# Aplikasi untuk Identifikasi Short Message Service (SMS) Spam Berbasis Android

*An Application for Identification of Short Message Service (SMS) Spam Based on Android*

**Zaid Romegar Mair*[1], Ahmad Ashari[2]**
Jurusan Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta
Jurusan Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta
e-mail: *[1]mair@mail.ugm.ac.id, [2]ashari@ugm.ac.id

## Abstrak

Penggunaan smartphone sebagai salah satu teknologi komunikasi yang murah, mudah dan cepat, sering dimanfaatkan oleh sebagian orang yang tidak bertanggung jawab untuk melakukan tindak kejahatan melalui SMS. Maraknya penipuan melalui SMS dapat menyebabkan terjadinya kejahatan berupa spamming SMS, sehingga dalam penelitian ini dibangun sebuah sistem tool berbasis web untuk Melakukan identifikasi pada smartphone android dalam pengkategorian spam SMS, yang diimplementasikan dengan menggunakan algoritma Naive Bayessian Filter untuk menentukan nilai probabilitas isi SMS dari smartphone Android Samsung Galaxy Young Kernel NAND.

Metode yang digunakan adalah model proses forensik, yang terdiri dari tahap pemeliharaan, pengumpulan, pemeriksaan, analisis dan pelaporan. Pengambilan data dilakukan dengan cara mengektrak pesan yang ada dalam memori internal (ROM) dari handphone Android menggunakan tool AFLogical-OSE dengan memilih data SMS yang disimpan dalam /mnt/sdcard/forensics dan kemudian ditransfer ke komputer via kabel data.

Berdasarkan hasil penelitian yang dilakukan pengujian terhadap 65 jenis sms, 43 sms yang diidentifikasi sebagai spam dan 22 sms bukan spam. Analisis data menggunakan pendekatan kata (n-gramar) maka diperoleh hasil yang sebagian besar adalah spam.

**Kata kunci**: Model Proses Forensik, *Naive Bayessian Filter, spam.*

## Abstract

The use of smartphones as a communication technology that is cheap, easy and fast, is often used by some people who are not responsible for committing crimes through SMS. Rampant fraud through SMS may cause crime in the form of SMS spam, so in this study constructed a system of web-based tool to conduct analysis on smartphone android to identify categorization of SMS, which is implemented using Naive Bayessian Filtering algorithm for determining the value of the probability SMS contents of the smartphone Android Samsung Galaxy Young Kernel NAND.

The method used is a forensic process model, which consists of the maintenance Preservation, Acquisition, Examination, Analysis and Reporting. Data were collected by means extraction messages from internal memory (ROM) of Android phones using AFLogical-OSE tool by selecting the data stored in the / mnt / sdcard / forensics and then transferred to a computer via a data cable.

Based on the results of research conducted tests on 65 types of sms, 43 sms, identified as spam and 22 sms identified as not spam. Data analysis approach word (n-grammar) of the obtained results are mostly spam.

**Keywords**: Forensics Process Model, Naive Bayessian Filter, spam.

## 1. Pendahuluan

Perkembangan teknologi yang semakin pesat, dapat menimbulkan permasalahan bagi pengguna teknologi itu sendiri, semakin maju kehidupan masyarakat, maka kejahatan juga ikut semakin maju, hal tersebut dapat dilihat dari aplikasi yang ada pada *smartphone* Kejahatan juga menjadi sebagian dari hasil budaya itu sendiri, ini berarti bahwa semakin tinggi tingkat budaya dan semakin modern suatu bangsa, maka semakin modern pula kejahatan itu dalam bentuk, sifat dan cara pelaksanaannya. *Handphone* merupakan salah satu bentuk teknologi utama yang digunakan oleh orang untuk berkomunikasi dengan sesamanya dan tidak perlu menghabiskan waktu untuk bertemu secara fisik, salah satu teknologi komunikasi yang murah, mudah dan cepat yang digemari masyarakat sebagai layanan publik adalah *Short Message Service* (SMS). Maraknya penipuan melalui SMS dapat menyebabkan terjadinya kejahatan berupa *spamming* SMS, baik yang diberitakan media maupun tidak, sudah banyak korban dari spamming SMS (Hoog, 2011)

