Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik...
Vol. x, No. x, April 2014, hlm. x-x p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 51/E/KPT/2017 e-ISSN: 2528-6579

DETEKSI SPAM BERBAHASA INDONESIA BERBASIS TEKS MENGGUNAKAN MODEL BERT

(Naskah masuk: dd mmm yyyy, diterima untuk diterbitkan: dd mmm yyyy)

Abstrak

Spam pada SMS dan *Email* menyebabkan pengalaman kurang menyenangkan bagi pengguna dalam pemanfaatan teknologi. Spam secara umum merupakan sebuah tindakan mengirim pesan yang tidak diinginkan atau tidak diminta kepada sejumlah besar orang. Spam kini dapat ditemui dalam berbagai bentuk, seperti web maupun multimedia. Klasifikasi berbasis model IndoBERT dan MultilingualBERT digunakan untuk mendeteksi spam berbahasa Indonesia pada pesan SMS dan *Email*. Model yang sudah dipilih kemudian dilatih untuk mengidentifikasi perbedaan antara pesan spam dan bukan spam. Hasil evaluasi pada percobaan menggunakan dataset SMS dan *Email* memiliki nilai akurasi sebesar 98% pada model IndoBERT dan 95% pada model MultilingualBERT, yang menunjukkan tingkat akurasi yang tinggi. Hasil ini menunjukkan bahwa model BERT efektif dalam mendeteksi pesan spam dalam Bahasa Indonesia.

Kata kunci: *spam, deteksi spam, pemrosesan bahasa alami, BERT, text mining, klasifikasi teks*

TEXT-BASED INDONESIAN SPAM DETECTION USING THE BERT MODEL

Abstract

Spam on SMS and Email causes an unpleasant experience for users in using technology. Spam in general is the act of sending unwanted or unsolicited messages to a large number of people. Spam can now be found in various forms, such as web and multimedia. Classification based on the IndoBERT and MultilingualBERT models is used to detect Indonesian language spam in SMS and Email messages. The selected model is then trained to identify the differences between spam and non-spam messages. Evaluation results in experiments using SMS and Email datasets have an accuracy value of 98% in the IndoBERT model and 95% in the MultilingualBERT model, which shows a high level of accuracy. These results indicate that the BERT model is effective in detecting spam messages in Indonesian.

Keywords: *spam, spam detection, natural language processing, BERT, text mining, text classification*

1. PENDAHULUAN

Populasi penduduk di Indonesia pada tahun 2019 tercatat telah mencapai lebih dari 250 juta jiwa. Dengan banyaknya jumlah penduduk tersebut, Indonesia memiliki potensi pasar yang sangat signifikan dari berbagai teknologi. Pada tahun 2018, tercatat bahwa jumlah pengguna aktif perangkat seluler di Indonesia telah melampaui 100 juta pengguna. Dengan jumlah pengguna sebesar ini, Indonesia menduduki peringkat keempat sebagai negara dengan jumlah pengguna aktif perangkat seluler terbanyak di dunia, setelah Cina, India, dan Amerika Serikat (Rahmayani, 2019).

Seperti koin yang memiliki dua sisi, pertumbuhan pesat dalam penggunaan teknologi, termasuk perangkat seluler, juga membawa tantangan yang besar. Dengan peningkatan jumlah pengguna perangkat seluler, akan menciptakan lingkungan yang ideal untuk berbagai praktik penipuan, khususnya melalui media pesan singkat, yang sering disebut dengan *Short Message Service* (SMS).

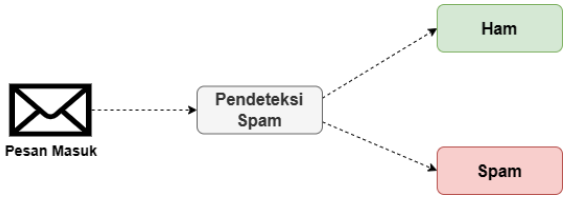
SMS atau layanan pesan singkat merupakan salah satu bentuk komunikasi jarak jauh yang masih sering digunakan saat ini. Namun, seiring dengan perkembangan penggunaan layanan pesan singkat ini, muncul banyak dampak negatif berupa serangan pada perangkat seluler, yang dikenal sebagai SMS spam. SMS spam merujuk pada pesan singkat yang tidak diinginkan oleh penerima, seperti iklan dan upaya penipuan (*fraud*) (Ma, 2016). Pesan spam yang banyak beredar meliputi informasi perbankan, pengumuman promosi dan diskon toko, tarif baru penyedia layanan komunikasi, atau pesan-pesan yang tidak memiliki makna dan mengganggu lainnya (Uysal, dkk., 2012). SMS spam tidak hanya dapat menimbulkan kerugian karena penipuan, tetapi juga dapat mengganggu kotak masuk dan merusak pengalaman pengguna (*user experience*) dalam menggunakan perangkat seluler.

Selain SMS, *Email* pada awalnya juga merupakan tujuan dari penyerangan spam. *Email* atau yang dikenal juga dengan surat elektronik merupakan bentuk komunikasi elektronik yang memungkinkan pengiriman pesan dan dokumen melalui jaringan

internet. *Email* digunakan untuk berkomunikasi secara tertulis antara pengguna yang memiliki alamat *Email*. Dalam sebuah *Email*, biasanya terdapat komponen-komponen seperti alamat pengirim, alamat penerima, subjek, dan isi pesan. *Email* memungkinkan untuk mengirim pesan teks, gambar, video, dan berbagai jenis dokumen dalam format digital. Spam *Email* merupakan pesan *Email* yang tidak diminta atau tidak diinginkan yang dikirimkan secara massal ke ribuan atau jutaan alamat *Email*. Spam *Email* biasanya mengandung iklan atau promosi produk, *phishing*, *virus* atau *malware*, dan pesan palsu lainnya. Spam *Email* bisa sangat mengganggu dan merugikan, karena dapat menguras waktu dan sumber daya komputer pengguna (Hartono, dkk., 2023).

Spam secara umum merupakan sebuah tindakan mengirim pesan yang tidak diinginkan atau tidak diminta kepada sejumlah besar orang. Spam kini dapat ditemui dalam berbagai bentuk, seperti web maupun multimedia. Spam sendiri adalah upaya untuk menyalahgunakan atau memanipulasi suatu sistem tekno-sosial dengan membuat atau menyuntikkan konten yang tidak diminta dan atau tidak diinginkan yang bertujuan untuk mengarahkan perilaku manusia atau sistem demi keuntungan jangka panjang maupun jangka pendek dari *spammer* baik secara langsung maupun tidak langsung (Lutfiyani & Retnowati, 2021).

