BAB I

PENDAHULUAN

I.1 Latar Belakang

Salah satu layanan internet yang paling banyak digunakan adalah surat elektronik, atau email. Email adalah media korespondensi yang sederhana, cepat, dan mudah digunakan Pengiriman korespondensi melalui media elektronik, yang juga dikenal sebagai e-mail, merupakan praktik yang umum dilakukan dalam kehidupan sehari-hari. Namun, penyalahgunaan email dalam komunikasi juga meningkat seiring dengan meningkatnya penggunaan email. Munculnya email komersial yang tidak diharapkan, yang juga dikenal sebagai spam, adalah salah satunya. Perkembangan spam sangat mengganggu klien email karena dapat memperluas penggunaan transmisi data asosiasi web, dan akan berubah menjadi sampah yang terkumpul sehingga mengurangi batas kapasitas. Beberapa spam digunakan untuk menyebarkan pesan yang mengandung unsur cabul atau media penyebar infeksi, oleh karena itu diperlukan media yang dapat membedakan dan menyalurkan spam, sehingga dapat mengisolasi antara spam dan email. (Mubarikah et al., 2021)

Spam adalah kegiatan yang dilakukan lebih dari satu kali, dapat juga dikatakan bahwa pengirim data melakukan spam (spammer) dengan sengaja untuk melakukan suatu perbuatan yang tidak baik atau spammer tersebut tidak disangka-sangka sehingga

tidak menyadari bahwa ia telah melakukan spam. Pesan spam terus meningkat dengan cepat. Saat ini, lebih dari 85% email yang masuk adalah spam. Email spam dapat dikategorikan sebagai "email massal spontan", yaitu email yang dikirimkan kepada sejumlah besar penerima yang sebagian besar berisi substansi yang tidak aman dan berbahaya. Kehadirannya sangat mengganggu klien email, terutama organisasi atau pekerja perorangan yang menggunakan email kantor setiap hari untuk membantu pekerjaan mereka. Pesanan email spam sangat penting dalam mengelola hal ini. (Siddique et al., 2021)

Selama beberapa dekade terakhir, teknologi telah memperoleh kecepatan yang cepat dalam perkembangannya mempermudah komunikasi. Mempertimbangkan beberapa mode komunikasi, email (surat elektronik) adalah cara terbaik untuk percakapan informal dan formal. Beberapa juga menggunakan email untuk menyimpan dan berbagi informasi penting dalam bentuk teks, gambar, dokumen, dan lain lain, antara orang yang menggunakan perangkat elektronik. Selain itu, beberapa orang menggunakan secara tidak benar alat komunikasi ini dengan mengirimkan email yang tidak berguna atau tidak diinginkan secara massal, yaitu spam email yang dapat mengakibatkan penggunaan memori yang tidak proporsional di kotak surat. ada banyak pendekatan yang disarankan dalam praktik yang dapat mengidentifikasi email spam dari kotak surat menggunakan metode pembelajaran mesin. Makalah ini terutama berkaitan dengan analisis komparatif pendeteksian Email Spam dengan berbagai metodologi pembelajaran mesin bersama dengan yang diusulkan metodologi. Mempertimbangkan berbagai metrik evaluasi seperti *accuracy, error, evaluation*

waktu, efisiensi, dan sebagainya untuk evaluasi model. Dokumen ini menggambarkan kontrasnya kelebihan, kekurangan, dan keterbatasan dari beberapa teknik yang ada yang menggunakan pendekatan pembelajaran mesin untuk mendeteksi email spam. Metode pembelajaran mesin lebih lanjut dari pendekatan yang tidak melibatkan spesifikasi dari setiap instruksi. Mempertimbangkan berbagai metrik evaluasi seperti akurasi, error, waktu evaluasi, efisiensi, dan sebagainya untuk evaluasi model. Berbagai akurasi yang didapatkan pada *framework* ini adalah KNN sebanyak 93.20%, Naïve Bayes sebanyak 92.46% dan SVM sebanyak 93,90% (Madhavan et al., 2021)

