**PROJECT AKHIR**

“Analisis Terhadap SMS untuk Mengklasifikasikan Pesan Normal, Penipuan dan Promo Menggunakan Metode Naive Bayes ”



Disusun Oleh :

Syarief Noor Permadi 165150200111034

Farhan Setya Dhitama 165150201111285

Mahendra Okza Pradhana 165150200111079

Ricky Satria Topaa 165150200111049

Program Studi Teknik Informatika

Fakultas Ilmu Komputer

Universitas Brawijaya

2018

**Latar Belakang**

Dilatarbelakangi oleh perkembangan teknologi informasi dan komunikasi, berdampak pada maraknya tindak kejahatan yang memanfaatkan bidang teknologi informasi dan komunikasi. Salah satunya melalui media komunikasi SMS atau Short Message Service, merupakan media komunikasi yang paling sering digunakan oleh pengguna handphone maupun smartphone. Namun selain tindak kejahatan, SMS juga bisa memberikan informasi mengenai pesan yang bermanfaat mengenai promo yang sedang berlangsung.

Tindak kejahatan yang biasanya dilakukan melalui SMS yaitu berupa pesan penipuan yang mengatasnamakan seseorang agar korban mau mengirimkan uang kepada pelaku. Pelaku akan mengirimkan pesan penipuan ke seluruh nomor telepon secara acak dan berharap penerima pesan sedang dalam kondisi yang sesuai dengan pesan yang dikirimkan agar percaya pada pesan penipuan. Biasanya SMS penipuan yang marak terjadi yaitu pesan mama minta pulsa, pesan korban kecelakaan, pesan sedang ditilang polisi ,pesan keluarga yang masuk rumah sakit, dan yang lainnya.

Pesan promo yang masuk melalui SMS biasanya memberikan pemberitahuan sedang berlangsungnya promo pada suatu produk. Pesan promo tentunya sangat bermanfaat bagi penerima SMS karena membantu memberikan informasi promo. Namun terdapat masalah dimana pengguna sulit membedakan manakah yang pesan penipuan atau pesan promo yang asli, oleh karena itu kami melakukan pengujian untuk membedakan pesan mana yang merupakan pesan normal, pesan promo dan pesan penipuan dengan menggunakan metode Naive Bayes.

**Rumusan Masalah**

Dari latar belakang masalah diatas maka dapat diambil beberapa rumusan masalah seperti :

1. Bagaimana mengklasifikasikan pesan SMS merupakan pesan normal atau pesan promo atau pesan penipuan.
2. Bagaimana tingkat akurasi dari hasil pengujian terhadap analisis SMS yang telah dilakukan dengan menggunakan metode Naive Bayes.

**Perhitungan Manual**

Sumber Data dan Preprocessing Data

Sumber data diambil dari data set yang telah diberikan oleh dosen yang berupa pesan SMS yang berjumlah 1143 pesan dan telah diklasifikasikan sebagai pesan normal, pesan promo dan pesan penipuan. Dimana 80% dari jumlah data digunakan sebagai data latih yaitu 914 data dan 20% dari jumlah data digunakan sebagai data uji yaitu 229 data.

**Pre-processing Data**

1. **Parsing**

Parsing merupakan proses pemecahan struktur dokumen menjadi komponen-komponen yang terpisah. Pada studi kasus ini terdapat total 4 dokumen, dimana 1 dokumen berisi 80% data training dari kelas 0, 80 % dari kelas 1 dan 80% dari kelas 2, dan dokumen terakhir adalah dokumen untuk diuji.

Dokumen 1 : 80% dari tiap kelas.

Dokumen 2 : 20% dari tiap kelas.

1. **Tokenisasi**

Tokenisasi merupakan awal dari proses pre-processing berupa teks. Proses yang dilakukan yaitu memisahkan kalimat menjadi beberapa bagian kata, proses ini dilakukan secara berulang kali hingga semua kalimat terpisah. Contoh pada kasus ini adalah sebagai berikut :

Sebelum proses tokenisasi (dari kelas 0 atau pesan normal):

- Segan baa lo ko wqwq hai di, gila kurus banget skrg diii

- Sekalian aku mau nanya alur dari rawat inap ke lab sama radiologi. Kalau obat kan ga ada yang ngolah.

