Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data *Content* Berita SMS

Windu Gata[1], Purnomo[2]
[1,2]*Program Pasca Sarjana Ilmu Komputer* ***STMIK Nusa Mandiri*** *Jakarta*
STMIK Nusa Mandiri*, Menara Salemba (A5) Jl. Salemba Raya No.5 Jakarta*
[1]windu_gata@yahoo.com
[2]purnomo@antara.co.id

Abstract **– Since 2004, LKBN ANTARA still using SMS-Gateway application as media delivery of news from the field, due to Internet network conditions in the Territory of Indonesia is still limited. SMS-GATEWAY also be used for sending notifications or activities in the workplace from the Leader, editor, Bureau Chief, reporter or IT team. Indicated in case of misuse or SMS sending application that does not comply with labor. Due to these conditions, the research carried out by manual methods as well as methods of K-Nearest Neigbour to know about it. Results of research conducted to get the results on the accuracy of the prediction selection YA number 772 does not fit properly and a number of 32, so that the precision is 96.02%. While predictions TIDAK have the results of a number of 0 and 14 correct errors in the prediction of YA. So that the accuracy of the results obtained number of 96.15%. So the use of SMS-Gateway application used ANTARA environment in accordance with the needs during November 2015. As well as can help ANTARA within the classification SMS content with klasifika YA and TIDAK.**

Keywords: SMS-Gateway, Text Mining, Preprocessing K-NN, News, SMS

Abstrak **– Sejak tahun 2004, LKBN ANTARA masih menggunakan aplikasi SMS-Gateway sebagai media pengiriman berita dari lapangan, dikarenakan kondisi jaringan Internet di Wilayah Indonesia yang masih terbatas. SMS-GATEWAY juga digunakan untuk melakukan pengiriman notifikasi atau kegiatan di lingkungan kerja dari Pimpinan, redaktur, Kepala Biro, Pewarta ataupun tim IT. Diindikasikan terjadi penyalahgunaan aplikasi atau pengiriman SMS yang tidak sesuai dengan kerja. Dikarenakan hal tersebut maka penelitan dilakukan dengan metode manual maupun metode K-Nearest Neigbour untuk mengetahui hal tersebut. Hasil penelitian yang dilakukan mendapatkan hasil pada akurasi seleksi prediksi YA sejumlah 772 benar dan tidak sesuai sejumlah 32, sehingga presisinya adalah 96.02%. Sedangkan prediksi TIDAK mempunyai hasil sejumlah 0 kesalahan dan 14 benar dalam prediksi TIDAK. Sehingga hasil akurasi yang didapatkan sejumlah 96.15%. Sehingga penggunaan dari aplikasi SMS-Gateway yang digunakan dilingkungan LKBN ANTARA telah sesuai dengan kebutuhan selama bulan November 2015. Serta dapat membantu LKBN ANTARA dalam melakukan klasifikasi isi SMS dengan klasifikasi YA dan TIDAK.**

Kata Kunci: SMS-Gateway, Text Mining, Preprocessing K-NN, News, **SMS**

I. PENDAHULUAN

Penggunaan *Short Message Service* (SMS) untuk pengiriman berita, mengirimkan informasi, ataupun agenda liputan masih tetap dibutuhkan di era sekarang. Hal tersebut dikarenakan tidak semua wilayah di Indonesia mempunyai akses internet yang baik, terlebih lagi di daerah-daerah luar kota besar dan terpencil.

Sebagai Kantor Berita Nasional, LKBN ANTARA, yang harus mendapatkan berita dari seluruh pelosok nasional di Indonesia telah menggunakan aplikasi berbasis web serta surat elektronik(*email*), tetapi tidak dapat meninggalkan teknologi SMS tersebut. SMS masih dapat digunakan selama jaringan perangkat genggam terkoneksi walaupun dengan sinyal yang sangat rendah.

Sistem SMS yang telah dibangun sejak tahun 2005 tersebut, tidak hanya digunakan oleh wartawan untuk mengirimkan berita, tetapi dapat juga digunakan oleh kantor pusat, yaitu: Pempinan; Sekretariat Redaksi (Sekred); serta Kelapangan Departemen Pemberitaan, untuk mengirimakan pesan yang berisikan agenda liputan ke semua wartawan pusat maupun seluruh wartawan di seluruh Indonesia ataupun sebaliknya.

Keguanaaan SMS lainnya adalah memberikan konfirmasi penerimaan berita yang dikirim oleh wartawan yang menggunakan teknologi surat elektronik secara otomatis. Hal tersebut membuat kebutuhan atas SMS pada LKBN ANTARA cukup tinggi dan menjadi prioritas pada proses pemberitaan.

Tim sistem informasi dan infrastruktur juga menggunakan aplikasi SMS untuk memberikan informasi kepada semua wartawan dan pengguna komputer di LKBN ANTARA seperti pemberian perubahan akses komputer dalam bentuk *password* langsung ke perangkat genggam pengguna yang bersifat rahasia (*confidential*), informasi

gangguan akses, sosialisasi informasi dan rencana kerja divisi Teknik Informatika(IT) yang mempunyai dampak pemeliharaan perangkat keras, aplikasi pemberitaan, dan jaringan komputer.

Semua aktivitas yang menggunakan aplikasi SMS-Gateway tercatat dengan baik pada database. Catatan dari aktivitas tesebut terinci, berupa: pengguna pengirim, waktu akses, jabatan, isi SMS, dan nomor yang dikirim.

Sistem SMS-Gateway dapat digunakan oleh pengguna komputer selama 24 jam atau satu hari tanpa pembatasan oleh sistem. Batasan atau jumlah karakter dari SMS yang dikirimkan sebanyak 320 karakter, tetapi batasan karakter untuk pengiriman dari wartawan di lapangan tak terbatas, karena sistem aplikasi tersebut dapat otomatis menyambungkan data berita menjadi satu kesatuan utuh dan disimpan ke dalam aplikasi pemberitaan.

Transaksi SMS yang menggunakan sistem SMS-Gateway yang tersimpan pada *database* mempunyai jumlah perbarisnya sejumlah 3.000 sampai dengan 5.000 SMS perbulan. Sedangkan jumlah pengguna sistem SMS-Gateway sejumlah 400 akun pengguna.

Setiap pengiriman dari SMS yang dikirimkan oleh sistem aplikasi SMS-Gateway tidak terdapat konfirmasi, sehingga sifat dari pengiriman SMS adalah satu arah (*one-way*).

Penggunaan sistem aplikasi yang telah digunakan oleh pengguna di lingkungan LKBN ANTARA diindikasikan disalahgunakan oleh pengguna dengan mengirimkan SMS yang tidak menyangkut dengan kegiatan proses pemberitaan maupun kegiatan yang operasional yang ada di lingkungan LKBN ANTARA.

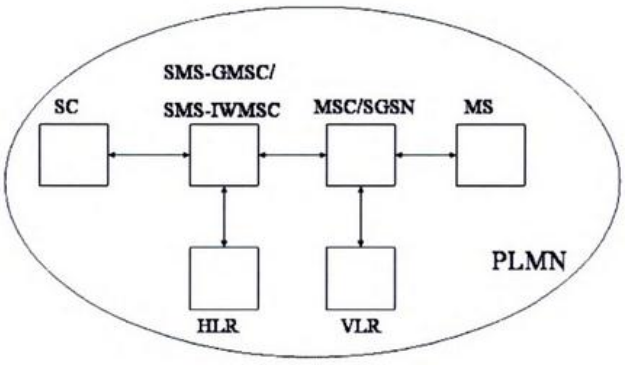
Berdasarkan hal tersebut di atas, maka diadakan penelitian menggunakan metode *Text Mining* dengan algoritma K-Nearest Neighbour, untuk mengetahui seberapa besar penggunaan SMS yang berkaitan dengan pekerjaan maupun penggunaan yang tidak terkait dengan pekerjaan di lingkungan LKBN ANTARA. Penelitian mengambil data bulan November tahun 2015 sejumlah 3.120 SMS.

