Perbandingan Metode Spam Classifier Untuk SMS

Muhammad Insan Al-Amin (23517009), Sigit Kariagil Bimonugroho (23517032)

Institut Teknologi Bandung, Indonesia

Penelitian ini membandingkan beberapa metode algoritma machine learning dan teknik preprocessing untuk klasifikasi spam pada kumpulan sms. Data latih yang digunakan 5000 sms, data tes yang digunakan 574 sms.

*Index Terms*— Natural language processing, Spam classifier

# Analisis Dataset

T

his document is a Microsoft *Word* template for IEEE Transactions on Magnetics. Use of this document as a template is optional.

If you have a question about formatting your paper, or a suggestion on improving these instructions, please contact [r.goldfarb@ieee.org](mailto:r.goldfarb@ieee.org).

# Pemrosesan Dataset

Tahapan umum pre processing yang dilakukan dalam pemrosesan dataset adalah pembersihan partikel-partikel non alfanumerik yang tidak dikehendaki menggunakan regex, penyingkatan spasi, normalisasi kata, stemming, lemmatization, dan pembersihan stop words, kemudian tokenization.

Adapun tahapan khusus yang dilakukan adalah memisahkan angka 0 dengan spasi disekitarnya agar menjadi token, karena dalam pengamatan banyak spam yang mengandung angka 0.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SKENARIO** | **A** | **B** | **C** | **D** | **E** |
| Stemming | x | o | x | x | X |
| Lemmatization | o | x | o | o | O |
| Token Nol | x | o | o | o | O |
| Stopwords | o | o | o | o | X |
| Normalization | x | x | x | o | o |

Kesemua proses ini dijadikan lima jenis skenario dengan untuk menguji pada kondisi optimal mana kah akurasi tertinggi didapat ketika proses pengujian. Tiap skenario memiliki tahapan-tahapan pre processing tersendiri.

Selanjutnya dilakukan proses Feature Extraction mengubah token-token dari tiap row dataset, serta fitur-fitur kata unik dari seluruh dokumen menjadi bentuk document-term matrix yang berisi nilai TF-IDF untuk tiap fitur kata di setiap data-set.

# Skenario Pengujian

Pengujian dilakukan dengan menggunakan delapan jenis teknik klasifikasi yaitu Rule Based, Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine (SVM), Neural Network, K-Nearest Neighbour (KNN), dan Stochastic Gradient Descent (SGD)(1).

Untuk setiap teknik klasifikasi yang digunakan, dicoba lima jenis skenario pre processing A, B, C, D, E. Untuk masing-masing skenario dilakukan training terhadap teknik klasifikasi, lalu dilakukan testing terhadap data tes, dan data training. Setelah itu dicatat waktu training untuk tiap jenis teknik klasifikasi.

# Hasil Eksperimen dan Analisis

Dari data percobaan terlihat bahwa SGD dan Neural Network pada skenario pre processing D memiliki akurasi paling tinggi. Sementara akurasi paling rendah adalah Rule Based skenario pre processing A.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **TEST DATA = TEST DATA** | | | | | |
| **Scenarios** | **A** | **B** | **C** | **D** | **E** | **ALL SCENARIOS TIME** |
| Rule based | 0.77 | 0.878 | 0.878 | 0.864 | 0.905 | 21.472 |
| Decision tree | 0.777 | 0.78 | 0.771 | 0.771 | 0.777 | 228.007 |
| Naïve bayes | 0.861 | 0.855 | 0.862 | 0.86 | 0.86 | 39.127 |
| Random forest | 0.979 | 0.977 | 0.982 | 0.976 | 0.977 | 56.512 |
| SVM | 0.871 | 0.871 | 0.871 | 0.871 | 0.871 | 831.416 |
| Neural network | 0.989 | 0.988 | 0.987 | 0.991 | 0.986 | 656.161 |
| KNN | 0.923 | 0.927 | 0.925 | 0.925 | 0.928 | 1869.915 |
| SGD | 0.984 | 0.991 | 0.989 | 0.991 | 0.986 | 15.121 |
|  |  |  |  |  |  |  |
|  | **TEST DATA = TRAINING DATA** | | | | | |
| **Scenarios** | **A** | **B** | **C** | **D** | **E** | **ALL SCENARIOS TIME** |
| Rule based | 0.997 | 0.998 | 0.998 | 0.998 | 0.998 | 21.472 |
| Decision tree | 0.767 | 0.767 | 0.767 | 0.767 | 0.767 | 228.007 |
| Naïve bayes | 0.921 | 0.92 | 0.921 | 0.92 | 0.92 | 39.127 |
| Random forest | 0.998 | 0.997 | 0.997 | 0.998 | 0.998 | 56.512 |
| SVM | 0.865 | 0.865 | 0.865 | 0.865 | 0.865 | 831.416 |
| Neural network | 1 | 1 | 1 | 1 | 1 | 656.161 |
| KNN | 0.921 | 0.922 | 0.922 | 0.922 | 0.927 | 1869.915 |
| SGD | 0.998 | 0.998 | 0.998 | 0.999 | 0.994 | 15.121 |

Teknik klasifikasi dengan waktu training tercepat berturut turut adalah SGD, Rule Based, Naïve Bayes, dan Random Forest. Ketiganya memiliki waktu training rata-rata untuk tiap skenario dibawah 15 detik.

KNN memerlukan waktu training rata-rata paling lama yaitu 374 detik untuk satu skenario, lebih dari dua kali waktu training SVM, diatas seratus kali waktu training SGD.

Untuk tiap skenario, SVM memiliki nilai akurasi yang sama, 0.871, dengan varians terendah dari semua teknik klasifikasi yaitu 1.54 \* 10-32 ~ 0. Sedangkan pada Rule Based, nilai akurasi skenario A dan skenario E berbeda 0.135, varians tertinggi dari semua teknik klasifikasi yaitu 2.696 \* 10-3.

# Kesimpulan

SGD dan Random Forest memiliki waktu latih rata-rata tercepat, 3.02 dan 11.3 detik dengan akurasi di atas 0.95. Sementara untuk akurasi tertinggi diperoleh dari teknik klasifikasi SGD dan Neural Network untuk skenario pre processing C.

Referensi

1. Shafi’i Muhammad Abdulhamid, “A Review on Mobile SMS Spam Filtering Techniques” Federal University of Technology, Minna PMB 65, Nigeria, IEEE, Agustus 2017.