Hasil proses tokenisasi :

|  |  |
| --- | --- |
| Kalimat 1 | Segan baa lo ko wqwq hai di, gila kurus banget skrg diii |
| Tokens | Segan |
| baa |
| lo |
| ko |
| wqwq |
| hai |
| di, |
| gila |
| kurus |
| banget |
| skrg, |
| diii |

|  |  |
| --- | --- |
| Kalimat 2 | Sekalian aku mau nanya alur dari rawat inap ke lab sama radiologi. Kalau obat kan ga ada yang ngolah. |
| Tokens | Sekalian |
| aku |
| mau |
| nanya |
| alur |
| dari |
| rawat |
| inap |
| ke |
| lab |
| sama |
| radiologi. |
| Kalau |
| obat |
| kan |
| ga |
| ada |
| yang |
| ngolah. |

1. **Cleansing**

Cleansing merupakan tahap kedua setelah tokenisasi, dimana dalam langkah ini menerima masukan hasil dari proses tokenisasi yang akan dihilangkan komponen yang tidak perlu seperti spesial karakter (#,/,@, dll), link url, dan tanda baca yang tidak perlu.

|  |  |
| --- | --- |
| Kalimat 1 | Hasil cleansing |
| Tokens | Segan |
| baa |
| lo |
| ko |
| wqwq |
| hai |
| di |
| gila |
| kurus |
| banget |
| skrg |
| diii |

|  |  |
| --- | --- |
| Kalimat 2 | Hasil cleansing |
| Tokens | Sekalian |
| aku |
| mau |
| nanya |
| alur |
| dari |
| rawat |
| inap |
| ke |
| lab |
| sama |
| radiologi |
| Kalau |
| obat |
| kan |
| ga |
| ada |
| yang |
| ngolah |

1. **Case Folding**

Case folding merupakan tahap ketiga setelah cleansing, dimana dalam langkah ini menerima masukan dari proses sebelumnya yaitu cleansing, kemudian hasil keluarannya merubah seluruh teks masukan menjadi huruf kecil semua.

|  |  |
| --- | --- |
| Kalimat 1 | Hasil case folding |
| Tokens | segan |
| baa |
| lo |
| ko |
| wqwq |
| hai |
| di |
| gila |
| kurus |
| banget |
| skrg |
| diii |

|  |  |
| --- | --- |
| Kalimat 2 | Hasil case folding |
| Tokens | sekalian |
| aku |
| mau |
| nanya |
| alur |
| dari |
| rawat |
| inap |
| ke |
| lab |
| sama |
| radiologi |
| kalau |
| obat |
| kan |
| ga |
| ada |
| yang |
| ngolah |

1. **Filterisasi**

Filterisasi merupakan tahap keempat setelah case folding, dimana dalam langkah ini menerima masukan dari proses sebelumnya yaitu case folding dimana seluruh teks telah diubah menjadi huruf kecil. Proses filterisasi yaitu akan menghilangkan kata-kata yang sering muncul tetapi tidak memiliki pengaruh dalam proses klasifikasi. Proses ini akan melakukan pengecekan tiap kata yang ada dalam daftar stop list, maka kata tersebut akan dihapus. Kelompok kami memakai stop list dari sastrawi 1.0.1 yang dibuat oleh Hanif Amal Robbani

|  |  |
| --- | --- |
| Kalimat 1 | Hasil filterisasi |
| Tokens | segan |
| baa |
| lo |
| ko |
| wqwq |
| hai |
| gila |
| kurus |
| banget |
| skrg |
| diii |

|  |  |
| --- | --- |
| Kalimat 2 | Hasil filterisasi |
| Tokens | nanya |
| alur |
| rawat |
| inap |
| lab |
| sama |
| radiologi |
| kalau |
| obat |
| kan |
| ga |
| ada |
| ngolah |

1. **Stemming**

Stemming merupakan tahap kelima setelah filterisasi, proses ini dimulai ketika menerima masukan hasil dari proses filterisasi. Stemming akan melakukan proses mengubah kata yang mempunyai imbuhan dan menguraikannya dalam bentuk kata dasar.

