JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.9864

RANCANG BANGUN SISTEM DETEKSI TOXIC DAN SPAM BERBASIS FINITE AUTOMATA DENGAN ALGORITMA AHO-CORASICK

Alya Namira[1*], Zulfahmi Indra[2], Adinda Soleha[3], Bryant Tinambunan[4]

[1,2,3,4]Prodi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan, Jalan Willem Iskandar, Pasar V, Medan Estate, Kecamatan Percut Sei Tuan, Kabupaten Deli Serdang, Sumatera Utara, 20371

Keywords:
Finite Automata, Aho-Corasick, Toxic Detection, Spam Detection, Pattern Matching.

Corespondent Email:
alyanamira3010@gmail.com

Abstrak. Perkembangan media sosial meningkatkan risiko penyebaran konten toxic dan spam yang dapat menurunkan kualitas interaksi pengguna. Penelitian ini bertujuan merancang dan mengimplementasikan sistem deteksi konten toxic dan spam berbahasa Indonesia berbasis Finite Automata menggunakan algoritma Aho-Corasick. Dataset yang digunakan terdiri dari dataset_kata_kasar_idn sebanyak 3.111 data dan dataset_spam_idn sebanyak 2.636 data. Dari kedua dataset tersebut diekstrak 89 kata toxic dan 83 kata/frasa spam yang digunakan sebagai kamus keyword sistem. Sebelum proses deteksi, teks melalui tahap preprocessing berupa case folding dan normalisasi karakter berulang. Selanjutnya, keyword dimasukkan ke dalam struktur trie dan dibangun failure function untuk membentuk automata Aho-Corasick. Pengujian dilakukan terhadap 48 kalimat uji yang terdiri atas kategori Toxic, Spam, Normal, dan Toxic+Spam. Hasil penelitian menunjukkan bahwa sistem mampu mencapai akurasi sebesar 91,7% dengan nilai precision rata-rata 0,925, recall 0,939, dan F1-score 0,931. Hasil tersebut menunjukkan bahwa pendekatan Finite Automata dengan algoritma Aho-Corasick efektif digunakan untuk mendeteksi konten toxic dan spam secara cepat pada teks berbahasa Indonesia.

Abstract. *The growth of social media has increased the spread of toxic and spam content that may reduce the quality of user interactions. This study aims to design and implement an Indonesian toxic and spam detection system based on Finite Automata using the Aho-Corasick algorithm. The datasets used consist of 3,111 entries from dataset_kata_kasar_idn and 2,636 entries from dataset_spam_idn. From these datasets, 89 toxic keywords and 83 spam words or phrases were extracted to build the system dictionary. Prior to detection, input text undergoes preprocessing through case folding and repeated-character normalization. The extracted keywords are then organized into a trie structure and enhanced with failure functions to construct the Aho-Corasick automaton. The system was evaluated using 48 test sentences categorized as Toxic, Spam, Normal, and Toxic+Spam. Experimental results achieved an accuracy of 91.7%, with an average precision of 0.925, recall of 0.939, and F1-score of 0.931. These findings demonstrate that the Finite Automata approach combined with the Aho-Corasick algorithm is effective for fast and accurate detection of toxic and spam content in Indonesian text.*

1. PENDAHULUAN

Perkembangan media sosial dan teknologi komunikasi digital telah meningkatkan intensitas interaksi antar pengguna dalam berbagai platform daring. Di balik kemudahan tersebut, muncul berbagai bentuk komunikasi

negatif yang dapat mengganggu kenyamanan pengguna, salah satunya adalah konten *toxic*. Konten *toxic* umumnya ditandai dengan penggunaan kata-kata kasar, penghinaan, ujaran kebencian, maupun bentuk komunikasi lain yang berpotensi merugikan pengguna lain. Keberadaan konten semacam ini dapat memicu konflik, menurunkan kualitas interaksi, serta menciptakan lingkungan digital yang kurang sehat bagi masyarakat [1].

Selain konten *toxic*, penyebaran pesan *spam* juga menjadi permasalahan yang sering ditemukan pada media sosial dan platform komunikasi digital. Pesan *spam* dapat mengurangi kualitas informasi yang diterima pengguna karena berisi konten yang tidak relevan atau tidak diinginkan. Untuk mengatasi permasalahan tersebut, diperlukan sistem yang mampu melakukan identifikasi secara otomatis terhadap pola-pola kata tertentu. Salah satu pendekatan yang banyak digunakan adalah *pattern matching*, yaitu teknik yang digunakan untuk mencari dan mengenali pola karakter atau kata tertentu dalam sekumpulan data secara efisien sehingga sesuai diterapkan pada proses deteksi konten *toxic* dan *spam* [2].