II. LANDASAN TEORI DAN METODE

1. *Short Message Service* (SMS)

Menurut Narang, *Short Message Service* atau yang dikenal dengan SMS merupakan pesan teks dengan jumlah terbatas yang menggunakan pelanggan perangkat bergerak. Dalam satu SMS dapat berisikan pesan sejumlah 140 kata[5].

Sedangkan menurut Gomes, SMS adalah pesan teks yang dikirim dari sebuah perangkat bergerak ke perangkat gerak lainnya atau perangkat yang terkoneksi pada sebuah jaringan menggunakan SMS center (SC atau SMSC). Maksud dari hal tersebut adalah SMS tidak terkirim secara langsung dari pengirim ke penerima, melainkan melalui SC, yang berarti setiap jaringan perangkat bergerak yang mendukun SMS mempunyai satu atau lebih messeging centers untuk menangani dan mengatur pesan singkat[2].



Gambar 1. Arsitektur Jaringan SMS

Gomes juga menggambarkan arsitektur dari SMS yang tergambarkan pada gambar 1 dimana SC bertanggung jawab atas penyampaian, penyimpan, serta pengiriman dari pesan pendek yang berlangsung dari pengiriman dan penerimaan SMS (MS). Satu SC memungkinkan terkoneksi dengan beberapa *Public Land Mobile Network* (PLMN) yang juga mungkin terkoneksi beberapa *Mobile Switching Center* (MSC) atau *Gateway Message Center* (SMS-MSC) atau *SMS Interworking MSC* (SMS-IWMSC) yang terinstalasi dalam satu PLMN. SC juga dapat terkoneksi dengan banya SMS-GMSC atau SMS-IWMSC.

SMS-GMSC merupakan sebuah fungsi dari MSC yang berfungsi sebagai penerima dari pesan pendek yang dikirimkan oleh SC, kemudian diidentifikasikan oleh *Home Location Register* (HLR) untuk mengetahui informasi pengiriman dan pengiriman dari pesan pendek ke MSC atau *Serving GPRS Support Node* (SGSN) dari penerima MS.

SMS-IWMSC merupakan fungsi dari MSC yang menerima pesan pendek dari MSC atau SGSN serta menyesuaikan dengan alamat SC dan mengirimkan kepada SC terkait. Jika konfirmasi pengiriman terkait dengan pesan pendek yang diterima dengan SC, maka SMS-IWMSC bertanggung jawab untuk mengirimkan konfirmasi ke MSC atau SGSN.

Penghentian pesan pendek, SMS-GMSC menerima pesan dari SC, jika terdapat error atau kesalahan di dalam parameter dari pesan tersebut, maka akan terdapat pesan yang berisikan informasi kesalahan tersebut kepada SC

sebagai laporan kesalahan, dengan kata lain HLR akan mengirimkan informasi kemungkinan kesalahan dalam pesan dan dikirimkan kembali ke MSC (atau SGSN) pengirim pesan pendek. Jika SMS-GSMC mengirimkan dua alamat, maka SMS-GMSC akan mencoba menggunakan jalur kedua jika SMS pertama tidak sukses.

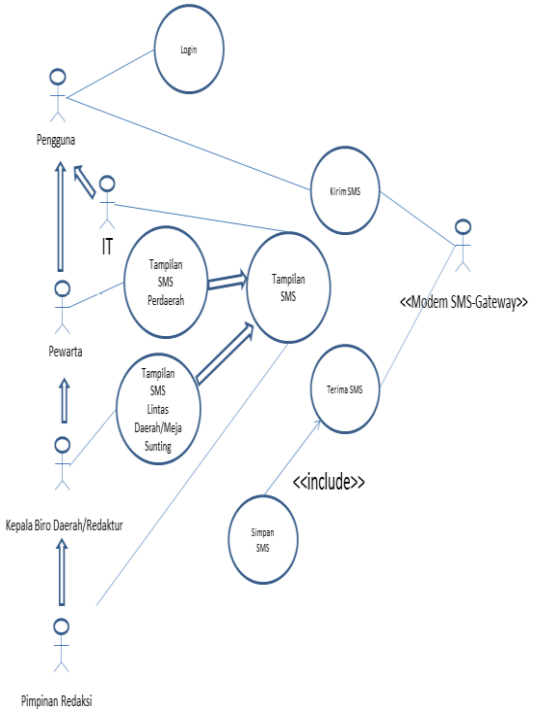
Ketika MSC menerima pesan pendek yang diterima oleh MS dari *Visitor Location Register* (VLR), contoh lokasi area, dan jika tidak terdapat kesalahan yang terindikasi dari VLR, maka pesan pendek akan dikirimkan kepada MS. Di sisi lain, jika SGSN menerima pesan pendek, SGSN mengirimkan pesan pendek ke MS secara langsung. Kedua elemen tersebut, VLR dan SGSN, dapat mendeteksi kesalahan prosedur pengiriman pesan pendek. Jika itu terjadi, maka informasi kesalahan tersebut dikembalikan kepada SMS-GMSC di laporan kegagalan pengiriman pesan pendek, dan pengiriman dibatalkan. ketika menerima konfirmasi bahwa pesan telah diterima oleh MS, maka elemen-elemen tersebut mengirimkan konfirmasi kepada SMS-GSMC di dalam laporan pengiriman.

2. Aplikasi SMS-Gateway

Penggunaan sistem SMS-Gateway pada LKBN ANTARA mempunyai tata cara yang harus dilakukan, adalah sebagai berikut:

a) Pimpinan Redaksi Pemberitaan (Pemred, Wapemred, Ombudsman) dapat menggunakan fasilitas pada aplikasi SMS ke semua pengguna pusat maupun daerah serta biro dan kontributor luar negeri.
b) Divisi teknologi informasi memiliki hak yang sama dengan pimpinan tertinggi Redaksi untuk pengiriman SMS yang berhubungan dengan pengelolaan system informasi di lingkungan LKBN ANTARA kepada semua pengguna.
c) Redaktur berita yang terdiri dari Redaktur Lintas Departemen serta Redaktur bertanggung jawab (RBT) juga memiliki hak yang sama dalam penggunaan sistem SMS dengan Pimpinan Pemberitaan.
d) Kepala Departemen dan Kepala Biro Daerah dapat melakukan komunikasi berupa SMS kepada pengguna atau wartawan di wilayahnya dan dapat pula mengirimkan SMS kepada Pimpinan lain.
e) Pengguna wartawan dapat mengirimkan SMS di wilayah yang ditugaskan kepada mereka serta pimpinan atau redaktur di biro terkait.

