KLASIFIKASI DATA SMS CENTER BUPATI PAMEKASAN MENGGUNAKAN NAÏVE BAYES DENGAN MAD SMOOTHING

Badar Said[1] dan Yuliana Melita Pranoto[2]
[1]Teknik Informatika Universitas Madura
[2]Teknik Informatika Sekolah Tinggi Teknik Surabaya
badarsaid@unira.ac.id dan ymp@stts.edu

ABSTRAK

Penelitian ini fokus untuk melakukan klasifikasi data SMS (Short Message Service) masyarakat Kabupaten Pamekasan yang ditujukan kepada Bupati Pamekasan. Data SMS yang akan diklasifikasikan berasal dari database aplikasi SMS Center Bupati Pamekasan. Klasifikasi merupakan suatu proses pengelompokkan berdasarkan kelas yang telah ditentukan sebelumnya. Dalam penelitian ini data SMS diklasifikasikan dalam 15 kelas yaitu Pendidikan, Kesehatan, Infrastruktur, Kriminalitas, Pelayanan Administrasi, Olahraga, Pemerintahan, Pertanian, UKM, Ketertiban, Ekonomi Lemah, keagamaan, Seni dan Budaya, Bencana Alam, dan Lain-lain.

Sebelum melakukan proses klasifikasi terlebih dahulu dilakukan preprocessing seperti menyamakan karakter, penghapusan tanda baca, mengembalikan singkatan, terjemah bahasa daerah (Bahasa Madura), penghapusan angka, penghapusan kata yang tidak penting dalam SMS, dan stemming untuk mengubah kata menjadi kata dasar. Penelitian ini menggunakan algoritma Naïve Bayes dengan Modified Absolute Discounting (MAD) Smoothing. Pembuatan aplikasi menggunakan bahasa pemograman PHP dan Mysql sebagai database.

Dalam beberapa uji coba yang telah dilakukan mendapatkan rata-rata akurasi sebesar 76.83%, dengan hasil akurasi salah satu uji coba mencapai 82,68%. Dan metode MAD Smoothing terbukti dapat meningkatkan akurasi Naïve Bayes sebesar 6,6%.

Kata Kunci: *Klasifikasi, SMS, naïve bayes, MAD Smoothing*

ABSTRACT

This research focuses on classification of SMS (Short Message service) in Pamekasan Regency society sent to Pamekasan Major. SMS that will be classified is from the database of SMS Center Pamekasan application. Classification is a process to categorize data from the previous class. In this research, SMS is classified into 15 classes which is Education, Health, Infrastructure, Criminality, Administrative services, Sport, Governance, Agricultural, small and intermediate enterprise, Orderliness, Poor economy, religion, Art and Culture, Natural disasters, and Others.

Preprocessing done in this research is casefolding, removing punctuation, converting word, vernacular translation (Madura language), removing number, removing stopwords, and stemming. This research utilizes Naïve Bayes algorithm with Modified Absolute discounting (MAD) Smoothing. This application utilizes PHP language programming and Mysql as database.

From the trials, the results show that the average accuracy is 76.83%, one of the results show the average accuracy can reach 82,68%. Therefore, MAD Smoothing method can be used to increase the accuracy of Naïve Bayes until 6,6%.

Key word: Classification, SMS, naïve bayes, MAD Smoothing

I. PENDAHULUAN

Pada era globalisasi saat ini perkembangan teknologi telekomunikasi sangatlah berkembang pesat. Perkembangan tersebut salah satunya dipengaruhi oleh manfaat yang didapatkan oleh para pengguna serta beraneka ragamjenis teknologi yang disajikan.Salah satu teknologi telekomunikasi yang populer adalah SMS (Short Message Service). Teknologi ini sangat praktis dan hanya membutuhkan biaya sangat murah, karena hanya dengan perangkat Handphone dengan harga yang sangat terjangkau teknologi SMS sudah dapat dinikmati. Dengan kepraktisan teknologi ini tidak sedikit perusahaan atau instansi yang memanfaatkan untuk peningkatan kinerja. Tentunya dalam hal penyampaian informasi.

