Membuat Model Spam Filter Menggunakan Algoritma Klasifikasi Naïve Bayes

# Latar Belakang

Berdasarkan data dari beberapa sumber yang kami kaji hampir 45 persen dari akun email adalah akun spam ditambah sekitar 14,5 milyar spam email dikirimkan tiap harinya, sering kali kejahatan di dunia maya bermula dari sebuah email atau pesan spam. Bahkan di tengah pandemi covid-19 ini, kejahatan yang berkedok spam semakin berkembang, melalui spam atau phising yang menargetkan orang-orang yang menunggu penigiriman barang. Oleh karena itu, kami rasa spam baik melalui media sosial seperti Instagram, facebook, email, ataupun melalui media seperti pada aplikasi messenger konvensional sudah seharusnya ditindaklanjuti lebih tegas karena selain menjengkelkan, ternyata terdapat celah yang dapat disusupi oleh scammers.

Kita dulu terlalu “bodoh” menganggap sepele spam seperti “mama minta pulsa”, kita menganggap itu hanya sebuah tipuan yang hanya akan menjebak segelintir orang dan akhirnya jebakan itu tidak akan relevan lagi seiring dengan waktu. Pemikiran ini terbukti salah, spam yang sudah kita ketahui sejak dulu dan kita pikirkan tidak akan menyebabkan potensi kejahatan yang lebih parah, ternyata sekarang menjad cikal bakal perkembangan kejahatan di dunia maya, untuk itu laporan ini ingin membuat sebuah ide dan gagasan untuk mengatasi permasalahan ini.

# Tujuan dan Manfaat

Dari tinjauan sebelumnya kita sudah tau bahwa kejahatan dengan kedok spam semakin berkembang, untuk terealisasinya tujuan kami menemukan solusi itu maka kami akan membuat sebuah model sederhana yang kami harapkan dapat mengurangi dan menjadi sebuah solusi utama memerangi kejahatan tersebut. Untuk itu terdapat beberapa manfaat yang kami jabarkan sebagai tujuan terciptanya model ini :

1. Membuat model yang mampu mendeteksi apakah sebuah pesan merupakan spam atau bukan
2. Membuat model dengan akurasi yang tinggi (diatas 90 persen) agar proses deteksi dan klasifikasi bisa dipercaya dan valid
3. Jika model sudah layak maka diharapkan bisa membantu dan menjadi solusi utama untuk permasalahan persebaran spam
4. Seturut perkembangannya kami harapakan teknologi model kami dapat digunakan dan disematkan ke dalam sistem sebuah software seperti pada media sosial ataupun pada aplikasi pesan yang konvensional sekalipun, agar tujuan awal kami semakin nyata dan meluas.

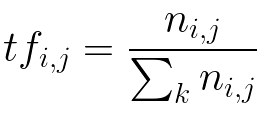
# Data

Dataset yang kami dapatkan bersumber dari <https://www.kaggle.com/uciml/sms-spam-collection-dataset>. Dengan deskirpsi datanya sebagai berikut :

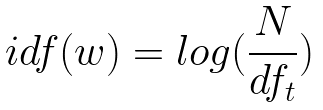
|  |  |
| --- | --- |
| v1 | Berisi data berupa object class yang menandakan apakah pesan adalah sebuah spam. Ditandai dengan kategori spam sebagai data yang dilabeli spam atau label ham yang berarti sebuah data bukan merupakan kategori spam. |
| v2 | Berisi data text yang merupakan sebuah pesan yang ingin kita label kan spam atau bukan spam |

# Metodologi Penelitian

Metode yang kami gunakan dalam penelitian ini adalah menggunakan Naïve Bayes Algorithm. Dataset diinput menggunakan library pandas. Lalu, diolah menggunakan model sebagai berikut. Dataset memasuki proses tokenisasi, yaitu membagi teks menjadi kata-kata. Stopwords juga dihilangkan berdasarkan bahasa dari dataset itu sendiri. Dataset yang kami ambil menggunakan bahasa Inggris. Karakter spesial juga dihilangkan. Lalu lemmatisation diterapkan pada dataset dengan tujuan untuk mengubah kata-kata yang memiliki imbuhan atau bentuk lain kembali menjadi kata dasarnya. Setelah itu dataset siap memasuki preprocessing dan training.