Penggunaan aplikasi SMS-Gateway dapat digambarkan penggunaan dengan menggunakan diagram *Use Case* pada gambar 2, sebagai berikut:



Gambar 2. Sistem Aplikasi SMS-Gateway

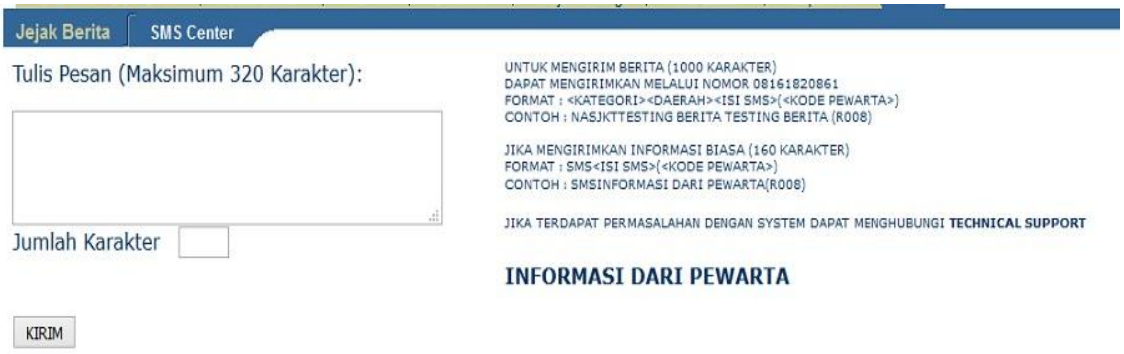
Aplikasi SMS-Gateway merupakan aplikasi sub dari aplikasi pemberitaan LKBN ANTARA, dimana setiap pengguna aplikasi pemberitaan dapat menggunakan sub aplikasi Aplikasi SMS-Gateway dengan cara memasukkan *user* dan *password* pada halaman *login*.

Pengguna yang berasal dari divisi IT merupakan turunan dari pengguna aplikasi dan dapat membuka halaman tampilan SMS yang menampilkan seluruh pengguna dalam aplikasi dan dapat mengirimkan SMS kepada seluruh pengguna aplikasi.

Pengguna Pewarta atau wartawan hanya dapat menggunakan tampilan SMS yang berisikan Kepala Biro Daerah dan wartawan dalam satu daerah yang dapat dikirimkan SMS.

Kepala Biro Daerah dan Redaktur berita dapat menggunakan tampilan SMS yang berisikan pengguna di daerah atau meja sunting yang terkait juga pengguna setingkat Pimpinan atau Kepala Biro lainnya yang dapat dikirimkan SMS.

Setiap SMS yang dikirimkan oleh pengguna dalam aplikasi SMS-Gateway akan dikirimkan melalui modem SMS-Gateway yang telah terinstalasi pada server.



Gambar 3. Aplikasi Pengiriman SMS

Penerimaan berita yang berasal dari wartawan di lapangan diterima oleh modem SMS-Gateway dan disimpan ke dalam Aplikasi pemberitaan dan ditampilkan pada layar Redaktur seperti yang tergambarkan pada Gambar 4. Konfirmasi SMS yang secara otomatis terkirim seperti pada Gambar 5.

Gambar 4. Editor Berita LKBN ANTARA

Gambar 5. Konfirmasi SMS

3. Data Mining

Turban mengungkapkan Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terakit dari berbagai proses besar[7].

Menurut Maimon, *Data Mining* (DM) adalah inti dari proses *Knowledge Discovery in Database* (KDD), yang melibatkan algoritma dalam mengeksplorasi data, mengembangkan model dan menemukan pola yang sebelumnya tidak diketahui. Model ini digunakan untuk memahami fenomena dari data, analisis dan prediksi. *Knowledge*

Discovery in Database (KDD) adalah proses terorganisir untuk mengidentifikasi pola yang valid, baru, berguna, dan dapat dimengerti dari sebuah data set yang besar dan kompleks[4].

Istilah *data mining* dan *Knowledge Discovery in Database* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda tetapi berkaitan satu sama lain. Tahapan dalam keseluruhan proses KDD adalah *data mining*. Proses KDD secara garis besar dijabarkan oleh Fayyad[1], adalah sebagai berikut :

1). Data seleksi (*Data selection*)

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam Knowledge Discovery in Database KDD dimulai. data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari data operasional.

2). *Pre-processing/cleaning*

Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus Knowledge Discovery in Database KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalah cetak (tipografi). Juga dilakukan proses enrichment yaitu proses memperkaya data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk *Knowledge Discovery in Database* KDD seperti data atau informasi eksternal.

3). Transformasi (*Transformation*)

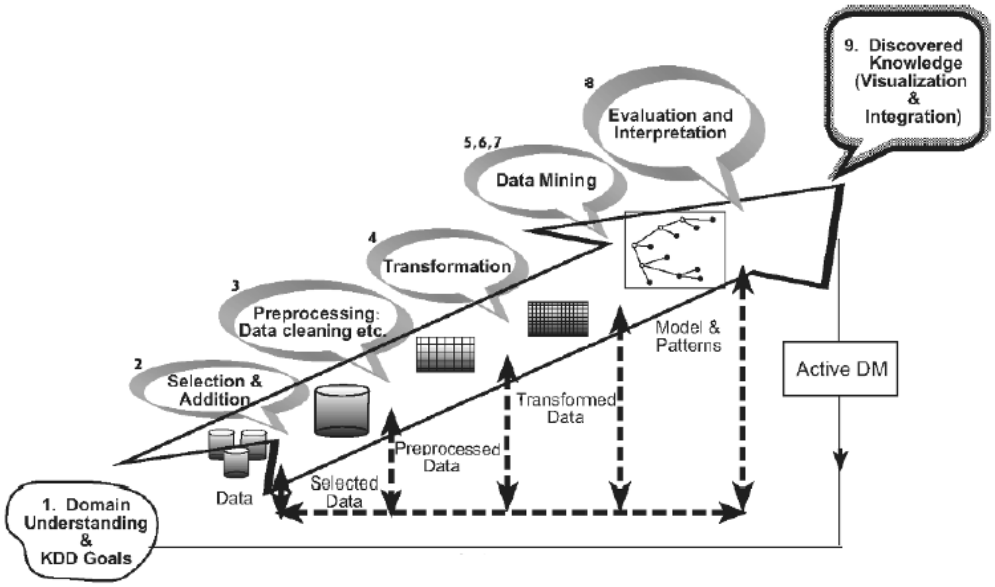
Coding adalah proses transformasi pada data yang telah dipilih sehingga data tersebut sesuai untuk proses data mining. Proses *coding* dalam *Knowledge Discovery in Database* KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4). *Data mining*

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *Knowledge Discovery* in Database KDD secara keseluruhan.

5). *Interpretation/evaluation*

Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses *Knowledge Discovery in Database* KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.



Gambar 6. Proses *Knowledge Discovery in Database (KDD)*

Menurut Hemawati mengenai teknik dan sifat data mining[4], adalah sebagai berikut:

1). *Classification*

Klasifikasi adalah menentukan sebuah *record data* baru ke salah satu dari beberapa kategori (atau klas) yang telah didefinisikan sebelumnya atau yang *sering* disebut dengan *Supervised Learning*.

2).*Clustering*

Mempartisi data-set menjadi beberapa sub-set atau kelompok sedemikian rupa sehingga elemen-elemen dari suatu kelompok tertentu memiliki set properti yang di-*share* bersama, dengan tingkat similaritas yang tinggi dalam satu kelompok dan tingkat similaritas antar kelompok yang rendah. klusterisasi disebut juga dengan unsupervised learning.

3). *Association Rule*

Mendeteksi kumpulan attribut-attribut yang muncul bersamaan (*co-occur*) dalam frekuensi yang sering, dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut.

4). *Sequential Pattern Discovery*

Mencari sejumlah event yang secara umum terjadi bersama-sama.