Stemming yang digunakan adalah sastrawi 1.0.1 yang dibuat oleh Hanif Amal Robbani

|  |  |
| --- | --- |
| Kalimat 1 | Hasil stemming |
| Tokens | segan |
| baa |
| lo |
| ko |
| wqwq |
| hai |
| gila |
| kurus |
| banget |
| skrg |
| diii |

|  |  |
| --- | --- |
| Kalimat 2 | Hasil stemming |
| Tokens | nanya |
| alur |
| rawat |
| inap |
| lab |
| sama |
| radiologi |
| kalau |
| obat |
| kan |
| ga |
| ada |
| ngolah |

**Perhitungan Naïve Bayes**

Untuk rumus metode Naïve Bayes yaitu:

Dimana

* P(c|d) disebut posterior atau peluang kelas c diberikan dokumen d
* P(c) disebut prior, atau peluang awal munculnya kategori c
* ∏1\_(𝑤∈𝑑)▒〖𝑃(𝑤|𝑐)〗 adalah perkalian dari conditional probability masing-masing fitur atau kata yang terdapat pada dokumen d

Untuk rumus Prior Probability yaitu:

Dimana

* Nc = Jumlah dokumen pada data latih yang masuk pada kategori c
* N = Jumlah dokumen pada data latih

Untuk menghitung Conditional Probability, disini kita memilih untuk menggunakan Multinomial Model. Untuk menghindari adanya probabilitas nol, biasanya dilakukan add-one atau laplace smoothing. Ada penambahan +1 pada pembilang, dan |V| pada penyebut. Maka dari itu, rumus dituliskan sebagai berikut:

Dimana

* 𝑐𝑜𝑢𝑛𝑡(𝑤,𝑐) = jumlah kemunculan kata w pada kategori c
* 𝑐𝑜𝑢𝑛𝑡(𝑐) = jumlah total kemunculan semua kata pada kategori c
* |𝑉| = jumlah term unik atau fitur

Alasan kita menggunakan Multinomial Model karena terbukti secara umum lebih akurat dalam klasifikasi teks dibandingkan dengan model lain

Untuk dokumen-dokumen kita ambil 5 sms dari data training dan 1 sms dari data testing, jadi N = 5.

D1, Kelas 2= Anda tlh menukarkan poin 95 poin, Tukarkan terus poinnya siapa tahu anda pemenang hadih LG G3 32gb, pemenang adalah yang menukarkan poin tertinggi s/d 30 November 15

D2, Kelas 2 = Isi ulang Kuota++ 3GB berhasil, dan kuota tlh ditambahkan. Nikmati juga BONUS NELPON ke semua Tri Rp10rb, cek bonus di \*111\*2# atau BimaTri.

D3, Kelas 1 = ANEKA SHOP Berbagi Promo Type Blackberry DAKOTA 2,4 jt, TOURCH 1,7 jt, ONIX.2 1,5 jt, GEMINI 800 rb, U/Pemesanan 085325557999 Pin BB-2A96DE82 [www.aneka999.blogspot.com](http://www.aneka999.blogspot.com)

D4, Kelas 1 = Belikan dulu mama pulsa simpati 20 ribu ini nomor barunya mama 081223052854. kalau bisa kirim sekarang secepatnya penting. Ini hpnya orang mama pinjam

D5, Kelas 0 = Ada mba, harganya 1,6jt/bln. Fasilitasnya air hangat, kamar mandi dalam, tv, tv kabel, wifi, meja, tempat tdr, dan lemari.

D6, Kelas ? = Selamat, Anda telah membeli Paket Combo Kenyang Internet As. Cek bonus di \*889#.

Fitur-fitur yang telah ditentukan adalah sebagai berikut

|  |  |
| --- | --- |
| Fitur Unik | |
| 1 | selamat |
| 2 | anda |
| 3 | beli |
| 4 | mama |
| 5 | nikmat |
| 6 | bonus |
| 7 | bagi |
| 8 | promo |
| 9 | fasilitas |
| 10 | lemari |

Menentukan Prior Probability berdasarkan rumus yang telah ditentukan:

