**Proyek Tengah Semester**
**CSIE604284: Analitika Media Sosial – Genap 2021/2022**
[Document Version 1.0 – update 2022-03-25, 5.30 PM]

# *Demographics Prediction in Twitter*

Timeline:
*Project description release*: Jumat, 25 Maret 2022
*Group registration deadline*: Kamis, 31 Maret 2022
*Leaderboard open*: Jumat, 8 April 2022
*First model submission deadline*: Jumat, 15 April 2022, jam 17.00
*Mentoring deadline*: Rabu, 20 April 2022, jam 17.00
*Leaderboard closed* (*model building deadline*): Jumat, 22 April 2022, jam 16:00
*Report deadline*: Senin, 25 April 2022, jam 17.00

Bentuklah kelompok dengan maksimal anggota 3 orang. Beri nama kelompok. Daftarkan anggota kelompok pada *link* yang ada di SCeLE dengan menuliskan nama, NPM, dan email setiap anggota kelompok.

## A. Motivasi

Informasi demografis adalah atribut yang sangat penting dalam *data analytics*. Contohnya, suatu perusahaan bisa jadi tertarik untuk merancang *marketing plan* yang lebih terpersonalisasi berdasarkan atribut standar ekonomi, atau analisis dukungan publik pada calon tertentu saat kontestasi politik akan memberikan *insight* lebih kaya jika dipetakan berdasarkan atribut jenis kelamin atau lokasi.

Media sosial, seperti Twitter, mengakomodasi *user* untuk mengisi atribut demografi pada akun profilnya. Akan tetapi, tidak ada keharusan untuk mengisi atribut demografis dengan lengkap dan tepat. Penelitian menunjukkan tidak lebih dari 50% *user* (dihitung dari jumlah sampel tertentu) yang mengisi atribut *gender*, jenis pekerjaan, dan sebagainya. Sebagian dari *user* yang mengisi atribut tersebut mengisinya dengan *fake data*. Oleh karena itu, memakai data atribut langsung dari hasil *crawling* Twitter bisa menyebabkan hasil analisis demografi yang tidak akurat.

## B. Deskripsi Tugas

Pada proyek tengah semester ini, peserta kuliah dalam kelompok diminta untuk mengembangkan *analytics* model untuk memprediksi atribut demografi jika diberikan akun *user*. Peserta kuliah akan membangun model *predictive analytics* untuk menebak nilai atribut demografi pada *unlabeled test dataset*. Terdapat dua atribut yang akan diprediksi, yakni *gender* dan *occupational area*.

1. *Gender*
   *Gender* merepresentasikan jenis kelamin dari pemilik akun Twitter pada dataset. Atribut ini pada dataset merupakan atribut *binary* di mana terdapat dua kelas, yaitu **pria** dan **wanita**.

2. *Occupational*
   *Occupational area* merepresentasikan kategori pekerjaan yang ditekuni oleh pemilik akun Twitter pada dataset ini. Atribut ini merupakan atribut *multiclass* di mana terdapat 10 kelas atau kemungkinan nilai, yaitu:
   - pendidikan dan penelitian
   - sains dan teknologi
   - kesehatan
   - ekonomi dan bisnis
   - sosial kemasyarakatan
   - media
   - *hospitality & tourism*
   - olahraga
   - hiburan
   - seni

Berikut adalah contoh *username* beserta atributnya

1. Username : @NajwaShihab
   *Gender* : wanita
   *Occupation* : media

2. Username : @DennyCagur
   *Gender* : pria
   *Occupation* : hiburan

Untuk masing-masing atribut demografi tersebut, setiap kelompok diminta mengerjakan sekurang-kurangnya dua model. Model pertama menggunakan data terstruktur dari profil *user*. Model kedua menggunakan data Tweet.

Akan disediakan sejumlah data latih dan data uji. Sebagian data uji sudah disediakan lengkap mencakup data profil dan Tweet, sedangkan sebagian data uji hanya disediakan *username.*

Output dari model prediksi dikumpulkan ke *online grader* yang akan diumumkan di SCeLE. Output yang di*submit* **harus** merupakan hasil luaran program. Kelompok yang melakukan kecurangan dalam submisi *output* (misalnya melakukan *manual tagging*) atau misalnya *assign random output* akan memperoleh nilai 0 untuk proyek tengah semester ini. *Online grader* akan mengevaluasi output prediksi.

Peserta kuliah disarankan untuk submit seluruh model yang diimplementasikan ke *online grader*, termasuk model *baseline* yang performanya mungkin tidak terlalu baik. Mencapai performa paling tinggi bukan satu-satunya tujuan tugas ini. Harapannya, peserta kuliah juga belajar bagaimana merancang eksperimen, membandingkan performa antar model, serta menganalisis hasil prediksi model.

Sesi Mentoring

Dalam pengerjaan proyek tengah semester ini, setiap kelompok harus mengikuti sesi mentoring minimal sekali dengan asisten. Setiap kelompok bebas mengatur waktu dengan asisten untuk mentoring, selama rentang 1 April – 20 April. Di sesi mentoring, setiap kelompok membahas idenya dalam pengerjaan tugas (lebih ideal jika peserta kuliah sudah pernah submit model yang diimplementasikan) dan asisten akan memberikan masukan. Peserta kuliah dibolehkan untuk meminta mentoring dengan asisten lebih dari sekali.

Di luar mentoring dengan asisten, dosen juga terbuka jika ada kelompok yang ingin diskusi (tapi mentoring dengan asisten harus tetap dilakukan, karena adalah bagian dari penilaian).

## C. Komponen Penilaian

Nilai proyek tengah semester *Demography Prediction* ini memiliki skema penilaian dengan *breakdown* (tentatif) sebagai berikut

- 28%: metode, *feature extraction*, rancangan eksperimen
- 15%: skor *online evaluation* (*leaderboard*)
- 12%: *mentoring* (pemahaman, keaktifan, *eager to learn*) dan *project management* (kerjasama tim dan aspek non teknis lainnya)
- 10%: metodologi, data *collection*, data *preprocessing*
- 10%: evaluasi dan *error analysis*
- 10%: kode, *script*, dan model (beserta dokumentasi)
- 10%: penulisan laporan
- 5%: studi literatur

Nilai yang diperoleh anggota kelompok bisa berbeda dengan nilai kelompok, tergantung penilaian *peer review*.