5). *Regression*

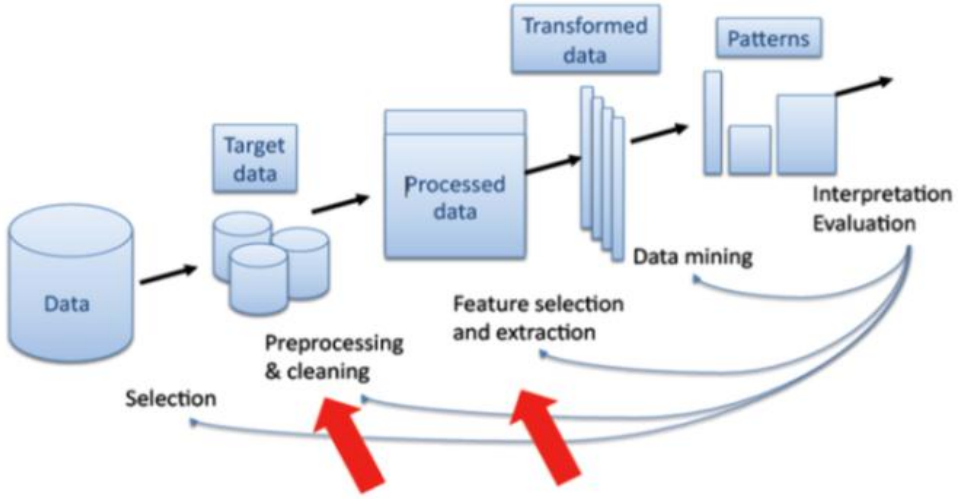
Regresi adalah memprediksi nilai dari suatu variabel kontinyu yang diberikan berdasarkan nilai dari variabel yang lain, dengan mengasumsikan sebuah model ketergantungan *linier* atau *nonlinier*. Teknik ini banyak dipelajari dalam statistika, bidang jaringan saraf tiruan (*neural network*).

6). *Deviation Detection*

Deviation Detection melakukan deteksi anomali secara otomatis dengan tujuan untuk mengidentifikasi kebiasaan suatu entitas dan menetapkan sejumlah *norm* melalui *pattern discovery*.

4. ***Text Mining***

Proses pengolahan data dalam penelitian ini menggunakan metode yang digunakan oleh Yan Y[8]. seperti pada gambar 7.



Gambar 7: Proses *Data Mining*

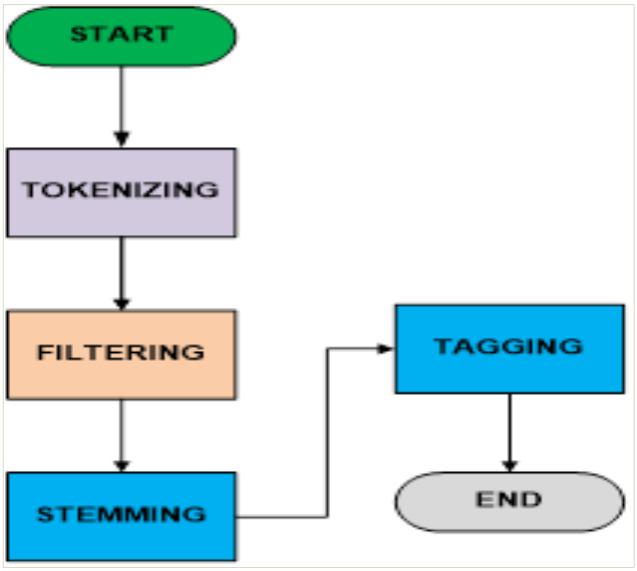
Pada **Tahap Pertama**: Melakukan persiapan data dalam dari log sistem yang ada dalam aplikasi server pemberitaan. Data dipilih dari sekian banyak data log khusus bagian log transaksi sistem sms.

Tahap Kedua: Melakukan pre-prosesing dengan text mining yang meliputi *tokenizing, filtering, steaming* dan *stopward* sehingga data siap diolah ke proses berikutnya.

Tahap Ketiga: Melakukan pengolahan data dengan menggunakan algoritma data mining yang terdiri dari, estimasi, prediksi, klasifikasi, cluster dan asosiasi.

Tahap Keempat menentukan dictionary manual terkait dengan pemisahan content isi SMS terdiri dari Label Kerja dan tidak Kerja.

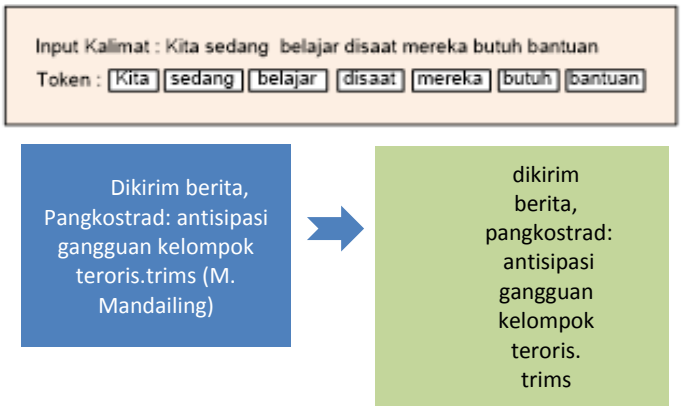
Tahap Kelima: Melakukan evaluasi komparasi dari hasil untuk mengukur tingkat akurasi kerja proses text mining dengan KNN terhadap proses manualnya.



Gambar. 8: Tahapan proses pre-processing.

Tahap *Preprocessing* data merupakan bagian dari data/*text mining* bertujuan untuk mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Preproses text mining dilakukan langkah tokenizing, filtering, stemming, tagging.

Tahap tokenizing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunya, disini text sms yang dikirimkan dilakukan pemotongan kata untuk dipisahkan dalam rangka proses lebih lanjut.



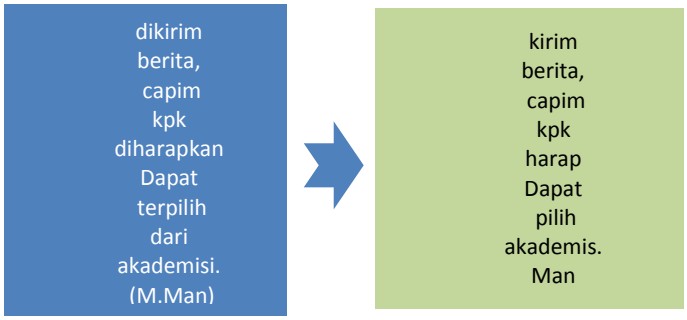
Gambar 10. Proses *Tokenizing*.

Tahap *Filtering* adalah tahap mengambil kata-kata penting disini yang telah dipotong kata per kata dilakukan filter. Pada tahap ini dilakukan pembuangan kata-kaat yang dianggap tidak penting seperti tanda hubung dan lainnya.

Gambar 11. Tahap *Filtering*

Tahap *Stemming* menurut Manning adalah tahap mencari root kata dari tiap hasil filtering disini dapat diambil contoh bahwa kata liputan maka pasca stemming menjadi liput, peresmian menjadi kata stemming resmi dalam bahasa lain stemming mencari kata dasar utama. Dokumen atau artikel dibuat oleh rangkaian kalimat-kalimat yang saling berhubungan. Pada setiap kalimat tersebut jelas menggunakan pola dasar kalimat dasar minimal berbasis SPO (Subjek Predikat Objek). Secara alasan gramatikal muncul kata-kata yang berbeda arti tetapi memiliki satu dasar kata yang sama[5].

