PENGEMBANGAN MODEL DEEP LEARNING BERBASIS TRANSFORMER UNTUK MENDETEKSI SPAM EMAIL

DEEP LEARNING PROJECT IS794-C

Kelompok 6

Nikolas Lyen Agung
00000098199

Jonathan Chandra
00000094067

Alvin Clarence
00000098317

Darren Chikal Setiawan
00000081930

Jonahnes Sebastian W.A
00000098319

***Abstract*—** Email merupakan media komunikasi digital yang sangat penting dalam aktivitas pribadi dan profesional. Peningkatan penggunaan email diikuti oleh meningkatnya spam email yang mengganggu kenyamanan pengguna dan menimbulkan risiko keamanan seperti phishing dan penyebaran malware. Metode deteksi spam tradisional berbasis aturan dan kata kunci sering gagal beradaptasi terhadap pola spam yang terus berkembang. Kondisi ini mendorong penggunaan pendekatan deep learning yang mampu memahami konteks dan struktur bahasa secara lebih mendalam. Penelitian ini bertujuan mengembangkan dan mengevaluasi model deep learning untuk mendeteksi spam email secara otomatis. Dataset yang digunakan adalah SpamAssassin Public Corpus yang terdiri dari kategori spam, easy ham, dan hard ham sehingga mencerminkan tantangan klasifikasi pada kondisi nyata. Data diproses melalui tahapan pembersihan, penanganan encoding, ekstraksi konten email, tokenisasi, padding, serta pembagian data latih dan data uji dengan rasio 80 banding 20. Penelitian ini membandingkan beberapa arsitektur deep learning, yaitu RNN, LSTM, BiLSTM, GRU, dan CNN. Hasil eksperimen menunjukkan bahwa model BiLSTM dan CNN memberikan kinerja terbaik dengan akurasi pelatihan dan validasi mendekati 1.00 serta nilai loss yang rendah dan stabil. Model LSTM dan GRU menunjukkan performa yang baik namun masih berada di bawah BiLSTM dan CNN, sedangkan RNN dasar memiliki keterbatasan dalam memahami dependensi jangka panjang. Hasil penelitian ini menunjukkan bahwa BiLSTM dan CNN merupakan arsitektur yang paling efektif untuk deteksi spam email berbasis teks.

***Keywords*—** *Deep Learning, Spam Email detection, BilSTM, CNN, SpamAssasin dataset*

I. INTRODUCTION

Dalam Era komunikasi digital saat ini, email telah menjadi salah satu sarana utama dalam pertukaran informasi, baik dalam konteks pribadi maupun profesional. Namun, meningkatnya volume penggunaan email juga diikuti oleh peningkatannya spam email yang merupakan sebuah pesan yang dikirim secara massal tanpa diminta, bersifat tidak relevan, atau bahkan berbahaya. Pesan-pesan spam ini tidak hanya mengganggu kenyamanan pengguna dengan memenuhi kotak masuk, tetapi juga menimbulkan ancaman serius terhadap keamanan data, seperti serangan phishing, penyebaran malware, dan kebocoran informasi sensitif [1]. Oleh karena itu, pengembangan sistem deteksi spam yang efektif menjadi kebutuhan penting dalam menjaga keamanan dan efisiensi komunikasi digital [2].

Metode deteksi spam tradisional, seperti sistem berbasis rule-based atau pencocokan kata kunci, sering kali tidak mampu beradaptasi dengan pola spam yang terus berkembang. Para pengirim spam semakin kreatif dalam menyamarkan pesan mereka dengan menggunakan teks yang dimodifikasi, tautan menyesatkan, atau lampiran berbahaya untuk menghindari penyaringan otomatis. Kondisi ini menuntut adanya model yang lebih cerdas dan adaptif untuk dapat membedakan antara email yang sah dan email spam secara akurat [3].

Seiring kemajuan teknologi Deep Learning munculah berbagai solusi yang menjanjikan terhadap permasalahn ini. Dengan memanfaatkan arsitektur jaringan saraf seperti Transformer yang kami gunakan untuk project kami, model deep learning mampu mempelajari pola kompleks dan hubungan semantik dari kumpulan data email dalam jumlah besar. Dibandingkan metode machine learning tradisional, model deep learning memiliki kemampuan lebih baik dalam memahami konteks dan nuansa linguistik yang menjadi ciri khas pesan spam [4].

