Knowledge Discovery Pada Email Box Sebagai Penunjang Email Marketing

Knowledge Discovery In The Email Box For Support Email Marketing

Gusti Ngurah Mega Nata [1)], Putu Pande Yudiastra [2)]
STMIK STIKOM BALI [1,2]
Jln Raya Pupuran-Renon, telp: (0361) 244445
e-mail: **mega@stikom-bali.ac.id, yudiastra87@gmail.com**

Abstrak

Memanfaatkan email untuk pemasaran adalah salah satu strategi marketing yang sangat efektif dan murah. Walaupun demikian, email marketing tidak boleh dikirim secara sembarangan dan terus – menerus. Pengiriman email yang berlebihan serta tidak sesuai dengan minat atau bisnis dari orang yang menerima akan diabaikan atau bahkan dianggap spam 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 berkaitan dengan penerima. Informasi tentang minat atau kesukaan dari client dapat ditemukan dari kumpulan email yang pernah mereka kirim ke inbox kita. Namun menemukan minat client dari dokumen email perlu teknik khusus dalam mengolah data teks. Maka dari itu, Pada penelitian ini fokus melakukan studi dan implementasi text mining pada dokumen email. Dokumen email dianalisis dengan memperhitungkan jumlah kata yang muncul pada email tersebut. Representasi teks yang digunakan yaitu single word. Single word tersebut kemudian menjadi masukan dari proses dokumen clustering dengan algoritma K-mean pada fase data mining. Hasil dari clustering inilah yang dapat digunakan oleh pihak marketing sebagai penunjung kegiatan promosi produk. Dari hasil perancangan, pengembangan serta pengujian aplikasi terhadap dokumen email didapatkan hasil yaitu sistem sudah dapat melakukan preprocessing teks menggunakan teknik parsing, stopword removal, dan stemming sehingga menghasilkan kumpulan kata dasar (bag of word). Pada proses pencarian minat dengan term yang dimasukkan, sistem sudah dapat menemukan email client dengan arah minat tertentu menggunakan teknik TF-IDF. Pada proses pengelompokkan dokumen email menggunakan algoritma K-mean clustering pada dokumen email yang sudah dilabelkan sebelumnya memberikan akurasi 63,63% dari 11 dokumen email yang sudah digabung dari setiap pengirim. Dari hasil pengujian sistem sudah dapat digunakan sebagai dasar pemilihan calon penerima email marketing sesuai dengan term / key word yang dimasukkan serta berdasarkan kemiripan dari isi email marketing.

Kata kunci: email marketing, email box, *preprocessing* email, *text mining*, K-mean

Abstract

Email marketing is one of the most effective and inexpensive marketing strategies. However, email marketing should not be sent indiscriminately. Excessive email marketing submissions, and inappropriate to the person receiving them will be ignored or even considered spam by the recipient. As a result, the business relationship between the company and the client or prospective client becomes unfavorable. So before sending a product promotion email it would be better if the marketing department finds out interest and some important information related to the recipient. Information about the interests or client information can be found from the collection of emails they have sent to the email inbox. However, finding the client interest in the email document needs to process a text document. Therefore, in this study focused on the study and implementation of text mining on email documents. The email document is analyzed by taking into account the number of words that appear in the email. The text representation used is bag of word. bag of word is then an input of the process of grouping documents with K-mean algorithm in the data mining phase. This clustering result can be used by marketing as promoter of product promotion activity. the results of text mining application testing on an email document whose system is capable of preprocessing text using parsing, stop-word removal, and stemming techniques resulting in a collection of basic words (word bags). the system can already find email clients with particular interest using the TF-IDF technique. In the process of grouping clients based on email

communications, the system can also group email documents using the K-mean clustering algorithm. From the results of testing the system can be used as the basis of the selection of prospective email marketing in accordance with the terms / keywords entered and based on similarity of email marketing content.

Key word : email marketing, email box, preprocessing email, text mining, K-mean

1. Pendahuluan

Meningkatnya pengguna internet di masyarakat telah mengubah cara komunikasi dan promosi dari perusahaan. Banyaknya layanan *online* untuk berkomunikasi antar pengguna internet adalah alasan kenapa internet sangat digemari. Salah satu layanan jasa komunikasi yaitu surat elektronik yang lebih dikenal dengan istilah e-mail (*electronic mail*). Dalam dunia bisnis email bukan hanya sebagai media komunikasi antar pelaku bisnis tapi juga sebagai media marketing. Dalam kegiatan marketing email digunakan sebagai media promosi suatu produk atau jasa secara langsung ke setiap calon *client* (maxmanroe.com). Memanfaatkan email *list* untuk pemasaran adalah salah satu strategi marketing yang sangat efektif karena penawaran bisa langsung masuk ke *inbox* email calon pelanggan dan dibaca oleh mereka. Selain itu, email marketing dianggap lebih efektif karena pemasaran dengan cara ini dapat menjangkau calon pelanggan yang potensial, ditambah lagi biasanya penjualan dengan menggunakan media email konversinya sangat baik (maxmanroe.com). Walaupun demikian, email marketing tidak boleh dikirim ke semua client atau calon client tanpa mengetahui minat atau ketertarikan penerima terhadap produk yang ditawarkan. Penawaran produk melalui email marketing kepada penerima yang tidak tertarik serta dilakukan berulang – ulang kali akan berimbas tidak baik terhadap hubungan bisnis antar client dengan perusahaan. Akibat lain yang juga berpengaruh dari pengiriman email marketing yang tidak sesuai dengan minat penerima yaitu setiap email marketing yang dikirim akan dianggap *spam* bahkan dapat diblok (*black list*) oleh penerima, dan hubungan dengan client atau calon *client* menjadi tidak terjalin dengan baik. Maka, upaya yang harus dilakukan bagian marketing sebelum mengirim email promosi harus tahu dulu minat dan bisnis yang dikelola setiap penerima email.