Algoritma Aho-Corasick merupakan salah satu algoritma pencocokan banyak pola (multiple pattern matching) yang mampu menemukan sejumlah kata kunci sekaligus dalam satu kali proses penelusuran teks. Algoritma ini dikembangkan dengan membangun struktur automata yang memungkinkan pencarian dilakukan secara efisien terhadap sekumpulan pola yang telah ditentukan sebelumnya. Aho-Corasick banyak digunakan pada sistem penyaringan web, *spam filter, antivirus*, serta sistem deteksi intrusi jaringan karena kemampuannya dalam melakukan pencarian pola secara cepat pada data berukuran besar [3].

Konsep dasar yang digunakan dalam algoritma Aho-Corasick adalah Finite Automata (FA), yaitu model komputasi yang terdiri dari sejumlah *state* dan *transisi* yang digunakan untuk mengenali pola tertentu dalam suatu *string*. Finite Automata banyak diterapkan pada berbagai sistem pemrosesan teks karena mampu melakukan pengenalan pola secara sistematis dan memiliki kompleksitas waktu yang efisien. Integrasi antara Finite Automata dan algoritma Aho-Corasick memungkinkan proses deteksi banyak kata sekaligus dilakukan secara *real-time* tanpa harus melakukan pencarian berulang terhadap setiap pola yang ada [4].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk merancang dan mengimplementasikan sistem deteksi *toxic* dan *spam* berbasis Finite Automata menggunakan algoritma Aho-Corasick pada media sosial. Sistem yang dikembangkan diharapkan mampu mengidentifikasi kata atau frasa yang termasuk kategori *toxic* maupun *spam* secara cepat dan akurat sehingga dapat membantu proses moderasi konten serta meningkatkan kualitas interaksi pengguna dalam lingkungan media sosial.

2. TINJAUAN PUSTAKA

2.1. Konten Toxic pada Media Sosial

Konten *toxic* merupakan bentuk komunikasi negatif yang banyak ditemukan pada berbagai platform media sosial. Konten ini umumnya ditandai dengan penggunaan bahasa kasar, penghinaan, pelecehan verbal, provokasi, maupun ungkapan yang dapat menimbulkan ketidaknyamanan bagi pengguna lain. Dalam lingkungan digital, perilaku *toxic* sering muncul sebagai ekspresi emosi yang tidak terkontrol dan dapat memengaruhi kualitas interaksi antar pengguna. Penggunaan kata-kata seperti penghinaan, ejekan, maupun umpatan berpotensi memicu konflik dan mendorong terbentuknya lingkungan komunikasi yang tidak sehat di media sosial [5].

Salah satu bentuk konten *toxic* yang paling sering dijumpai adalah *hate speech* atau ujaran kebencian. *Hate speech* merupakan ungkapan yang mengandung penghinaan, hasutan, diskriminasi, atau permusuhan yang ditujukan kepada individu maupun kelompok tertentu berdasarkan atribut seperti ras, etnis, agama, jenis kelamin, maupun karakteristik lainnya. Keberadaan *hate speech* di media sosial semakin meningkat seiring dengan tingginya kebebasan pengguna dalam menyampaikan pendapat secara daring. Selain berpotensi merusak hubungan sosial, ujaran kebencian juga dapat memicu perilaku agresif serta menimbulkan dampak psikologis bagi korbannya [6].

Bentuk lain dari konten *toxic* adalah *cyberbullying*, yaitu tindakan agresif yang dilakukan melalui media digital dengan tujuan

merendahkan, mengintimidasi, mempermalukan, atau menyakiti individu lain secara berulang. *Cyberbullying* dapat terjadi melalui komentar, pesan pribadi, unggahan gambar, maupun penyebaran informasi yang bersifat merugikan korban. Berbeda dengan perundungan konvensional, *cyberbullying* dapat terjadi tanpa batas ruang dan waktu sehingga dampaknya cenderung lebih luas dan berlangsung lebih lama [7].

Dampak dari konten *toxic*, terutama *hate speech* dan *cyberbullying*, tidak hanya memengaruhi kenyamanan pengguna media sosial tetapi juga dapat berdampak pada kondisi psikologis korban. Beberapa penelitian menunjukkan bahwa korban *cyberbullying* memiliki risiko lebih tinggi mengalami stres, kecemasan, depresi, penurunan kepercayaan diri, isolasi sosial, hingga gangguan kesehatan mental lainnya. Selain itu, paparan konten *toxic* secara terus-menerus juga dapat menormalisasi perilaku agresif dan memperburuk kualitas komunikasi di ruang digital [6][7].

Oleh karena itu, diperlukan mekanisme moderasi konten yang mampu mendeteksi dan mengidentifikasi kata atau frasa yang mengandung unsur *toxic* secara otomatis. Deteksi dini terhadap konten *toxic* dapat membantu menciptakan lingkungan media sosial yang lebih aman, nyaman, dan kondusif bagi seluruh pengguna.