Proyek ini bertujuan untuk mengembangkan dan mengevaluasi model berbasis deep learning untuk mendeteksi spam email secara otomatis. Tujuan utamanya adalah membangun sistem classification yang mampu membedakan antara email spam dan non-spam dengan tingkat akurasi tinggi melalui penerapan teknik representasi teks modern seperti transformer-based encoders [5]. Dengan adanya pendekatan ini, diharapkan sistem yang dikembangkan dapat meningkatkan efektivitas penyaringan email, mengurangi kesalahan deteksi (false positives), serta menciptakan lingkungan komunikasi digital yang lebih aman dan efisien.

II. LITERATURE STUDY

Pengembangan sistem deteksi spam email telah beralih dari metode pencocokan pola sederhana menuju algoritma Deep Learning yang kompleks. Bagian ini membahas teori dasar dan meninjau penelitian terkini untuk menempatkan kontribusi penelitian ini dalam konteks yang tepat.

A. Landasan Teori

Email telah menjadi sarana komunikasi vital secara global, namun dominasinya diikuti oleh peningkatan volume pesan spam yang signifikan. Spam tidak hanya membebani infrastruktur jaringan, tetapi juga membawa ancaman serius seperti penipuan, phishing, dan distribusi malware yang dapat merugikan individu maupun organisasi secara finansial,. Metode deteksi konvensional yang mengandalkan pencocokan kata kunci atau algoritma Machine Learning tradisional, seperti Naive Bayes dan SVM, sering kali gagal beradaptasi dengan taktik spammer yang dinamis karena ketergantungannya pada rekayasa fitur manual yang kaku,. Oleh karena itu, pendekatan berbasis Deep Learning (DL) semakin diminati karena kemampuannya mempelajari representasi fitur secara otomatis dari data mentah melalui lapisan jaringan saraf yang kompleks, sehingga lebih efektif dalam menangani dimensi tinggi dan struktur non-linear pada data teks.

Dalam domain klasifikasi teks, arsitektur Convolutional Neural Network (CNN) dan Recurrent Neural Network (RNN) telah menjadi standar yang banyak digunakan. CNN, yang awalnya dikembangkan untuk pengolahan citra, terbukti efisien dalam mengekstrak fitur lokal yang penting dari teks, seperti frasa atau pola kata kunci spesifik yang menjadi indikator spam. Sementara itu, untuk menangani sifat sekuensial dari bahasa, varian RNN seperti Long Short-Term Memory (LSTM) dan Bidirectional LSTM (Bi-LSTM) dikembangkan guna mengatasi masalah vanishing gradient dan menangkap ketergantungan jangka panjang antar kata. Bi-LSTM memiliki keunggulan khusus karena memproses teks dari dua arah (maju dan mundur), memungkinkan model memahami konteks masa lalu dan masa depan secara bersamaan. Keandalan arsitektur ini telah dikonfirmasi melalui eksperimen awal yang dilakukan oleh Kelompok 6 pada dataset SpamAssassin, di mana model Bi-LSTM dan CNN berhasil mencapai akurasi validasi mendekati sempurna, mengungguli varian RNN standar maupun GRU.

Meskipun model sekuensial seperti Bi-LSTM menunjukkan kinerja yang sangat baik, arsitektur ini memproses kata secara berurutan yang dapat membatasi efisiensi komputasi dan terkadang gagal menangkap nuansa semantik global pada dokumen yang panjang. Untuk mengatasi keterbatasan ini, arsitektur Transformer diperkenalkan dengan mekanisme self-attention, yang memungkinkan model menimbang relevansi setiap kata terhadap kata lainnya dalam kalimat secara simultan tanpa bergantung pada urutan waktu,. Salah satu implementasi Transformer yang paling mutakhir adalah BERT (Bidirectional Encoder Representations from Transformers). Berbeda dengan model sebelumnya, BERT dilatih menggunakan masked language model pada korpus teks yang sangat besar, memungkinkannya menghasilkan representasi kata yang sangat kontekstual (contextual embedding). Kemampuan ini memungkinkan model berbasis Transformer untuk memahami ambiguitas bahasa dengan lebih baik dibandingkan model DL konvensional, menjadikannya solusi yang menjanjikan untuk mendeteksi spam yang menggunakan teknik manipulasi teks yang canggih.