Kekurangan pada penelitian Mangena Venu Madhavan dan kawan kawan yang menggunakan algoritma *machine learnin*g adalah Pertama, ketergantungan pada data pelatihan dimana *machine learning* membutuhkan data pelatihan yang representatif dan bervariasi untuk menghasilkan model yang baik. Jika data pelatihan tidak cukup mewakili variasi spam email yang ada, model dapat menjadi tidak efektif dalam mengklasifikasikan spam dengan akurasi yang tinggi. Kedua, sensitivitas terhadap perubahan dalam spam taktik. Para pengirim spam terus mengembangkan teknik baru untuk menghindari deteksi spam email. *Machine learning* cenderung kurang adaptif terhadap perubahan ini dan memerlukan pembaruan dan pelatihan ulang secara teratur untuk tetap efektif dalam mendeteksi spam terbaru. Ketiga, *overfitting*, jika model *machine learning* terlalu rumit atau jika data pelatihan terlalu sedikit atau tidak seimbang, model dapat mengalami *overfitting*, Hal ini mengimplikasikan bahwa model tersebut ternyata terlalu berpusat pada informasi persiapan dan tidak dapat menyimpulkan dengan baik informasi yang belum pernah dilihat sebelumnya.

Keempat, tugas labeling data yang memakan waktu. Proses pelabelan data untuk melatih model *machine learning* dapat memakan waktu dan membutuhkan sumber daya manusia yang cukup. Data yang digunakan untuk melatih model spam email harus diklasifikasikan secara manual sebagai spam atau bukan spam, yang dapat menjadi tugas yang memakan waktu dan memerlukan konsistensi dalam penilaian.

Spam adalah email spontan yang berisi penawaran barang, hiburan erotis, infeksi, dan substansi yang tidak relevan, yang dikirim dari banyak orang. Masalah spam dapat diatasi dengan adanya aplikasi pengelompokan email, yaitu aplikasi yang secara otomatis mengenali sebuah email, terlepas dari apakah email tersebut spam atau bukan. Innocent Bayes adalah strategi pengelompokan langsung yang dapat digunakan untuk mengkarakterisasi informasi berdasarkan standar tertentu. Strategi ini menggunakan hipotesis kemungkinan, yang mencari peluang yang paling jelas, dengan mengantisipasi probabilitas masa depan dengan melihat data masa lalu. Tujuan utama dari eksplorasi ini adalah untuk berkonsentrasi pada pemanfaatan strategi Innocent Bayes untuk memutuskan email spam dan email ham. Pengujian aplikasi terhadap lima email, dua di antaranya adalah spam dan tiga di antaranya bukan, mengungkapkan bahwa algoritma naive bayes mengklasifikasikan email dengan benar dengan akurasi 60%. (Wibisono et al., 2020)

Kekurangan pada penelitian Aria Wibisono dan kawan kawan ini adalah: Pertama, sensitivitas terhadap *fitur teks* yang tidak relevan. *Naïve Bayes* dapat dipengaruhi oleh *fitur teks* yang tidak relevan dalam email. Fitur-fitur tersebut dapat memberikan bobot yang tinggi dalam perhitungan probabilitas, meskipun sebenarnya tidak memiliki

kontribusi yang signifikan dalam membedakan spam dan email bukan spam. Hal ini dapat mempengaruhi akurasi model. Kedua, ketidakseimbangan kelas. Dalam kasus *filtering* spam email, biasanya terjadi ketidakseimbangan antara jumlah email spam dan email bukan spam. *Naïve Bayes* cenderung memiliki kesulitan dalam mengatasi ketidakseimbangan ini, karena probabilitas prior yang dihitungnya mungkin tidak mewakili distribusi kelas yang sebenarnya. Hal ini dapat mengakibatkan performa yang tidak seimbang dalam mengklasifikasikan spam dan email bukan spam. Ketiga, keterbatasan dalam memahami konteks. *Naïve Bayes* hanya memperhitungkan frekuensi kemunculan kata-kata dalam email, ini berarti metode ini tidak sepenuhnya memahami konteks atau makna di balik kata-kata tersebut. Hal ini dapat menyebabkan kesalahan klasifikasi jika email spam menggunakan strategi penipuan atau manipulasi kata-kata yang lebih canggih.

