**Klasifikasi Jenis Pesan pada Data SMS Menggunakan Metode Klasifikasi Naïve Bayes**

**Laporan Tugas Akhir**

***Natural Language Processing***

****

**FERO RESYANTO 1301154318**

**DHUKI DWI RACHMAN 13011543XX**

**Program Studi Sarjana S1 Teknik Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**Bandung**

**2018**

# **ABSTRAK**

SMS (Short Message Service) atau pesan singkat merupakan layanan yang disediakan oleh suatu jasa telekomunikasi bagi para penggunanya untuk bertukar pesan. Dengan banyaknya pengguna layanan SMS dewasa ini, banyak SMS yang kurang penting dan mengganggu penggunanya, yaitu dengan adanya SMS sampah atau biasa disebut SPAM. Tidak sedikit juga SMS yang berisi promo-promo dari berbagai pihak bahkan dari penyedia layanan SMS itu sendiri. SMS promo atau SMS spam tersebut dianggap kurang penting bagi kebanyakan pengguna layanan telekomunikasi. Oleh karena itu, pada penelitian tugas akhir ini, penulis melakukan pengklasifikasian terhadap SMS kedalam 3 kelas yaitu kelas normal, spam dan promo. Pengklasifikasian SMS tersebut menggunakan algoritma klasifikasi Naïve Bayes. Data Set yang digunakan berjumlah 1143 pesan. *Preprocessing* yang digunakan adalah *casefolding* sedangkan Feature Selection yang digunakan *tokenize,* *punctuations handling* dan *spell correction*. Pengujian dilakukan dengan membagi SMS menjadi data latih dan data uji. Hasil akurasi dari pengujian dengan menggunakan klasifikasi Naïve Bayes yaitu 73,3% pada skenario pertama dan 93% pada skenario kedua.

Kata kunci: Klasifikasi, Naïve bayes, SMS.

# **PENDAHULUAN**

Pada tugas akhir ini, penulis akan melakukan klasifikasi terhadap pesan atau SMS (Short Message Service) kedalam 3 kelas, yaitu kelas normal, spam dan promo. SMS normal merupakan SMS yang dikirimkan sesama pengguna ( person to person ) layanan telekomunikasi. Kemudian sebuah SMS dikatakan SMS promo jika dalam SMS tersebut terdapat kata-kata yang berhubungan dengan suatu promo layanan atau produk. Kata-kata tersebut seperti kata promo, bonus, layanan, download dan sebagainya. Untuk sebuah SMS spam biasanya merupakan pesan yang dikirim oleh sistem layanan telekomunikasi atau terdapat alamat URL didalam SMS tersebut. Jika tidak terdapat ciri-ciri dari SMS promo ataupun SMS spam maka SMS tersebut termasuk kelas normal.

Input yang digunakan pada sistem yaitu sekumpulan SMS yang disimpan pada file .csv kemudian akan diklasifikasi berdasarkan model *training* pada sistem yang telah dibuat sebelumnya. Keluaran dari sistem yang dibuat yaitu nilai akurasi klasifikasi yang dihasilkan oleh sistem menggunakan klasifikasi Naïve Bayes.

# **DATASET**

Dataset yang digunakan pada penelitian ini yaitu dataset SMS spam Bahasa Indonesia v1.0 yang diterbitkan oleh yudi wibisono pada halaman wordpressnya. Dataset ini merupakan dataset yang dibuat oleh salah satu mahasiswa beliau pada penelitiannya dengan lisensi creative commons. dataset tersebut disimpan pada file dengan ekstensi .csv. Jumlah dataset yang digunakan yaitu 1143 SMS. Dataset SMS spam Bahasa Indonesia ini dapat didownload secara gratis untuk kepentingan pembelajaran dan penelitian pada halaman URL : <http://nlp.yuliadi.pro/dataset> .

Berikut ini beberapa contoh dataset :