Salah satu solusi untuk mengatasi masalah pesan spam adalah dengan menerapkan teknik klasifikasi menggunakan pembelajaran mesin (*machine learning*) yang secara khusus termasuk dalam bidang pemrosesan bahasa alami (*natural language processing*) untuk secara otomatis menyaring pesan-pesan spam tersebut. Klasifikasi adalah proses yang bertujuan untuk menemukan suatu model atau fungsi yang dapat mengidentifikasi karakteristik dari dua kategori yang berbeda, yaitu spam dan bukan spam (*ham*). Klasifikasi pada teks telah diterapkan dalam beberapa hal misalnya filterisasi email, filterisasi berita, prediksi kecenderungan user, kategorisasi teks dalam web, dan pengorganisasian dokumen. Dengan harapan bahwa model pembelajaran mesin yang dikembangkan dapat dengan akurat mengenali ciri-ciri pesan spam dan memisahkannya dari pesan yang bukan spam, seperti yang dapat dilihat pada gambar (1). Pendekatan ini akan menjadi sangat efektif dalam mengatasi permasalahan pesan spam.



Gambar 1. Pendeteksian Pesan Spam

Mengklasifikasikan teks ke dalam kategori tertentu merupakan isu sentral dalam pemrosesan bahasa alami. Langkah-langkah yang penting dalam proses ini melibatkan perancangan arsitektur saraf dan penciptaan representasi data dengan menggunakan *word embedding*. Representasi bahasa yang mendalam ini selalu menjadi faktor penting untuk kategorisasi teks yang efisien (Paul & Saha, 2020). Selama beberapa tahun terakhir, BERT telah menjadi model representasi yang sangat populer dan efektif, menghasilkan kinerja terdepan dalam tugas-tugas tingkat kalimat dan token-level, bahkan melampaui banyak arsitektur yang dirancang khusus untuk tugas tertentu (Devlin, dkk., 2019).

Berdasarkan uraian penjelasan sebelumnya, penulis terdorong untuk mengajukan penelitian yang berjudul Deteksi Spam Berbahasa Indonesia Berbasis Teks Menggunakan Model BERT. Pada penelitian ini, sistem akan mengidentifikasi sekumpulan data pesan SMS dan *Email* para pengguna, kemudian membedakan apakah pesan tersebut termasuk ke dalam kategori spam atau bukan spam (*ham*). Melalui pendekatan berbasis model BERT, penelitian ini akan berfokus pada pengembangan metode yang dapat menghasilkan hasil yang lebih akurat dalam deteksi spam berbahasa Indonesia, dengan tujuan untuk meningkatkan keamanan dan pengalaman berkomunikasi pengguna di lingkungan digital.

2. KAJIAN PUSTAKA

2.1. *Text Mining*

Text mining adalah salah satu bidang khusus dalam data mining yang memiliki definisi menambang data berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen (Mooney, 2006). *Text mining* termasuk tipe *natural languange processing* atau pengolahan bahasa alami yang menguraikan istilah (berupa kata dan frasa) dari dokumen tertentu (Gegick dkk., 2009). *Text mining* merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi, dimana text mining merupakan variasi dari *data mining* yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar (Feldman & Sanger, 2007).

Kehadiran *Text Mining* telah membuka jendela besar bagi peneliti, analis data, dan profesional di berbagai bidang untuk memahami lebih dalam isi teks yang ada, tujuan *text mining* adalah untuk menghasilkan inovasi yang membantu orang untuk mengerti akan suatu sistem dengan menggunakan gudang dokumen (Kumar, 2009).

Salah satu aspek paling menarik dari *Text Mining* adalah kemampuannya dalam mengungkapkan pengetahuan tersembunyi dalam data teks. Ini bukan hanya tentang mengekstraksi kata-kata atau frasa kunci, tetapi juga tentang mengidentifikasi pola yang mungkin tersembunyi di antara kata-kata tersebut. Dalam konteks pengambilan keputusan, ini memiliki implikasi besar. Sebab, ketika data teks digali secara

efisien, kita dapat mengungkapkan wawasan yang dapat menjadi dasar bagi pengambilan keputusan yang lebih baik dan lebih terinformasi. *Text Mining* tidak hanya menghadirkan teknologi canggih dalam pemrosesan teks, tetapi juga membawa dampak positif dalam upaya menyaring dan menyajikan informasi yang relevan dalam berbagai konteks, memungkinkan pengambilan keputusan yang lebih cerdas dan efektif. (Putri & Setiadi, 2015)

2.2. *Natural Language Processing*

Natural Language Processing (NLP) adalah penerapan ilmu komputer, khususnya linguistik komputasional (*computational linguistics*), untuk mengkaji interaksi antara komputer dengan bahasa (alami) manusia (Amber & James, 2012). Konsep ini membuka jalan bagi komputer untuk menjalin interaksi yang semakin mendalam dan bermakna dengan bahasa manusia, yang seringkali kompleks, ambigu, dan penuh dengan nuansa.

Basis dari NLP adalah penerapan ilmu komputer dalam mengurai dan memproses bahasa manusia dalam segala bentuknya, termasuk tulisan, ucapan, atau pesan-pesan teks. Konsep ini telah memberikan dorongan signifikan dalam berbagai aspek kehidupan sehari-hari. Sebagai contoh, analisis sentimen, yang melibatkan penilaian subjektif terhadap teks, telah menjadi lebih tepat dan efisien berkat peran NLP. Kemampuan untuk menerjemahkan teks dari satu bahasa ke bahasa lain juga adalah salah satu aplikasi NLP yang sangat berguna. Penerapan lainnya termasuk pemrosesan bahasa alami, yang digunakan untuk berkomunikasi dengan komputer dalam bahasa yang lebih intuitif.

Dalam lingkup bahasa Indonesia, salah satu contoh penerapan NLP yang sangat relevan adalah dalam deteksi spam. Pada dasarnya, NLP adalah elemen kunci dalam upaya ini. Melalui analisis bahasa dan pola komunikasi, NLP membantu mengklasifikasikan pesan-pesan sebagai spam atau bukan. Ini melibatkan pemindaian kata-kata kunci yang sering digunakan dalam pesan spam, pengenalan pola yang mencurigakan, dan identifikasi fitur lain yang dapat membedakan pesan spam dari yang sah.

