjek dan bagian konten email yang berupa teks.
3. Aplikasi *text preprocessing* yang dibangun hanya dapat mengasilkan kumpulan kata dasar dari dokumen email yang dimasukkan.
4. Algoritma *porter stemmer* bahasa Indonesia dapat digunakan untuk melakukan *stemming* pada dokumen email berbahasa Indonesia.
5. Kumpulan kata dasar (*bag of word*) dari dokumen email dapat digunakan sebagai dasar dokumen *classtering*, pencarian dokumen, atau analisis *data mining* lebih lanjut.

Daftar Pustaka

[1] Asian, J., Williams, H. E., Tahaghoghi, S.M.M. *Stemming Indonesian*. Australian Computer Society Inc. 2005.

[2] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M.M., Williams, H.E. *Stemming Indonesian : A Confix-Stripping Approach*. Transaction on Asian Langeage Information Processing. 2007. Vol. 6, No. 4, Artikel 13. Association for Computing Machinery : New York

[3] Fadillah Z. Tala. *A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia*. Netherland, Universiteit van Amsterdam, 2002, http://ucrel.lancs.ac.uk/acl/P/P00/P00-1075.pdf

[4] Bambang kurniawan, syahril effendi, opim. Klasifikasi konten berita dengan metode text mining. *Jurnal Dunia Teknologi Informasi*.2012. Vol.1, No.1, Hal 14-19.

[5] M.Sukanya, S. Biruntha. *Techniques on Text Mining*. International conference on advanced communication control and computing technologies (ICACCCT), 2012.

[6] Feldman, R & Sanger, J. The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press : New York. 2007.

[8] Berry, M.W. & Kogan, J. Text Mining Aplication and theory. WILEY : United Kingdom. 2010.