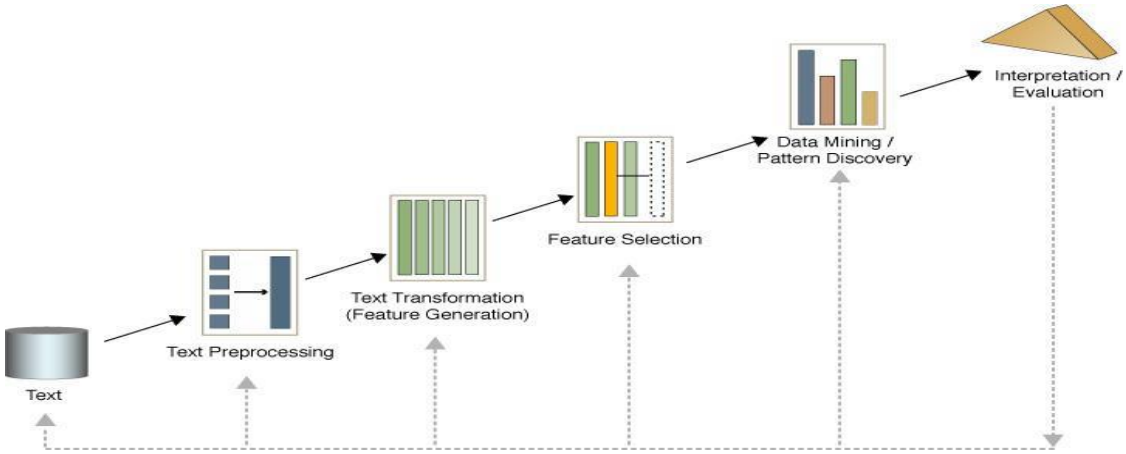
Mengetahui informasi minat dari calon penerima email akan sulit ketika jumlah *client* atau calon *client* yang akan dianalisis sangat banyak [1]. Teknik yang diusulkan yaitu dengan menggali email *box* yang pernah menerima email dari *client*, calon *client* atau *partner*. Maka dari itu, Pada penelitian ini fokus melakukan studi dan implementasi *text mining* pada email *box* yang pernah menerima email dari *client*, calon *client* atau partner sebagai penunjung kegiatan promosi produk melalui email marketing.

Proses *text mining* dari setiap email akan memunculkan kata – kata yang sering digunakan dalam email. Kata – kata tersebut tentunya akan memiliki arti pembicaraan apa saja yang pernah terjadi antara penerima email dengan seseorang pengirim email tersebut seperti membicarakan suatu produk, jasa, bisnis, keluhan, gaya hidup dan berbagai informasi lainnya. Kata – kata yang dieksplor akan dicari jumlah kemunculannya dalam dokumen email tersebut. Metode yang digunakan untuk menemukan jumlah penggunaan kata akan menggunakan TF-IDF yang sebelumnya sudah melewati proses *preprocessing* data dari teks yang sebelumnya tidak terstruktur menjadi bentuk terstruktur. Selain dilakukan perhitungan jumlah kata, juga dilakukan pengelompok dokumen email berdasarkan *bag of word* dari masing – masing dokumen email menggunakan algoritma *k-mean clustering*.

2. Tinjauan Pustaka

2.1. *Text Mining*

Text Mining adalah proses eksplorasi dan analisis data teks dalam jumlah besar untuk menggali pengetahuan baru baik secara otomatis maupun semi otomatis [8]. Pengetahuan baru yang dimaksud adalah informasi penting yang didapat dari hasil eksplorasi yang sebelumnya tidak diketahui.



Gambar 1. Tahapan *text mining* [8].

Tahapan dalam *data mining* ada 5 (lima) yaitu sebagai berikut [8].

1) *Text pre-processing*
Pada tahap ini dilakukan upaya pengolahan data text yang tidak terstruktur (*unstruktur*) menjadi terstruktur dan menghilangkan bagian - bagian yang tidak diperlukan dalam proses mining. Pada tahap *text* pre-prosessing juga terdapat beberapa tahapan untuk mendapatkan data yang bersih dan standar. Berikut adalah tahapan yang dilakukan pada phase text pre-processing [5].
1. *Parsing / Tokenizing*,
2. *Stopword* removal
3. *Stemming*.

2) *Text transformation*
Text transformation atau *feature generation* adalah tahapan untuk menghasilkan / membangkitkan *feature* dari data teks yang diolah. *Feature* adalah pola menarik yang dianggap dapat mewakili data sebenarnya. Feature merupakan bentuk representative dari data. Salah satu bentuk feature dari teks adalah bentuk *sequence of word*.