B. Penelitian Terdahulu

Berbagai macam penelitian terdahulu telah di published yang membahas mengenai model deep learning transformasi untuk mendeteksi spam email. Berikut merupakan list penelitian terdahulu yang membahas tentanhg hal ini;

Peneliti	Model	Analisis
AbdulNabi & Yaseen (2021)	BERT (Fine-tuned) vs. Bi-LSTM vs. KNN	Menunjukkan superioritas BERT (+2%) atas Bi-LSTM karena kemampuan attention menangkap konteks semantik lebih baik daripada model sekuensial.
Guo et al. (2022)	BERT (Feature Extraction) + Machine Learning (LogReg, SVM)	Menggabungkan fitur kaya dari BERT dengan klasifikasi sederhana terbukti efektif, namun pendekatan hibrida ini mungkin tidak sekuat end-to-end fine-tuning.
Zavrak & Yilmaz (2023)	HAN (Hierarchical Attention Network) + CNN + GRU	Menekankan pentingnya representasi hierarkis dan attention mechanism, namun model ini masih berbasis RNN (GRU) yang memproses secara sekuensial.
Tusher et al. (2025)	Review Deep Learning (CNN, LSTM, Bi-LSTM)	Mengonfirmasi bahwa DL mengungguli metode klasik, namun tantangan tetap ada pada imbalanced dataset dan serangan spam yang terus berevolusi.

Tabel 1. Komparasi antara penelitian terdahulu mengenai model deep learning transformasi untuk mendeteksi spam email

III. METHODOLOGY

Penelitian metodologi ini dirancang untuk mengembangkan dan mengevaluasi sistem deteksi spam email yang mempunyai berbasis deep learning secara sistematis dan

terstruktur. Pendekatan yang digunakan sebagian besar mencangkup adanya tahapan dalam pengumpulan dan prapemrossesan data, perancangan arsitektur neural network, implementasi model, serta evaluasi dan optimasi kinerja. Dataset SpamAssassin Public Corpus digunakan sebagai sumber data yang utama, sementara itu berbagai arsitektur jaringan saraf diuji dan membandingkan untuk mengidentifikasi sebuah model yang paling efektif dalam mengklarifikasi email spam dan non-spam. Seluruh proses penelitian dilakukan dengan mempertimbangkan aspek akurasi, kemampuan generalisasi, serta adanya reprodisibilitas hasil eksperimen.

1. Neural Network Design

Pada tahap awal penelitian, melakukan perancangan dan pemeliharaan arsitektur neural network yang sesuai dengan klasifikais spam email. Beberapa model yang digunakan dalam deep learning yang diterapkan dan dibandingkan, yaitu Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (Gru), Bidirectional LSTM (BiLSTM), dan Convolutional Neural Network (CNN). Berdasarkan hasil eksplorasi dan evaluasi dalam peforma pada data yang memberikan pelatihan dan validasi, arsitektur BiLSTM dan CNN dipilih sebagai model utama karena secara konsisten menunjukkan kinerja terbaik. Arsitektur model secara umum terdiri dari embedding layer yang berfungsi sebagai sesuatu yang mengubah token numerik menjadi vektor berdimensi tetap untuk mengektraksi pola penting dari teks email, dimana model BiLSTM memproses adanya urutan teks secara dua arah untuk memahami konteks secara menyeluruh, sementara CNN memanfaatkan convolutional layer untuk mengektraksi pola lokal seperti n-gram yang pada umumnya ditemukan pada email spam. Untuk meningkatkan adanya kemampuan generalisasi model, diterapkan ntuk meningkatkan adanya kemampuan generalisasi model, diterapkan Dropout Layer setelah lapisan BiLSTM dan CNN. Teknik ini berfungsi untuk mencegah *overfitting* dengan cara mematikan sebagian neuron secara acak selama proses pelatihan, sehingga model tidak hanya menghafal data tetapi benar-benar mempelajari fitur-fitur pembeda antara email spam dan ham (email normal).

2. Implementasi Model Berbasis Transformer (BERT)

Berdasarkan evaluasi pada tahap sebelumnya menunjukkan bahwa model baseline BiLSTM dan CNN mampu menghasilkan akurasi yang tinggi, kedua arsitektur tersebut memproses teks secara sekuensial atau lokal, yang membatasi kemampuan model dalam menangkap hubungan semantik jarak jauh secara simultan. Untuk mengatasi keterbatasan ini dan meningkatkan kemampuan generalisasi sistem, penelitian ini berlanjut pada penerapan arsitektur Transformer, secara spesifik menggunakan model BERT (Bidirectional Encoder Representations from Transformers). Berbeda dengan model RNN yang membaca teks secara urut dari kiri ke kanan atau sebaliknya, BERT menggunakan mekanisme self-attention yang memungkinkannya memahami konteks setiap kata dengan melihat keseluruhan kalimat sekaligus dari dua arah (bidirectional), menghasilkan representasi fitur yang jauh lebih kaya dan kontekstual.