Pemerintah daerah kabupaten Pamekasan merupakan salah satu instansi pemerintah yang telah memanfaatkan teknologi telekomunikasi yaitu dengan adanya Aplikasi SMS Center Bupati. Teknologi ini digunakan untuk mempermudah dan mempercepat penyampaian informasi dari masyarakat kepada Bupati Pamekasan baik berupa pengaduan, pertanyaan, saran ataupun kritik. Sehingga Pemerintah Daerah Kabupaten Pamekasan dapat memberikan pelayanan yang lebih baik. Pesan yang diterima, langsung dijawab oleh Asisten Bupati, tetapi akan menunggu apabila permasalahan tersebut perlu dikomunikasikan dengan Satuan Kerja Perangkat Daerah (SKPD) yang terkait.

Setelah satu tahun aplikasi SMS Center Bupati ini dijalankan, SMS dari masyarakat tersimpan didalam database dalam jumlah besar dan dibiarkan tampa manfaat. Oleh sebab itu penulis bermaksud untuk merancang sebuah aplikasi untuk mengelompokkan atau mengklasifikasikan data SMS tersebut ke dalam beberapa kategori atau kelas, sehingga dapat mengetahui prosentase jumlah SMS untuk setiap kelas dan dapat dipergunakan sebagai bahan untuk evaluasi dan proyeksi.

Untuk mengklasifikasikan data ada beberpa metode yang dapat digunakan seperti SVM, Naïve Bayes, KNN dan lain sebagainya. Untuk penelitian ini penulis memilih metode Naïve Bayes karena teknik ini dikenal sebagai teknik yang paling baik dalam hal waktu komputasi dibandingkan teknik algoritma data mining lainnya . Dalam implementasi metode ini terdapat beberapa metode smoothing yang dilakukan diantaranya Jelinek-Mercer (JM), Dirichlet (Dir), Absolute discounting (AD) dan Two-stage (TS) . Untuk memaksimalkan performa dari Naïve Bayes dalam penelitian ini penulis menggunakan Modified Asolute Discounting (MAD) Smoothing.

II. TUJAN DAN MANFAAT

Penelitian ini merupakan kontribusi pemanfaatan maka perlu disampaikan tujuan yang akan dicapai dan manfaat yang diperoleh dari penelitian ini. Adapun beberapa tujuan yang ingin dilakukan dalam penelitian ini antara lain sebagai berikut :

1. Mengklasifikasikan semua SMS dari masyarakat pada aplikasi SMS Center Bupati Pamekasan mulai bulan Juli sampai Desember tahun 2013 menggunakan Naïve Bayes dengan Modified Absolute Discont Smoothing.
2. Menghitung prosentase jumlah SMS untuk setiap kelas
3. Membandingkan akurasi klasifikasi menggunakan Naïve Bayes dengan MAD Smoothing dan Naïve Bayes tampa Smoothing.

Manfaat yang diharapkan dari penelitian ini adalah memudahkan dalam pengklasifikasian SMS pada aplikasi SMS Center Bupati Pamekasan, serta Bupati dapat mengetahui prosentase jumlah SMS untuk setiap kelas sehingga dapat dijadikan bahan untuk evaluasi dan proyeksi.

III. RUANG LINGKUP

Dalam melakukan penelitian ini melibatkan beberapa data dan proses yang harus dilakukan. Untuk itu perlu disampaikan batasan-batasan mengenai data dan proses tersebut. Beberapa batasan data dan proses diuraikan sebagai berikut:

1. Data Input

Data yang digunakan sebagai input adalah data SMS di database Aplikasi SMS Center Bupati Pamekasan. Aplikasi SMS Center ini mulai dioperasikan pada bulan Juli 2013 dan ditangani langsung oleh Asisten Pribadi Bupati Pamekasan, dengan tujuan menampung aspirasi masyarakat Pamekasan baik berupa pertanyaan, saran maupun kritik. Masyarakat Pamekasan sangat antusias dengan keberadaan Aplikasi SMS Center ini, hal ini terbukti dalam satu hari SMS yang diterima rata - rata 13 SMS dan sampai bulan Desember 2013 sudah mencapai 2134 SMS. Selain SMS dengan menggunakan bahasa Indonesia, juga terdapat SMS dengan bahasa daerah yaitu bahasa Madura walaupun jumlahnya tidak banyak. Data yang akan diklasifikasikan adalah semua data SMS mulai bulan Juli sampai Desember tahun 2013, baik SMS dengan bahasa Indonesia maupun bahasa Madura.

2. Daftar kelas

Dalam penelitian ini jumlah kelas untuk klasifikasi data SMS sebanyak 15 kelas yaitu Pendidikan, Kesehatan, Infrastruktur, Kriminalitas, Pelayanan Administrasi, Olahraga, Pemerintahan, Pertanian, UKM, Ketertiban, Ekonomi Lemah, keagamaan, Seni dan Budaya, Bencana Alam, dan kelas Lain-lain. Kelas yang terakhir yaitu kelas 'Lain-lain' merupakan pengelompokan SMS yang tidak relevan, seperti SMS dari Operator, SMS yang hanya berisi sapaan kepada Bupati Pamekasan dan lain sebagainya. Setelah dihitung jumlah SMS dengan kelas 'Lain-lain' sebanyak 103 SMS.

3. Pre-Processing

Sebelum dilakukan klasifikasi, yang perlu dilakukan terlebih dahulu adalah pre-processing untuk mempersiapkan data agar mudah untuk dilakukan klasifikasi. Ada beberapa tahap pre-processing yang harus dilakukan:

a. Remove Punctuation and Number

Tanda baca dan angka adalah hal yang tidak penting dalam penelitian ini dan harus di hilangkan karena akan mengganggu proses pengklasifikasian. Tanda baca dalam hal ini adalah karakter selain karakter a – z atau A – Z.

b. Casefolding

Untuk mempermudah proses pre-processing selanjutnya, proses ini juga perlu dilakukan yaitu mengubah semua huruf menjadi lowercase atau huruf kecil.

c. Convert Word

Konversi kata dilakukan apabila penulisan kata tersebut tidak baku. Sering sekali penulisan SMS menggunakan singkatan atau penulisan kata tidak lengkap. Hal ini dilakukan dengan menggunakan kamus yang sudah disediakan.

d. Translation

Dalam penelitian ini dibutuhkan proses menerjemahkan Bahasa Madura ke dalam Bahasa Indonesia, walaupun jumlah kata dengan Bahasa Madura tidak banyak namun dapat mempengaruhi akurasi klasifikasi. Hal ini juga dilakukan dengan menggunakan kamus yang sudah disediakan.

e. Remove stopword

Menghilangkan kata yang tidak penting seperti kata penghubung dan kata sapaan juga harus dilakukan untuk meminimalisir waktu yang dibutuhkan dalam proses klasifikasi. dan hal ini dilakukan dengan mendaftar kata – kata yang tidak penting untuk selanjutnya digunakan untuk pengecekan kata tidak penting dalam SMS.

f. Stemming

Menghilangkan imbuhan pada kata untuk mendapatkan kata dasar juga dilakukan untuk menambah akurasi hasil klasifikasi. dalam hal ini penulis menggunakan metode Enhanced Confix Stripping karena merupakan Algoritma stemming kata pada Bahasa Indonesia dengan performa yang paling baik .

Sebelum melakukan 6 tahapan pre-processing diatas semua data SMS harus diklasifikasikan secara manual atau labeling.