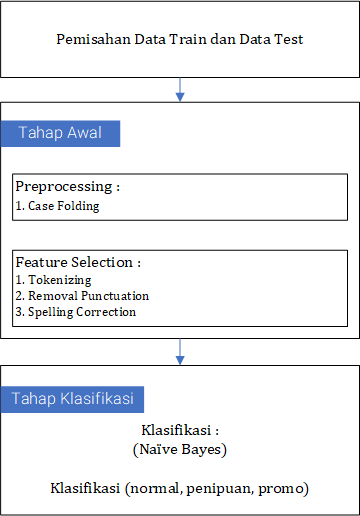
|  |  |  |
| --- | --- | --- |
| No | Teks | Label |
| 1 | [PROMO] Beli paket Flash mulai 1GB di MY TELKOMSEL APP dpt EXTRA kuota 2GB 4G LTE dan EXTRA nelpon hingga 100mnt/1hr. Buruan, cek di tsel.me/mytsel1 S&K | 2 |
| 2 | 2.5 GB/30 hari hanya Rp 35 Ribu Spesial buat Anda yang terpilih. Aktifkan sekarang juga di \*550\*905#. Promo sd 30 Nov 2015. Buruan aktifkan sekarang. S&K | 2 |
| 3 | INFO RESMI PT.INDOFOOD No.Anda t'dftr pemenang UNDIAN POP MIE edisi 2013 Mndptkn 1unit "NISSAN JUKE" PIN;a1c5h79 info,kunjungi;www.popmieget.webs.com | 1 |
| 4 | INFO PT.M-KIOS hsil pengundian berkat isi pulsa, Selamat No.HP 0852952xxxxx m-dpt Hadiah PIN Pemenang anda 94ad653 Info klik www.berkat-isiulang-pulsa.webs.com | 1 |
| 5 | Besok saya ada rapat dinas.. Selasa aja ya... | 0 |
| 6 | Ini hujan semalaman ruang tengah sama kamar kayanya gabisa di nyalain (:() harus nunggu kering baru bisa nyala, smg ujannya berenti | 0 |

Terdapat 2 atribut pada dataset yaitu atribut teks dan label. Atribut teks berisi teks SMS yang digunakan pada tugas akhir ini, sedangkan atribut label merupakan label dari teks SMS tersebut. Teks SMS tersebut telah diberi label sebelumnya secara manual untuk digunakan sebagai data latih pada proses klasifikasi menggunakan Naïve bayes. SMS yang diberi label 2 menunjukan bahwa SMS tersebut merupakan SMS promo, SMS yang diberi label 1 menunjukan bahwa SMS tersebut merupakan SMS spam dan terakhir SMS yang diberi label 0 menunjukan bahwa SMS tersebut merupakan SMS normal.

Jumlah total instance atau SMS pada dataset ini yaitu 1143 SMS, dimana terdapat 568 SMS dengan kelas SMS normal, 335 SMS dengan kelas SMS penipuan dan 239 SMS dengan kelas SMS promo. Dari 1143 SMS tersebut diambil 15 SMS secara acak untuk dijadikan data tes dimana 5 SMS normal, 5 SMS spam dan 5 SMS promo.

# **Analisis Fitur dan Metode Klasifikasi**

Berikut ini merupakan tahapan pengerjaan yang di lakukan untuk mengklasifikasikan teks sms menjadi 3 kelas yaitu kelas normal, penipuan, promo seperti pada gambar 1.1.

* Fitur Selection

Ada 3 fitur yang dipilih pada kasus ini, fitur pertama adalah *Tokenizing* fitur ini dipilih karena untuk menghitung probabilitas menggunakan naïve bayes di perlukan probability tiap kata sehingga tiap kalimat harus di pecah menjadi per kata.

Fitur kedua *Removal Punctuation* fitur ini digunakan karena ternyata setelah di *tokenize*, banyak sekali tanda baca yang menjadi fitur seperti : ‘.’ ’[’ sehingga kami menghapus fitur yang seperti itu karena tidak mencirikan suatu kelas.

Fitur ketiga adalah *Spelling Correction* fitur ini digunakan karena setelah mengamati data banyak sekali kata yang disingkat karena ini pesan sms sehingga kami mengkoreksi kata yang disingkat tersebut agar kata yang sama artinya tetapi berbeda penulisan tidak menjadi fitur yang berbeda.

* Klasifikasi Naïve Bayes

Klasifikasi Naïve Bayes digunakan pada tugas akhir ini karena menurut kami, pada penelitian sebelumnya, metode klasifikasi Naïve Bayes memiliki nilai akurasi yang cukup bagus dibandingkan dengan metode klasifikasi lainnya seperti *fuzzy* atau *Decision Tree* pada kasus pengklasifikasian teks.

* Libray yang digunakan

Library yang digunakan untuk memudahkan pengerjaan adalah sebagai berikut :

1. csv : Library ini digunakan untuk menconvert file .csv ke arraylist
2. nltk : Library ini digunakan untuk memabntu dalam men-tokenize kata
3. numpy : Library ini digunakan untuk memodifikasi array
4. string : Library ini digunakan untuk membantu *removal punctuation*
5. re : Library yang digunakan untuk membantu membaca file .txt
6. collections : Library yang digunakan untuk melakukan perhitungan pada file .txt
7. corpus sastrawi : corpus ini digunakan untuk memfilter kata baku dan tidak baku

* Preprocessing

Preprocessing yang dilakukan adalah *case folding* karena apabila kita tidak menggunakan preprocessing tersebut maka satu kata yang mempunyai karakter yang sama persis tetapi beda huruf besar atau kecilnya maka kedua kata tersebut akan berbeda sehingga menyebabkan *redundancy*.