Hingga akhir tahun 2011, jumlah pelanggan telekomunikasi selular di Indonesia mencapai 250 juta pelanggan. Angka tersebut melampaui jumlah penduduk Indonesia yang mencapai sekitar 240 juta penduduk. Trend kejahatan di Indonesia yang melibatkan perangkat digital ini mengalami peningkatan signifikan sejak tahun 2010 dan tahun 2011. Pada awalnya hanya 3 kasus kejahatan pada tahun 2006 dan tahun 2007, kemudian naik menjadi 7 dan 15 kasus ditahun 2008 dan tahun 2009. Pada tahun 2010 terjadi peningkatan yang signifikan baik jumlah kasus maupun barang bukti elektronik, yaitu menjadi 52 kasus dengan jumlah barang bukti 214 item. Pada awal desember 2011 barang bukti elektronik yang diperiksa dan dianalisis berjumlah lebih dari 400 item yang berasal dari 60 kasus, hingga Oktober 2012 diperkirakan telah masuk lebih 400 kasus. Pada saat *handphone* yang digunakan seseorang sebagai alat untuk mengorganisasikan kejahatan maka *handphone* tersebut dapat disita oleh aparat penegak hukum sebagai salah satu barang bukti. sehingga ketika ada barang bukti *handpone* yang disita dari pelaku kejahatan, maka dapat diperiksa secara benar sesuai dengan prinsip-prinsip dasar *digital* forensik (Quick dan Alzaabi, 2011; Lessard dan Kessler, 2010)(

## 2. Metodologi Penelitian

Metode yang digunakan dalam menyelesaikan permasalahan ini adalah model proses forensik (*The Forensic Process Model*) yang terdapat pada Gambar 1 yaitu ada lima fase metode dalam penanganan kasus digital forensik yaitu :

1 Tahap *Preservation* : Pada tahap ini semua barang bukti digital yang ada dijaga secara khusus agar tidak berubah. Penjagaan dapat dilakukan berupa penjagaan fisik atau penjagaan dengan cara digital. Penjagaan barang bukti dilakukan dengan membuat suatu area tanpa sinyal, dengan menggunakan kantong *faraday* (suatu kantong yang terbuat dari bahan tertentu untuk meminimalisasi sinyal frekuensi radio *handphone)* atau *jammer* (suatu peralatan untuk mengacak sinyal frekuensi radio). Dilakukan fotografi terhadap barang bukti, dokumentasi dengan pencatatan merek, model dan hal-hal lain yang berkaitan dengan *handphone.* Implemantasi tahap *preservation* pada topik penelitian ini yaitu dengan mempersiapkan objek penelitian berupa *smartphone* android dan *tool* untuk ekstrak data pada smartphone.

2 Tahap pengumpulan *(Acquisition*) : Data SMS yang dikumpulkan adalah semua data yang ada pada *smartphone* android dan dilakukan *capture* data terhadap *provider* yang ada, dengan mencari bukti-bukti penyusuapan yang dilakukan oleh sistem untuk

mendeteksi *spam* SMS dengan cara mengenali isi pesan terhadap kamus Bahasa Indonesia.

3 Tahap pengujian (*Examination)* : Pada tahap ini, dilakukan pencarian informasi yang tersembunyi dan mengungkapkan dokumentasi yang relevan. Pemeriksaan dilakukan pada SMS dan pengambilan itu diperiksa dengan menggunkan metode algoritma bayes 1 dan bayes 2.

4 Analisa *(Analysis)* : Pada saat data diekstrak, penyidik melakukan analisis terhadap data yang berisi informasi, sehingga penyidik dapat merumuskan kesimpulan dalam menggambarkan data dari file ekstrak. Informasi apa yang disampaikan, siapa yang melakukan, kapan pesan tersebut disampaikan, dimana pesan tersebut disampaikan. Pemeriksaan terhadap sms yang ada, bgaimana pesan terkirim, dan mengapa ini terjadi. Tahap analisis ini menggunakan algoritma bayes.

5 Perhitungan probabilitas berdasarkan algoritma bayesian

Pada awalnya, Bayesian filter ini harus di-training terlebih dahulu menggunakan sejumlah spam dan sejumlah ham . Bayesian filter akan menghitung probabilitas lokal dari suatu kata, misalnya kata “adult”, untuk muncul di kelompok spam sms. Probabilitas lokal ini dapat dirumuskan sebagai berikut :

$$P_{local-spam} = N_{spam} / (N_{spam} + N_{non\text{-}spam})$$

dimana : $P_{local-spam}$ = probabilitas suatu kata “x” terdapat pada spam-mail $N_{spam}$ = jumlah spam sms dengan kata “x” di dalamnya

N non-spam = jumlah non-spam sms dengan kata “x” di dalamnya

6 Laporan (*Reporting)* Penulisan laporan mulai dari tahap pengambilan data hingga analisis melalui data yang diperoleh dari penyelidikan *mobile forensic*. Pengambilan laporan terhadap analisis terhadap spam.