Deep Neural Network telah menjadi solusi yang menjanjikan untuk menyuntikkan *Artifial Inteligence (AI)* ke dalam kehidupan kita sehari-hari. *Deep Neural Network* sering digunakan untuk akurasi karena sifatnya yang adaptif di bidang penelitian pengklasifikasian secara otomatis. Ada arsitektur canggih dan jaringan yang telah dilatih sebelumnya dan dapat diubah serta disesuaikan untuk menangani tugas klasifikasi yang lebih baru. *Deep Neural Network* ini biasanya terdiri dari beberapa lapisan yang terhubung satu sama lain dengan banyak parameter masing-masing. Dengan menggunakan set data pelatihan, fase pelatihan menghitung parameter ini untuk menangani klasifikasi data yang lebih baru. (Ridwan et al., 2020)

Deep Neural Network memiliki kapabilitas untuk mengekstraksi fitur yang lebih

kompleks, mampu menangani dataset yang lebih besar, dan dapat memberikan performa yang lebih baik pada data dengan struktur yang kompleks sedangkan *K-Nearest Neighbors* lebih sederhana, mudah dipahami, dan efektif pada dataset kecil hingga menengah. (Madhavan et al., 2021)

Long Short-Term Memory (LSTM) yang merupakan desain tingkat tinggi dari Organisasi Otak Berselang (RNN) yang memiliki komponen pintu masuk termasuk sel memori. LSTM memiliki pemikiran utama untuk mengingat nilai bobot dengan menggunakan memori sel. Dalam penerapannya, LSTM memiliki masalah kerumitan yang sangat tinggi. (Mubarikah et al., 2021)

Long Short-Term Memory merupakan adalah semacam rekayasa dalam organisasi otak berulang (RNN) yang dalam banyak kasus digunakan untuk menangani informasi sebagai rangkaian, misalnya, teks atau deret waktu sementara K-Closest Neighbors bekerja dengan sorotan informasi sebagai vector numerik. Fitur-fitur tersebut dapat berupa numerik atau kategorikal, dan perlu dilakukan normalisasi jika fitur memiliki skala yang berbeda. Dalam kesimpulannya, KNN cocok untuk kasus-kasus sederhana yang melibatkan fitur-fitur numerik atau kategorikal dan tidak ada ketergantungan temporal yang signifikan dalam data sedangkan LSTM cocok untuk kasus-kasus yang melibatkan urutan data seperti teks atau deret waktu, di mana hubungan temporal antara data memiliki peran penting dalam klasifikasi atau prediksi. (Siddique et al., 2021)

Saat ini penelitian terkait Deep Neural Network (DNN) dan Long Short-Term Memory (LSTM) dalam kesadaran manusia terus berkembang seiring dengan kemajuan mekanis dan kebutuhan aplikasi yang semakin kompleks.

Berdasarkan penelitian sebelumnya (Madhavan et al., 2021) dan (Wibisono et al., 2020) tentang spam email menggunakan metode *Naïve Bayes classifier* yang kurang akurat dalam membedakan spam dan yang bukan spam pada email maka penelitian ini bermaksud untuk meningkatkan penelitian sebelumnya dengan mengimplementasikan algoritma *Deep Neural Network* dan *Long Short Term Memory* untuk klasifikasi spam dan non-spam serta membandingkan dari kedua algoritma tersebut algoritma mana yang lebih akurat dalam melakukan filtering pada spam email dengan mengangkat judul penelitian "**Perbandingan Algoritma *Deep Neural Network (DNN)* dan *Long Short Term Memory (LSTM)* Dalam *Filtering* Spam Email**"

I.2 Rumusan Masalah

Berdasarkan penjelasan di latar belakang, maka peneliti merumuskan beberapa permasalahan, yaitu :

1. Bagaimana cara *clustering* spam pada email menggunakan algoritma *Deep Neural Network* dan *Long Short Term Memory*?
2. Seberapa nilai akurasi, recall, presisi dan F1-Score dari algoritma *Deep Neural Network* dan *Long Short-Term Memory (LSTM)* dalam *clustering* spam email?
3. Bagaimana cara kerja D*eep Neural Network* dan *Long Short-Term Memory* dalam *clustering* spam email?
4. Apa saja *output* yang dihasilkan dari *clustering* spam pada email menggunakan algoritma *Deep Neural Networkn* dan *Long Short Term Memory* ?