2.2. Spam pada Media Sosial

Spam merupakan pesan yang dikirim secara massal kepada banyak pengguna tanpa permintaan atau persetujuan dari penerima. Pesan spam umumnya berisi promosi produk, iklan, penawaran tertentu, tautan mencurigakan, maupun upaya penipuan yang bertujuan memperoleh keuntungan bagi pengirim. Seiring berkembangnya teknologi digital, *spam* tidak hanya ditemukan pada *email*, tetapi juga pada layanan pesan singkat (*SMS*), media sosial, aplikasi perpesanan, dan berbagai platform komunikasi daring lainnya [8].

Karakteristik utama pesan *spam* adalah dikirim dalam jumlah besar, tidak relevan dengan kebutuhan penerima, serta sering menggunakan kata atau frasa yang bersifat persuasif untuk menarik perhatian pengguna. Beberapa contoh pola yang umum ditemukan pada *spam* antara lain penawaran hadiah, promosi, diskon, tautan tertentu, maupun informasi yang mengarahkan pengguna untuk melakukan tindakan tertentu. Selain itu, *spam* sering memanfaatkan kata-kata yang berulang dan dirancang agar dapat menjangkau sebanyak mungkin pengguna dalam waktu singkat [9].

Keberadaan *spam* dapat menurunkan kualitas informasi yang diterima pengguna karena memenuhi ruang komunikasi dengan konten yang tidak diinginkan. Selain mengganggu kenyamanan pengguna, *spam* juga berpotensi menjadi sarana penyebaran penipuan, *phishing*, maupun *malware* yang dapat membahayakan keamanan data. Oleh karena itu, diperlukan mekanisme deteksi otomatis yang mampu mengidentifikasi pola-pola *spam* secara cepat dan akurat untuk menjaga kualitas komunikasi di lingkungan digital [10].

2.3. Finite Automata

Finite Automata (FA) merupakan model matematika yang digunakan untuk merepresentasikan sistem yang memiliki jumlah keadaan (*state*) terbatas serta mampu berpindah dari satu keadaan ke keadaan lainnya berdasarkan masukan tertentu. FA banyak digunakan untuk memodelkan alur kerja suatu sistem karena mampu menggambarkan proses pengambilan keputusan secara terstruktur. Dalam implementasinya, FA menerima masukan berupa simbol atau karakter dan menghasilkan keluaran berdasarkan aturan transisi yang telah ditentukan [11].

Secara formal, Finite Automata terdiri atas beberapa komponen utama, yaitu himpunan *state* (Q), himpunan simbol masukan atau *input alphabet* (Σ), fungsi transisi (δ), *initial state*, dan *final state*. Setiap simbol yang diterima akan menyebabkan perpindahan dari satu *state* ke *state* lainnya sesuai fungsi transisi yang telah didefinisikan. Berdasarkan karakteristik transisinya, FA dibedakan menjadi Deterministic Finite Automata (DFA) dan Non-Deterministic Finite Automata (NFA), di mana DFA hanya memiliki satu kemungkinan transisi untuk setiap masukan, sedangkan NFA dapat memiliki lebih dari satu kemungkinan transisi [12].

Dalam bidang pengolahan teks, Finite Automata banyak dimanfaatkan untuk proses *pattern matching* atau pencocokan pola. FA bekerja dengan membaca rangkaian karakter

secara berurutan dan mencocokkannya dengan pola yang telah didefinisikan melalui serangkaian *state* dan transisi. Pendekatan ini memungkinkan sistem mengenali kata, frasa, maupun pola tertentu secara efisien sehingga sering diterapkan pada proses pencarian kata kunci, validasi masukan, transliterasi teks, serta deteksi kata-kata tertentu dalam dokumen digital [13].

2.4. Algoritma Aho-Corasick

Algoritma Aho-Corasick merupakan algoritma pencocokan banyak pola (*multiple pattern matching*) yang dikembangkan untuk menemukan seluruh kemunculan beberapa pola kata sekaligus dalam sebuah teks. Berbeda dengan metode pencarian yang memproses setiap pola secara terpisah, algoritma ini menggabungkan seluruh pola ke dalam satu struktur pencarian sehingga proses pencocokan dapat dilakukan dalam satu kali pembacaan teks. Kemampuan tersebut *menjadikan* Aho-Corasick banyak digunakan pada sistem deteksi kata kunci, kamus digital, *intrusion detection system*, serta aplikasi moderasi teks yang membutuhkan pencarian banyak kata secara bersamaan [14].