Strategi yang diterapkan dalam penelitian ini adalah Transfer Learning, yaitu memanfaatkan model BERT yang telah dilatih sebelumnya (pre-trained) pada korpus teks masif seperti Wikipedia dan BookCorpus. Pendekatan ini memungkinkan model untuk memiliki pemahaman linguistik dasar yang kuat tanpa perlu melatihnya dari awal (scratch), yang kemudian akan diadaptasi (fine-tuned) khusus untuk tugas klasifikasi email spam. Pada tahap input, teks email akan melalui proses pra-pemrosesan khusus agar sesuai dengan format yang dibutuhkan BERT. Setiap kalimat akan ditambahkan token spesial [CLS] di awal sebagai representasi klasifikasi seluruh urutan dan [SEP] di akhir kalimat sebagai pemisah. Selain itu, proses tokenisasi menggunakan WordPiece tokenizer yang memecah kata menjadi sub-kata untuk menangani kosakata yang jarang muncul, dengan panjang urutan yang diseragamkan, misalnya menjadi 300 token, untuk memastikan konsistensi dimensi input.

Arsitektur untuk klasifikasi akhir dibangun di atas lapisan encoder BERT. Output vektor dari token [CLS], yang merepresentasikan konteks semantik dari keseluruhan email, akan diteruskan ke lapisan klasifikasi tambahan. Arsitektur ini terdiri dari lapisan Dropout untuk mencegah overfitting dengan mematikan sebagian neuron secara acak selama pelatihan, diikuti oleh lapisan Fully Connected (Dense). Lapisan terakhir menggunakan fungsi aktivasi (seperti Log Softmax atau Sigmoid) untuk menghasilkan probabilitas prediksi biner, yaitu menentukan apakah email tersebut termasuk kategori spam atau ham (bukan spam).

3. Konfigurasi pelatihan dan optimasi Hyperparameter

Proses pelatihan model berbasis Transformer memerlukan konfigurasi hyperparameter yang presisi untuk menjaga stabilitas bobot pre-trained saat mempelajari fitur baru dari dataset SpamAssassin. Berdasarkan studi komparatif terdahulu yang dilakukan oleh AbdulNabi dan Yaseen, konfigurasi yang terbukti efektif untuk tugas ini melibatkan penggunaan optimizer **AdamW (Adam with Weight Decay)**. Algoritma ini dipilih karena kemampuannya dalam menangani pembaruan bobot secara efisien pada model besar. Tingkat pembelajaran (learning rate) diatur pada nilai yang sangat kecil, umumnya pada rentang $2e^{-5}$ hingga $4e^{-5}$, untuk mencegah perubahan drastis pada bobot yang telah dipelajari sebelumnya oleh BERT. Selain itu, pelatihan dilakukan dengan ukuran batch (batch size) sebesar 32 sampel per iterasi. Mengingat efisiensi metode transfer learning, jumlah epoch yang diperlukan relatif sedikit, yaitu sekitar 2 hingga 4 epoch, yang sudah cukup untuk mencapai konvergensi optimal tanpa menyebabkan overfitting pada data latih. Strategi ini memastikan bahwa model dapat beradaptasi dengan nuansa spesifik dari dataset email tanpa melupakan pengetahuan bahasa umumnya.

4. Skenario Pengujian dan Evaluasi

Setelah pelatihan, kinerja model Transformer akan dibandingkan kembali dengan model baseline (BiLSTM dan CNN) yang telah dikembangkan sebelumnya. Evaluasi dilakukan menggunakan data uji yang terpisah (20% dari total dataset). Mengingat karakteristik dataset spam yang sering kali tidak seimbang, metrik akurasi saja tidak cukup.

Oleh karena itu, kinerja dievaluasi menggunakan empat metrik utama:

1. Akurasi (Accuracy): Rasio prediksi benar terhadap keseluruhan data.
2. Presisi (Precision): Mengukur seberapa akurat model dalam memprediksi email sebagai spam (meminimalkan False Positive agar email penting tidak masuk ke folder spam).

$$\text{Pr}\,e\,cision = \frac{TP}{TP + TP}$$

3. Recall (Sensitivitas): Mengukur kemampuan model menemukan seluruh email spam yang ada (meminimalkan False Negative).

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score: Rata-rata harmonik antara Presisi dan Recall, memberikan gambaran keseimbangan kinerja model pada dataset yang tidak seimbang.