3) *Feature selection*
Seluruh *feature* yang dihasilkan harus relevan dan bermanfaat untuk proses – proses selanjutnya, sehingga pada tahap ini dilakukan seleksi untuk menghasilkan feature yang benar – benar bermanfaat.

4) *Data Mining*
Tahap *data mining* adalah tahap *mining* yang sebenarnya. Pada tahap ini akan dilakukan pencarian pola menarik dari sekumpulan data yang telah dibersihkan dan diseleksi sebelumnya.

5) *Interpretation / evaluation*
Hasil mining yang didapat pada tahap sebelumnya akan dievaluasi lebih lanjut.

2.2. Term Frequency dan Invers Document Frequency (TF-IDF)

Metode TF-IDF merupakan metode untuk menghitung jumlah bobot *term* / kata dalam dokumen teks. Metode tersebut sudah umum digunakan dalam proses *text mining* dan *information retrival.* Metode ini memiliki dua tahapan perhitungan yaitu pertama menghitung *Term Frequency* (TF) dan kedua menghitung *Inverse document Frequency* (IDF) pada setiap token (kata) pada setiap dokumen teks dalam korpus. Berikut adalah tahapan dan rumus dari masing – masing pembobotan:

1. **Term Frequency (TF),** merupakan perhitungan yang digunakan untuk mengetahui jumlah suatu term dalam suatu dokumen yang dianalisis dan kemudian dilogaritmik agar mengurangi nilai bilangan. Lagaritmik akan mengurangi digit bilangan. Rumus TF adalah sebagai berikut:

$$W_{t,d} = \begin{cases} 1+log_{10}\ tf_{t,d}, & if\ tf_{t,d}>0 \\ 0, & if\ tf_{t,d}=0 \end{cases}$$

2. **Inverse Document Frequency (IDF),** digunakan untuk mengurangi bobot suatu term (kata) jika kemunculanya banyak tersebar didalam dokumen. Dengan kata lain IDF adalah untuk mengetahui seberapa unik suatu term, dengan kata semakin banyak term tersebar dalam banyak dokumen maka nilai IDF semakin kecil dan begitu sebaliknya. Rumus dari IDF adalah sebagai berikut:

$$IDF(t) = \log (N/df(t))$$

N adalah jumlah dokumen dan df(t) adalah jumlah dokumen yang mengandung term yang bersangkutan. Gabungan dari kedua hasil perhitungan TF dan IDF akan memberikan diskripsi kecenderungan isi kata dalam dokumen dan tingkat keunikannya pada setiap dokumen. Berikut adalah pengabungan dari kedua rumus diatas:

$$TF\text{-}IDF(d.t)=TF(d.t) * IDF(t)$$

2.3. K-mean Clustering

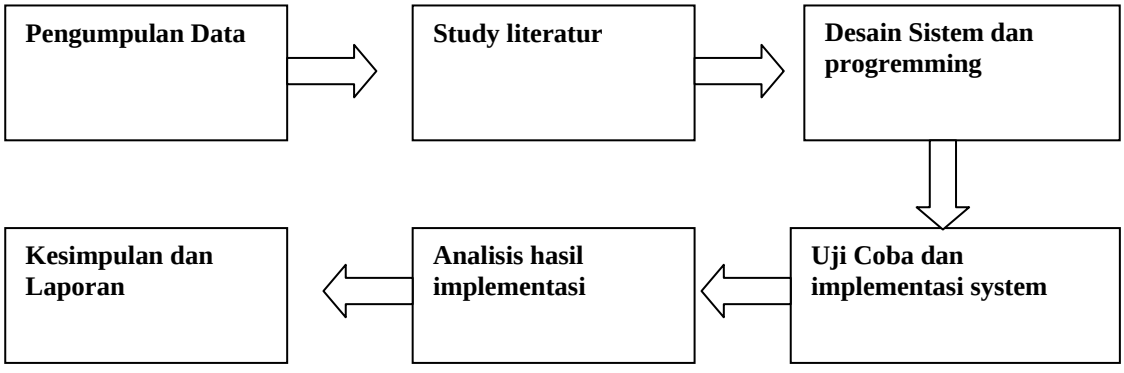
K-mean adalah salah satu metode *clustering non hirarki* yang berusaha mempartisi data yang ada ke dalam beberapa *cluster* [10]. Hasil dari pengklasteran data yaitu dimana data dalam satu klaster akan memiliki karakteristik yang mirib dan akan memilki karakteristik yang beda dengan klaster yang beda. Secara umum algoritma dasar dari k-means clustering adalah sebgai berikut:

1. Tentukan jumlah cluster
2. Tentukan titik pusat cluster secara acak
3. Hitung jarak pusar obyek dengan centroid
4. Perbaharui nilai titik centroid
5. Ulangi langkah 3 dan 4 sampai nilai dari titik centroid tidak lagi berubah

Peroses pengelompokkan data ke dalam suatu cluster dapat menggunakan perhitungan jarak seperti *minkowski* atau *Euclidean distance*.