Struktur utama yang digunakan oleh algoritma Aho-Corasick adalah *trie*, yaitu struktur data berbentuk pohon yang menyimpan sekumpulan pola berdasarkan awalan (*prefix*) yang sama. Setiap simpul (*node*) pada *trie* merepresentasikan karakter tertentu, sedangkan jalur dari akar hingga simpul akhir menunjukkan sebuah pola yang dicari. Penggunaan *trie* memungkinkan penyimpanan pola menjadi lebih efisien karena bagian awalan yang sama hanya disimpan satu kali sehingga dapat mengurangi redundansi data dan mempercepat proses pencarian [15].

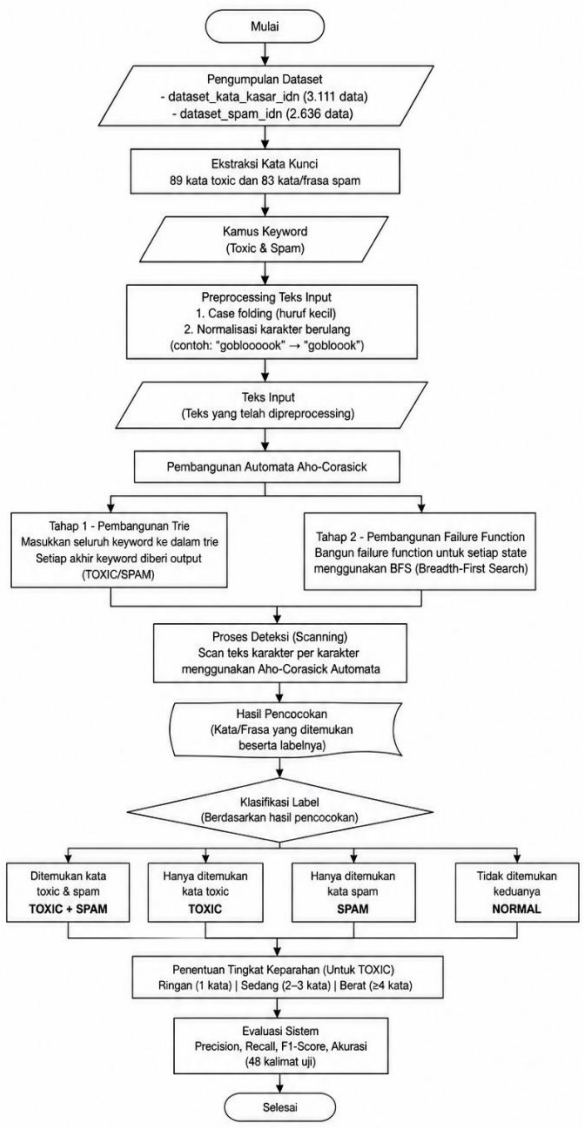
Untuk meningkatkan efisiensi pencocokan, algoritma ini memanfaatkan mekanisme *failure link*. Ketika terjadi ketidaksesuaian karakter saat proses pencarian, sistem tidak perlu kembali ke awal teks, melainkan langsung berpindah ke simpul lain yang merepresentasikan awalan terpanjang yang masih mungkin menghasilkan kecocokan. Mekanisme ini memungkinkan proses pencarian tetap berjalan secara berkelanjutan tanpa melakukan pemeriksaan ulang terhadap karakter yang telah diproses sebelumnya. Selain itu, setiap simpul juga dapat memiliki *output function* yang digunakan untuk menandai pola yang berhasil ditemukan [14][16].

Keunggulan utama algoritma Aho-Corasick terletak pada kompleksitas waktunya yang efisien. Proses pembangunan automata membutuhkan waktu yang sebanding dengan total panjang seluruh pola, sedangkan proses pencarian berjalan dalam kompleksitas O(n), dengan n merupakan panjang teks yang diperiksa. Kompleksitas tersebut relatif tidak dipengaruhi oleh jumlah pola yang dicari karena setiap karakter pada teks hanya diproses satu kali. Oleh karena itu, algoritma Aho-Corasick sangat sesuai digunakan pada aplikasi yang memerlukan pencarian cepat terhadap banyak kata, termasuk sistem deteksi kata *spam* dan kata berunsur toksik pada media digital.

3. METODE PENELITIAN

3.1. Rancangan Penelitian

Penelitian ini menggunakan pendekatan *rule-based text classification* dengan memanfaatkan Finite Automata dan algoritma Aho-Corasick untuk mendeteksi konten *toxic* dan *spam* berbahasa Indonesia. Tahapan penelitian meliputi pengumpulan dataset, ekstraksi kata kunci, preprocessing teks, pembangunan automata menggunakan algoritma Aho-Corasick, proses deteksi kata pada teks masukan, klasifikasi label, serta evaluasi performa sistem.



Gambar 1. Alur penelitian sistem deteksi *toxic* dan *spam* berbasis Finite Automata menggunakan algoritma Aho-Corasick.