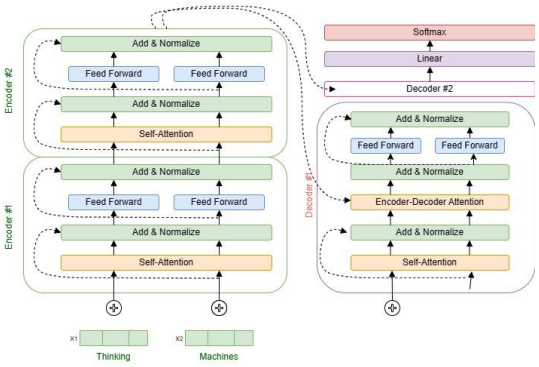
Sebagai kesimpulan, NLP adalah suatu disiplin ilmu yang membuka pintu ke berbagai aplikasi yang membantu kita berinteraksi dengan bahasa manusia secara lebih efektif dan cerdas. Dalam berbagai konteks, dari analisis teks hingga komunikasi dengan komputer, peran NLP semakin penting dan relevan (Amber & James, 2012).

2.3. BERT

BERT (*Bidirectional Encoder Representations from Transformer*) merupakan suatu model yang dirancang untuk melatih representasi dua arah yang mendalam dari teks tak berlabel yang termasuk ke dalam model *deep learning* yang berguna untuk merepresentasikan kata-kata secara kontekstual dalam pemrosesan bahasa alami (NLP) pada tahap pra-pelatihan. Model *pre-trained* BERT bisa di *fine tuning* hanya bermodalkan tambahan 1 *layer* yang menghasilkan model untuk mengerjakan berbagai tugas, seperti menjawab pertanyaan. Model BERT dikembangkan oleh Google pada tahun 2018. Selama proses pelatihan, kata-kata disesuaikan dengan *Masked Language Model* (MLM) dan Transformer dua arah (Devlin, dkk., 2018).

BERT menggunakan struktur model transformer yang terdiri dari beberapa lapisan *encoder*. Struktur model BERT ini pada penerapannya hanya menggunakan tumpukan *encoder* dalam transformer dan tidak menggunakan tumpukan *decoder* (McMahan & Rao, 2019). Penerapan BERT dalam NLP sangat berguna terutama dalam pemahaman konteks kata-kata dalam teks. Dalam deteksi spam berbahasa Indonesia, BERT dapat digunakan untuk menganalisis teks pesan dan mengidentifikasi apakah pesan tersebut merupakan spam atau bukan. Salah satu keunggulan BERT adalah kemampuannya untuk menciptakan representasi vektor kata yang kaya dengan konteks. Hal ini memungkinkan deteksi untuk mengenali kata-kata yang sering muncul dalam pesan spam.

Dalam MLM, beberapa token secara acak disembunyikan dari masukan, dan tujuannya adalah untuk meramalkan kata asli yang telah disembunyikan berdasarkan konteks sekitarnya. Berbeda dengan pem-pre-training model bahasa yang hanya bergerak dari kiri ke kanan, tujuan MLM seperti pada BERT ini memungkinkan representasi untuk mengintegrasikan konteks dari kedua arah. Konsep ini memungkinkan BERT untuk melakukan pem-pre-training yang mendalam dengan sifat ganda (Devlin, dkk., 2019), seperti yang ditunjukkan oleh arsitektur transformer model BERT pada gambar (2).



Gambar 2. Arsitektur Transformer dari Model BERT

2.4. IndoBERT

IndoBERT adalah suatu inovasi terkini dalam dunia pemrosesan bahasa alami, sebuah model yang berakar pada konsep yang terkenal diperkenalkan oleh BERT-Base (Koto, dkk., 2020), suatu model bahasa revolusioner yang lahir dari penelitian Devlin dan rekan-rekannya pada tahun 2019. Akan tetapi, apa yang membuat IndoBERT menjadi sesuatu yang

unik adalah fokusnya yang sangat khusus dalam proses pelatihan, di mana model ini didesain dan dioptimalkan sepenuhnya untuk bahasa Indonesia (Pires, dkk., 2019), menjadikannya sebagai alat yang sangat efisien dalam memahami dan menghasilkan teks dalam bahasa tersebut. Dalam beberapa tahun terakhir model bahasa menggunakan *pre-trained* telah menunjukkan terobosan besar dalam NLP dan pada Devlin dkk membuat BERT yaitu sebuah arsitektur untuk melatih model bahasa yang lebih cepat yang mehilangkan recurrences dengan menambahkan multi-head attention layer (Devlin dkk., 2018).

Proyek pengembangan IndoBERT juga memanfaatkan platform terkemuka di dunia pemrosesan bahasa alami, yaitu *Hugging face*. Penggunaan platform ini memberikan kemudahan akses dan integrasi, serta memfasilitasi komunitas peneliti dan pengembang untuk mengadopsi dan menggunakannya dengan lebih efektif. Ini membuka pintu untuk pemrosesan bahasa alami yang lebih maju dalam bahasa Indonesia dan memperluas berbagai aplikasi di berbagai sektor, mulai dari kecerdasan buatan hingga analisis teks yang lebih canggih. Dengan IndoBERT, pengguna dapat menjelajahi potensi tak terbatas dalam analisis teks, pemahaman dokumen, dan bahkan pengembangan aplikasi yang diperkaya oleh kecerdasan buatan, semuanya dengan tingkat kemudahan dan efisiensi yang tinggi (Pires, dkk., 2019).

2.5. MultilingualBERT

Multilingual BERT merupakan suatu terobosan signifikan dalam dunia pemrosesan bahasa alami. Dalam dunia yang semakin terhubung dan multikultural, kebutuhan untuk memiliki alat yang efektif dalam memahami dan menghasilkan teks dalam berbagai bahasa menjadi semakin mendesak. Dalam pandangan ini, Multilingual BERT hadir sebagai varian terkemuka dari model BERT yang memikirkan lebih jauh, mendukung beberapa bahasa sekaligus dengan tingkat akurasi dan kualitas yang tinggi.

Satu hal yang menjadi sorotan utama yang membedakan Multilingual BERT dari sebagian besar model BERT lainnya adalah keragaman yang luar biasa dalam data pelatihannya. Model ini telah melibatkan diri dalam memahami tidak kurang dari 104 bahasa, termasuk bahasa Indonesia (Pires, dkk., 2019), dan ini benar-benar mencerminkan konsep multibahasa yang tak terbatas. Dalam proses pelatihan yang mencakup bahasa-bahasa dari berbagai rumpun dan konteks budaya, Multilingual BERT telah mengembangkan kemampuan yang luar biasa dalam hal pemahaman teks.

