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

Badar Said

Program Studi Teknik Informatika, Fakultas Teknik, Universitas Madura
badarsaid@unira.ac.id

Abstrak

Pemerintah Daerah Kabupaten Pamekasan memiliki data keluhan dan kritik dari masyarakat dalam bentuk SMS yang berasal dari Aplikasi SMS Center Bupati. Data tersebut dapat digunakan sebagai bahan untuk melakukan evaluasi, yaitu dengan mengelompokkan atau mengklasifikasikan data SMS tersebut kedalam beberapa kategori seperti Pendidikan, Kesehatan, Infrastruktur, Kriminalitas, Pelayanan Administraasi, Olahraga, Pemerintahan, Pertanian, Usaha Kecil Menengah, Ketertiban, dan lain-lain. Kemudian dilakukan proses analisis sentimen untuk setiap kategori. Dari hasil klasifikasi dapat diketahui prosentase jumlah SMS untuk setiap kategori serta prosentase SMS positif dan negatif untuk masing-masing kategori, sehingga dapat dievaluasi sektor-sektor yang masih banyak permasalahan dan dapat dilakukan proyeksi untuk memperbaiki sektor tersebut.

Dalam penelitian ini metode Naïve Bayes digunakan untuk proses klasifikasi karena teknik ini dikenal sebagai teknik yang paling baik dalam hal waktu komputasi dibandingkan teknik algoritma data mining lainnya . dan untuk metode smoothing yang sertakan adalah Modified Asolute Discounting (MAD) dengan tujuan untuk memperbaiki kinerja dari metode Naïve Bayes.

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%. Kesalahan klasifikasi sering disebabkan oleh tidak seimbangnya jumlah SMS di setiap kelas pada data latih. Dan persentase SMS Positif hanya 0,52 % dari total SMS sebanyak 2134 SMS.

Kata kunci: Klasifikasi, Analisis Sentimen, SMS, Naïve Bayes, MAD Smoothing

1. Pendahuluan

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 di jawab oleh Asisten Bupati, tetapi akan menunggu apabila permasalahan tersebut perlu di komunikasikan 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. Pada penelitian sebelumnya telah dilakukan klasifikasi terhadap data SMS Center Bupati Pamekasan dalam 15 kategori, sehingga dapat diketahui prosentase jumlah SMS pada setiap kategori. Namun dari hasil klasifikasi tersebut prosentase jumlah SMS yang besar pada beberapa kategori belum tentu merepresentasikan bahwa beberapa kategori tersebut harus mendapatkan perhatian lebih, karena belum diketahui apakah isi SMS tersebut berisi keluhan dan kritikan dari masyarakat atau berisi sanjungan dan ucapan terimakasih. Oleh sebab itu penulis bermaksud untuk merancang sebuah aplikasi untuk mengelompokkan atau mengklasifikasikan serta analisis sentiment pada data SMS tersebut, sehingga dapat mengetahui prosentase jumlah SMS untuk setiap kelas serta dapat diketahui jumlah SMS positif dan negatif pada setiap kategori. Sehingga diharapkan dapat dipergunakan sebagai bahan untuk evaluasi dan proyeksi.

(Dwi Widiastuti:2011) Untuk mengklasifikasikan data ada beberapa 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. (Quan Yuan, Gao Cong, Nadia M. Thalmann:2012) Dalam implementasi metode ini terdapat beberapa metode smoothing yang dilakukan diantaranya Jelinek-Mercer (JM), Dirichlet (Dir), Absolute discounting (AD) dan Two-stage (TS). Diantara beberapa metode smoothing tersebut Absolute discounting merupakan yang terbaik untuk klasifikasi teks pendek. (Astha Chharia, R.K. Gupta:2013) Untuk lebih memaksimalkan kinerja dari Naïve Bayes dalam penelitian ini penulis menggunakan Modified Asolute Discounting (MAD).