4. Subsistem Model Analisis Klasifikasi

Data SMS yang telah melalui proses labeling dan 6 tahapan pre-processing akan yang digunakan sebagai data latih dan data uji untuk membentuk model analisis klasifikasi. Pada penelitian ini dipergunakan metode pembelajaran mesin Naïve Bayes. Naïve Bayes dipilih karena sudah teruji di beberapa penelitian mampu menghasilkan akurasi yang baik. Untuk memaksimalkan performa dari Naïve Bayes dalam penelitian ini, penulis menggunakan Modified Asolute Discounting (MAD) Smoothing yang terbukti dalam penelitian sebelumnya meningkatkan akurasi ketepatan Naïve Bayes dalam klasifikasi.

Naïve Bayes merupakan metode pembelajaran mesin yang memiliki model dalam bentuk probabilitas atau peluang. Data latih yang berupa pasangan SMS dan kelas dijadikan sebagai sumber pembentukan model analisis. Setiap fitur yang merepresentasikan SMS dihitung probabilitas kemunculannya di setiap kelas. Formula untuk menentukan kelas dari SMS ditunjukkan oleh persamaan berikut ini:

$$P(w_k|C_i) = \frac{\max(count(w_k,C_i)-delta,0)+delta\left(N_{uC_i}\right)f(w_k)}{\sum_{w\in V} count(w,C_i)} \quad \ldots\ldots\ldots\ldots \quad (1)$$

Dengan, N_{uC_i}=jumlah kata unik pada *Ci*

$$f(w_k) = P_{unif}(w_k)\sum_{j=1}^{m} count(w_k,C_j)$$

$$P_{unif}(w_k) = \frac{1}{|V|} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

Untuk penelitian ini implementasi Naïve Bayes dengan MAD Smoothing menggunakan delta=0,1. Perhitungan Naïve Bayes tanpa Smoothing juga akan dilakukan untuk mengetahui peningkatan kinerja Naïve Bayes dengan MAD Smoothing. Aplikasi dibangun menggunakan bahasa pemrograman PHP dan Mysql sebagai database.

5. Data Output

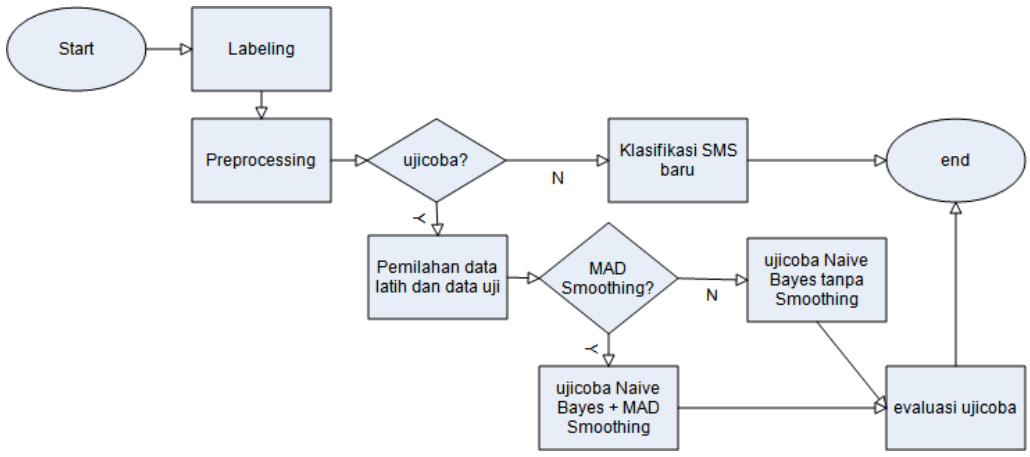
Beberapa hasil ujicoba klasifikasi akan disajikan untuk mengetahui perbandingan akurasi antara Naïve Bayes dengan MAD Smoothing dan Naïve Bayes tanpa Smoothing. Selain itu sistem akan mengklasifikasikan seluruh SMS dan akan disajikan dalam bentuk persentase untuk setiap kelas.