Menghitung raw TF tiap fitur pada tiap dokumen di data training

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fitur Unik | | D1(2) | D2(2) | D3(1) | D4(1) | D5(0) |
| 1 | selamat | 0 | 0 | 0 | 0 | 0 |
| 2 | anda | 2 | 0 | 0 | 0 | 0 |
| 3 | beli | 0 | 0 | 0 | 1 | 0 |
| 4 | mama | 0 | 0 | 0 | 3 | 0 |
| 5 | nikmat | 0 | 1 | 0 | 0 | 0 |
| 6 | bonus | 0 | 1 | 0 | 0 | 0 |
| 7 | bagi | 0 | 0 | 1 | 0 | 0 |
| 8 | promo | 0 | 0 | 1 | 0 | 0 |
| 9 | fasilitas | 0 | 0 | 0 | 0 | 1 |
| 10 | lemari | 0 | 0 | 0 | 0 | 1 |
| **wTF** | | **2** | **2** | **2** | **4** | **2** |

Conditional Probability tiap fitur pada tiap kelas

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Fitur Unik | | P(w|2) | P(w|1) | P(w|0) |
| 1 | selamat | 0,071 | 0,062 | 0,083 |
| 2 | anda | 0,214 | 0,062 | 0,083 |
| 3 | beli | 0,071 | 0,125 | 0,083 |
| 4 | mama | 0,071 | 0,250 | 0,083 |
| 5 | nikmat | 0,143 | 0,062 | 0,083 |
| 6 | bonus | 0,143 | 0,062 | 0,083 |
| 7 | bagi | 0,071 | 0,125 | 0,083 |
| 8 | promo | 0,071 | 0,125 | 0,083 |
| 9 | fasilitas | 0,071 | 0,062 | 0,167 |
| 10 | lemari | 0,071 | 0,062 | 0,167 |

Menghitung posterior tiap kelas

D6 = Selamat, Anda telah membeli Paket Combo Kenyang Internet As. Cek bonus di \*889#.

Kata-kata yang tidak diperhitungkan karena tidak termasuk fitur yaitu: “telah”, “membeli”, “Paket”, “Combo”, “Kenyang”, “Internet”, “As”, “Cek”, “di”, “\*889#”.

Kata-kata yang diperhitungkan yaitu: “selamat”, “anda”, “beli”, “bonus”

(2│𝑑6)=𝑃(2)∗𝑃(selamat│2)∗𝑃(anda│2) ∗𝑃(beli│2) ∗𝑃(bonus│2)

(2│𝑑6)=0.4∗0.071∗0.214∗0.071∗0.143=0,0000617058728

(1│𝑑6)=𝑃(1)∗𝑃(selamat│1)∗𝑃(anda│1) ∗𝑃(beli│1) ∗𝑃(bonus│1)

(1│𝑑6)=0.4∗0.062∗0.062∗0.125∗0.062=0,0000119164

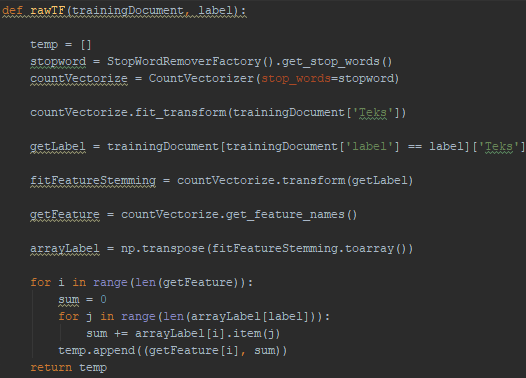
(0│𝑑6)=𝑃(0)∗𝑃(selamat│0)∗𝑃(anda│0) ∗𝑃(beli│0) ∗𝑃(bonus│0)

(0│𝑑6)=0.2∗0.083∗0.083∗0.083∗0.083=0,0000094916642

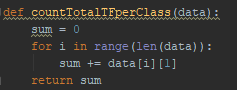
Karena P(2|d6) > P(1|d6) > P(0|d6), maka D6 termasuk kelas 2

**Analisa Hasil**

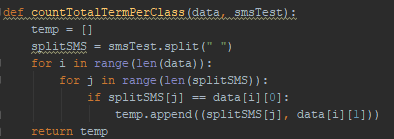
Pertama, kita membuat metode untuk mengambil rawTF pada tiap fitur yang ada pada data training.