2. Tinjauan Pustaka

2.1. Naïve Bayes

(Junaedi Widjojo:2012) Merupakan pengklasifikasian dengan metode probabilitas dan statistikyang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitumemprediksi peluang di masa depan berdasarkan pengalaman di masasebelumnya sehingga dikenal sebagai teorema Bayes. Secara sederhana,Naïve Bayes menggunakan kemiripan fitur antara data training dan datatesting dimana nantinya akan diambil class yang paling mirip dari datatraining tersebut. Dalam penilaian, algoritma ini dikenal sebagai algoritmayang sederhana, cepat dan berakurasi tinggi.Teorema tersebut dikombinasikan dengan "naive" dimana diasumsikan kondisi antar atribut saling bebas. Pada sebuah dataset, setiap baris/dokumen diasumsikan sebagai vector dari nilai-nilai atribut <x1,x2,...,x3> dimana tiap nilai-nilai menjadi peninjauan atribut Xi(i∈[1,n])). Setiap baris mempunyai label kelas ci ∈ {c1,c2,...,ck} sebagainilai variabel kelas C, sehingga untuk melakukan klasifikasi dapat dihitungnilai probabilitas p(C=ci|X=xj), dikarenakan pada Naïve Bayesdiasumsikan setiap atribut saling bebas, maka persamaan yang didapat adalah sebagai berikut:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Keterangan :
c : sebuah kelas
d : sebuah dokumen

Dimana setiap peluang pada masing-masing kelas akan dikalikan dan akanmenghasilkan nilai Naïve Bayes pada masing-masing rumus tersebut.Nilai tertinggi pada klasifikasi ini akan menjadi hasil klasifikasi dari NaïveBayes tersebut. Berikut adalah rumusnya:

$$C_{map} = max_{c \in C} \hat{P}(c|d) = max_{c \in C} \hat{P}(c) \prod_{1 \le k \le nc} P(t_k|c)$$

$\hat{P}(c|d)$ adalah estimasi probabilitas Kategori *c* terhadap dokumen *d*
$\hat{P}(c)$ adalah estimasi probabilitas *prior* dari dokument yang muncul di kateori *c*
$P(t_k|c)$ adalah probabilitas bersyarat dari term $t_k$yang muncul di kategori *c*

2.2. Metode Smoothing pada Naïve Bayes

(Gilang Jalu Selo W.T, Budi Susanto, Rosa Delima:2013) Smoothing perlu dilakukan untuk menghindari nilai 0 pada hasil perhitungan probabilitas bersyarat dari term tertentuyang muncul di sebuah kategori. (Quan Yuan, Gao Cong, Nadia M. Thalmann:2012) Beberapa metode smoothing yang dapat digunakan adalah :

a. Laplace smoothing

$$p(w|c_i) = \frac{1 + c(w, c_i)}{|V| + \sum_{w' \in V} c(w', c_i)}$$

b. Jelinek-Mercer (JM) smoothing

$$p_\lambda(w|c_i) = (1 - \lambda) \frac{c(w, c_i)}{\sum_{w' \in V} c(w', c_i)} + \lambda\, p(\mathrm{w}|\mathrm{C})$$

c. Dirichlet (Dir) smoothing

$$p_\mu(w|c_i) = \frac{c(w, c_i) + \mu\, p(w|C)}{\sum_{w' \in V} c(w', c_i) + \mu}$$

d. Absolute Discounting (AD) smoothing

$$p_\delta(w|c_i) = \frac{\max(c(w, c_i) - \delta, 0) + \delta |c_i|_u p(w, C)}{\sum_{w' \in V} c(w', c_i)}$$

Dengan
$\delta \in [0,1]$dan $|c_i|_u$adalah jumlah kata unik pada c_i

e. Two-stage (TS) smoothing

$$p_{\lambda,\mu}(w|c_i) = (1 - \lambda) \frac{c(w, c_i) + \mu\, p(w|C)}{\sum_{w' \in V} c(w', c_i) + \mu} + \lambda\, p(\mathrm{w}|\mathrm{C})$$

2.3. Modified Absolute Discounting Smoothing pada Naïve Bayes

(Astha Chharia, R.K. Gupta:2013) MAD Smoothing merupakan modifikasi dari Absolute Discounting Smoothing, perbedaannya tidak melakukan perhitungan $P(w_k)$. Metode ini tidak mempertimbangkan kemungkinan kata dalam model koleksi, melainkan menganggapnya sebagai fungsi dari kata, yang merupakan probabilitas distribusi seragam dikalikan dengan terjadinya kata dalam model pengumpulan. Dengan rumus:

$$f(w_k) = P_{unif}(w_k) \sum_{j=1}^{m} count(w_k, C_j) \quad \ldots\ldots 2.2$$

Dengan, $P_{unif}(w_k) = \frac{1}{|V|}$

Sehingga, $P(w_k|C_i)$ dihitung dengan :

