Data Analysis (IS388-B)

Assignment Week 2

Anggota Kelompok 3:

Christian Alexander	(00000054292)
Christopher Darren	(00000054804)
Azzahra Shaffira Wijaya	(00000055376)
Michele Stepanie	(00000055477)
Meisha Geovanni Mulin	(00000055487)

Business Understanding

Dari ketiga topik yang diberikan, kami memilih topik mengenai AirBNB New User Booking. Permasalahan yang kami temukan:

- Para pengguna baru AirBnb banyak yang kebingungan untuk melakukan pemesanan sehingga waktu pemesanan yang dibutuhkan cukup lama
- Mempunyai kekurangan untuk memprediksi di mana pengguna akan menginap.
- Mempunyai Kekurangan untuk dapat memprediksi permintaan dari pelanggan.
- Konten yang diberikan kurang sesuai dengan komunitas atau pengguna mereka.

AirBNB juga memiliki tujuan yaitu:

- Untuk memprediksi di mana pengguna baru dapat memesanan tempat penginapan mereka secara akurat, sebagai bentuk memberikan *travel experience* kepada pengguna baru agar merasa puas.
- User dapat melakukan pemesanan tanpa membutuhkan waktu yang lama.
- Dapat memprediksi permintaan pelanggan dengan lebih akurat.
- Membagikan konten menarik yang sesuai dengan komunitas mereka.

Akan berhasil bila:

- Terdapat sebuah peningkatan dalam pemesanan kamar.
- Mendapatkan banyak respon positif dari pelanggan terhadap pelayanan yang diberikan.

Data Collection

- Data yang digunakan adalah data yang bersumber dari kaggle yang merupakan salah satu open source. Jenis data yang digunakan adalah data terstruktur.
- Data yang tersedia adalah data umur, negara, sample submission, session, test user, dan train user

- Data berisikan daftar-daftar pengguna beserta demografi mereka, catatan sesi web, dan beberapa ringkasan statistik
- Data digunakan untuk memprediksi negara mana yang akan menjadi tujuan pemesanan pertama pengguna baru. Semua pengguna dalam kumpulan data ini berasal dari AS.
- Data-data tersebut dapat diperoleh baik melalui website maupun dari aplikasi yang digunakan pengguna dalam melakukan pendaftaran dan pemesanan kamar AirBnb

(TWITTER)

Business Understanding

- Banyaknya pengguna twitter yang menggunakkan kata yang kurang pantas sehingga dapat menjadi suatu permasalahan dikarenakan banyak pengguna yang masih dibawah umur
- Perlu diadakannya fitur banned agar kata kata yang kurang pantas dapat dihapus
- Twitter hanya mempunyai batasan usia 12+ untuk mengakses aplikasi nya

Dalam penelitian ini digunakan sebuah dataset Twitter mengenai ujaran kebencian serta bahasa kasar dari beberapa penelitian sebelumnya. Pengambilan data dibagi menjadi dua fase, yaitu :

- fase anotasi yang pertama, dikumpulkannya sebanyak 16.500 tweet dari proses crawling untuk dianotasi dilakukan oleh 30 annotator.
- lalu, pada fase anotasi yang kedua, kami membuat anotasi sebanyak 5.700 tweet ujaran kebencian dan dianotasi oleh tiga annotator terbaik dari fase anotasi pertama untuk membubuhi keterangan target, kategori, serta tingkat ujaran kebencian yang terdapat di Twitter.

Data Collection

Pengelompokkan data dilakukan seperti ini:

- HS : label ujaran kebencian dan kekerasan;
- Abusive : label bahasa yang kasar atau bahasa yang tidak layak digunakan;
- HS_Individual : ujaran kebencian yang ditujukan kepada individu yang bersifat pribadi;
- HS_Group : ujaran kebencian yang ditujukan kepada suatu grup atau kelompok;
- HS_Religion : ujaran kebencian yang berkaitan dengan agama/keyakinan;
- HS_Race : ujaran kebencian terkait ras/etnis;
- HS_Physical : ujaran kebencian terkait fisik/cacat;
- HS_Gender : ujaran kebencian terkait gender/orientasi seksual;
- HS_Gender : kebencian terkait makian/fitnah lainnya;
- HS_Weak : ujaran kebencian yang lemah, tidak menimbulkan dampak yang besar;
- HS_Moderate : ujaran kebencian sedang, dapat menimbulkan dampak kekerasan;
- HS_Strong : ujaran kebencian yang kuat, menimbulkan kekerasan.

Hasil anotasinya menunjukkan bahwa :

- 7.608 tweet bukan ujaran kebencian dan 5.561 tweet ujaran kebencian
- Dari total 5.561 tweet ujaran kebencian, sebagian besar tweet pidato kebencian diarahkan pada individu (3.575 tweet ditargetkan untuk seorang individu dan 1.986 tweet ditargetkan ke kelompok).
- Tweet ujaran kebencian tersebut terdiri dari beberapa kategori ujaran kebencian yaitu 793 tweet terkait agama/keyakinan, 566 tweet terkait ras/suku, 323 tweet terkait fisik/cacat, 306 tweet terkait gender/seksual orientasi, dan 3.740 tweet terkait makian/fitnah lainnya.
- Untuk label tingkat ujaran kebencian, terdiri dari 3.383 ujaran kebencian yang lemah, 1.705 ujaran kebencian yang moderat, dan 473 ujaran kebencian yang kuat.

Sumber:
https://www.kaggle.com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text