Dalam era globalisasi dan interkoneksi, memiliki model seperti Multilingual BERT menjadi sangat penting. Ini tidak hanya memungkinkan perusahaan dan peneliti untuk menjalankan analisis teks yang lebih lanjut dan mendalam di berbagai bahasa, tetapi juga membuka pintu bagi pengembangan aplikasi yang mendukung komunikasi lintas budaya dan berbagai konteks bisnis. Dengan demikian, Multilingual BERT adalah suatu terobosan penting yang membantu memperkuat pemahaman dan kerjasama di dunia yang semakin terhubung ini.

2.6. Relevansi dengan Penelitian Terdahulu

Berdasarkan hasil penelitian sebelumnya yang relevan, studi ini berhasil menemukan hasil penting dalam menganalisis sentimen dalam teks spam. Hasil analisis menggunakan data uji dan data validasi dalam Bahasa Indonesia menunjukkan bahwa model BERT mampu mengklasifikasikan kasus perundungan siber dalam Bahasa Indonesia dengan tingkat akurasi sebesar 81%. Dengan menggunakan algoritma BERT, penulis dapat mengidentifikasi teks yang mengindikasikan perundungan siber dan spam, serta mengklasifikasikannya ke dalam kategori yang telah ditentukan sebelumnya. Sistem ini juga memberikan persentase terkait dengan kategori yang diberikan oleh program tersebut (Niluh, dkk., 2023).

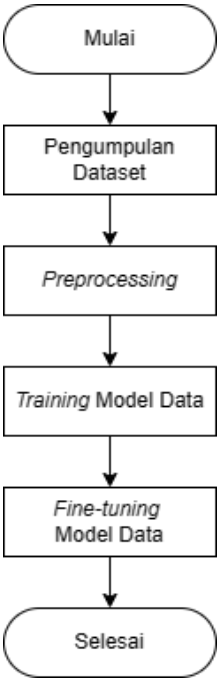
Sebuah penelitian telah dilakukan untuk mengidentifikasi tweet yang merupakan spam dan yang bukan spam di platform Twitter menggunakan metode klasifikasi. Salah satu metode yang digunakan dalam data mining untuk melakukan ini adalah Naïve Bayes. Naïve Bayes sering digunakan karena kesederhanaan algoritmanya dan kemudahan dalam penerapannya. Penelitian ini mengumpulkan data tweet yang mencurigakan sebagai spam dari Twitter, kemudian membaginya menjadi dua bagian: 70% data digunakan untuk pelatihan, dan 30% digunakan untuk pengujian menggunakan metode klasifikasi Naïve Bayes. Data Twitter yang digunakan dalam penelitian ini seringkali berisi kata-kata yang tidak formal, sehingga perlu dilakukan pra-pemrosesan data, yang mencakup tokenisasi, penyaringan, normalisasi kata, dan pemotongan kata. Hasil klasifikasi menunjukkan tingkat akurasi sebesar 95.57% dalam membedakan tweet yang merupakan spam dan yang bukan spam (Wahyuningtyas, dkk., 2022).

Kemudian dari penelitian lainnya telah dilakukan penelitian deteksi bot spammer di Twitter dengan menggunakan data API Twitter dari 18 akun bot dan 14 akun yang sah, masing-masing dengan 1.000 tweet. Hasil terbaik yang ditemukan dalam hal ketepatan, kelengkapan, dan metrik f-measure adalah 100%. Hasil ini menunjukkan bahwa alat yang disebut Glove dan Time Interval Entropy sangat efektif dalam mendeteksi bot spammer. Selain itu, temuan penelitian menunjukkan bahwa penggunaan hashtag juga berperan penting dalam meningkatkan deteksi bot spammer (Priyatno & Arif, 2019).

3. METODOLOGI PENELITIAN

Penelitian ini menggunakan model BERT untuk mengklasifikasikan teks pada pesan yang dapat

digolongkan sebagai spam ataupun bukan spam (*ham*) dari dataset SMS dan *Email* berbahasa Indonesia. Penelitian ini dilakukan dengan pendekatan metodologi penelitian yang sistematis untuk mengembangkan dan mengevaluasi model BERT. Alur penelitian dapat dilihat pada gambar (3).



Gambar 3. Diagram Alur Penelitian

3.1. Pengumpulan Dataset

Terdapat 2 dataset yang digunakan dalam penelitian ini, yang bertujuan untuk memperkuat hasil dari yang didapatkan. Dataset pertama yang digunakan adalah kumpulan dari pesan teks SMS berbahasa Indonesia dengan panjang teks yang beragam. Dataset ini tersedia pada tautan https://yudiwbs.wordpress.com/2018/08/05/dataset-klasifikasi-bahasa-indonesia-sms-spam-klasifikasi-teks-dengan-scikit-learn/, yang terdiri dari 1143 pesan SMS (Rahmi & Wibisono, 2016). Dataset ini terdiri dari dua kolom yaitu kolom teks dan kolom kelas. Pada kolom kelas, data-data dikategorikan menjadi tiga yakni SMS normal, SMS penipuan, dan juga SMS promosi. Kategori SMS penipuan dan SMS promosi ini nantinya dapat dilabelkan sebagai spam dengan jumlah sebanyak 574 pesan, sedangkan SMS normal dapat dilabelkan sebagai *ham* dengan jumlah sebanyak 569 pesan.

Contoh SMS yang dikategorikan sebagai spam adalah "Plg Yth: Simcard anda mendapatkan bonus poin plus-plus 555 dr:PT.INDOSAT pin anda:277fg49 u/info klik di www.indosat-555.blogspot.com atau Hub:021-3338-0074." dan contoh SMS yang dikategorikan sebagai *ham* adalah "Iya ih ko sedih sih gtau kapan lg ke bandung :(".

Dataset kedua adalah kumpulan *email* berbahasa Indonesia yang tersedia pada tautan *github* https://raw.githubusercontent.com/gevabriel/dataset/main/indo_spam_5.csv. Dataset berasal dari *website Kaggle* yang memiliki 5 kolom (Kumar, 2022) dengan menggunakan bahasa Inggris, yang penulis sederhanakan menjadi 2 kolom, yaitu kolom pesan dan kategori, serta data pada kolom pesan diterjemahkan menjadi bahasa Indonesia. Kolom pesan terdiri dari data *Email*, dan kolom kategori terdiri dari data label yang memberikan label "spam" ataupun "*ham*" pada data Email di kolom pesan. Dataset ini terdiri dari 2636 *Email* dengan 1368 *Email* dilabeli sebagai spam dan 1268 *Email* dilabeli dengan *ham*.