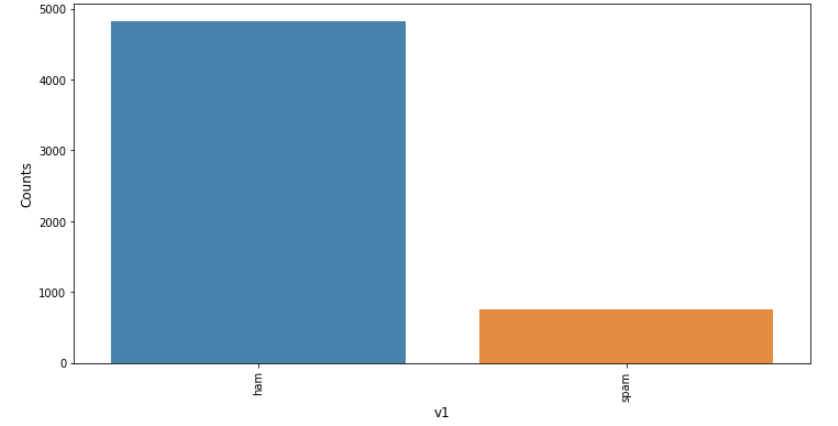
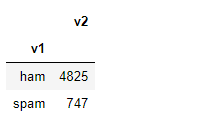
Pada proses ini kami menggunakan bantuan pipeline. String-string pada dataset diubah menjadi integer. Kemudian model mengkonversi jumlah bilangan bulat menjadi skor TF-IDF. TF-IDF berasal dari Term Frequency (TF) dikalikan dengan Inverse Data Frequency (IDF). Term Frequency merupakan nilai dari jumlah berapa kali kata tersebut muncul pada dataset dibagi dengan jumlah seluruh kata pada dataset. 

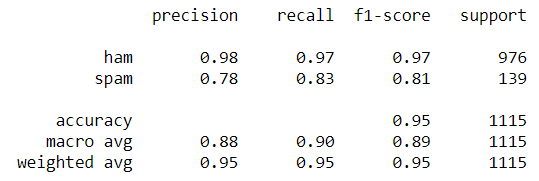
Inverse Data Frequency merupakan nilai dari log jumlah dokumen dibagi dengan jumlah dokumen yang memiliki kata tertentu. IDF berfungsi untuk menentukan seberapa jarangnya seluruh kata pada dataset digunakan.



Kemudian dilakukan proses training pada dataset menggunakan Naïve Bayes Algorithm. Dataset tersebut lalu displit menjadi train data dan test data. Setelah itu dilakukan evaluasi model untuk melihat hasil prediksi berdasarkan data seperti precision call dan lain-lain menggunakan bantuan classification report dan accuracy score.

# Hasil Eksperimen dan Analisis



Dari data itu kita dapat melihat bahwa kebanyakan pesan pada dataset ini dilabeli ham, yang berarti data ini baik untuk kita lakukan permodelan sederhana untuk deteksi dan classifikasi spam.

Seperti yang bisa kita lihat pada classification report, hasil precision dan recall yang didapatkan cukup tinggi yaitu 98% dan 97% untuk recall. Model dapat dibilang baik ketika hasil dari f1-score bernilai tinggi dan hasil yang kami dapatkan adalah 97% yang sudah dapat dibilang model ini sudah berfungsi dengan baik.

# Kesimpulan

Bedasarkan hasil yang diperoleh, model spam filter yang kami buat berhasil mendeteksi pesan spam atau phising dengan akurasi 95,06% yang tentunya sudah termasuk sangat akurat dalam memilah pesan spam ataupun phising. Kami berharap bahwa dengan adanya model spam filter ini dapat memicu developer untuk mengembangkan dan mengaplikasikan kedalam software berbasis messaging untuk mengurangi pesan spam atau phising yang masuk.

# Referensi

1. Contoh Code-1: <https://www.kaggle.com/mohitr/simple-spam-filter/notebook>
2. Contoh Code-2: <https://www.kaggle.com/astandrik/simple-spam-filter-using-naive-bayes>
3. Contoh Code-3: <https://www.ritchieng.com/machine-learning-multinomial-naive-bayes-vectorization/>
4. Contoh Code-4: <https://towardsdatascience.com/multinomial-naive-bayes-classifier-for-text-analysis-python-8dd6825ece67>
5. Dataset: <https://www.kaggle.com/uciml/sms-spam-collection-dataset>
6. Scikit Library: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html>
7. Berita: <https://teknologi.bisnis.com/read/20200506/84/1237008/pandemi-covid-19-waspada-ancaman-penipuan-delivery-malware-melalui-spam-dan-phising>
8. Facts about spam message: <https://www.propellercrm.com/blog/email-spam-statistics>