## 3. Hasil dan Pembahasan

### 3.1 Membaca SMS dari Perangkat Mobile

Pembacaan SMS dilakukan pada perangkat *smartphone* Samsung Galaxy Young GT-S5360 dengan menggunakan aplikasi AFLogical-OSE yang menghasilkan file berektensi csv yang berisi record pesan pada memori internal *handphone*. Uraian dari tabel ekstrak data pada smartphone android samsung galaxy young meliputi _id, thread_id, address, person, date, protocol, read, status, type, replay_path, subject, body, service_center, locked, error_code, seen, deletable dan delivery_date. Data yang diambil adalah data-data yang dibutuhkan dalam proses analisis yaitu _id, address, date dan body dan disimpan dalam field yang berbeda yaitu id, nomor, tanggal dan pesan. Jumlah SMS yang digunakan dalam penelitian ini adalah 65 pesan, dengan rincian sesuai pada Tabel 1.

Dilakukan import data SMS dengan tahapan-tahapan memasukkan data hasil dari capture yang disimpan dengan nama SMS.csv ke dalam database untuk memudahkan pengujian isi pesan dan mendapatkan informasi detail SMS, kemudian memilih dan mengkonversi data yang akan digunakan pada tahap analisis. Setiap record data yang diperoleh dari beberapa kolom dipisahkan dengan tanda koma (,) yang kemudian disimpan dalam tabel sms. Tanggal pesan direpresentasikan dalam format Y-m-d H:i:s (tahun-bulan-tanggal jam:menit:detik) untuk memudahkan pembacaan oleh pengguna. Teks pesan hasil pembacaan disaring menggunakan fungsi bawaan php mysql_real_escape_string untuk memformat isi pesan agar bisa disimpan dengan baik dan fungsi get_string_between untuk

mengganti tanda petik (“) di awal dan akhir teks pesan dengan tag html untuk menebalkan huruf yaitu  yang berguna untuk menandai awal dan akhir suatu teks pesan.

Hasil import data disajikan dalam bentuk tabel menggunakan perintah dasar html dan skrip php untuk menampilkan data per baris agar mudah dibaca dan diidentifikasi oleh pengguna. Pengguna juga diberi wewenang untuk menghapus data sms yang tidak dibutuhkan dalam proses analisis.

**Tabel 1**. Beberapa SMS yang digunakan dalam penelitian

| ID | Nomor | Waktu Kirim | Isi Pesan |
|---|---|---|---|
| 6731 | +6287745631362 | 2013-08-29 22:13:42 | Uangnya dikirim saja kenomor rekening ini: 1530396-548, Bank BCA a/n DADANG ARSETO |
| 6730 | +6287745631362 | 2013-08-29 22:13:38 | Maaf, dengan tidak mengurangi rasa hormat kami, saya Ibu Hj. ANISAH yg sudah liat mobil Bpk/Ibu yang mau dijual untuk nego harga hubungi suami saya Pak H. GUNAW |
| 6728 | +6287745631362 | 2013-08-29 22:13:29 | Slmt!!No anda Mendpt hadiah Rp.75jt.dari TELKOMSELpoin Diundi tadi mlm Pukul.23:30.wib Di RCTI.info Hub: 085217777509 Drs.H.SRIANTO Pengirim: +777 |
| 6727 | +6287745631362 | 2013-08-29 22:12:52 | INFO RESMI !! Selamat Anda Terpilih sbgi Pemenang' semarak Undian PT.INDOFOOD INDONESIA No pin: d377hi8 Info Pin&hadiah Klik: www.semarak-undianpopmie.webs.com |
| 6726 | +6287745631362 | 2013-08-29 22:12:46 | Plg Yth! No' Anda 081310xx m-dpt Hadiah All new avanza dr PT.TELKOMSEL PIN Anda:b89c7h9 Info klik: www.tsel-resmi777.jimdo.com A/hub;0016282333306308 |
| 6725 | +6287745631362 | 2013-08-29 22:12:41 | TOGEL SGP EDISI SENIN 10/06/2013 ANGKA JITU 40,76,65,53,89 di jamin tbs,stlh tbs anda harus krim plsa 100rb sebagai maharnya dkrm ke no; 082329969711 KI KARSO |
| 6681 | +6285273155339 | 2013-08-27 06:39:19 | Dik cak kok tgh d pondok meri jam 2 jn bingi |
| 6678 | 6616 | 2013-08-24 23:09:43 | Welcome to Telkomsel GPRS Service! Your service has been activated. Please visit www.telkomsel.com for more information |
| 6677 | 543202 | 2013-08-24 23:09:28 | Telkomsel wants to send handset configuration. Please reply with YES to this message to reveive the configuration. |