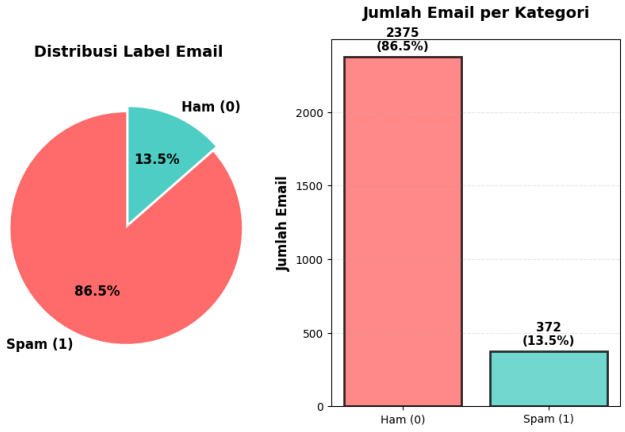
$$F1 = 2 \times \frac{\text{Pr}\,e\,cision \times Recall}{\text{Pr}\,e\,cision + Recall}$$

Selain itu, ROC-AUC (Receiver Operating Characteristic - Area Under Curve) juga akan dianalisis untuk melihat kemampuan model membedakan kelas spam dan ham pada berbagai ambang batas klasifikasi.

IV. Result and Discussion

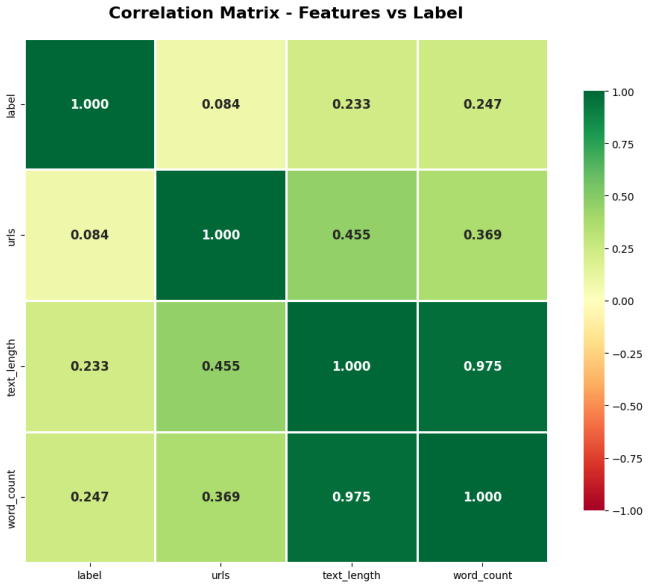
A. *Exploratory Data Analysis (EDA)*

Pada proyek ini kami sudah melakukan berbagai macam Exploratory Data Analysis (EDA) untuk memahami karakteristik, pola, serta distribusi informasi yang terkandung dalam dataset email spam dan ham. Proses ini krusial untuk memastikan bahwa data yang akan diproses oleh model *deep learning* memiliki fitur yang dapat dibedakan secara signifikan. Salah satu aspek utama yang dianalisis adalah karakteristik struktural dari email, seperti panjang subjek, yang sering kali menjadi indikator awal dalam klasifikasi spam.



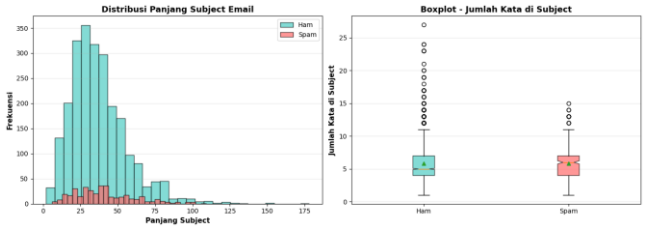
Gambar 1.1 Pie & Bar Chart Label Distribution

Berdasarkan grafik pie chart beserta Bar Chart yang menunjukan distribusi antara email spam dan ham (email normal). Berdasarkan grafik di atas memunculkan bahwa dataset ini memiliki data yang mendominasi kelas spam sebesar 86.5%, disisi lain grafik ini juga menunculkan bahwa kelas ham hanya 13.5% dari total dataset.



Gambar 1.2 Correlation Matrix

Pada visualisasi Correlation Matrix ini menunjukkan bahwa terdapat korelasi positif yang sangat kuat sebesar 0.975 antara text_length dan word_count. Hal ini logis dikarenakan jumlah kata berbanding lurus dengan panjang karakter teks. Disisi lain nilai 0.247 menunjukkan adanya korelasi positif lemah yang berarti seiring bertambahnya jumlah kata dalam subjek, ada kecenderungan probabilitas email tersebut dikategorikan sebagai spam meningkat.