Contoh *Email* yang dikategorikan sebagai spam adalah "Mencerahkan gigi itu membuat gigi Anda putih cerah sekarang! Sudahkah Anda mempertimbangkan pemutihan gigi profesional? Jika demikian, Anda tahu biasanya harganya antara $ 300 dan $ 500 dari dokter gigi setempat! Kunjungi situs kami untuk mempelajari cara memutihkan gigi Anda secara profesional, menggunakan sistem pemutih yang sama persis yang digunakan dokter gigi, dengan sedikit biaya! Inilah yang Anda dapatkan: Kami akan menunjukkan kepada Anda apa yang harus dicari dalam sistem pemutih! Kami akan menunjukkan kepada Anda perbandingan semua produk yang tersedia saat ini, termasuk biayanya! Kami tahu produk kami adalah yang terbaik di pasaran, dan kami mendukungnya dengan jaminan uang kembali 30 hari! Klik di sini untuk mempelajari lebih lanjut! Anda menerima *Email* ini sebagai anggota jaringan afiliasi internet. Jika Anda tidak lagi ingin menerima promosi khusus melalui *Email* dari jaringan afiliasi internet, lalu klik di sini untuk berhenti berlangganan"

Sementara itu, contoh *Email* yang dikategorikan sebagai *ham* adalah "Jadwal wawancara untuk Jeff Lei yang dilampirkannya, silakan temukan paket wawancara untuk orang yang dirujuk di atas. Wawancara akan terjadi Jumat 14 Juli 2000. Harap cetak ketiga dokumen untuk copy hard Anda. Jika Anda memiliki pertanyaan, atau konflik jadwal, jangan ragu untuk menghubungi saya. Sean 58701".

3.2. *Preprocessing*

Dataset yang akan digunakan perlu melalui proses *cleaning data* terlebih dahulu agar hasil *training* dapat maksimal. Proses *cleaning* yang diterapkan adalah *str.lower()* untuk membuat semua huruf pada teks yang ada pada dataset menjadi huruf kecil, dan menggunakan library *RegEx* untuk membersihkan huruf-huruf yang tidak digunakan seperti simbol, tanda baca, dan spasi. Dataset juga dibersihkan dari *mention, link, hashtag*, dan URL yang tidak lengkap.

3.3. *Fine-tuning* Model data

Parameter yang digunakan untuk mendeteksi pesan spam menggunakan model BERT antara lain adalah jumlah kata, frekuensi kata, struktur kalimat,

emosi, dan panggilan tindakan. Parameter-parameter ini digunakan untuk melatih model BERT guna membedakan antara pesan biasa dan pesan spam. Model BERT akan mempelajari hubungan antara parameter-parameter ini dan membaginya ke dalam kelas pesan spam atau bukan spam (*ham*).

Berdasarkan jumlah kata, pesan spam pada SMS biasanya lebih pendek daripada SMS biasa karena pengirim spam ingin menghemat biaya. Berdasarkan frekuensi kata, pesan spam sering menggunakan kata-kata dan frasa yang umum digunakan dalam spam, seperti "menangkan hadiah", "diskon besar", dan "tekan di sini". Kata-kata dan frasa tersebut dapat digunakan untuk mendeteksi pesan spam dengan menggunakan teknik analisis leksikal. Kemudian berdasarkan struktur kalimat, pesan spam sering menggunakan struktur kalimat yang tidak natural, seperti kalimat yang terlalu pendek atau terlalu panjang. Hal ini dapat disebabkan oleh pengirim spam yang biasanya menggunakan *template* atau skrip.

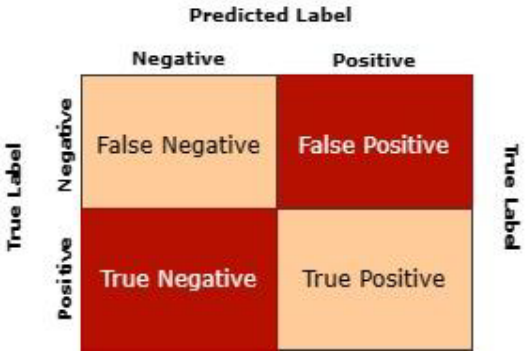
Selanjutnya berdasarkan emosi, pesan spam seringkali menggunakan emosi negatif seperti ketakutan atau kegembiraan untuk menarik perhatian penerima. Hal ini dapat dilakukan untuk membuat penerima lebih mungkin untuk membuka dan membaca pesan tersebut. Terakhir, berdasarkan panggilan tindakan, di mana pesan spam sering menyertakan panggilan tindakan. Contohnya meminta penerima untuk mengklik tautan atau memberikan informasi pribadi. Hal ini dilakukan untuk mengarahkan penerima ke situs web atau aplikasi yang dapat digunakan untuk menipu penerima.

3.4. Evaluasi Kinerja Model

Dalam menilai hasil kinerja dari model yang sudah dilatih, perlu sebuah parameter dalam sebuah laporan klasifikasi yang memastikan penilaian tersebut agar hasilnya dapat dilihat secara objektif. Keluaran dari suatu model klasifikasi pada algoritma *machine learning* dapat dipetakan menjadi 4 poin:

A. *True Positive* (TP): Klasifikasi dengan hasil positif pada data asli yang positif.
B. *True Negative* (TN): Klasifikasi hasil negatif pada data asli yang negatif.
C. *False Positive* (FP): Klasifikasi dengan hasil positif pada data asli yang negatif.
D. *False Negative* (FN): Klasifikasi dengan hasil negatif pada data asli yang positif.

Dengan 4 poin tersebut, akan terbentuk *confusion matrix* seperti pada gambar (4), yang digunakan untuk menghitung akurasi dari sebuah model yang telah dibuat nantinya (Harikrishnan, 2019).



Gambar 4. *Confusion Matrix*

Accuracy yang didapat berdasarkan 4 poin pemetaan didapat menggunakan persamaan (1). Perlu dicatat, bahwa *accuracy* bukan suatu tolak ukur yang bagus jika data yang tersedia tidak seimbang (Harikrishnan, 2019).

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \quad (1)$$

Parameter lain yang digunakan adalah *precision* (*positive predictive value*) dengan menggunakan persamaan (2). *Precision* seharusnya idealnya bernilai 1 (tinggi) untuk sebuah pengklasifikasi yang baik, dengan mencapai nilai 1. ketika pembilang dan penyebutnya sama, yaitu TP = TP + FP. Hal ini juga berarti bahwa FP (*False Positive*) sama dengan nol. Ketika FP meningkat, nilai penyebut menjadi lebih besar dari pembilang, dan nilai *precision* akan turun (Harikrishnan, 2019).