* Parameter yang digunakan

Kami menggunakan klasifikasi Naïve Bayes dengan leplace smoothing. Leplace smoothing merupakan suatu cara untuk menangani nilai dengan probabilitas 0 (nol). Pada setiap perhitungan datanya akan ditambah 1 (satu) dan tidak akan membuat perbedaan pada estimasi probilitas sehingga dapat menghindari probabilitas 0 (nol).

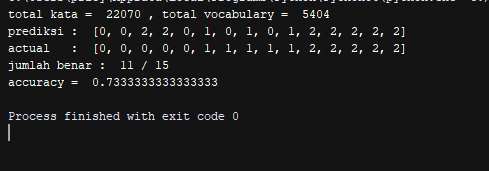


Pada rumus diatas, kami menggunakan 3 parameter yaitu vocabulary, class\_word dan unique\_vocab. Vocabulary merupakan kata unik pada kelas tertentu sedangkan class\_word merupakan total seluruh kata pada kelas tertentu dan yang terakhir yaitu unique\_vocab merupakan total seluruh kata unik yang ada pada dataset.

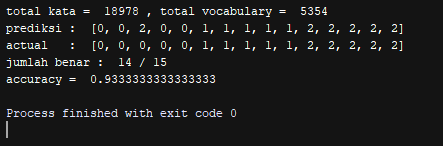
# **Evaluasi dan Analisis Hasil Klasifikasi**

Pada tugas ini kami melakukan percobaan klasifikasi terhadap pesan sms menjadi 3 kelas yaitu kelas normal, penipuan, dan promo. Kami menggunakan klasifikasi naïve bayes karena pada paper sebelumnya yang sudah melakukan klasifikasi, dinilai sangat cocok dengan menggunakan klasifikasi tersebut dengan tingkat akurasi mencapai 90% [1]. Dari percobaan yang kami lakukan terdapat perbedaan yaitu dimana kami langsung mengklasifikasikan data uji ke tiga kelas tersebut sedangkan pada paper sebelumnya mereka membagi 2 kali klasifikasi (spam, notspam) dan (penipuan, promo). Selain itu kami juga membuat 2 skenario yang berbeda untuk proses klasifikasi dimana scenario tersebut adalah :

1. Skenario 1 : Kami melakukan klasifikasi dengan :
2. Preprocessing : Case folding
3. Feature Selection : Tokenize



1. Skenario 2 : Kami melakukan klasifikasi dengan :
2. Preprocessing : Case folding
3. Feature Selection : Tokenize, Removal Punctuation, Spelling Correction



Dari hasil percobaan tersebut didapatkan hasil klasifikasi yang lebih baik didapatkan pada scenario 2 dengan akurasi 93%. Spelling Correction dan Removal Punctuation menurut kami cocok dijadikan feature selection pada kasus ini. Karena menurut kami *Punctuation* tidak penting apabila dijadikan 1 fitur utuh tersendiri sedangkan Spelling Correction sangat cocok digunakan pada untuk data – data yang mempunyai format baku.

# **Kesimpulan**

Feature Selection Spelling Correction terutama, sangat cocok digunakan pada kasus ini tetapi kami mengimplementasikannya hanya terhadap 2 kelas yaitu promo dan penipuan karena format penulisannya, setelah kami amati memiliki format penulisan yang rapih, tertata, dan jarang sekali *typo* maka kami mengkoreksi kata – kata yang sering dituliskan tetapi terdapat kesalahan dalam penulisannya sebaliknya untuk kelas kata yang normal setelah kami amati kata – kata yang tidak baku dan typo sering kami jumpai dan setelah kami coba Spelling Correction pada kata – kata *typo* tersebut kami mendapatkan akurasi 63% pada skenario kedua tersebut. Maka dari itu pada kasus kami kata - kata yang tidak baku menjadi ciri dari kelas kata normal sedangkan kata yang baku menjadi ciri kelas kata penipuan dan khusunya promo.

Kendala yang dialami adalah kami kesulitan untuk memfilter kata yang mempunyai struktur seperti : itu... , sayaaaa, …bapak, wihh. Kata – kata tersebut yang sulit kami proses untuk menjadi kata utuh tanpa punctuation karena kami juga memerlukan kata yang mempunyai punctuation seperti \*363#, <http://tsel.me/tse>, 1,5GB menjadi fitur. Sehingga kami biarkan kata tersebut menjadi fitur utuh.

# **Referensi**

|  |  |
| --- | --- |
| [1] | R. Ferdila, "APLIKASI SMS SPAM FILTERING PADA ANDROID MENGGUNAKAN ALGORITMA NAÏVE BAYES. S1 skripsi, Universitas Pendidikan Indonesia.," 2016. |