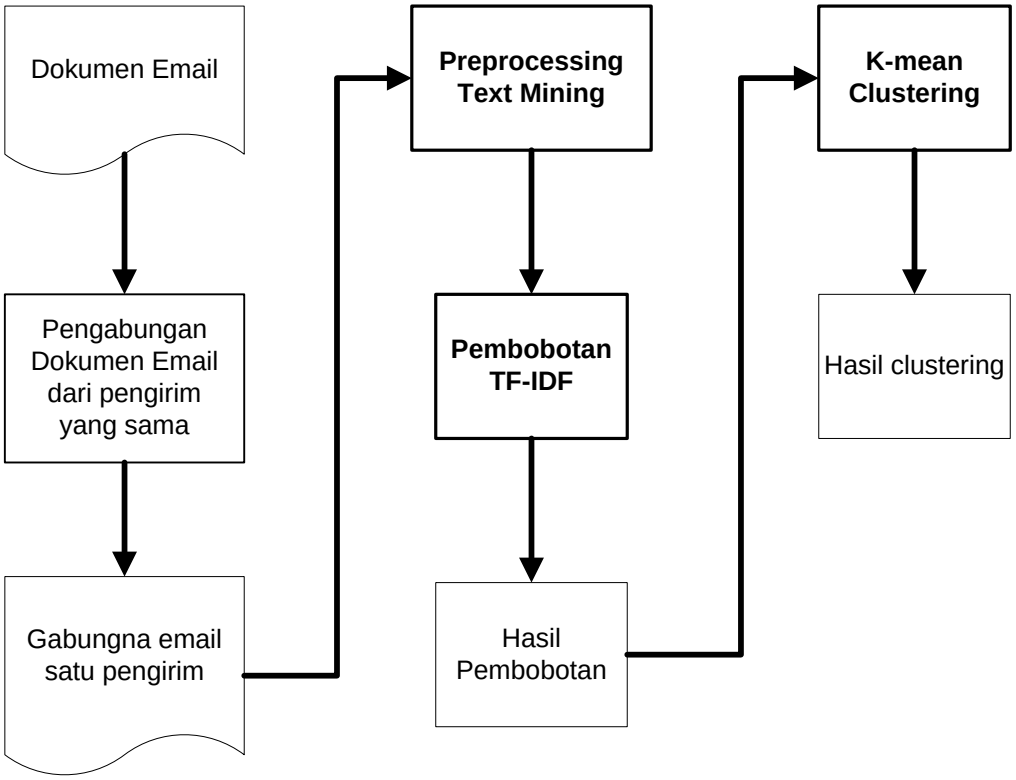
3. Metode Penelitian

Tahapan – tahapan penelitian terdiri dari enam bagian yaitu pengumpulan data, studi literalur, desain sistem dan *programming*, uji coba, analisis hasil, dan yang paling terahir adalah pembuatan kesimpulan dan pelaporan. Jika digambarkan dalam sebuah *flow cha*r maka seperti pada gambar berikut:



Gambar 2. Alur Penelitian

Perancangan design dari sistem *knowledge discovery* terdiri dari beberapa tahapan penting yaitu proses pengabungan dokumen email, preprocessing text mining, pembobotan TF-IDF serta dokumen *clustering* dengan algoritma k-mean. Secara alur proses sistem dapat dilihat pada gambar alur sistem berikut:



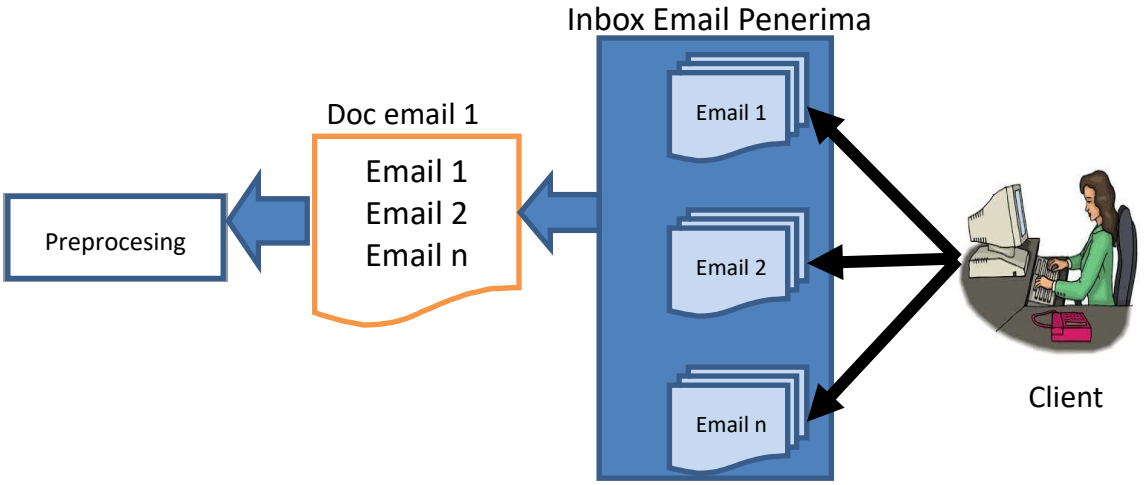
Gambar 3. Alur proses sistem

Proses dimulai dari mendownload terlebih dahulu email yang pernah diterima pada inbox email kemudian format email dirubah menjadi dalam format .txt. Dokumen Email yang telah didownload dari satu pengirim kemudian harus digabung menjadi satu dokumen email. Pengabungna email dari pengirim yang sama akan meningkatkan bobot dokumen untuk menjelaskan minat atau kecenderungan komunikasi dari pengirim dan penerima email. Gabungan email dari satu pengirim kemudian disimpan dalam format .txt yang selanjutnya menjadi inputan dalam proses *preprocessing text mining*. Dalam proses preprocessing text mining dokumen email tersebut merubah dokumen yang bersifat tidak terstruktur (*unstructured*) menjadi semi terstruktur dalam representasi teks yaitu *bag of word.* Dalam proses *preprocessing text mining* terdapat proses *parsing*, *stop word removal*, dan *stemming* menggunakan algoritma CS sehingga mengasilkan kumpulan kata dasar atau *bag of word*. Representasi teks *bag of word* kemudian menjadi inputan pada proses pembobotan TF-IDF untuk membuat indek kata dari semua dokumen email yang telah di *preprocessing* sebelumnya. Langkah terakhir dari proses aplikasi adalah pengelompokkan dokumen email dari setiap pengirim berdasarkan kemiriban kata dalam Indek kata dasar.berikut adalah penjelasan dari masing – masing tahapan dalam mengembangkan aplikasi *knowledge discovery* pada dokumen email.

A. Pengabungan Dokumen Email Pengirim

Dokumen email memiliki kontak email pengirim, subjek, isi atau konten dan berkas (*attacement*). Email pada dasarnya adalah komunikasi yang menggunakan tulisan, dimana pengirim dan penerima saling mengirim pesan dalam membahas suatu hal. Namun didalam inbox mail server data fisik email yang pernah kita terima dari satu pengirim dibuat terpisah. Jadi seolah-olah dokumen email dari satu pengirim berbeda komunikasi. Dalam penelitian ini email – email yang pernah dikirim oleh satu pengirim email akan digabung menjadi satu dokumen email. Teknik seperti ini pernah dilakukan pada penelitian sebelumnya yaitu pada paper [1] dan konsepnya digunakan pada penelitian ini. Seperti dijelaskan pada

paper tersebut bahwa pengabungan dokumen email dari satu pengirim karena email – email tersebut adalah komunikasi yang akan memberikan satu informasi yang utuh untuk satu orang pengirim [1]. Maka pada penelitian ini semua email yang pernah dikirim oleh satu alamat email dijadikan satu dokumen teks email. Berikut adalah ilustrasi pengabungan email-email dari satu pengirim menjadi satu dokumen teks email.

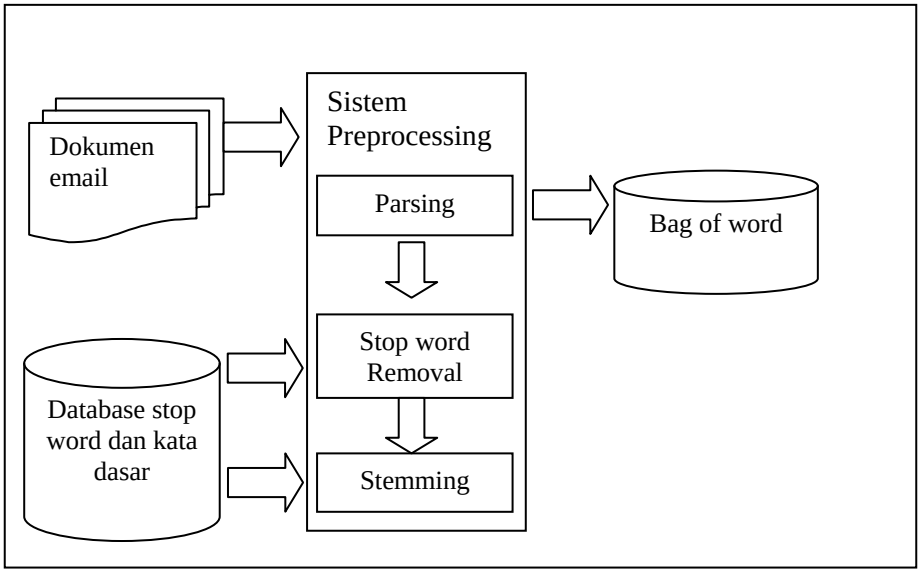


Gambar 4. Ilustrasi pengabungan dokumen email pengirim [1]

B. Preprocessing *text mining*

1) Gambaran Umum Sistem

Gambaran umum sistem dari penelitian ini digambarkan dengan block diagram. Dalam gambaran umum sistem ini terdapat inputan data dokumen dalam format .txt, proses *preprocessing* yang melakukan *parsing, stop word removal* dan *stemming*. Proses *preprocessing* terkoneksi dengan database karena *stop word* dan kata dasar tersimpan di dalam database. Hasil atau output dari sistem berupa *bag of World* atau kumpulan kata dasar namun kata yang masuk dalam *list stop word* hilang. *Bag of word* yang dihasilkan dari setiap dokumen disimpan dalam tabel database relasional dan dalam format .txt untuk perbandingan. Berikut adalah arsitektur dari aplikasi preprocessing:



Gambar 5 Gambaran Umum Sistem *Text-Processing*

Hasil pengujian tahap awal sudah mendapatkan hasil seperti berikut. Hasil yang didapat merupakan langkah awal untuk melanjutkan penelitian yaitu proses data mining.