## 3.2 Analisis Pesan

Proses analisis pesan dimulai dari penentuan n-grammer. N-grammer diperoleh dari pemecahan pesan menjadi bagian-bagian yang lebih kecil yaitu kata/ frase. Apabila potongan kata tersebut terdapat spasi (“ “) di dalamnya, maka karakter spasi diubah menjadi karakter kosong (“”). Kemudian hasilnya diubah menjadi huruf kecil menggunakan fungsi php `strtolower` dan disimpan dalam buffer yang selanjutnya dihitung frekuensi masing-masing n-grammer.

## 3.3 Optimalisasi Database

Proses optimalisasi dilakukan dengan menghapus data n-grammer yang frekuensinya kurang dari atau sama dengan 1 dan jangkauan prosentase probabilitasnya (dihitung dari nilai maksimum probabilitas dikurangi nilai minimum probabilitas) lebih dari 0.3. Perhitungan ini dimaksudkan agar penentuan nilai probabilitas spam lebih akurat. langkah

optimalisasi *database* dengan cara membuat tabel sementara (*temporary table*) untuk menampung data yang jumlah ngrammer lebih dari 1. Kemudian menghapus data dari table knowledge_base yang mempunyai prosentase yang tidak digunakan untuk perhitungan.

**Tabel 2**. Beberapa n-grammer yang diperoleh dari hasil perhitungan

| n-Grammer | Kepemilikan | Frekuensi Muncul | Probabilitas |
|---|---|---|---|
| Rumah | Spam | 2 | 1 |
| Tanah | Spam | 2 | 1 |
| Yg | Spam | 3 | 1 |
| Jual | Spam | 1 | 1 |
| Ibu | Spam | 4 | 1 |
| Ayu | Spam | 1 | 1 |
| Sdh | Spam | 2 | 1 |
| Survei | Spam | 1 | 1 |
| Dan | Spam | 5 | 1 |
| Berminat | Spam | 1 | 1 |
| U | Spam | 16 | 1 |
| Membeli | Spam | 1 | 1 |
| Mslh | Spam | 1 | 1 |
| Hrg | Spam | 2 | 1 |

### 3.4 Perhitungan Probabilitas Frase Spam dan Non Spam

Probabilitas frase spam dan non spam dihitung menggunakan fungsi dari *library* Naïve Bayessian Filtering yaitu `$spam->isItSpam($text,'spam')/100` dan untuk frase non spam digunakan fungsi yaitu `$spam->isItSpam($text,'ham')/100`. Dalam penelitian ini, penulis menggunakan algoritma Naïve Bayessian Filtering karena hasil yang diperoleh lebih akurat. perhitung probabilitas menggunakan algoritma Naïve Bayessian Filtering versi 1 (Mahmoud dan Mahfouz, 2012; Mukhtidi, 2012).

**Tabel 3**. Beberapa hasil identifikasi isi SMS

| No | Pengirim | Pesan | Awal | Bayes v.1 |
|---|---|---|---|---|
| 1 | +6287745631362 | RUMAH&TANAH yg mau di jual,sy IBU AYU sdh SURVEI dan berminat u/membeli Mslh hrg Mhn hub suami saya 085282157222 Dr.HJ.SOPIAN | Spa m | spam (98.64 of spam)% |
| 2 | +6287745631362 | Sy Dr. Hj FATIMA sdh cocok mengenai Rumah dan Tanah Bpk/Ibu, tlg jgn ditawarkan lg ke yg lain. Soal Hrg negosiasi ke suami sy di 082375833300 Dr. H. HENDRA IRAW | Spa m | spam (95.92 of spam)% |
| 3 | +6287745631362 | Beliin dulu mama pulsa As/20ribu di nomor barunya mama ini nomornya 085211428218 secepatnya penting, saya tunggu ya dan ini Nomornya orang aku pinjam | Spa m | spam (90.64 of spam)% |
| 4 | +6287745631362 | SELAMAT! Simcard anda mendapatkan 1 unit AVANZ | Spa m | ham (76.33 of spam)% |
| 5 | +6287745631362 | Bapak tukokno pulsa As ndisek 50 ribu iki no anyar bapak, 082330528392. Cepat saiki, tak enteni penting, bapak ada masalah di kantor polisi, jgn telpon/sms dulu | Spa m | ham (29.86 of spam)% |