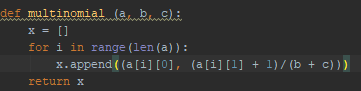
Menghitung jumlah total TF pada tiap kelas



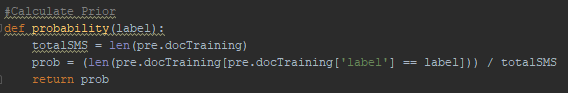
Menghitung jumlah total term pada tiap kelas



Selanjutnya mendefinisikan rumus multinomial model



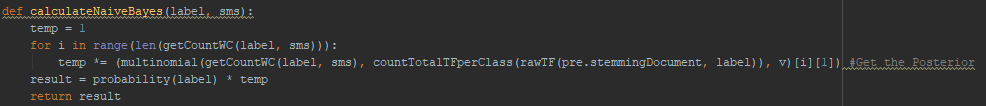
Lalu menghitung prior probability



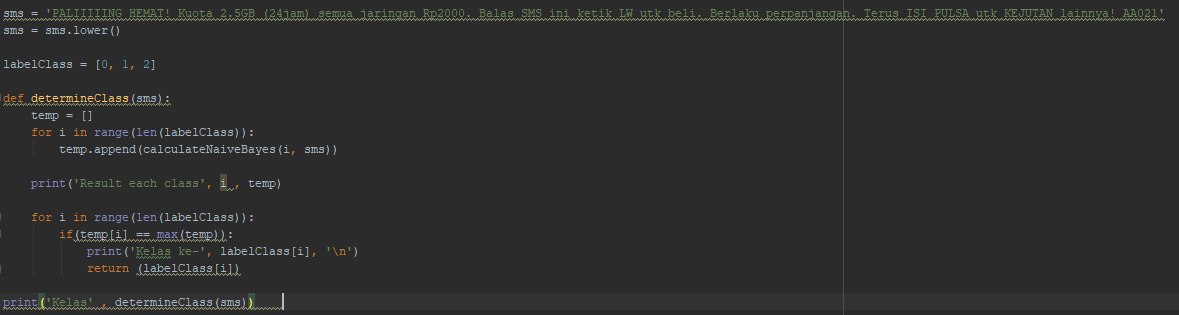
Menghitung conditional probability



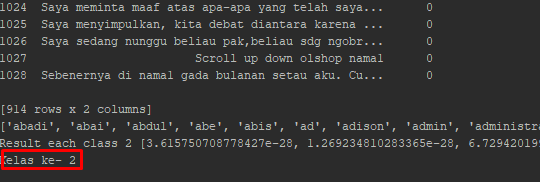
Lalu hitung keseluruhan dengan menggunakan naïve bayes



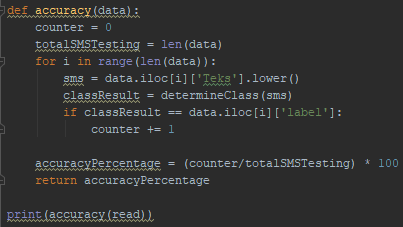
Berikutnya kita menentukan kelas pada tiap sms yang akan digunakan. Data yang diambil adalah data testing yang telah dibuat sebelumnya. Dan disini kita coba untuk mengambil salah satu sms dari <https://codeshare.io/2pmMe6>.