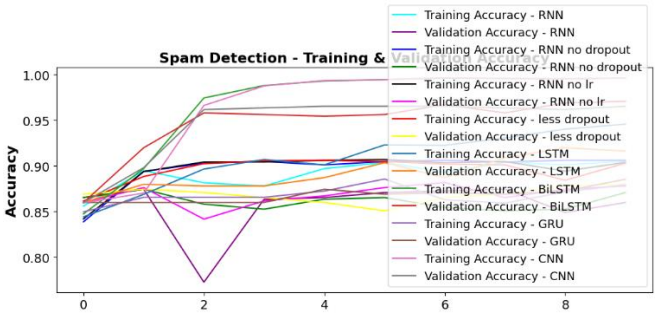


Gambar 1.3 Histogram & Boxpot ubject Length analysis

Dari visualisasi diatas memilik 2 visualisasi yang membandingkan sebaran panjang subjek email antara kategori Ham dan Spam. Setelah kami analisis memunculksan bahwa jumlah email Ham jauh lebih banyak, karakteristik panjang subjek (baik dalam jumlah karakter maupun jumlah kata) antara Ham dan Spam terlihat cukup serupa. Disaat kami melihat pola distribusinya kedua kategori memiliki distribusi yang miring ke kanan (right-skewed). Artinya, sebagian besar email memiliki subjek yang relatif pendek, namun ada beberapa email yang memiliki subjek sangat panjang.

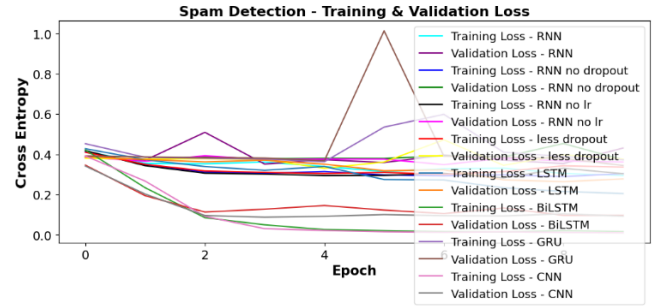
B. *Hasil Analysis Deep Learning*

Pada proyek ini kami sudah melakukan berbagai macam analisis dengan menggunakan berbagai model-model Deep Learning yang membantu kami dalam pendekataan analisis mengenai pendeteksian antar spam emails. Model-Model deep learning yang membantu kami dalam melakukan analisis ini terdiri dari RNN, LSTM, BLSTM, GRU, dan CNN. Kami menganalisis dataset ini dengan banyak model agar dapat membawakan hasil akurasi terbaik yang akan dibandingkan dengan model satu sama lain.
Berikut merupakan visualisasi-visualisasi berdasarkan hasil modelling yang sudah kami lakukan:



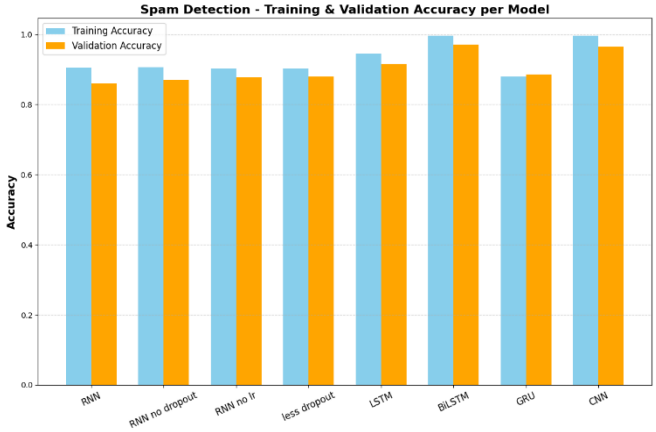
Gambar 1.4 Akurasi Spam Detection Training & Validation

Berdasarkan grafik Akurasi Pelatihan dan Validasi, model CNN dan BiLSTM menunjukkan kinerja terbaik dalam deteksi spam, dengan akurasi pelatihan mendekati 100% dan akurasi validasi stabil di sekitar 96%. Model lain seperti RNN, LSTM, dan GRU umumnya memiliki tingkat akurasi yang lebih rendah dan mengalami plateau, menunjukkan keterbatasan dalam menangkap pola data dibandingkan dengan CNN dan BiLSTM. Terdapat juga perbedaan antara akurasi pelatihan dan validasi, terutama pada model CNN, yang menunjukkan gejala overfitting yang ringan.