IV. PERANCANGAN SISTEM

Dari analisis data dan deskripsi sistem secara umum perlu dirancang alur proses mulai dari awal sampai didapatkan hasil klasifikasi. Terdapat beberapa proses yang harus dilakukan yaitu pelabelan, preprocessing, pemilahan data ujicoba, ujicoba klasifikasi menggunakan Naïve Bayes dengan MAD Smoothing, ujicoba klasifikasi menggunakan Naïve Bayes tanpa Smoothing, evaluasi hasil ujicoba, dan klasifikasi SMS baru.

Proses pelabelan dilakukan dengan semi manual dengan memilih kelas yang telah disediakan pada setiap tampilan SMS, data hasil labeling menjadi input untuk proses preprocessing dengan enam tahap yaitu casefolding, remove punctuation, convert word, translation, remove number, remove stopword, dan stemming. Data hasil preprocessing akan dipilah menjadi data latih dan data uji dengan ketentuan lima bulan sebagai data latih dan satu bulan sebagai uji, hal itu dilakukan dengan beberapa variasi. Dari beberapa variasi data ujicoba dilakukan klasifikasi menggunakan Naïve Bayes dengan MAD Smoothing dan Naïve Bayes tanpa Smoothing. Hasil ujicoba langsung dievaluasi untuk mengetahui perbandingan akurasi antar kedua ujicoba.

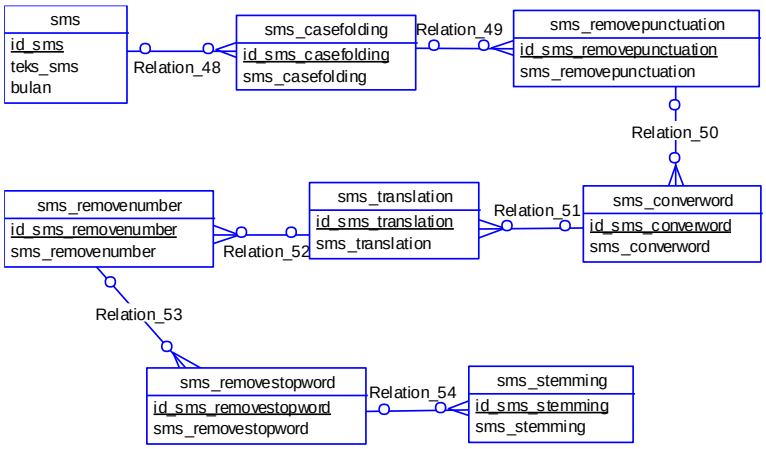
Selain melakukan ujicoba sistem juga dapat melakukan proses klasifikasi SMS baru dari masyarakat Kabupaten Pamekasan dengan data latih berasal dari seluruh data SMS hasil preprocessing. Dan hasil klasifikasi SMS baru ditambahkan menjadi data latih untuk klasifikasi SMS baru selanjutnya. Diagram alur atau flowchart sistem ditunjukkan pada gambar 1.



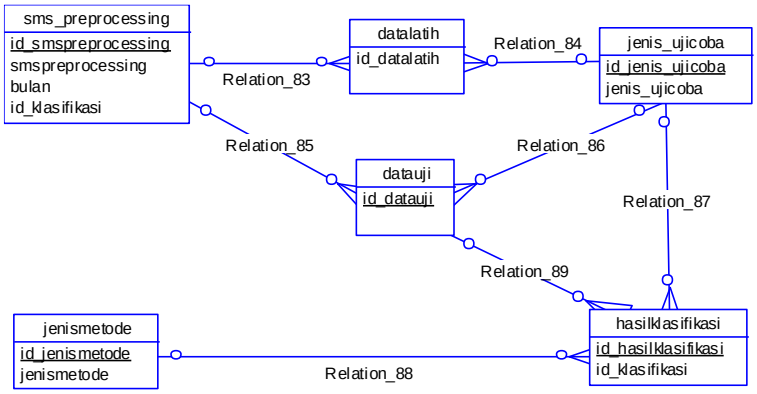
Gambar 1. Flowchart Sistem

Diagram Alur diatas menggambarkan tahapan proses secara umum dari keseluruhan sistem. Dari beberapa proses masih terdapat beberapa subproses yang masih belum diuraikan seperti preprocessing yang masih terdiri dari beberapa tahap.