Berdasarkan Gambar 1, penelitian diawali dengan pengumpulan dua dataset, yaitu *dataset_kata_kasar_idn* dan *dataset_spam_idn*. Dari kedua dataset tersebut dilakukan proses ekstraksi kata kunci sehingga diperoleh kamus kata *toxic* dan *spam* yang digunakan sebagai pola pencarian. Selanjutnya, teks masukan melalui tahap *preprocessing* berupa *case folding* dan normalisasi karakter berulang untuk mengurangi variasi penulisan.

Setelah proses *preprocessing*, sistem membangun automata menggunakan algoritma Aho-Corasick yang terdiri atas pembentukan struktur *trie* dan pembangunan *failure function*. Automata yang telah terbentuk kemudian digunakan untuk melakukan proses pencocokan pola (*pattern matching*) terhadap teks masukan. Hasil pencocokan digunakan untuk menentukan label keluaran berupa TOXIC, SPAM, TOXIC+SPAM, atau NORMAL. Selain itu, sistem juga melakukan klasifikasi tingkat keparahan konten *toxic* berdasarkan jumlah kata *toxic* yang terdeteksi. Tahap terakhir adalah evaluasi performa sistem menggunakan metrik *Precision*, *Recall*, *F1-Score*, dan akurasi.

3.2. Dataset dan Preprocessing

Penelitian ini menggunakan dua dataset berbahasa Indonesia, yaitu *dataset_kata_kasar_idn* yang berisi 3.111 data berlabel emosi dan *dataset_spam_idn* yang berisi 2.636 data pesan spam dan non-spam. Pada *dataset_kata_kasar_idn*, data dengan label *Disgust* dan *Anger* dikategorikan sebagai *TOXIC*, sedangkan data berlabel *Neutral* dikategorikan sebagai *NORMAL*. Pada *dataset_spam_idn*, data berlabel *spam* dikategorikan sebagai *SPAM* dan data berlabel *ham* dikategorikan sebagai *NORMAL*.

Dari kedua dataset tersebut dilakukan proses ekstraksi kata kunci sehingga diperoleh 89 kata *toxic* dan 83 kata atau frasa *spam* yang digunakan sebagai kamus utama sistem deteksi.

Sebelum proses deteksi dilakukan, teks masukan terlebih dahulu melalui tahap preprocessing untuk mengurangi variasi penulisan yang dapat memengaruhi hasil pencocokan pola. Tahap pertama adalah *case folding*, yaitu mengubah seluruh karakter menjadi huruf kecil. Tahap kedua adalah normalisasi karakter berulang dengan membatasi jumlah kemunculan karakter yang sama secara berturut-turut. Sebagai contoh, kata "gobloooook" akan dinormalisasi menjadi "goblok".

3.3. Pembangunan Automata Aho-Corasick

Pembangunan automata dilakukan menggunakan algoritma Aho-Corasick yang terdiri atas dua tahap utama, yaitu pembentukan *trie* dan pembangunan *failure function*.

Pada tahap pertama, seluruh kata kunci *toxic* dan *spam* dimasukkan ke dalam struktur *trie*. Setiap karakter direpresentasikan sebagai sebuah *state* dan setiap akhir kata diberi penanda keluaran (*output*) sesuai label kategorinya, yaitu TOXIC atau SPAM.

Tahap kedua adalah pembangunan *failure function* menggunakan algoritma Breadth-First Search (BFS). Fungsi ini digunakan untuk menentukan perpindahan *state* ketika terjadi ketidaksesuaian karakter selama proses pencarian. Dengan adanya *failure function*, automata tidak perlu kembali ke *state* awal sehingga proses pencarian menjadi lebih efisien.

3.4. Proses Deteksi dan Klasifikasi

Setelah automata terbentuk, teks masukan yang telah dipreprocessing dipindai karakter demi karakter menggunakan algoritma Aho-Corasick. Setiap kecocokan terhadap kata atau frasa yang terdapat pada kamus akan dicatat beserta label kategorinya.

Kompleksitas proses pencarian mengikuti karakteristik algoritma Aho-Corasick yaitu:

$$O(n + m + z)$$

dengan n menyatakan panjang teks masukan, m menyatakan total panjang seluruh kata kunci, dan z menyatakan jumlah kecocokan yang ditemukan.

Berdasarkan hasil pencocokan, sistem memberikan label keluaran sesuai kondisi yang terdeteksi. Jika ditemukan kata *toxic* dan *spam* secara bersamaan maka sistem menghasilkan label TOXIC+SPAM. Jika hanya ditemukan kata *toxic*, sistem menghasilkan label TOXIC. Jika hanya ditemukan kata *spam*, sistem menghasilkan label SPAM. Apabila tidak ditemukan keduanya maka sistem menghasilkan label NORMAL.