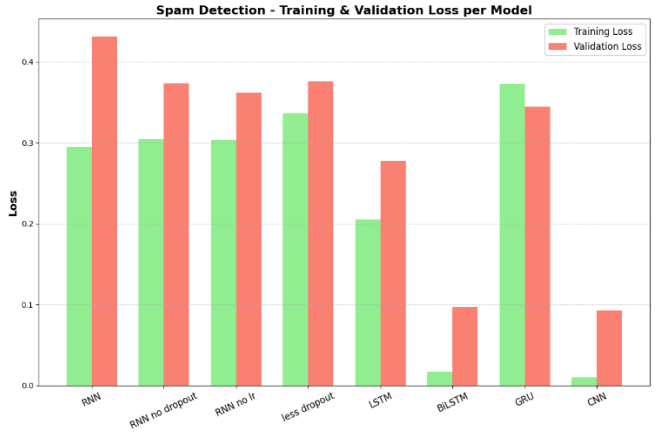
Gambar 1.5 Validation Loss & Training dalam Spam Detection

Pada grafik training dan validation loss, menunjukkan kemampuan terbaik dalam meminimalkan kesalahan karena kerugian pelatihan turun sangat cepat mendekati nol dan kerugian validasi tetap stabil pada nilai terendah dibandingkan dengan model lain. Namun, terdapat selisih yang cukup jelas antara kerugian pelatihan dan validasi pada kedua model tersebut, yang menunjukkan gejala overfitting ringan. Di sisi lain, model RNN, LSTM, dan GRU mengalami penurunan kerugian yang lambat dan cenderung tetap pada nilai yang relatif tinggi, sehingga menunjukkan keterbatasan mereka dalam mempelajari pola data. Selain itu, ketidakstabilan model RNN tanpa penyesuaian laju pembelajaran menegaskan pentingnya laju pembelajaran dan dropout dalam menjaga stabilitas dan kemampuan generalisasi model.



Gambar 1.6 Perbandingan Akurasi antara Model

Dari visualisasi grafik deteksi spam diatas menunjukkan, bahwa CNN dan BiLSTM adalah model dengan kinerja terbaik karena mampu mencapai akurasi pelatihan mendekati 100% dan akurasi validasi di atas 95%, serta memiliki nilai kerugian terendah yang turun sangat cepat mendekati nol. Model lain seperti RNN, LSTM, dan GRU menunjukkan akurasi yang lebih rendah dan cenderung stagnan, disertai dengan nilai kerugian yang relatif tinggi, sehingga kurang efektif dalam mempelajari pola data.



Gambar 1.7 Perbandingan Training & Validation Loss antara Model

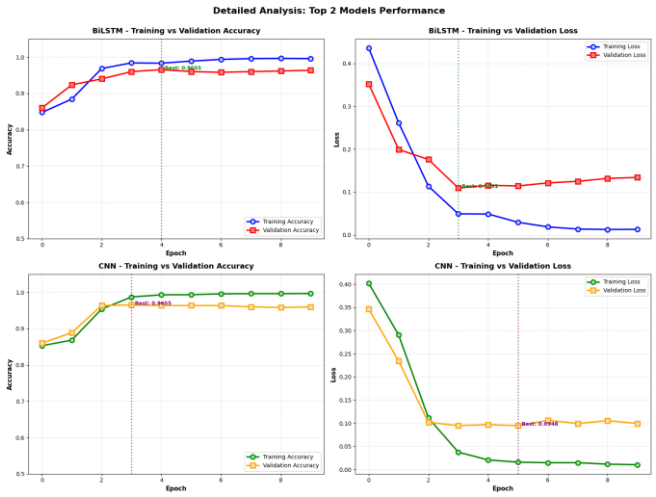
Pada grafik training dan validation loss per model, hasilnya menunjukkan bahwa CNN dan BiLSTM terbukti menjadi model yang paling unggul karena mampu mencapai akurasi pelatihan mendekati 100% dan akurasi validasi di atas 95%, serta memiliki nilai kerugian terendah yang menunjukkan tingkat kesalahan yang sangat kecil. Model-model berurutan lainnya seperti RNN, LSTM, dan GRU menunjukkan kinerja yang lebih rendah dengan akurasi validasi yang stagnan di sekitar 85-90% dan nilai kerugian yang relatif tinggi, yang menunjukkan bahwa model-model ini memiliki keterbatasan dalam mempelajari pola data secara optimal.

Model	Train Accuracy	Validation Accuracy
RNN (Adam, Dropout=0.5)	0.9053	0.8600
RNN tanpa Dropout	0.9062	0.8709
RNN tanpa Adam LR	0.9040	0.8782
RNN Dropout=0.4	0.9030	0.8800
LSTM	0.9458	0.9164
BiLSTM	0.9964	0.9709
GRU	0.8794	0.8855
CNN	0.9964	0.9655

Gambar 1.8 Perbandingan hasil akurasi antar Model

Berdasarkan hasil perbandingan yang kami dapatkan dapat diambil kesimpulan bahwa CNN dan BiLSTM merupakan model terbaik dalam eksperimen deteksi spam untuk mencapai akurasi pelatihan tertinggi (CNN 0,9964; BiLSTM 0,9964), akurasi validasi terbaik di atas 96%, serta nilai kerugian terendah yang turun tajam dan stabil. Di sisi lain, RNN, LSTM, dan GRU menunjukkan kinerja yang lebih rendah dengan akurasi validasi yang masih naik turun di sekitar 85–89% dan kerugian validasi di atas 0,30, sehingga terbatas dalam mengekstrak fitur teks.