2) Teks pre-processing

Dokumen *email* merupakan dukumen teks yang memerluklan proses *preprocessing* sebelum di *mining* lebih lanjut [1]. Preprocessing dokumen email yang berupa kumpulan kata memiliki Tujuan yaitu menyederhanakan dan mempermudah dalam proses pembobotan term / kata. Tahapan yang digunakan dalam proses preprocessing dokumen email dalam penelitian ini yaitu sebagai berikut:

1. ***Parsing / Tokenizing***
   Teks processing diawali dengan memotong setiap kata yang ada dalam teks tersebut menjadi perkata. Process pemotongan teks menjadi kata – kata terpisah disebut *parsing / tokenizing*. Pemotongan teks yang digunakan dalam penelitian ini yaitu titik (.), koma (,), kutip satu (’), kutif dua (“) operator aritmatika (+,/,*,-), tanda baca seperti tanda kalimat seru (!), tanda kalimat tanya (?), angka 0 sampai 9, dan juga karakter lain yang tertera di keyboard computer (@,#,$,%,^,&,(,),~,`,:,;<,>) serta jenis kurung kata.

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*. Dalam penelitian ini jenis stop word yaitu : yang, mampu, tentang, setelah, semua, hampir, juga. am, antara, dan , ada, seperti, jadi, karena, tetapi, oleh, bisa, tidak, lakukan, memang, lain, setiap, untuk, dari, kepada, yth, saya, kamu, anda, to, for, from, dan banyak lagi kata – kata umum yang terdapat dalam email yang jumlahnya 1037 kata. Semua *stop word* yang digunakan disimpan di dalam database.

3. ***Stemming***
   Stemming adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya [4,5,7]. Proses *stemming* untuk setiap Bahasa berbeda dengan Bahasa yang lain misal, proses *stemming* Bahasa Inggris dengan Bahasa Indonesia tentunya berbeda karena perbedaan pembentukan dan perubahan kata menjadi bentuk kata lain [1,2,4]. 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 [3,5]. Berikut adalah arti dan contoh dari imbuhan dalam Bahasa Indonesia [5]:
   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”*)

   Algoritma Stemming atau tool stemmer untuk Bahasa Indonesia sudah banyak dikembangkan diantaranya: Nazief dan Adriani dari Universitas Indonesia pada tahun 1996, Vega dari Univertas nasional singapura tahun 2001, Arifin dan setiono dari Institut teknologi sepuluh November 2002, *Porter Stemmer for Bahasa Indonesia* dikembangkan oleh Fadillah Z. Tala pada tahun 2003 [5]. Pada penelitian ini algoritma yang digunakan yaitu algoritma yang dikembangkan oleh Nazief dan Adriani pada paper [4]. Pada paper tersebut algoritmanya disebut Algoritma CS, dimana algoritma CS tersebut bekerja dengan cara menerapkan urutan dan pengetahuan dasar dari Bahasa Indonesia sebagai dasar pendekatan yang digunakan dalam stemming. Algoritma ini didasarkan pada aturan morfologi komprehensif [4]. Algoritma tersebut menggunakan 3 komponen yaitu :

1. Pengelompokan imbuhan
2. Urutan aturan penggunaan affix (dan pengecualian)
3. Kamus kata dasar

Kamus digunakan untuk mengecek kata hasil dari proses stemming, jika kata hasil stemming ada didalam kamus, maka stemming berasil. Dasar pendekatan yang digunakan yaitu Urutan dan pengetahuan dasar dari Bahasa Indonesia [3]

1. Kata – kata dari 3 karakter atau lebih sedikit tidak dilakukan stemming.
2. Afik tidak pernah diulang, sehinga stemmer harus menghapus hanya satu dari sekumpulan afiks.
3. Menggunakan pembatasan konfiks (*confix*) selama *stemming* untuk meyingkirkan kombinasi imbuhan (affix) tidak valid.
4. Ketika mengembalikan karakter setelah menghapus prefix, dilakukan recoding jika perlu.
   Berikut adalah cara kerja algoritma CS [4]:

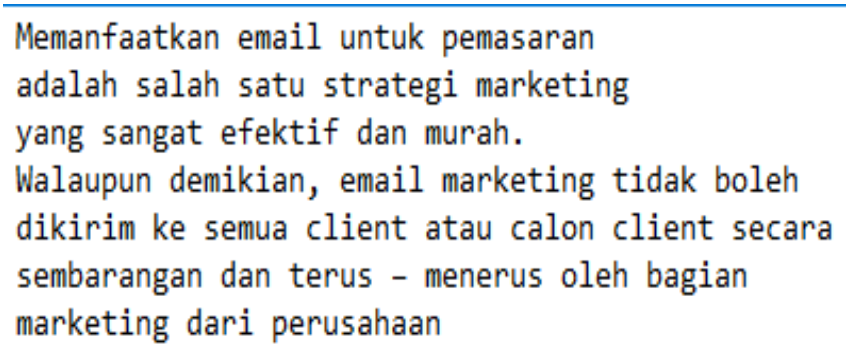
1. Periksa Kata pada kamus kata dasar
2. Hapus Akhiran infeksi (infection suffixes) [P] kemudian [PP]
3. Hapus setiap derivational suffixes [DS]
4. Hapus derivational prefixes [DP]
   - Proses berhenti jika
     - Ditemukan awalan-akhiran yang tidak valid
     - Awalan (prefix) yang di identifikasi sama (identical) dengan awalan yang dihapus sebelumnya.
     - Tiga awalan sudah dihapus tapi belum ditemukan root-word.
   - Mengidentifikasi jenis awalan dan disambiguate jika perlu
   - JIka kata / root-word tidak ditemukan dalam kamus, maka lanjutkan kembali ke step 4.
5. Recording (menggunakan tabel Prefix Disambigu, pada slide berikutnya)
6. Jika root-word belum ditemukan, kembalikan kata menjadi sebelum di stemming.