Selain klasifikasi utama, sistem juga menentukan tingkat keparahan konten *toxic* berdasarkan jumlah kata *toxic* unik yang terdeteksi. Tingkat keparahan dibagi menjadi tiga kategori, yaitu Ringan untuk satu kata *toxic*, Sedang untuk dua hingga tiga kata *toxic*, dan Berat untuk empat kata *toxic* atau lebih.

3.5. Evaluasi Sistem

Evaluasi dilakukan menggunakan 48 kalimat uji yang terdiri atas empat kategori, yaitu kalimat TOXIC, SPAM, NORMAL, dan TOXIC+SPAM. Hasil klasifikasi kemudian dibandingkan dengan label aktual untuk memperoleh nilai performa sistem.

Metrik evaluasi yang digunakan meliputi *Precision*, *Recall*, *F1-Score*, dan Akurasi. Nilai *Precision* digunakan untuk mengukur ketepatan prediksi sistem terhadap suatu kelas, *Recall* digunakan untuk mengukur kemampuan sistem dalam menemukan seluruh data yang relevan, sedangkan *F1-Score* digunakan untuk menyeimbangkan nilai *Precision* dan *Recall*. Akurasi digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data uji.

4. HASIL DAN PEMBAHASAN

4.1 Implementasi Sistem

Sistem dibangun menggunakan algoritma Aho-Corasick berbasis Finite Automata yang diimplementasikan dalam bahasa pemrograman Python. Sebelum proses pencocokan pola (*pattern matching*) dimulai, sistem menerima input teks bahasa Indonesia dan melakukan proses *preprocessing* dengan normalisasi karakter berulang, misalnya "goblooook" dinormalisasi menjadi "goblok".

Dataset yang digunakan berasal dari dua sumber: *dataset_kata_kasar_idn*, yang mengandung 3.111 data berlabel emosi, dan *dataset_spam_idn*, yang mengandung 2.636 data. Dari masing-masing sumber, 89 kata berbahaya dan 83 kata/frasa spam diekstrak, yang kemudian dimasukkan ke dalam struktur *trie* pada automata. Antarmuka sistem dapat diakses secara lokal melalui *browser* berkat *framework Streamlit.*

4.2 Hasil Pengujian Sistem

Pengujian dilakukan terhadap 48 kalimat uji yang terbagi menjadi empat kategori, yaitu 15 kalimat TOXIC, 15 kalimat SPAM, 14 kalimat NORMAL, dan 4 kalimat TOXIC+SPAM. Karena keterbatasan ruang artikel, Tabel 1 hanya menampilkan sebagian hasil pengujian yang mewakili setiap kategori data.

No.	Kalimat Uji	Label Aktual	Prediksi Sistem
1.	"dasar anjing kamu!"	TOXIC	TOXIC
2.	"goblok banget sih orang itu"	TOXIC	TOXIC
3.	"bangsat lo ngomong apa"	TOXIC	TOXIC
4.	"idiot beneran dah"	TOXIC	TOXIC
5.	"mampus aja lo"	TOXIC	TOXIC

6.	"penipu! bohong terus!"	TOXIC	TOXIC
7.	"kurang ajar banget"	TOXIC	NORMAL
8.	"Gratis! Klik link sekarang juga!"	SPAM	SPAM
9.	"Pinjaman cepat cair hari ini tanpa jaminan"	SPAM	SPAM
10.	"Selamat anda menang, klaim hadiah sekarang!"	SPAM	SPAM
11.	"Daftar slot online dan raih jackpot besar!"	SPAM	SPAM
12.	"Obat herbal terbukti ampuh 100%!"	SPAM	SPAM
13.	"Modal kecil untung besar, terbukti!"	SPAM	SPAM
14.	"Halo, apa kabar hari ini?"	NORMAL	NORMAL
15.	"Terima kasih sudah membantu saya"	NORMAL	NORMAL
16.	"Indonesia adalah negara yang indah"	NORMAL	TOXIC
17.	"Itu bukan salah siapa-siapa"	NORMAL	TOXIC
18.	"Goblok! Mau pinjaman cepat gratis?"	TOXIC +SPAM	TOXIC+SPAM
19.	"Anjing! Klik link sekarang idiot!"	TOXIC +SPAM	TOXIC+SPAM
20.	"Besok ada rapat jam 9 pagi"	NORMAL	NORMAL

Tabel 1. Contoh Hasil Pengujian Sistem Deteksi (20 dari 48 kalimat uji)

Dari 48 kalimat uji, sistem berhasil mengklasifikasikan 44 kalimat dengan benar sehingga diperoleh akurasi keseluruhan sebesar 91,7%.

4.3 Evaluasi Performa

Evaluasi performa sistem dilakukan menggunakan tiga metrik utama yaitu *precision, recall,* dan *F1-score* untuk masing-masing kelas. Hasil evaluasi disajikan pada Tabel 2.