[prefix1]+[prefix2]+root+[suffix]+[possessive_pronoun]+[particle]



Gambar 12. Tahap *Stemming*

Tahap *Tagging* adalah tahap mencari bentuk awal/root dari tiap kata lampau atau kata hasil stemming seperti pada beberapa literature yang pernah ada bahwa proses Tagging tidak dipakai pada proses ini karena dalam bahasa Indonesia tidak mengenal bentuk kta lampau.

Tahap *Analizing*, Pada tahap ini dilakukan proses perhitungan pembobotan **(w)** dari dokumen yang didapat agar diketahui seberapa jauh tingkat similaritas/kesesuaian antara keyword yang dimasukkan dengan dokumennya.

5. *Algoritma K-Nearest Neighbor*

Pengolahan *Text Mining data* SMS pada aplikasi SMS-Gateway pada penelitian ini menggunakan algoritma K-Nearest Neighbor digunakan untuk mengklasifikasi SMS yang telah dilakukan pada preproses data.

K-Nearest Neighbour (KNN) merupakan suatu pendekatan klasifikasi yang mencari semua data latih yang relatif mirip dengan data uji .

Teknik klasifikasi ini disebut *lazy learning* karena teknik ini tidak membangun model klasifikasi terlebih dahulu, seperti: pohon keputusan (*decision tree*), klasifikasi berbasis aturan (rule-based).

Algoritma KNN diawali dengan menentukan nilai k, yaitu banyak SMS yang memiliki jarak terdekat. Nilai k biasanya gasal/ganjil. Kemudian, hitung semua jarak antara SMS uji dengan semua SMS pada data latih. k merupakan SMS yang memiliki nilai jarak paling dekat.

Setelah itu, hitung nilai skor kategori dari k SMS terpilih. Jika SMS dari k SMS yang terpilih tadi memiliki kategori yang sama, nilai skor kategori adalah penjumlahan semua nilai kemiripan SMS dengan SMS uji.

Dengan mengurutkan hasil skor tiap kategori, SMS uji mendapatkan label dari kategori yang memiliki skor tertinggi. Secara matematis, aturan tersebut digambarkan dalam persamaan dengan menggunakan perhitungan jarak Euclidean.

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Dimana matriks *d(x,y)* adalah jarak skalar dari kedua vektor x dan *y* dari matriks dengan ukuran *d* dimensi.

***Pembobotan**:* Pembobotan merupakan faktor penentu dalam kategori kerja. Hasil perhitungan dari perbandingan SMS, frekuensi/jumlah conten SMS sesuai dengan klasifikasinya. Perhitungan bobot berasal dari logika setelah mengamati penggunaan SMS yang memiliki kata kata dalam kamus kerja. Bobot dari perbandingan isi konten sms yang memiliki kesesuaian minimal diatas 10% terhadap kamus kata dapat dikategorikan berkaitan SMS tersebut berkaitan dengan pekerjaan sedangkan dibawah dari prosentase tersebut tidak termasuk dalam pengkategorian SMS kerja.

6. *Tinjauan Studi*

Beberapa penelitian yang menggunakan metode Text Mining, adalah sebagai berikut:

1) Perancangan dan Pembuatan Perangkat Lunak Pendeteksi Selingkuh Menggunakan Metode K-Nearest Neighbor berbasis Android. Thaufan Ardi Arafat NRP. 5108 100 053.

   Hasil yang dilakukan dari penelitian ini adalah pengembangan aplikasi berbasis android untuk mendeteksi perselingkuhan melalui SMS. Adapun hasil percobaan menghasilkan pendeteksian perselingkuhan secara baik.

2) E-ISSN 2442-5168 Volume 1, Nomor 2, Juli-Desember 2015. Klasifikafi Dokumen Temu Kembali Informasi dengan *K-Nearest Neghbour Information Retrieval Document Classified* with K-Nearest Neighbor Endah Purwanti1 Fakultas Sains dan Teknologi Universitas Airlangga.

Sistem temu kembali klasifikasi jurnal ini dilakukan dalam 3 tahap. Tahap pertama adalah pengumpulan data dan informasi berupa jurnal. Tahap kedua adalah analisis sistem yang meliputi proses pengolahan dengan menggunakan text mining, pembobotan pada setiap token dengan menggunakan term frequency-inverse *document frequency* (TF-IDF), menghitung tingkat kesamaan atau *similarity* dari tiap dokumen dengan menggunakan *cosinus similarity* dan melakukan klasifikasi dengan menggunakan *k-Nearest Neighbor*. Tahap ketiga adalah implementasi sistem berbasis desktop dengan menggunakan Netbeans dan MySql. Tahap keempat adalah evaluasi sistem dengan membandingkan nilai *F-Measures* dari nilai k yang digunakan.

Berdasarkan hasil uji coba, dengan menggunakan total dokumen sebanyak 180 dokumen yang terdiri dari 40 dokumen training untuk setiap kategori yaitu Physical Sciences and Engineering, Life Sciences, Health Sciences, dan Social Sciences and Humanities dan 20 dokumen testing menghasilkan nilai Recall sebesar 0.539, Precision sebesar 0.501, dan F-Measures sebesar 0.5193 pada nilai k = 43.

3) Edisi Juni 2015 Volume IX No. 1 ISSN 1979-8911. Penggunaan KNN *(K-NEARST NEIGHBOR*) Untuk Klasifikasi Text Berita yang tidak terkelompokkan pada saat pengklasteran oleh STC (Sufix Tree Clustering) Jumadi, Edi Winarko.

Berita-berita yang dipublikasi oleh para jurnalis melaui Twitter terkadang kurang nyaman untuk dibaca oleh para pembaca berita. Karena berita-berita tersebut ditampilkan secara tersusun beruntun ke bawah pada halaman web tersebut. Tetapi setelah tweet-tweet yang ada dikelompokkan secara tematik jadi semakin menarik karena pembaca dapat memilih berita-berita tertentu yang telah dikelompokkan oleh Algoritma Suffix Tree Clustering (STC). Tetapi pada algoritma ini, masih tetap menghasilkan dokumen-dokumen yang tidak memiliki kelompok. Pada Penelitian ini, dokumen-dokumen tersebut mencoba untuk di klasifikasikan ke dalam kelompok yang ada dengan menggunakan Algoritma K-Nearset Neighbor(KNN).

4) Jurnal Teknologi Informasi, Volume 10 Nomor 1, April 2014, ISSN 1414-9999. Edisi KLASIFIKASI BIDANG KERJA LULUSAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR Nursalim, Suprapedi, dan H. Himawan Pascasarjana Teknik Informatika Universitas Dian Nuswantoro.

Dari hasil analisis menggunakan confusion matrix dan ROC Curve dapat disimpulkan bahwa algoritma data mining k-nearest neighbor memiliki kinerja terbaik untuk klasifikasi bidang kerja lulusan dengan nilai accuracy yaitu 83,33% dan nilai Area Under The Curve (AUC) adalah 0,900.