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

2.4. Evaluasi Kinerja Klasifikasi

Menurut Han dan Kamber (2011:365) *Confusion matrix* merupakan alat yang digunakan untuk menganalisa seberapa baik klasifikasi mengenali *tuple* dari kelas yang berbeda. *True Positif* dan *True Negatif* memberikan informasi ketika klasifikasi benar, sedangkan *False Positif* dan *False Negatif* memberikan informasi ketika klasifikasi salah. Contoh *confusion matrix* sebagai berikut:

Tabel 2.1 Contoh *Confusion matrix*

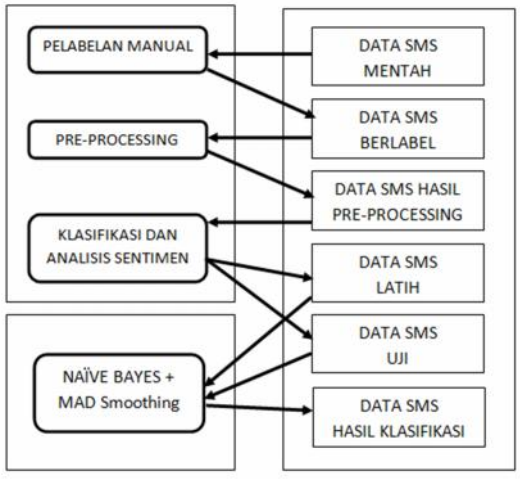
		Predicted class	
		Yes	No
Actual class	Yes	TP	FN
	No	FP	TN

Akurasi dapat dihitung dengan membagi jumlah hasil klasifikasi benar dengan jumlah seluruh data, dengan rumus :

$$Akurasi = \frac{TP + TN}{TP + FP + TN + TN}$$

3. Metode Penelitian

Adapun untuk diagram alir dan uraian tahapan penelitian ini adalah sebagai berikut:



Gambar 3.1 Diagram Alir Penelitian

Agar tercapai sasaran dan tujuan dari penelitian ini, maka dilakukan langkah analisis terhadap berbagai kebutuhan yang diperlukan yaitu sebagai berikut :

1. Analisis Kebutuhan Data

Data yang digunakan sebagai input adalah *data S*MS dari Aplikasi SMS Center Bupati Pamekasan. Aplikasi SMS Center ini mulai dioperasikan pada bulan Juli 2013 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. Analisis Kebutuhan Kategori

Setelah melakukan konsultasi dengan pihak terkait daftar kategori dalam pengklasifikasian adalah sebagai berikut : Pendidikan, Kesehatan, Infrastruktur, Kriminalitas, Pelayanan Administraasi, Olahraga, Pemerintahan, Pertanian, Usaha Kecil Menengah, Ketertiban, Ekonomi Lemah, Keagamaan, Seni Budaya dan Lain-lain. Kategori yang terakhir yaitu 'Lain-lain' merupakan pengelompokan SMS yang tidak relevan, seperti SMS dari Operator, SMS yang hanya berisi sapaan kepada Bupati Pamekasan dan lain sebagainya. Untuk proses analisis sentiment diklasifikasikan pada dua kategori yaitu SMS positif dan SMS negatif.

3. Analisis Kebutuhan Preprocessing

Sebelum dilakukan pengklasifikasian perlu dilakukan preprocessing dengan tujuan untuk mempersiapkan data agar layak untuk dilakukan klasifikasi. Beberapa preprocessing yang akan dilakukan adalah penghapusan tanda baca dan angka, mengubah teks ke dalam bentuk teks kapital, memperbaiki singkatan, menterjemahkan teks dari bahasa madura menjadi bahasa indonesia, menghilangkan kata-kata yang tidak berpengaruh dalam proses klasifikasi (kata penghubung) dan menghilangkan imbuhan.

4. Hasil dan Pembahasan

Pemilahan data dilakukan untuk melakukan 6 kali ujicoba klasifikasi. Pemilahan pertama data SMS yang diterima bulan Juli sebagai data uji dan data SMS yang diterima bulan Agustus – Desember sebagai data latih, Pemilahan kedua data SMS yang diterima bulan Agustus sebagai data uji dan data SMS yang diterima bulan Juli, September – Desember sebagai data latih, Pemilahan ketiga data SMS yang diterima bulan September sebagai data uji dan data SMS yang diterima bulan Juli – Agustus dan Oktober – Desember sebagai data latih, Pemilahan keempat data SMS yang diterima

bulan Oktober sebagai data uji dan data SMS yang diterima bulan Juli – September dan Nopember – Desember sebagai data latih, Pemilahan kelima data SMS yang diterima bulan Nopember sebagai data uji dan data SMS yang diterima bulan Juli – Oktober, Desember sebagai data latih, Pemilahan keenam data SMS yang diterima bulan Desember sebagai data uji dan data SMS yang diterima bulan Juli – Nopember sebagai data latih.
Hasil 6 ujicoba yang telah dilakukan untuk 15 kelas yanag telah ditentukan sebagai berukut:

Tabel 4.1 Hasil Ujicoba Klasifikasi

No	Ujicoba	Klasifikasi benar	Klasifikasi salah
1	*Pertama*	296	62
2	*Kedua*	349	111
3	*Ketiga*	371	132
4	*Keempat*	243	70
5	*Kelima*	258	80
6	*Keenam*	121	41

Rata-rata akurasi dari semua ujicoba adalah sebagai berikut:

Tabel 4.2 Akurasi Setiap Ujicoba Klasifikasi

Ujicoba	Akurasi (%)
1	82,68
2	75,87
3	73,76
4	77,64
5	76,33
6	74,69
Rata-rata	76,83

Hasil ujicoba klasifikasi SMS positif dan negatif sebagai berikut:

Tabel 4.3 Hasil klasifikasi SMS positif dan negatif pada 6 ujicoba

Uji coba	Jml SMS	Negatif		Positif	
		Jumlah	%	Jumlah	%
1	358	357	99.7	1	0.3
2	460	459	99.8	1	0.2
3	503	501	99.6	2	0.4
4	313	310	99	3	1
5	338	335	99.1	3	0.9
6	162	161	99.4	1	0.6

Dari data pada tabel 4.3 jumlah SMS Positif hanya 11 SMS yaitu 0,52 % dari total SMS sebanyak 2134 SMS.

Untuk melakukan sentiment analisis dilakukan ujicoba klasifikasi untuk mengetahui SMS positif dan negatif pada masing-masing kelas kecuali kelas ‘lain-lain’. Data SMS dalam kelas ‘Lain-lain’ tidak diperhitungkan karena hanya berisi sapaan kepada Bupati dan pemberitahuan dari operator kartu seluler. Dari proses klasifikasi diketahui jumlah dan persentase SMS positif dan negatif pada masing-masing kelas sebagai berikut :

Tabel 4.4 Hasil Klasifikasi SMS Positif dan Negatif

Kelas	Jml SMS	Negatif		Positif	
		Jml	%	Jml	%
Pendidikan	457	457	100	0	0
Kesehatan	75	67	89.33	8	10.67
Infrastruktur	469	468	99.79	1	0.21
Kriminalitas	44	44	100	0	0
Pelayanan Administrasi	45	45	100	0	0
Olahraga	94	93	98.94	1	1.06
Pemerintahan	265	264	99.62	1	0.38
Pertanian	39	39	100	0	0
UKM	30	30	100	0	0
Ketertiban	227	227	100	0	0
Ekonomi Lemah	148	148	100	0	0
keagamaan	94	94	100	0	0
Seni dan Budaya	35	35	100	0	0
Bencana Alam	20	20	100	0	0
Total		**2031**		**11**	
Persentase (%)		**99.46**		**0.54**	

Dari table 4.4 dapat diketahui SMS Positif terbesar pada pada kelas Kesehatan yaitu 8 SMS atau 10.67 % dari 75 SMS pada kelas tersebut.

5. Kesimpulan

Berdasarkan hasil beberapa ujicoba dapat disimpulkan beberapa hal 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%.
2. Kesalahan klasifikasi sering disebabkan oleh tidak seimbangnya jumlah SMS di setiap kelas pada data latih.
3. Persentase SMS Positif hanya 0,52 % dari total SMS sebanyak 2134 SMS. Tersebar pada kelas Kesehatan, Infrastruktur, Olahraga, dan Pemerintahan. Hal ini menyatakan bahwa mayoritas data SMS berisi SMS Negatif yaitu sebesar 99,48%.

Daftar Pustaka

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

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

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

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

Karl-Michael Schneider, 201, "*Techniques for Improving the Performance of Naive Bayes for Text Classification*", citeseerx.

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

Shruti Aggarwal, Devinder Kaur, 2005, "*Naïve Bayes Classifier with Various Smoothing Techniques for Text Documents*", IJCTT.