Kelas	Precision	Recall	F1-Score
TOXIC	0,842	0,941	0,889
SPAM	1,000	1,000	1,000
NORMAL	0,933	0,875	0,903
Rata-Rata	0,925	0,939	0,931

Tabel 2. Evaluasi Performa

Berdasarkan Tabel 2, kelas SPAM memperoleh nilai sempurna pada ketiga metrik dengan *precision, recall,* dan *F1-score* masing-masing bernilai 1,000. Hal ini menunjukkan bahwa kata-kata *spam* dalam bahasa Indonesia memiliki pola yang sangat khas dan spesifik sehingga automata dapat mengenalinya secara konsisten tanpa kesalahan. Kelas TOXIC memperoleh nilai *recall* tertinggi sebesar 0,941 yang berarti sistem sangat sensitif dalam mendeteksi kata toxic, namun terdapat sedikit *false positive* dengan nilai *precision* 0,842. Kelas NORMAL memperoleh *F1-score* sebesar 0,903 yang tergolong baik. Nilai tersebut menunjukkan bahwa sistem mampu membedakan konten normal dari konten *toxic* maupun *spam* dengan tingkat kesalahan yang relatif rendah.

4.4 Klasifikasi Tingkat Keparahan Toxic

Selain mengklasifikasikan label utama, sistem juga mampu menentukan tingkat keparahan konten *toxic* berdasarkan jumlah kata *toxic* yang terdeteksi dalam satu kalimat. Klasifikasi tingkat keparahan disajikan pada Tabel 3.

Tingkat	Jumlah Kata Toxic	Contoh Kalimat
RINGAN	1 kata	"tolol banget sih"
SEDANG	2-3 kata	"goblok banget, anjing kamu!"
BERAT	≥4 kata	"bajingan keparat brengsek goblok!"

Tabel 3. Klasifikasi Tingkat Keparahan Toxic

Klasifikasi tingkat keparahan ini memberikan informasi tambahan yang berguna bagi moderator konten dalam menentukan prioritas penanganan. Konten dengan tingkat BERAT dapat diprioritaskan untuk ditangani lebih cepat dibandingkan konten dengan tingkat RINGAN.

4.5 Analisis Kesalahan

Dari 48 kalimat uji, empat di antaranya salah dikategorikan. Kesalahan pertama terjadi pada kalimat "kurang ajar banget", yang seharusnya berlabel TOXIC tetapi diprediksi NORMAL karena frasa tersebut tidak terdapat dalam kamus keyword sistem. Selain itu, ditemukan kesalahan berupa *false positive* pada kalimat seperti "Indonesia adalah negara yang indah" dan "Itu bukan salah siapa-siapa", karena sistem mendeteksi substring tertentu yang memiliki kemiripan dengan keyword *toxic*.

Keterbatasan ini muncul sebagai akibat dari metode penggabungan string tepat yang digunakan oleh Aho-Corasick, yang tidak mempertimbangkan konteks atau batas kata. Untuk memastikan bahwa pencocokan hanya terjadi pada kata utuh, bukan bagian dari kata lain, mekanisme pengenalan batas kata dapat ditambahkan.

5. KESIMPULAN

a. Sistem deteksi konten *toxic* dan *spam* berhasil diimplementasikan menggunakan algoritma Aho-Corasick berbasis Finite Automata dalam bahasa pemrograman Python, dengan antarmuka berbasis *framework Streamlit* yang dapat diakses melalui browser secara lokal.
b. Akurasi sistem mencapai 91,7%, di mana dari 48 kalimat uji yang mencakup kategori TOXIC, SPAM, NORMAL, dan TOXIC+SPAM, sebanyak 44 kalimat berhasil diklasifikasikan dengan benar.
c. Performa terbaik dicapai pada kelas SPAM dengan nilai *precision, recall,* dan *F1-score* sempurna (1,000), menunjukkan bahwa pola kata spam berbahasa Indonesia sangat khas dan konsisten dikenali oleh automata.
d. Deteksi konten *Toxic* menunjukkan *recall* yang tinggi (0,941), artinya sistem sangat sensitif dalam menangkap kata *toxic*, meskipun terdapat sedikit *false positive* (*precision* 0,842) akibat pencocokan *substring* yang tidak mempertimbangkan batas kata.
e. Sistem mampu mengklasifikasikan tingkat keparahan konten *toxic* ke dalam tiga level, yaitu Ringan (1 kata *toxic*), Sedang (2–3 kata *toxic*), dan Berat (≥4 kata *toxic*), yang berguna bagi moderator konten dalam menentukan prioritas penanganan.
f. Keterbatasan utama sistem terletak pada mekanisme *exact string matching* milik Aho-Corasick yang tidak memperhatikan konteks atau batas kata, sehingga kata seperti "Indonesia" atau "sia-sia" berpotensi terdeteksi sebagai *toxic* karena mengandung *substring* dari kamus kata berbahaya.
g. Pengembangan selanjutnya dapat difokuskan pada penambahan mekanisme pengenalan batas kata (*word boundary detection*) untuk mengurangi *false positive*, serta perluasan dataset agar sistem lebih *robust* terhadap variasi bahasa gaul dan konteks kalimat yang lebih kompleks.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada dosen pembimbing, Program Studi Ilmu Komputer Universitas Negeri Medan, serta seluruh pihak yang telah memberikan arahan, dukungan, dan bantuan selama proses penelitian dan penyusunan artikel ini.