C. *Top performing model*



Gambar 1.9 Top 2 performing model

Berdasarkan visualisasi diatas yang menunjukkan kedua best performing model yaitu BiLSTM dan CNN dalam mengklasifikasikan email, di mana keduanya menunjukkan bahwa sifat overfitting. Dalam model BilSTM menunjukkan overfitting yang ringan setelah epoch ke-3, lain halnya dalam model CNN yang juga mengalami overfitting namun dengan pola yang lebih stabil di akhir. Kedua model memiliki overfitting gap yang kecil antar kedua yaitu 0.0254 yang menunjukkann bahwa BilSTM merupakan model deep learning yang kami gunakan paling optimize untuk dataset kami.

D. *Refleksi*

Proses pengumpulan dan persiapan dataset SpamAssassin untuk proyek deteksi spam kami menghadapi beberapa tantangan signifikan yang berpotensi mempengaruhi kinerja dan kemampuan generalisasi model kami. Tantangan-tantangan ini terutama muncul dari karakteristik bawaan dataset itu sendiri dan kompleksitas. Ada hal seperti Data Imbalance, Missing values, dan sistem sistem pemrosesan data email yang terlalu kompleks.

E. *Strategi Mitigasi dan Pemecahan Masalah*

Mengatasi tantangan temporal, kami menerapkan teknik augmentasi data dan transfer learning yang memungkinkan model beradaptasi dengan pola spam baru. Dalam menangani ketidakseimbangan data, teknik sampling dan fungsi kerugian berberat dapat dipertimbangkan untuk iterasi model selanjutnya. Proses pra-pemrosesan yang ketat dengan penanganan enkoding multi dan validasi data komprehensif telah berhasil menciptakan dataset yang relatif bersih dan terstruktur.

Pengalaman dalam menangani dataset ini memberikan wawasan berharga bahwa kualitas data tidak hanya tentang kelengkapan, tetapi juga tentang relevansi temporal dan representativitas terhadap skenario dunia nyata. Ke depan, integrasi dengan dataset yang lebih kontemporer dan pendekatan pembelajaran berkelanjutan akan menjadi faktor kunci dalam mengembangkan sistem deteksi spam yang benar-benar tangguh dan adaptif terhadap evolusi teknik spam yang terus berlanjut.

V. Conclusion

Penelitian komprehensif ini, yang dilakukan melalui pendekatan pemodelan komparatif, telah memberikan dasar empiris yang kuat untuk memilih arsitektur jaringan syaraf tiruan yang paling efektif. Temuan ini tidak hanya memberikan wawasan tentang kinerja berbagai arsitektur deep learning, tetapi juga membuka peluang untuk pengembangan yang lebih optimal di masa depan. Dengan memanfaatkan kekuatan BiLSTM dalam memahami konteks dan CNN dalam mengekstrak pola lokal, diharapkan sistem deteksi spam dapat dikembangkan menjadi semakin akurat dan tangguh dalam menghadapi teknik-teknik spam yang terus berkembang.

References

[1] *I. AbdulNabi and Q. Yaseen, "Spam email detection using deep learning techniques," Procedia Computer Science, vol. 184, pp. 853–858, 2021, doi: 10.1016/j.procs.2021.03.107.*

[2] X. Liu, "Deciphering Spam Through AI: From Traditional Methods to Deep Learning Advancements in Email Security," in Proceedings of the 13th International Conference on Cloud Computing and Services Science (CLOSER), 2024, pp. 553–558, doi: 10.5220/0012958700004508.

[3] S. Zavrak and S. Yilmaz, "Email spam detection using hierarchical attention hybrid deep learning method," Expert Systems with Applications, vol. 233, p. 120977, Dec. 2023, doi: 10.1016/j.eswa.2023.120977.

[4] E. H. Tusher, M. A. Ismail, and A. F. M. Raffei, "Email spam classification based on deep learning methods: a review," Iraqi Journal for Computer Science and Mathematics, vol. 6, no. 1, 2025, doi: 10.52866/2788-7421.1236.

[5] Y. Guo, Z. Mustafaoglu, and D. Koundal, "Spam detection using bidirectional transformers and machine learning classifier algorithms,"

Journal of Computational and Cognitive Engineering, vol. 2, no. 1, pp. 5–9, 2022, doi: 10.47852/bonviewjcce2202192.