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall (*true positive rate*) pada persamaan (3) lebih mengarah kepada persentase total hasil relevan yang diklasifikasikan dengan benar oleh algoritma yang digunakan pada model *machine learning* (Saxena, 2018). Ketika FN meningkat, nilai penyebut menjadi lebih besar dari pembilang, dan nilai *recall* akan turun (Harikrishnan, 2019).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Dalam klasifikasi yang baik, dibutuhkan parameter *F1-score* pada persamaan (4) yang menerima nilai dari *precision* dan *recall*. *F1-score* bernilai baik jika nilai *precision* dan *recall* mendekati nilai 1, yang berarti FP dan FN juga mendekati nilai 0, sehingga parameter *F1-score* merupakan parameter yang lebih baik untuk mengukur hasil dari klasifikasi pada model daripada parameter *accuracy* (Harikrishnan, 2019).

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \quad (4)$$

3.5. *Tools* dan Perangkat Keras

Dalam penelitian ini, bahasa pemrograman Python dipilih sebagai instrumen utama, mengingat

keunggulannya yang mendukung implementasi *deep learning*. BERT yang merupakan model *deep learning* mendukung untuk merepresentasikan kata-kata secara kontekstual dalam pemrosesan bahasa alami. Penelitian ini dijalankan menggunakan layanan komputasi awan dari Google Colab. Layanan ini bersifat gratis serta menyediakan GPU Tesla T4 yang menggunakan 16GB VRAM dengan memori GDDR6 dan 2.560 CUDA *cores*.

4. HASIL DAN PEMBAHASAN

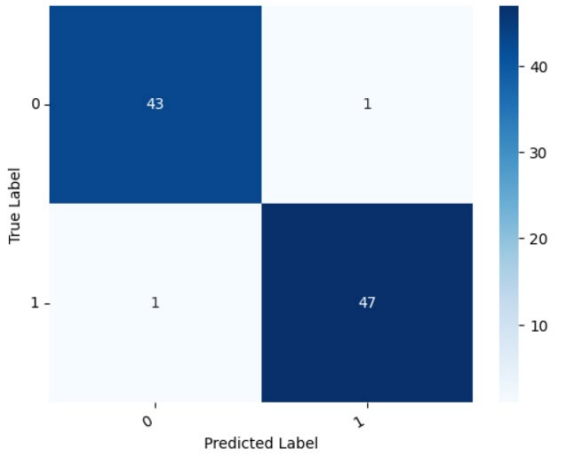
Sebelum melakukan proses *training,* dataset dibagi menjadi 3, yaitu data_train, data_eval, dan data_test dengan menggunakan *train-test-split*. Dataset pertama yaitu dataset SMS dengan jumlah 1143 data, dibagi menjadi perbandingan 80% untuk data_train, yang berjumlah 914 data, dan 20% untuk data_test, yang berjumlah 229 data. Dari 80% data_test tersebut, penulis bagi kembali menjadi perbandingan 60% untuk data_eval, yang berjumlah 137 data, dan 20% untuk data_test, yang berjumlah 92 data. penulis mengkategorikan label "0" sebagai *ham* (bukan spam) dan label "1" sebagai spam.

Sedangkan, untuk dataset kedua yaitu dataset *Email* dengan jumlah 2636 data, dibagi menjadi perbandingan 80% untuk data_train, yang berjumlah 2108 data, dan 20% untuk data_test, yang berjumlah 528 data. Dari 80% data_test tersebut, penulis bagi kembali menjadi perbandingan 60% untuk data_eval, yang berjumlah 316 data, dan 20% untuk data_test, yang berjumlah 212 data. penulis mengkategorikan label "0" sebagai *ham* (bukan spam) dan label "1" sebagai spam.

Proses *training* dilakukan dengan menggunakan model IndoBERT dari *pretrained* "indobenchmark/indobert-base-p2" dan model *Multilingual BERT* dari *pretrained "bert-base-multilingual-cased"*. Dua model tersebut bisa diakses menggunakan *library Transformers* (*Hugging Face*). Proses *training* dibantu dengan *library Accelerate* untuk menggunakan fungsi *TrainingArguments* dari *library transformers*, serta *library PyTorch* untuk melakukan proses *training* menggunakan GPU CUDA.

Dengan melakukan percobaan menggunakan 2 dataset dan 2 model, menghasilkan 4 *confusion matrix* yang mengarah kepada 4 laporan klasifikasi.

A. Dataset SMS dengan model IndoBERT



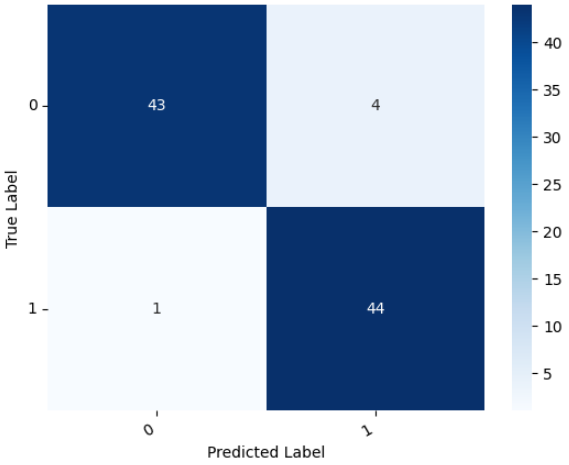
Gambar 5. Hasil *Confusion Matrix Dataset SMS* Dengan Model IndoBERT

Dapat dilihat pada gambar (5), bahwa *confusion matrix* dari data_test percobaan dengan dataset SMS menggunakan model IndoBERT, memiliki hasil *true positive* (TP) sebanyak 47 data, *true negative* (TN) sebanyak 1 data, *false positive* (FP) sebanyak 1 data, dan *false negative* (FN) sebanyak 43 data. Dari *confusion matrix* yang sudah terbentuk, maka dapat dibuat laporan klasifikasi dengan parameter *precision, recall* dan *f1-score* yang dapat dilihat pada tabel (1).