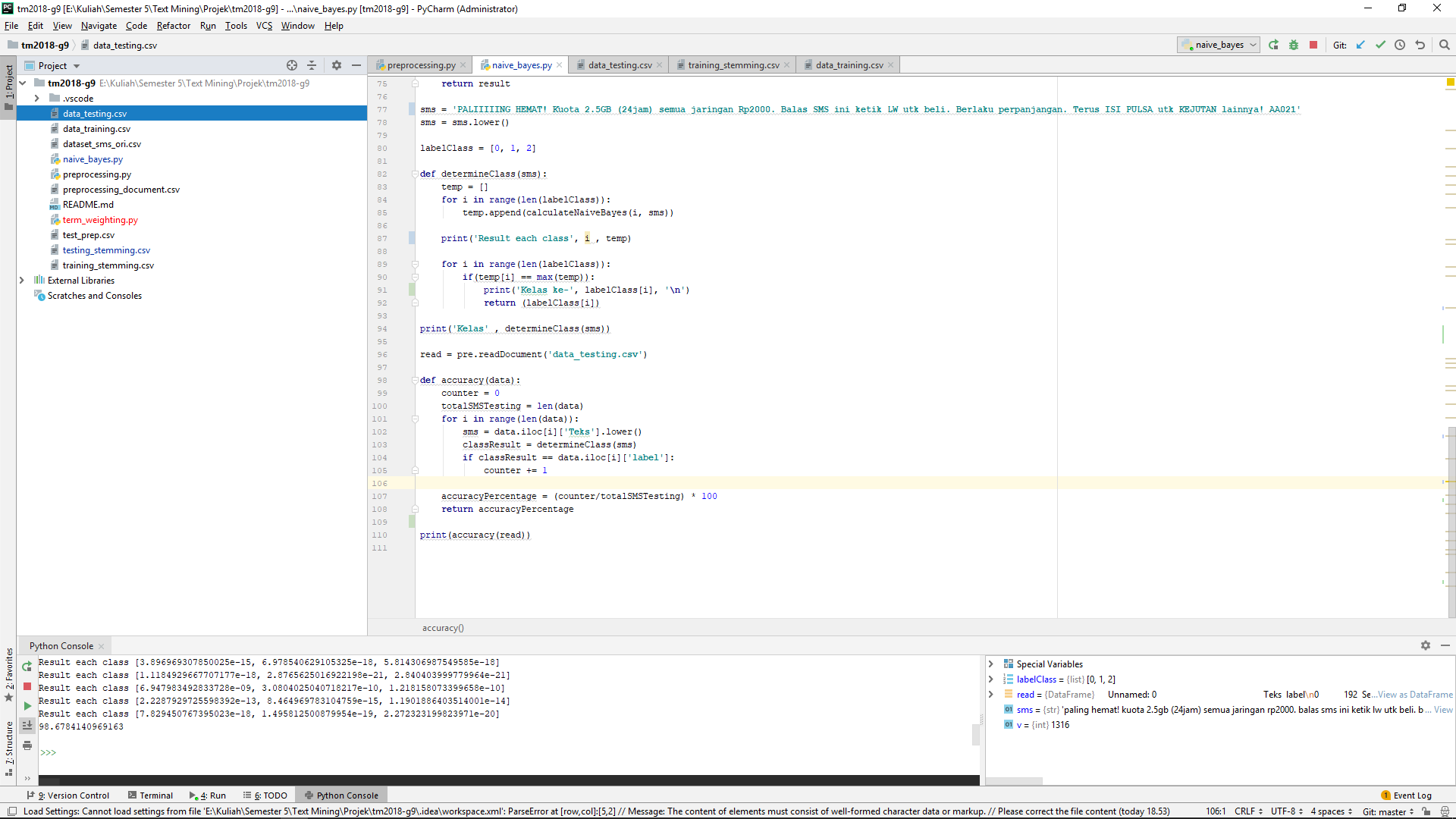
Dan hasil perhitungan yang kita dapatkan, label sesuai dengan kelas yang ditentukan, yaitu kelas 2



Setelah menentukan kelasnya, kita hitung tingkat akurasi untuk data testing. Hasil akurasi dihitung dengan rumus: (jumlah dokumen yang labelnya sesuai/ jumlah seluruh dokumen) \* 100



Hasil akurasi pada data testing yang kita gunakan yaitu 98,678% (screenshot diambil pada saat sebelum presentasi)



**Kesimpulan**

Untuk mengklasifikasikan sms apakah termasuk pesan promo, pesan fraud/ spam, atau pesan normal, kita melakukan preprocessing dahulu pada data training. Data training disini diambil dari 80% sms dari tiap label/ kelas, dan data testing diambil 20% sms dari tiap label/ kelas. Selanjutnya kita ambil sms/ dokumen dari data testing untuk mencari kelas tersebut. Pada data training, kita hitung rawTF untuk seluruh fitur yang telah ditentukan pada preprocessing sebelumnya. Lalu hitung conditional probability fitur tiap kelas dan output yang dihasilkan adalah model (nilai prior dan conditional probability). Pada data testing kita tentukan nilai posterior terbesar dari hasil hitungan posterior masing-masing kelas. Output yang dihasilkan nanti adalah kelas untuk masing-masing data testing

Pada data testing, sebelumnya sudah ditentukan kelas pada tiap sms/ dokumennya. Maka dari itu kita gunakan untuk menentukan seberapa besar akurasi hasil perhitungan untuk menentukan kelas tiap sms. Rumus yang kita gunakan untuk menentukan akurasinya adalah (jumlah dokumen yang kelasnya telah sesuai/ jumlah seluruh dokumen) \* 100. Dari hasil perhitungan akurasi, kita dapatkan sebanyak 98,678%.