III. PEKERJAAN DAN DISKUSI HASIL

Data SMS yang digunakan adalah data bulan November 2015 sejumlah 3.120 SMS. Adapun contoh data SMS tersebut dengan singkatan ID, *UserID*(UID), Jabatan(JB), Tanggal Kirim (TGLK), Pesan, *Hanphone*(HP), Nomor Terima (NT) serta Kaitan Kerja(KK) adalah sebagai berikut:

TABEL 1.
CONTOH DATA PESAN PENDEK PADA DATABASE

ID	UID	JB	TGLK	PESAN	HP	NT	KK
1	MMD	Pewarta	11/01/2015 12:34:57	dikirim berita, akademisi: pemerintah pertimbangkan hukuman kebiri kejahatan anak.trims	+628537xxxxx	user-red19	YA
2	LML	Kabiro	11/01/2015 15:08:16	Tmn2 harap kirim berita untuk mengisi rubrik hari minggu. Hari ini kita belum ada berita, sama seperti minggu lalu. Pewarta siaga harap dapat melaksanakan tugas dan tanggungjawabnya. Trims	+628521xxxxx	ya	YA
3	PNK	Support	11/01/2015 17:34:15	Angga, Tolong Cek Server 171 - c:/berita/	+628963xxxxx	user-ts01	YA
4	EFAS	Sekred	11/01/2015 19:15:49	Yth, Bapak/Ibu Mohon hadir pada Rapat Layout Korsip Inggris pada hari Senin, 2 Nopember 2015, pukul 13.00 WIB, di Ruang Rapat Lt. 20. Layout sementara sudah dikirim via email. Terima kasih.	+628138xxxx	efa	YA

ID	UID	JB	TGLK	PESAN	HP	NT	KK
5	user-ts04	Support	11/02/2015 09:33:38	bos ane balik ya ,biasa anter jemput anak ok (Technical Support - user-ts04)	+6281289640333	ip	TIDAK
6	BCH	Sekred	11/02/2015 11:21:24	Rekans, ada surat evaluasi jab fungsional 6 bln an. trims	+6285782932001	sg	YA
7	BCH	Sekred	11/02/2015 11:21:53	Imam B ada surat kenaikan jab fungsional. trims	+6285726083435	ib	YA
8	RNAS	Kadep	11/02/2015 12:35:30	yth pewarta n redaktur dn inggris, diminta kehadiran rekan2 dalam editorial clinic pada rabu (4/11) pkl 14.00 wib s/d selesai dgn pembicara, pak eliswan dan ibu fardah. seyogianya pelaksanaannya pekan lalu. terima kasih atas kehadiran rekan2. ada makan siang yg disiapkan pak jumroni.	+628129016901	yh	YA
9	BCH	Sekred	11/02/2015 18:03:42	LKP "Konf.Pers Hasil SUrvey-Satu Tahun Pem Jokowi-JK Kinerja Birokrasi Pelayanan Publik" Selasa 3/11 pkl 10.45 di Pulau Dua Rest (Komp.Taman Ria Senayan)	+6281385402424	mai	YA
10	AJ	Ombudsman	11/02/2015 15:31:52	Nomor KTP/NIK xxxx, AZ	+628157162445	aj	TIDAK

Dari keseluruhan data yang diambil, terdapat 177 konten yang tidak mengandung kontek kerja didalamnya, sedangkan yang mengandung kontek kerja sejumlah 2.942 konten.

Setelah melakukan pengambilan data, tahap berikutnya melakukan klasifikasi dengan membuat kamus manual, yaitu klasifikasi YA dan klasifikasi TIDAK.

Adapun hasil klasifikasi kerja manual untuk klasifikasi YA terlihat pada table 2.

TABEL 2.
HASIL SELEKSI KLASIFIKASI KERJA MANUAL KAMUS

Pekerjaan & Dideteksi Perkerjaan secara Manual	
SMS (…= kata selanjutnya)	**Hasil Klasifikasi**
Berita Irman Gusman …	YA
Password atas nama …	YA
Sekper Kadif hadir Pleno …	YA
Sekjen MPR terima tamu Kehormatan …	YA
Diminta hadir Rapat liputan …	YA
Berita rubrik minggu ini …	YA
Hadir tanda tangan kpi …	YA
Rapat agenda setting …	YA
Liputan Anugrah BI …	YA
Mengumpulkan Pumpunan …	YA

Sedangkan hasil klasifikasi kerja manual untuk klasifikasi TIDAK terlihat pada Tabel 3.

TABEL 3.
HASIL SELEKSI KLASIFIKASI TIDAK KERJA MANUAL KAMUS

Dideteksi Bukan Perkerjaan secara Manual	
SMS (…= kata selanjutnya)	**Hasil Klasifikasi**
Hay hay ….	TIDAK
Kuliner mana ….	TIDAK
Hujan Lebat ya ….	TIDAK
Rekening danamon …	TIDAK
Link berita Mathlaul….	TIDAK

Dideteksi Bukan Perkerjaan secara Manual	
Temens, mohon infonya ttg ukuran pakaian masing….	TIDAK

Setelah pengklasifikasian kamus manual, kedua hasil data baik secara manual maupun pengolahan pengolahan Text Mining dengan menggunakan algoritma K-NN akan dilakukan perbandingan.

Proses pengolahan data menggunakan algoritma K-NN dilakukan dengan menggunakan tools data mining dengan perangkat lunak rapidminer.

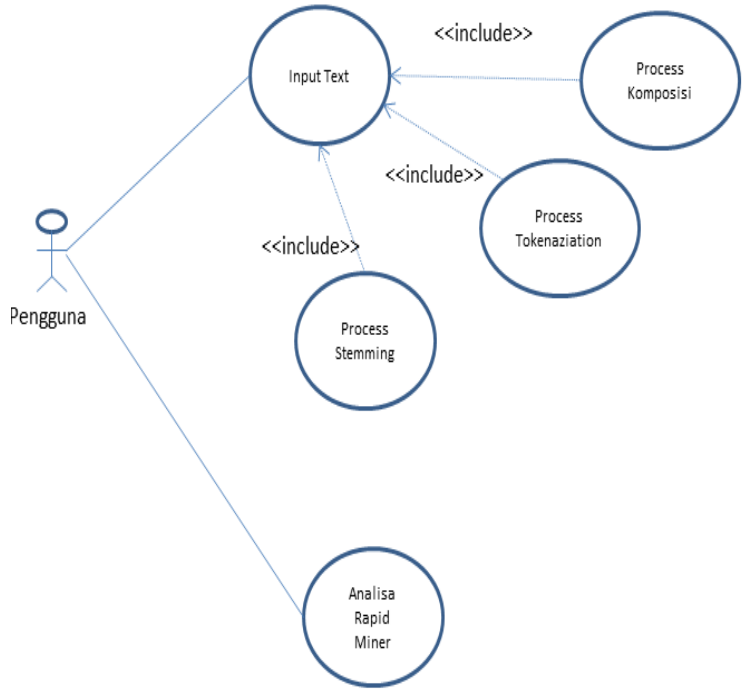
Proses pengolahan manual juga dibuat aplikasinya menggunakan bahasa pemrograman PHP berbasiskan web mining dalam pengolahan data SMS.

Desain aplikasi berbasis WEB tersebut didesain yang mempunyai desain *Use Case*, seperti disajikan pada Gambar 13.