I.3 Tujuan Penelitian

Berdasarkan rumusan masalah, tujuan penelitian adalah :

1. Penelitian ini bertujuan untuk mengetahui cara kerja algoritma *Deep Neural Network* dan *Long Short-Term Memory* dalam clustering spam email.
2. Penelitian ini bertujuan untuk mendapatkan algoritma yang paling akurat dalam memprediksi spam email. Algoritma yang digunakan algoritma *Deep Neural Network* dan *Long Short Term Memory (LSTM)*.
3. Mengimplementasikan algoritma *Deep Neural Network* dan *Long Short Term Memory* untuk clustering *spam* dan *non-spam.*

I.4 Manfaat Penelitian

Adapun manfaat hasil dari penelitian yang dilakukan baik secara teoretis maupun praktis sebagai berikut:

1. Manfaat penelitian ini adalah membantu para *user* untuk mengetahui apakah email yang diterima tersebut adalah spam atau bukan.
2. Manfaat teoritis dari penelitian ini yaitu diharapkan dapat memberikan sumbangan untuk pengembangan teori yang berkaitan dengan penerapan algoritma *Deep Neural Network* dan *Long Short Term Memory* untuk meningkatkan akurasi prediksi spam email.
3. Manfaat penelitian ini mengetahui penerapan algoritma *Deep Neural Network* dan

Long Short-Term Memory dalam *clustering* spam pada email.

4. Sebagai referensi bagi penelitian lain yang hendak melakukan penelitian sejenis

I.5 Batasan Masalah

Untuk mengindari meluasnya pembahasan, maka penelitian ini memiliki batasan masalah, yaitu:

1. Algoritma *clustering* yang digunakan adalah *Deep Neural Network* (DNN) dan *Long Short-Term Memory* (LSTM).
2. *Clustering* spam yang diteliti hanya pada Email.
3. Data yang digunakan adalah dataset *Spambase* dan *dataset Email.*
4. Data Spam email yang digunakan berbahasa Inggris.
5. *Output* yang dihasilkan yaitu klasifikasi spam dan bukan spam (ham)
6. Evaluasi yang digunakan adalah evaluasi akurasi, precision, recall, dan F1-score.
7. Dataset yang digunakan sebanyak 2200 data yang berbahasa inggris

I.6 Sistematika Pembahasan

Sistematika pembahasan penelitian ini adalah sebagai berikut:

BAB I : PENDAHULUAN

Latar belakang penulisan, identifikasi masalah, batasan dan rumusan masalah, tujuan penelitian, ruang lingkup, dan sistematika penulisan penelitian yang akan datang, semuanya tercakup dalam bab ini..

BAB II : TINJAUAN PUSTAKA

Pada bab ini berisi kajian Pustaka yang mengkaji dari landasan teori, rumus dari persamaan, tabel atribut yang digunakan, dan tabel konversi yang ada

BAB III : METODOLOGI PENELITIAN

Bagian ini berisi strategi penelitian yang membahas tentang jenis eksplorasi, teknik pengumpulan informasi, instrumen penelitian, pengujian informasi, dan teknik pemeriksaan.

BAB IV : HASIL PENELITIAN

Pada bab ini menjelaskan penjabaran dari teknik yang digunakan untuk pelaksaan metode yang diterapkan berdasarkan tingkat dari prioritasnya.

BAB V : PEMBAHASAN

Pada bab ini merupakan pemaparan dari pertanyaan pada permasalahan penelitian yang ada dan menjabarkan kontribusi yang keilmuan dari penelitian ini.

BAB VI : KESIMPULAN DAN SARAN

Bagian ini berisi tujuan berbeda yang bisa dibuat berdasarkan penggambaran yang sudah diselesaikan, serta gagasan untuk organisasi atau untuk spesialis tambahan.