Untuk merancang database dilakukan dengan pembuatan Conceptual Data Model dan Physical Data Model seperti gambar 2 dan gambar 3



Gambar 2. CDM untuk preprocessing



Gambar 3. CDM untuk ujicoba klasifikasi

Dengan melakukan generate Conceptual Data Model diatas menjadi Physical Data Model maka tahap perancangan database telah selesai dan dapat dilakukan generate kedalam mysql sebagai database. Pembuatan kode program dilakukan dengan bantuan tool dreamweaver untuk mempermudah desain interface dan pengetikan kode PHP.

V. HASIL DAN PEMBAHASAN

Dari pembuatan desain interface dan pengetikan kode PHP, diperlihatkan hasil implementasi proses labeling dengan memilih kelas yang telah ditentukan yang ditampilan di masing-masing SMS. Hal ini dilakukan satu persatu untuk semua data yaitu 2134 SMS. Dalam tahap labeling ini terkadang ditemui SMS yang menyampaikan lebih dari satu kategori atau kelas, untuk kasus tersebut penulis tetap memilih satu kelas

yang lebih dominan dari kelas yang lain berdasarkan jumlah kata yang berpengaruh dalam isi SMS.

Untuk semua tahapan preprocessing juga dilakukan otomatis oleh sistem. Selanjutnya proses ujicoba klasifikasi dengan enam variasi jumlah data latih dan data uji dengan ketentuan lima bulan sebagai data latih dan satu bulan sebagai data uji. Setelah proses ujicoba klasifikasi dilakukan evaluasi terhadap semua hasil ujicoba klasifikasi dengan memperhatikan waktu dan akurasi yang dihasilkan.

Untuk proses evaluasi akurasi klasifikasi dilakukan dengan memperhatikan confusion matrix.

Dari setiap confusion matrix setiap ujicoba klasifikasi ditemukan akurasi ujicoba klasifiasi masing-masing variasi ujocoba. Pada tabel 1 ditunjukan perbandingan akurasi serta rata-rata yang dihasilkan.

Tabel 1. Akurasi setiap ujicoba klasifikasi

Metode	Akurasi Ujicoba (%)						Rata - rata
	1	*2*	*3*	*4*	*5*	*6*	
NB dengan MAD Smoothing	82,68	75,87	73,76	77,64	76,33	74,69	76,83
NB tampa Smoothing	77,65	70,22	67,59	72,20	70,12	63,58	70,23
Selisih							6,6

Pada tabel 1 dapat diketahui rata-rata akurasi ujicoba klasifikasi menggunakan Naïve Bayes dengan MAD Smoothing sebesar 76,83%, sedangkan rata-rata akurasi ujicoba klasifikasi menggunakan Naïve Bayes tampa Smoothing sebesar 70,23%. Terbukti klasifikasi mengggunakan Naïve Bayes dengan MAD Smoothing dapat meningkatkan akurasi klasifikasi sebesar 6,6% apabila dibandingkan dengan klasifikasi menggunakan Naïve Bayes tampa Smoothing.

Selain analisa akurasi ujicoba klasifikasi dari hasil evaluasi setiap ujicoba klasifiasi juga dapat dianalisa waktu yang diperlukan dalam proses ujicoba klasifikasi, sebagaimana ditunjukan pada tabel 2 berikut :

Tabel 2. Waktu proses setiap ujicoba klasifikasi

Metode	Akurasi Ujicoba (%)						Rata - rata
	1	*2*	*3*	*4*	*5*	*6*	
NB dengan MAD Smoothing	1,2	3,23	3,35	2,46	3,09	1,51	2,47
NB tampa Smoothing	0,55	3,41	4,52	2,25	2,19	1,12	2,34
Selisih							0,13