Apabila hasil yang diperoleh dari versi 1 lebih dari 0.85 (85%) maka disimpulkan bahwa isi SMS tersebut diidentifikasi sebagai spam. Sebaliknya apabila hasil perhitungan probabilitas versi 1 diperoleh hasil kurang dari atau sama dengan 0.85 (85%), maka isi SMS

tersebut diidentifikasi sebagai ham (non spam). Tabel 3 berikut merupakan beberapa hasil identifikasi isi SMS.

### 3.5 Laporan Hasil Identifikasi Spamming SMS

Identifikasi spamming SMS dilakukan untuk apakah suatu SMS teridentifikasi sebagai spam atau non spam sehingga dapat dilakukan antisipasi terhadap penipuan yang marak terjadi di masyarakat. Pencarian jejak dari tindakan penipuan melalui spamming sms diperoleh dari file .csv. File tersebut diperoleh dari memori internal smartphone Android menggunakan tools AFLogical OSE atau tools lain yang memiliki fungsi sama. File .csv yang diperoleh tersebut diimport ke dalam database kemudian dianalisis untuk mengetahui untuk mengetahui probabilitas sebagai spam atau non spam dari isi pesan tersebut, untuk menganalisis data sms diperlukan adanya frase atau kata yang dicurigai sebagai spam dan berfungsi sebagai pembanding isi pesan. Jika hasil perbandingan diperoleh nilai probabilitas dengan threshold (nilai ambang) tertentu, maka sms tersebut dikategorikan sebagai spam.

## 4. Kesimpulan

1. Sistem yang dibangun merupakan aplikasi berbasiskan web dibuat dengan menggunakan bahasa pemrograman *Predefine Hypertext Preprocessor*, dengan probabilitas *threshold*.
2. Proses analisis pesan dimulai dari penentuan n-grammer. Berdasarkan pengujian terhadap 65 jenis sms dengan implementasi algoritma *naïve bayesian filtering*, maka terdapat 43 sms yang diidentifikasi sebagai spam dan 22 sms yang bukan spam.

## Ucapan Terimakasih

Kedua orang tua (Alm. Ayahanda dan Uminda), saudara ku ajo, cak, ayuk, kiai, atu, atin, kak cik. Bapak Ahmad Ashari, selaku pembimbing utama tesis, Bapak dan Ibu dosen Program Studi Ilmu Komputer FMIPA UGM, Teman-teman Magister Ilkom UGM, Pelmaha muja-muju Yogyakarta, Bapak Yusuf Yudi Prayudi, Bapak Ahmad Syakir Maksum, candidate my wife dan Pihak-pihak lain yang tidak bisa disebutkan satu persatu yang telah memberikan motivasi, perhatian dan bantuannya selama ini kepada penulis, hanya doa ku yang tulus pada Allah untuk membalas kebaikan teman, sahabat dan saudara semuanya.

## Daftar Pustaka

Hoog, A., 2011, Android Forensics Investigation, Analysis and Mobile Security for Google Android, United State of America.

Quick, D. and Alzaabi, M., 2011, Forensic Analysis of the Android File System YAFFS2, Conference Proceeding, pp.1-11.

Lessard, J. and Kessler G.C., 2010, Android Forensics : Simplifying Cell Phone Examination, Int. J. small scale digital device forensics., Vol. 4(1), pp.1–12.

Mahmoud, T.M. and Mahfouz, A.M., 2012, SMS Spam Filtering Technique Based on Artificial Immune System, Int. J. Computer Science Issues, Vol. 9, pp.589-597, Egypt.

Mukhtidi, K., 2012, Sistem SMS Spam Detector untuk SMS Berbahasa Indonesia pada Smartphone Android, Skripsi, Ilmu Komputer, Institut Pertanian Bogor.