Kelompok Anggota:

William Antony – 2201767775

Ronaldo Kenny Chandra – 2201763234

Limas Jaya Akeh – 2201763240

Andreas Aditya Alvaro Harryanto – 2201767655

Rio Nagano – 2201767232

Spam Detector – Machine Learning-Based Method

1. Motivasi Project

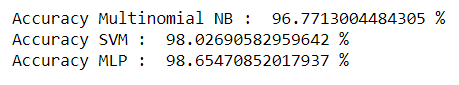
Project kami adalah membuat model spam filter menggunakan Machine Learning-based Method. Di zaman yang segalanya menggunakan Chat Messaging untuk pembicaraan sehari-hari, kerap kali terdapat pesan dari orang yang tidak dikenal mengenai iklan, maupun ancaman, yang umumnya dikenal sebagai “Spam”. Hal ini menganggu secara psikologis dikarenakan konten yang tidak senonoh, sehingga dengan diharapkannya terdapat Spam Filter ini, pengguna tidak dirugikan oleh kelompok-kelompok yang mengirimkan spam tersebut, contohnya adalah dengan menggunakan phising ataupun memberikan email yang berisi malware. Perusahaan juga perlu untuk memakai spam filter agar data-data yang penting dari perusahaan tersebut tidak dicuri oleh orang lain.

2. Metode yang digunakan

Model yang kami buat menggunakan beberapa Machine Learning model dengan metode Multinomial Naïve Bayes Classifiers, Support Vector Machine (SVM), dan Multi-layer Perceptron. Hal ini dilakukan untuk membandingkan performa dari tiap model Machine Learning terhadap dataset yang ada. Kami juga menggunakan beberapa library untuk membantu dalam Preprocessing data seperti SnowballStemmer dan WordNetLemmatizer. Tiap metode yang digunakan akan dimasukkan kedalam sebuah Pipeline dan menggunakan TF-IDF untuk melihat weight tiap kata dalam dokumen. Kemudian kami melihat akurasi dari Model yang kami buat menggunakan Cross-Validation dengan K-Fold dengan k = 4 untuk melihat apakah Model yang dibuat Overfitting atau tidak.

3. Pembahasan Hasil

Dari 3 Model tersebut, rata-rata akurasi Model sudah mencapai lebih dari 95%, sehingga sudah cukup baik digunakan. Terutama menggunakan model MLP dengan 2 Hidden layer, layer pertama dengan 14 neuron, dan layer kedua dengan 4 neuron.



Hasil ini sudah cukup memuaskan, terutama karena prosesnya yang jauh lebih cepat dan ringan dibandingkan menggunakan Deep Learning-approach (contohnya menggunakan BERT), akan tetapi, dengan akurasi yang sudah sangat tinggi (hingga 98%), dalam kasus ini lebih baik menggunakan Machine Learning-based method. keunggulan dari model yang kami buat adalah karena proses yang dibutuhkan tidak berat, sehingga dapat digunakan untuk keperluan sehari-hari. Diantara ketiga model tersebut, MLP memiliki akurasi tertinggi, tetapi proses yang sedikit lebih lama dibandingkan ketiganya, SVM memiliki akurasi yang cukup memuaskan tetapi proses cepat, dan Multinomial NB memiliki akurasi terrendah tetapi proses tercepat, akan tetapi, model Multinomial NB yang kami buat kemungkinan memiliki Overfitting, karena Cross-Validation score yang dimiliki ada di sekitar 93~%, sehingga akurasi sebenarnya seharusnya berada di sekitar 93%.



Untuk model lainnya, yaitu SVM dan MLP, memiliki Cross-Validation score yang cukup stabil yaitu memang di sekitar 96% dan 97%. Sehingga model yang paling cocok menurut kami digunakan untuk keperluan sehari-hari adalah SVM, karena proses yang cepat dan akurasi yang kurang lebih sama dengan model MLP.