Tabel 1. Laporan Klasifikasi

label	precision	recall	f1-score	support
0	0.98	0.98	0.98	44
1	0.98	0.98	0.98	48
accuracy			0.98	92
macro avg	0.98	0.98	0.98	92
weighted avg	0.98	0.98	0.98	92

B. Dataset SMS dengan model MultilingualBERT



Gambar 6. Hasil *Confusion Matrix* Dataset SMS Dengan Model MultilingualBERT

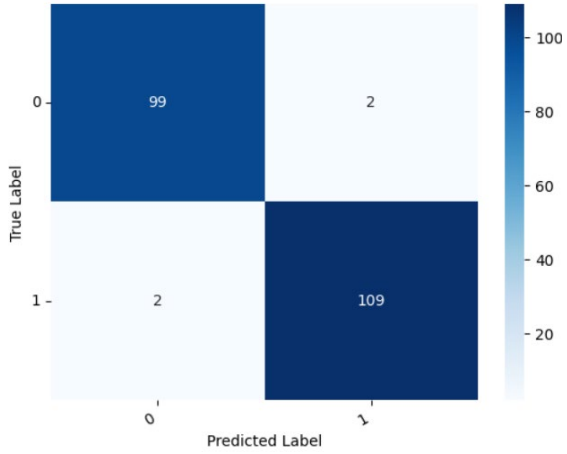
Dapat dilihat pada gambar (6), bahwa *confusion matrix* dari data_test percobaan dengan dataset SMS menggunakan model MultilingualBERT, memiliki hasil *true positive* (TP) sebanyak 44 data, *true*

negative (TN) sebanyak 1 data, *false positive* (FP) sebanyak 4 data, dan *false negative* (FN) sebanyak 44 data. Dari *confusion matrix* yang sudah terbentuk, maka dapat dibuat laporan klasifikasi dengan parameter *precision*, *recall* dan *f1-score* yang dapat dilihat pada tabel (3).

Tabel 2. Laporan Klasifikasi Dataset SMS Dengan Model MultilingualBERT

label	precision	recall	f1-score	support
0	0.98	0.91	0.95	47
1	0.92	0.98	0.95	45
accuracy			0.95	92
macro avg	0.95	0.95	0.95	92
weighted avg	0.95	0.95	0.95	92

C. Dataset *Email* dengan model IndoBERT



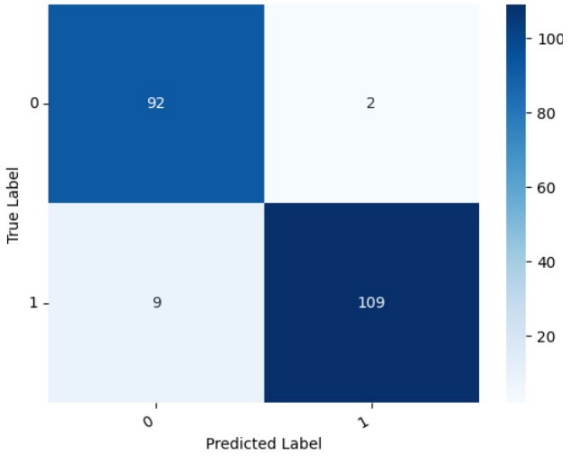
Gambar 7. Hasil *Confusion Matrix* Dataset *Email* Dengan Model IndoBERT

Dapat dilihat pada gambar (7), bahwa *confusion matrix* dari data_test percobaan dengan dataset *Email* menggunakan model IndoBERT, memiliki hasil *true positive* (TP) sebanyak 109 data, *true negative* (TN) sebanyak 2 data, *false positive* (FP) sebanyak 2 data, dan *false negative* (FN) sebanyak 99 data. Dari *confusion matrix* yang sudah terbentuk, maka dapat dibuat laporan klasifikasi dengan parameter *precision, recall* dan *f1-score* yang dapat dilihat pada tabel (3).

Tabel 3. Laporan Klasifikasi Dataset *Email* Dengan Model IndoBERT

label	precision	recall	f1-score	support
0	0.98	0.98	0.98	101
1	0.98	0.98	0.98	111
accuracy			0.98	212
macro avg	0.98	0.98	0.98	212
weighted avg	0.98	0.98	0.98	212

D. Dataset *Email* dengan model MultilingualBERT



Gambar 8. Hasil *Confusion Matrix* Dataset *Email* Dengan Model MultilingualBERT

Dapat dilihat pada gambar (8), bahwa *confusion matrix* dari data_test percobaan dengan dataset *Email* menggunakan model MultilingualBERT, memiliki hasil *true positive* (TP) sebanyak 109 data, *true negative* (TN) sebanyak 9 data, *false positive* (FP) sebanyak 2 data, dan *false negative* (FN) sebanyak 92 data. Dari *confusion matrix* yang sudah terbentuk, maka dapat dibuat laporan klasifikasi dengan parameter *precision, recall* dan *f1-score* yang dapat dilihat pada tabel (4).

Tabel 4. Laporan Klasifikasi Dataset *Email* Dengan Model MultilingualBERT

label	precision	recall	f1-score	support
0	0.91	0.98	0.94	94
1	0.98	0.92	0.95	118
accuracy			0.95	212
macro avg	0.95	0.95	0.95	212
weighted avg	0.95	0.95	0.95	212

Dengan hasil yang sudah didapat, penulis membuat perbandingan hasil yang telah didapat dengan penelitian terdahulu, yang dilakukan oleh Rahmi, F dan Wibisono, Y yang berjudul "Aplikasi SMS Spam Filtering pada Android menggunakan Naive Bayes" yang menggunakan model klasifikasi *Naive Bayes* (*MultinomialNB*) dan *Support Vector Machine* (SVM) dengan dataset SMS yang sama dengan yang penulis gunakan (Rahmi & Wibisono, 2016).

Tabel 5. Perbandingan Hasil Akurasi Pada Laporan Klasifikasi

model	akurasi
MultinomialNB	0.90
SVM	0.92
IndoBERT	**0.98**
MultilingualBERT	0.95

Dataset SMS

Penulis juga membuat perbandingan hasil yang telah didapat dengan penelitian terdahulu, yang dilakukan oleh Kumar, K yang berjudul "*Spam Email*

Classification using BERT" yang menggunakan model klasifikasi BERT dengan dataset *Email* yang sama dengan yang penulis gunakan (Kumar, 2022).

Tabel 6. Perbandingan Hasil Akurasi Pada Laporan Klasifikasi

model	akurasi
BERT	0.91
IndoBERT	**0.98**
MultilingualBERT	0.95

Dataset *Email*

5. KESIMPULAN

SMS dan *Email* merupakan beberapa contoh teknologi yang perkembangannya meningkat secara pesat, sehingga menimbulkan berbagai dampak. Dari sisi negatifnya, penggunaan SMS dan *Email* menjadi kurang efisien dikarenakan adanya serangan yang dikenal sebagai SMS atau *Email* spam. SMS spam merujuk pada pesan singkat yang tidak diinginkan oleh penerima, seperti iklan dan upaya penipuan (*fraud*), sedangkan *Email* spam merujuk pada surat elektronik yang tidak diinginkan oleh penerima yang isinya juga dapat berupa penipuan dan iklan.