Pada tabel 6.2 dapat diketahui rata-rata waktu yang dibutuhkan untuk proses ujicoba klasifikasi menggunakan Naïve Bayes dengan MAD Smoothing adalah 2,47 menit, sedangkan rata-rata waktu yang dibutuhkan untuk proses ujicoba klasifikasi menggunakan Naïve Bayes tampa Smoothing adalah 2,34 menit. Jadi waktu yang

dibutuhkan untuk proses ujicoba klasifikasi Naïve Bayes dengan MAD Smoothing lebih lama 0,13

Sebagai tindak lanjut dari penelitian ini disediakan fitur untuk klasifikasi SMS baru, dengan ketentuan yang menjadi data latih adalah seluruh data SMS hasil preprocessing. Setiap SMS baru yang akan diklasifikasikan terlebih dahulu melalui proses preprocessing seperti yang dilakukan terhadap data latih. SMS baru yang sudah diklasifikasikan otomatis oleh sistem dapat diperbaiki atau diubah apabila terjadi kesalahan klasifikasi, dan selanjutnya akan diakumulasikan dengan SMS hasil preprocessing sebagai data latih untuk proses klasifikasi SMS baru selanjutnya.

VI. PENUTUP

Berdasarkan fakta yang terjadi selama penelitian dan analisa dari hasil penelitian dapat ditarik beberapa kesimpulan sebagai berikut:

1. Pada penelitian ini rata-rata akurasi klasifikasi menggunakan Naïve Bayes dengan MAD Smoothing sebesar 76,83%, bahkan dalam salah satu ujicoba klasifikasi mencapai akurasi 82,68%. Sedangkan rata-rata akurasi klasifikasi menggunakan Naïve Bayes tampa Smoothing sebesar 70,23%.
2. Metode Modified Absolute Discounting Smoothing terbukti meningkatkan Akurasi klasifikasi menggunakan Naïve Bayes sebesar 6,6%. Dengan penambahan waktu 0,13 menit.
3. Kesalahan klasifikasi sering disebabkan oleh tidak seimbangnya jumlah SMS di setiap kelas pada data latih.

VII. DAFTAR PUSTAKA

[1] Gilang Jalu Selo W.T, Budi Susanto, Rosa Delima, *Implementasi Naïve Bayesian Classifier Untuk Kasus Filtering SMS Spam*,Universitas Kristen Duta Wacana, 2013.

[2] Q. Yuan, G. Cong, and N.M. Thalmann, *Enhancing Naive Bayes with Various Smoothing Methods for Short Text Classification*, WWW Companion, 2012.

[3] Astha Chharia, R.K. Gupta, *Enhancing Naïve Bayes Performance with Modified Absolute Discount Smoothing Method in Spam Classification*, IJARCSSE, 2013.

[4] Shruti Aggarwal, Devinder Kaur, *Naïve Bayes Classifier with Various Smoothing Techniques for Text Documents*, IJCTT, 2013.

[5] Karl-Michael Schneider, *Techniques for Improving the Performance of Naive Bayes for Text Classification*, citeseerx, 2005

[6] Junaedi Widjojo. *Prediksi Jenis Kelamin dan Usia untuk Blog Berbahasa Indonesia dengan Metode Klasifikasi Teks yang Dilengkapi dengan Pemilihan Fitur Terbaik*. iSTTS. 2012

[7] Dwi Widiastuti. *Analisa Perbandingan Algoritma Svm, Naive Bayes, Dan Decision Tree Dalam Mengklasifikasikan Serangan (Attacks) Pada Sistem Pendeteksi Intrusi. Jurusan Sistem Informasi*, Universitas Gunadarma. 2011

[8] Arifin, A.Z., I.P.A.K. Mahendra dan H.T. *Ciptaningtyas. Enhanced Confix Stripping Stemmer and Ants Algorithm for Classifying News Document in Indonesian Language*. ICTS. 2009 Komputindo.