Tugas 3 Analitika Media Sosial

Model Klasifikasi Jenis Kelamin Pengguna Twitter

Tugas ini dikerjakan secara berkelompok. Pembagian kelompok sama dengan **Tugas 2 kemarin**.

Anda diberikan sebuah dataset untuk *training* sebuah model klasifikasi, yang bernama **"train.gender.data"**. Penjelasan atribut/fitur pada dataset tersebut adalah sebagai berikut:

Atribut	Tipe	Deskripsi
name	String, **Categorical**	Nama account
fav_number	Integer	*Number of tweets the user has favorited*
retweet_count	Integer	*Number of times the user has retweeted (or possibly, been retweeted)*
tweet_count	Integer	*Number of tweets that the user has posted*
red1 (0-255)	0 – 255	Nilai red dari RGB link color yang ada di profile
green1 (0-255)	0 – 255	Nilai green dari RGB link color yang ada di profile
blue1 (0-255)	0 – 255	Nilai blue dari RGB link color yang ada di profile
red2 (0-255)	0 – 255	Nilai red dari RGB profile sidebar color
green2 (0-255)	0 – 255	Nilai green dari RGB profile sidebar color
blue2 (0-255)	0 – 255	Nilai blue dari RGB profile sidebar color
class	**String, Categorical**	**'female' atau 'male'**

Gunakan *training data* tersebut untuk membangun **model klasifikasi secara otomatis** terhadap jenis kelamin suatu pengguna Twitter. Untuk melakukan evaluasi terhadap model tersebut, Anda perlu mengujinya terhadap sebuah *testing dataset* yang belum diberi label. *Testing dataset* tersebut dapat Anda lihat pada file **"test.gender.nolabel.data"**.

Perhatikan bahwa fitur **"name"** tidak bisa digunakan langsung untuk membangun model karena nilainya bersifat *categorical* dan unik. Hal ini dapat merusak performa klasifikasi karena algoritma *machine learning* yang digunakan akan sulit menemukan pola keteraturan pada dataset Anda. Oleh karena itu, Anda perlu melakukan *pre-processing* terhadap fitur tersebut, atau mentransformasikan fitur **"name"** tersebut menjadi fitur-fitur tambahan lain yang siap digunakan untuk pengembangan model (biasanya bernilai numeric).

Untuk mengetahui nilai *accuracy* dari label yang dihasilkan oleh model Anda, Anda perlu *submit* **"jawaban"** atau **"tebakan"** ke sebuah sistem *submission online*:

https://dm-sma.000webhostapp.com/submit.php

Masukkan file text yang berisi jawaban/tebakan Anda terhadap semua *instance* yang ada di *testing data*, dan masukkan pula sebuah **"upload key"** yang merupakan kode unik dari kelompok Anda. Untuk mendapatkan **"upload key"**, Anda perlu mengirim email ke alfan@cs.ui.ac.id dengan menyertakan

informasi anggota kelompok Anda. Kami akan membalas email Anda dan sekaligus mengirimkan **"upload key"** untuk kelompok Anda.

Contoh format jawaban dapat merujuk kepada contoh **"contoh_ouput_siap_submit.txt"**.

Hasil evaluasi untuk model Anda beserta peserta dari kelompok lain dapat Anda lihat di link berikut:

https://dm-sma.000webhostapp.com/

Sistem evaluator tersebut dibuat sangat sederhana. Jadi, mohon untuk menjaga sistem tersebut dengan penuh tanggung jawab. Jangan melakukan hal-hal yang dapat merugikan peserta lain dan kuliah ini secara umum.

Tips Pengerjaan

Jangan terlalu sering melakukan *submission* ke sistem online. Lebih baik Anda melakukan eksperimen terlebih dahulu beberapa kali pada *training data* Anda. Jadi, training data Anda dapat di-*split* menjadi dua bagian. Bagian pertama untuk bangun model, dan bagian kedua untuk evaluasi model. Jika hasilnya dirasa sudah baik, silakan terapkan model tersebut pada *unlabeled testing data*, lalu hasilnya di-*submit*.

Anda dapat menggunakan API berikut untuk split training data:

http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Deliverables

- Dokumen (maksimal 5 halaman) yang berisi hasil eksperimen dan metode yang Anda gunakan untuk mencapai performa terbaik
    - Model klasifikasi yang Anda usulkan
    - *Feature Ablation Study*
    - Model *machine learning* yang Anda gunakan
    - Dan cerita metodologi yang lainnya.
- *Source code* yang Anda gunakan dalam eksperimen pengembangan model

Penilaian

- Penilaian sepenuhnya dilakukan berdasarkan proses pengerjaan yang tercermin dari Dokumen. **Jadi, jangan khawatir seandainya Anda menduduki peringkat paling bawah di klasemen *accuracy*.**
- Namun, Kelompok yang menduduki posisi atas (top-10) akan mendapatkan nilai **bonus** ☺

Submission

Dokumen dikumpulkan via SCELE hingga **Selasa, 21 Maret 2017, Pukul 23:00 malam**. Dokumen dan *source code* dikumpulan **oleh seorang anggota kelompok** dalam sebuah file terkompresi, dengan penamaan:

tugas3_<npm>_<nama depan>.zip