Maka dari itu, salah satu solusi untuk mengatasi masalah SMS dan *Email* spam adalah dengan menerapkan teknik klasifikasi untuk secara otomatis menyaring pesan-pesan spam tersebut. Tujuannya adalah untuk menemukan suatu model atau fungsi yang dapat mengidentifikasi karakteristik dari dua kategori pesan yaitu spam dan bukan spam (*ham*). Dengan ini, maka digunakan model IndoBERT dan MultilingualBERT untuk melakukan eksperimen terhadap dataset pesan SMS dan dataset *Email* berbahasa Indonesia.

Ketika dilakukan eksperimen, dataset dibagi menjadi data_train, data_eval, dan data_test. Dengan menggunakan data_test yang menghasilkan 4 *confusion matrix* dari 4 percobaan. Dari *confusion matrix* yang sudah terbentuk, pada percobaan menggunakan IndoBERT dapat terlihat hasil dari parameter *precision, recall* dan *f1-score* yang memiliki nilai akurasi sebesar 98% pada dataset SMS maupun pada dataset *Email*. Sedangkan, dengan menggunakan model MultilingualBERT, diperoleh nilai akurasi sebesar 95% pada dataset SMS maupun pada dataset *Email.* Data yang sudah dilatih kemudian disimpan dan digunakan untuk mendeteksi pesan SMS maupun *Email*, apakah pesan tersebut termasuk ke dalam kelas spam atau bukan spam (*ham*).

DAFTAR PUSTAKA

A. K. UYSAL, S. GUNAL, S. ERGIN & E. S. GUNAL, 2012. "The Impact of Feature Extraction and Selection on SMS Spam Filtering," in Elektronika ir Elektrotechnika (Electronics and Electrical Engineering).

AMBER, S. & JAMES, P., 2012. Natural Language Annotation for Machine Learning. California: O'Reilly.

DEVLIN, J., CHANG, M. W., LEE, K., & TOUTANOVA, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm), 4171–4186.

F. KOTO, A. RAHIMI, J. H. LAU, & T. BALDWIN, 2020. "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020, https://doi.org/10.48550/arXiv.2011.00677

FELDMAN, R & SANGER, J., 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press: New York

GEGICK, M., ROTELLA, P. & XIE, T., 2010. Identifying Security Bug Reports via Text Mining: An Industrial Case Study. IEEE.

HARIKRISHNAN, N. B., 2019. "Confusion Matrix, Accuracy, Precision, Recall, F1 Score, Binary Classification Metric". [online]. Tersedia di: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd.

HARTONO, M. B., DARMAWAN, A. K., & HORIYAH, H., 2023. "Komparasi Deep Learning Dan Traditional Machine Learning Untuk Email Spam Filtering". *Jurnal Minfo Polgan*, *12*(1), 636–643. https://doi.org/10.33395/jmp.v12i1.12474.

I. RAHMAYANI, 2015. "Atasi SMS Spam, Ini Langkah Operator Seluler.", Kominfo [online]. Tersedia di: https://kominfo.go.id/content/detail/6042/atasi-sms-spam-ini-langkah-operatorseluler/0/sorotan_media.

I. RAHMAYANI, 2019. "Indonesia Raksasa Teknologi Digital Asia.", Kominfo, [online]. Tersedia di: https://www.kominfo.go.id/content/detail/6095/indonesia-raksasa-teknologi-digital-asia/0/sorotan_media.

KUMAR, K., 2022. "Spam Email Classification using BERT". [online]. Tersedia di:

https://www.kaggle.com/code/kshitij192/spam-email-classification-using-bert.

KUMAR, V., 2009. Text Mining, Classification, Clustering, and Applications. CRC Press.

LUTFIYANI, R. S., & RETNOWATI, N., 2021. Implementasi Pendeteksian Spam Email Menggunakan Metode Text Mining Dengan Algoritma Naïve Bayes Dan Decision Tree J48. *Jurnal Komputer Dan Informatika*, *9*(2), 244–252. https://doi.org/10.35508/jicon.v9i2.5304

MA, J., ZHANG, Y., LIU, J., & YU, K., 2016. Intelligent SMS Spam Filtering Using Topic Model. 2016 International Conference on Intelligent Networking and Collaborative Systems, 380-383.

MCMAHAN, B. & D, RAO., 2019. Natural Language Processing with Pytorch. Gravenstein Highway North, Sebastopol: O'Reilly Media, Inc.

MOONEY, R. J., 2006. CS 391L Machine Learning Text Categorization. University of Texas, Austin.

PAUL, S., & SAHA, S., 2020. CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification. Multimedia Systems, 0123456789. https://doi.org/10.1007/s00530-020-0710-4.

PIRES, T., SCHLINGER, E., & GARRETTE, D. (2019). How multilingual is multilingual BERT? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4996--5001. https://doi.org/10.18653/v1/P19-1493

PRIYATNO & ARIF, M., 2019. "Deteksi bot spammer twitter berbasis time interval entropy dan global vectors for word representations tweet's hashtag." *Register: Jurnal Ilmiah Teknologi Sistem Informasi* 5.1: 37-46.

PUTRI, E.K., & SETIADI, T., 2014. Penerapan Text Mining Pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes: *Jurnal Sarjana Teknik Informatika*, Vol. 2(3), 73-83.

RAHMI, F. & WIBISONO, Y. , 2016. Aplikasi SMS Spam Filtering pada Android menggunakan Naive Bayes, Unpublished manuscript.

S. NILUH P. V. D., NOVANTO YUDISTIRA, N. & ADIKARA, P. P., 2023. "Analisis Sentimen terhadap Perundungan Siber pada Twitter menggunakan Algoritma Bidirectional Encoder Representations from Transformer (BERT)." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 7.2. 909-916.

SAXENA, S., 2018. "Precision vs Recall". [online]. Tersedia di: https://medium.com/@shrutisaxena0617/precision-vs-recall-386cf9f89488#:~:text=Precision%20means%20the%20percentage%20of,correctly%20classified%20by%20your%20algorithm.

WAHYUNINGTYAS, ANDITA, IMAS SUKAESIH SITANGGANG & HUSNUL KHOTIMAH. "Deteksi Spam pada Twitter Menggunakan Algoritme Naïve Bayes Spam Detection on Twitter using Naïve Bayes Algorithm." *vol* 7: 31-40.