4. Hasil dan Pembahasan

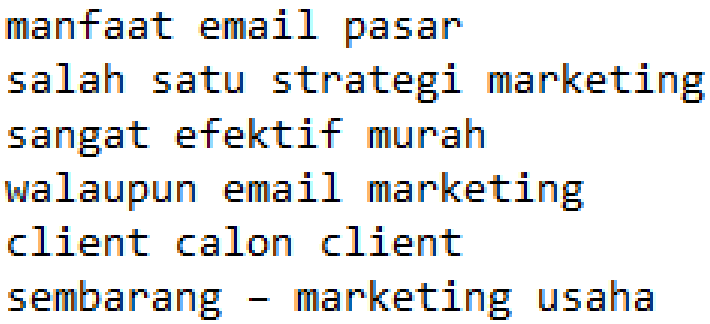
3) Hasil Uji Coba

Pengujian dilakukan pada setiap tahapan pengembangan sistem yaitu mulai dari pengujian proses *preprocessing* yang merubah dari data dukumen email menjadi data tersetruktur, pengujian TF-IDF serta pengujian proses clustering.

1. Pengujian pertama yaitu *preprocessing* dokumen email. Data terstruktur yang dihasilkan berupa tabel kata dasar dari dokumen email yang dianalisis. Jumlah email yang digunakan dalam proses pengujian yaitu 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. Berikut adalah pengujian preprocessing text. Pada pengujian ini, teks ditulis pada format .txt dan kemudian diproses dan hasilnya ditampilkan juga dalam format .txt:



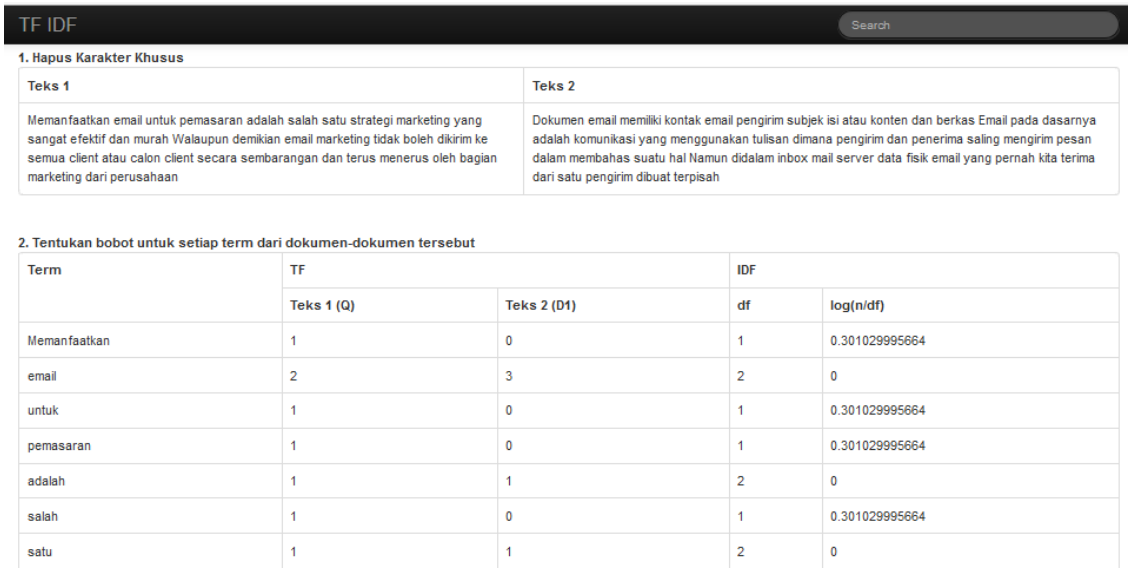
Gambar 6. Contoh teks yang akan di preprocessing



Gambar 7. Hasil text preprocessing

Setelah dilakukan preprocessing text menggunakan program yang telah dibangun mengasilkan seperti gambar diatas. Semua kata dalam teks yang diinput akan dirubah menjadi kata dasar.

2. Pengujian kedua yaitu pengujian proses pembobotan kata dengan teknik TF-IDF dan melakukan pengelompokkan data email menggunakan k-mean clustering. Scenario pengujian yaitu data email yang sudah di-preprocessing dimasukkan kedalam sistem untuk dihitung bobot dari setiap term yang disebut dengan indek. Bobot Indek dari setiap term pada setiap dokumen akan menjadi dasar dalam pencarian dokumen. Dari hasil pengujian sistem sudah dapat menentukan bobot dari setiap term menggunakan teknik TF-IDF. Jumlah term yang digunakan untuk mencarian yaitu 1 term, 2 term, dan multi term, kesimpulan yang dapat diambil yaitu semakin banyak term yang digunakan maka proses pencarian semakin lama dan jumlah dokumen yang ditemukan semakin sedikit. Berikut adalah contoh penghitungan menggunakan aplikasi TF-IDF yang telah dibangun:


TF IDF

Search

1. Hapus Karakter Khusus

Teks 1	Teks 2
Memanfaatkan email untuk pemasaran adalah salah satu strategi marketing yang sangat efektif dan murah Walaupun demikian email marketing tidak boleh dikirim ke semua client atau calon client secara sembarangan dan terus menerus oleh bagian marketing dari perusahaan	Dokumen email memiliki kontak email pengirim subjek isi atau konten dan berkas Email pada dasarnya adalah komunikasi yang menggunakan tulisan dimana pengirim dan penerima saling mengirim pesan dalam membahas suatu hal Namun didalam inbox mail server data fisik email yang pernah kita terima dari satu pengirim dibuat terpisah

2. Tentukan bobot untuk setiap term dari dokumen-dokumen tersebut

Term	TF		IDF	
	Teks 1 (Q)	Teks 2 (D1)	df	log(n/df)
Memanfaatkan	1	0	1	0.301029995664
email	2	3	2	0
untuk	1	0	1	0.301029995664
pemasaran	1	0	1	0.301029995664
adalah	1	1	2	0
salah	1	0	1	0.301029995664
satu	1	1	2	0

Gambar 8. Aplikasi penghitungan TF IDF

Pengujian proses clastering menggunakan algoritma k-mean baru dapat dilakukan jika indek dari setiap term / kata dari dokumen sudah didapatkan menggunakan rumus TF-IDF. Pada pengujian yang dilakukan jumlah claster yang ditentukan untuk mengelompokan data email yaitu 2 cluster. Dokumen email yang digunakan adalah hasil *preprocessing*. Jumlah dokumen yaitu 11 dokumen email hasil pengabungan. Untuk mengetahui tingkat akurasi clustering maka pada setiap dokumen email diberikan label berdasarkan jenis ketertarikan pengirim yang sudah diketahui sebelumnya dan diisi secara manual. Berikut adalah tabel hasil clustering.

Tabel 1. Hasil Clustering

Dokumen email	Cluster	Label	Akurasi
Doc_email_1	K1	K1	T
Doc_email_1	K2	K1	F
Doc_email_1	K2	K2	T
Doc_email_1	K2	K1	F
Doc_email_1	K1	K2	F
Doc_email_1	K2	K2	T
Doc_email_1	K1	K1	T
Doc_email_1	K2	K2	T
Doc_email_1	K2	K2	T
Doc_email_1	K1	K1	T
Doc_email_1	K2	K1	F

Pada hasil pengujian diatas terdapat dua centroid yang ditentukan. Dokumen email yang digunakan dalam pengujian sudah di beri label cluster. Setelah dilakukan pengujian clustering pada dokumen email yang sudah dilabeli tersebut ada beberapa dokumen yang masuk dalam cluster pada label lain. Jumlah akurasi clustering terdapat penentuan label secara manual yaitu 7 dokumen sesuai dengan labelnya dan 4 tidak sesuai dengan pelabelan. Secara persentase berarti akurasi ketepatan algoritma k-mean clustering pada dokumen email yaitu (7*100) /11 = 63.63%.

5. Simpulan

Setelah melakukan pengembangan sistem dan melakukan pengujian maka dapat disimpulkan seperti berikut:

1. *Knowledge discovery* atau pencarian pengetahuan di dalam dokumen email dalam dilakukan dengan teknik *text mining* dan data mining yaitu clustering. Teknik text mining dapat digunakan untuk merubah data email yang *unstructured* menjadi terstruktur dengan nilai bobot setiap kata dengan TF-IDF. Sedangkan proses clustering menggunakan k-mean clustering dapat menemukan kelompok dari pengirim.
2. 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.
3. Aplikasi *text preprocessing* yang dibangun hanya dapat mengasilkan kumpulan kata dasar dari dokumen email yang dimasukkan.
4. Hasil Aplikasi perhitungan TF-IDF yang dilakukan setelah proses *preprocessing text mining* yang berupa *index* bobot setiap term / kata dapat digunakan untuk mencari dokumen yang banyak mengandung *term* yang dicari. Pencarian term menggunakan perhitungan TF-IDF ini dapat digunakan untuk mencari email yang cenderung membahas produk atau jasa yang sedang ditawarkan oleh perusahaan sehingga dapat menentukan siapa yang harus dikirimkan email marketing yang bersangkutan.
5. Dari Hasil uji coba aplikasi *clustering* dengan algoritma k-mean dapat diambil kesimpulan bahwa sistem dapat digunakan untuk mengelompokkan pengirim email menjadi beberapa kelompok berdasarkan isi dokumen email yang mereka telah kirim ke perusahan kita namun akurasi dari pengelompokkan K-mean tidak terlalu bagus hanya 63.63% dari 11 dokumen email yang telah digabung.

Daftar Pustaka

[1] Mega Nata Gusti Ngurah, Yudiastra Putu Pande. Preprocessing Text Mining pada email box berbahasa Indonesia, Konferensi Nasional Sistem & Informatika (KNS&I) 2017

[2] Mega Nata Gusti Ngurah, Yudiastra Putu Pande. Stemming teks sor-singgih Bahasa Bali.

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

[4] 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

[5] 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

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

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

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

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

[10] Han Jiwai, Kamber, Pei., (2012), *Data Mining concepts and techniques third edition*. Morgan Kaufmann publishers.