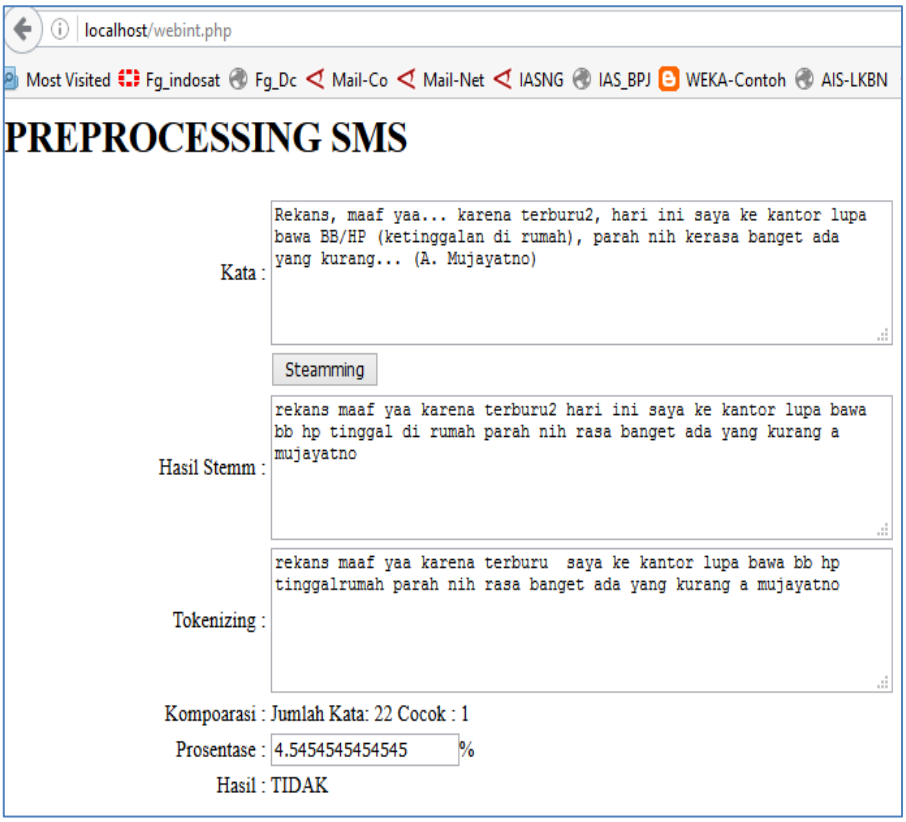
Berdasarkan desain *Use Case* pada Gambar 13, pengguna dapat memasukkan data ke dalam aplikasi. Setelah memasukkan, maka pengguna dapat menekan tombol proses, kemudian aplikasi akan melakukan *filtering* terhadap teks yang telah dimasukkan untuk mendapatkan kalimat yang telah di-*stemming* dan di-*tokenazation*, seperti disajikan pada Gambar 14 dan Gambar 15.

Hasil akhir dari teks yang telah di-*stemming* dan di-*tokenzation* dapat digunakan oleh Pengguna untuk diproses tingkat akurasinya lebih lanjut menggunakan aplikasi Rapidminer.

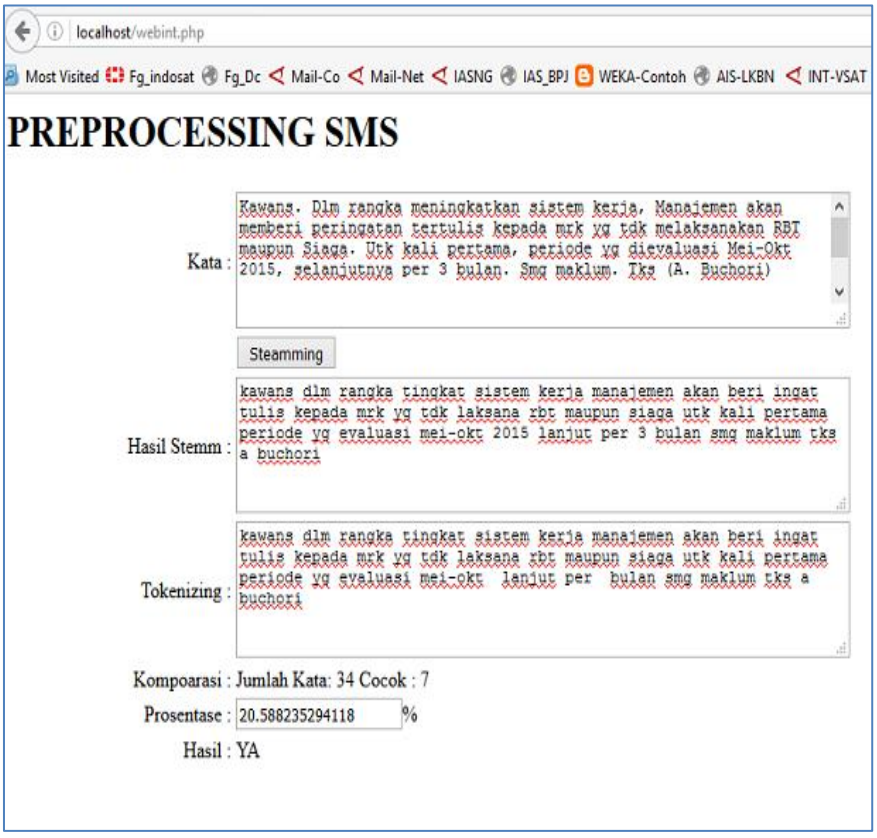
Didalam aplikasi PHP *Web Mining* yang dibuat juga dimasukkan untuk melakukan *filter Stemming* dan *Tokenizing* beserta komposisi dari tiap-tiap kata serta prosentasi kemungkinan serta hasil pengklasifikasiannya.



Gambar 13. Aplikasi *Text Mining filter Stemming* dan *Tokezition* berbasiskan Web.

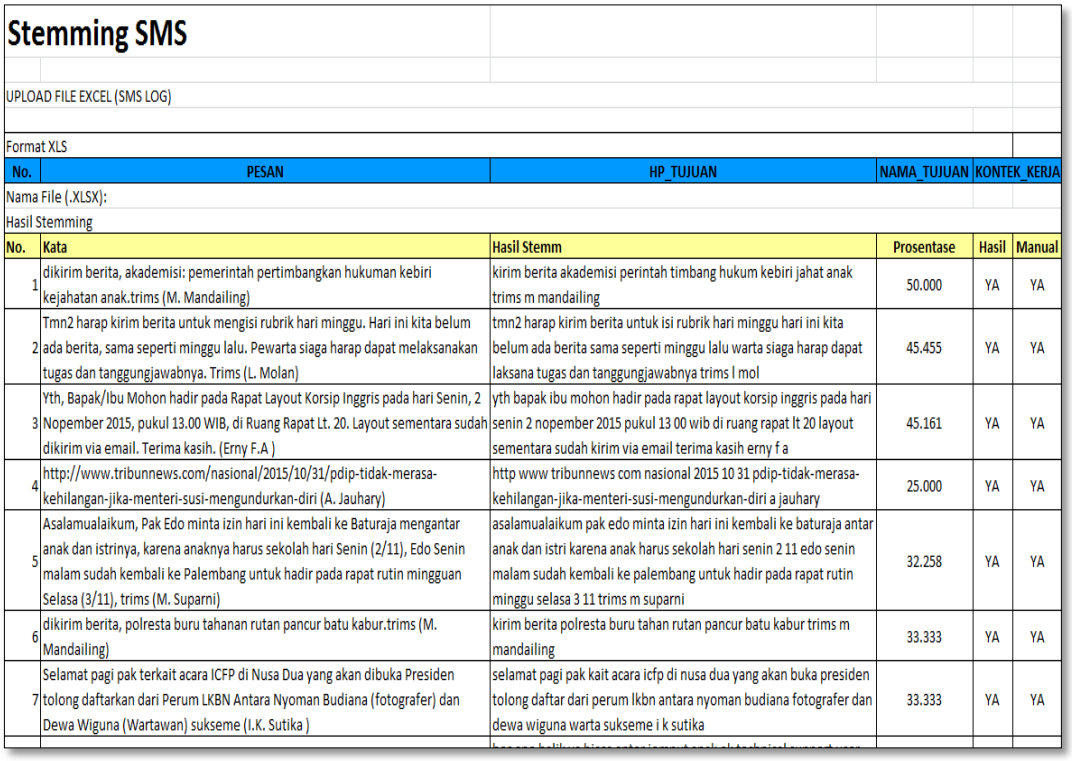


Gambar 14. Content SMS termasuk kategori kerja (YA)



Gambar 15. *Content SMS* bukan kategori kerja (TIDAK)

Setelah proses Stemming dan *Tokezination*, maka data yang telah diproses dimasukkan kembali ke dalam MS Excel untuk mengetahui prediksi secara manual dan perhitungan dari aplikasi untuk prediksi YA mapun TIDAK.



