# 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 *tokenize*, *casefolding*, *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%.

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.