Stemming SMS

UPLOAD FILE EXCEL (SMS LOG)

Format XLS

No.	PESAN	HP_TUJUAN	NAMA_TUJUAN	KONTEK_KERJA

Nama File (.XLSX):

Hasil Stemming

No.	Kata	Hasil Stemm	Prosentase	Hasil	Manual
1	dikirim berita, akademisi: pemerintah pertimbangkan hukuman kebiri kejahatan anak.trims (M. Mandailing)	kirim berita akademisi perintah timbang hukum kebiri jahat anak trims m mandailing	50.000	YA	YA
2	Tmn2 harap kirim berita untuk mengisi rubrik hari minggu. Hari ini kita belum ada berita, sama seperti minggu lalu. Pewarta siaga harap dapat melaksanakan tugas dan tanggungjawabnya. Trims (L. Molan)	tmn2 harap kirim berita untuk isi rubrik hari minggu hari ini kita belum ada berita sama seperti minggu lalu warta siaga harap dapat laksana tugas dan tanggungjawabnya trims l mol	45.455	YA	YA
3	Yth, Bapak/Ibu Mohon hadir pada Rapat Layout Korsip Inggris pada hari Senin, 2 Nopember 2015, pukul 13.00 WIB, di Ruang Rapat Lt. 20. Layout sementara sudah dikirim via email. Terima kasih. (Erny F.A)	yth bapak ibu mohon hadir pada rapat layout korsip inggris pada hari senin 2 nopember 2015 pukul 13 00 wib di ruang rapat lt 20 layout sementara sudah kirim via email terima kasih erny f a	45.161	YA	YA
4	http://www.tribunnews.com/nasional/2015/10/31/pdip-tidak-merasa-kehilangan-jika-menteri-susi-mengundurkan-diri (A. Jauhary)	http www tribunnews com nasional 2015 10 31 pdip-tidak-merasa-kehilangan-jika-menteri-susi-mengundurkan-diri a jauhary	25.000	YA	YA
5	Asalamualaikum, Pak Edo minta izin hari ini kembali ke Baturaja mengantar anak dan istrinya, karena anaknya harus sekolah hari Senin (2/11), Edo Senin malam sudah kembali ke Palembang untuk hadir pada rapat rutin mingguan Selasa (3/11), trims (M. Suparni)	asalamualaikum pak edo minta izin hari ini kembali ke baturaja antar anak dan istri karena anak harus sekolah hari senin 2 11 edo senin malam sudah kembali ke palembang untuk hadir pada rapat rutin minggu selasa 3 11 trims m suparni	32.258	YA	YA
6	dikirim berita, polresta buru tahanan rutan pancur batu kabur.trims (M. Mandailing)	kirim berita polresta buru tahan rutan pancur batu kabur trims m mandailing	33.333	YA	YA
7	Selamat pagi pak terkait acara ICFP di Nusa Dua yang akan dibuka Presiden tolong daftarkan dari Perum LKBN Antara Nyoman Budiana (fotografer) dan Dewa Wiguna (Wartawan) sukseme (I.K. Sutika)	selamat pagi pak kait acara icfp di nusa dua yang akan buka presiden tolong daftar dari perum lkbn antara nyoman budiana fotografer dan dewa wiguna warta sukseme i k sutika	33.333	YA	YA

Gambar 16. Pengolahan data *test bed*.

Data-data tersebut dimasukkan ke dalam perangkat bantu berupa perangkat lunak Rapidminer dan mendapat hasil Perbandingannya, seperti disajikan pada Tabel 4.

TABEL 4.
AKURASI SELEKSI

	true YA	**true TIDAK**	**class precision**
pred. YA	772	32	96.02%
pred. TIDAK	0	14	100.00%
class recall	100.00%	30.43%	

accuracy: 96.15% +/- 0.76% (mikro: 96.15%)

TABLE 5.
TABEL PRESISI SELEKSI TIDAK

	true YA	true TIDAK	class precision
pred. YA	772	32	96.02%
pred. TIDAK	0	14	100.00%
class recall	100.00%	30.43%	

recall: 31.00% +/- 17.44% (mikro: 30.43%) (positive class: TIDAK

Pada Tabel 4 dan Tabel 5, mempunyai hasil pada akurasi seleksi prediksi YA sejumlah 772 benar dan tidak sesuai sejumlah 32, sehingga presisinya adalah 96.02%. Sedangkan prediksi TIDAK mempunyai hasil sejumlah 0 kesalahan dan 14 benar dalam prediksi TIDAK. Sehingga hasil akurasi yang didapatkan sejumlah 96.15%.

IV. KESIMPULAN DAN SARAN

KESIMPULAN

Hasil penelitian yang dilakukan mendapatkan hasil hasil pada akurasi seleksi prediksi YA sejumlah 772 benar dan tidak sesuai sejumlah 32, sehingga presisinya adalah 96.02%. Sedangkan prediksi TIDAK mempunyai hasil

sejumlah 0 kesalahan dan 14 benar dalam prediksi TIDAK. Sehingga hasil akurasi yang didapatkan sejumlah 96.15%.

Sehingga penggunaan dari aplikasi SMS-Gateway yang digunakan dilingkungan LKBN ANTARA telah sesuai dengan kebutuhan selama bulan November 2015. Serta dapat membantu LKBN ANTARA dalam melakukan klasifikasi isi SMS dengan klasifikasi YA dan TIDAK.

SARAN

Penelitian dapat dikembangkan kembali dengan jumlah data yang lebih banyak, tidak hanya pada bulan November 2015. Melainkan data dua tahun dari bulan Januari 2015 hingga bulan Desember 2016, sehingga hasil yang digunakan lebih akurat dan lebih baik lagi.

REFERENSI

[1] Fayyad, Usama, 1996, "Advances in Knowledge Discovery and Data Mining", MIT Press.

[2] Gomes, G. dan Sanchez, R, End-toEnd Quality of Service over Celullar Networks - Data Services Performance and Optimization in 2G/3G

[3] Hermawati, Fajar Astuti. Data Mining. Yogyakarta:ANDI. 2013.

[4] Maimon, Oded & Lior Rokach. *Data Mining and Knowledge Discovery Handbook.* New York: Springer. 2010.

[5] Manning C. D. and Schutze H., Foundations of Statistical Natural Language Processing [M]. Cambridge:MIT Press. 1999.

[6] Narang, N. dan Kasera, S. 2G - Mobile Networks:GSM and HSCSD - Architecture, Protocols, Procedures and Service. New Delhi-Tata Mc Graw Hill Publishing, 2010.

[7] Turban, E., dkk. Decicion Support Systems and Intelligent Systems, Andi Offset. 2005.

[8] Yang Y. and Liu X., A Re-examination of Text Categorization Methods [A].In: Proceedings of 22nd Annual International ACM SIGIR Conference on Researchand Development in Information Retrieval [C]. 1999. Hal: 42-49