DAFTAR PUSTAKA

[1] R. W. Abie dan S. Rosmilawati, “PERILAKU TOXIC DALAM KOMUNIKASI VIRTUAL DI GAME ONLINE MOBILE LEGENDS: BANG BANG PADA MAHASISWA FAKULTAS ILMU SOSIAL DAN ILMU

POLITIK DI UNIVERSITAS MUHAMMADIYAH PALANGKARAYA," *Restorica: Jurnal Ilmiah Ilmu Administrasi Negara dan Ilmu Komunikasi*, vol. 9, no. 1, hlm. 44–48, Apr 2023, doi: 10.33084/restorica.v9i1.

[2] F. Yunita, R. D. Handayani, dan A. Mahmudi, "INTERPRETASI PERIBAHASA INDONESIA DENGAN METODE PATTERN MATCHING BERBASIS KATA KUNCI," *Jurnal TRANSFORMASI*, vol. 21, no. 2, hlm. 66–73, 2025, doi: https://doi.org/10.56357/jt.v21i2.467.

[3] S. Alshattnawi, A. Shatnawi, A. M. R. AlSobeh, dan A. A. Magableh, "Beyond Word-Based Model Embeddings: Contextualized Representations for Enhanced Social Media Spam Detection," *Applied Sciences (Switzerland)*, vol. 14, no. 2254, hlm. 1–25, Mar 2024, doi: 10.3390/app14062254.

[4] B. L. Sandeep, G. M. Siddesh, dan E. Naresh, "Suspicious behaviour detection in multilayer social networks using PF-KMA and SS-GAE techniques," *Soc. Netw. Anal. Min.*, vol. 14, no. 112, hlm. 1–15, Des 2024, doi: 10.1007/s13278-024-01265-2.

[5] M. B. S. Winardi dan K. Sinduwiatmo, "Representation of Toxic Messages in Windah Basudara Youtube Content," *Indonesian Journal of Cultural and Community Development*, vol. 15, no. 1, hlm. 6–12, Mar 2024, doi: 10.21070/ijccd.v15i1.1186.

[6] R. Abdillah *dkk.*, "STUDI PSIKOLOGI SIBER TENTANG DAMPAK HATE SPEECH BAGI PENGGUNA MEDIA SOSIAL," *SIBATIK JOURNAL*, vol. 2, no. 11, hlm. 3459–3472, 2023, doi: 10.54443/sibatik.v2i11.1478.

[7] S. A. Ni'mah, "Pengaruh Cyberbullying pada Kesehatan Mental Remaja," *Prosiding Seminar Nasional Bahasa, Sastra dan Budaya (SEBAYA)*, vol. 3, no. 3, hlm. 329–338, Jun 2023.

[8] M. B. M. Amin *dkk.*, "Deteksi Spam Berbahasa Indonesia Berbasis Teks Menggunakan Model Bert," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 6, hlm. 1291–1302, Des 2024, doi: 10.25126/jtiik.2024118121.

[9] F. A. Firmansyah, U. Enri, dan I. Maulana, "PENERAPAN ALGORITMA NAIVE BAYES DENGAN CHI-SQUARE UNTUK KLASIFIKASI SPAM EMAIL BERBASIS KATA DAN FREKUENSI," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 1, hlm. 78–87, Jan 2025, doi: 10.23960/jitet.v13i1.5506.

[10] E. S. Panggabean, C. I. Angkat, D. Sartika, D. Riswan, A. P. Utama, dan M. Iqbal, "Memahami Spam Terhadap Digitalisasi Masyarakat Desa Perkebunan Sei Balai Kecamatan Sei Balai Kabupaten Batubara," *Desember*, vol. 2, no. 2, hlm. 151–159, 2023, doi: https://doi.org/10.70340/japamas.v2i2.90.

[11] M. Ernawati, W. Gata, E. H. Hermaliani, L. Kurniawati, dan S. Rahayu, "IMPLEMENTASI KONSEP FINITE STATE AUTOMATA PADA DESAIN GAME EDUKASI JENIS HEWAN," *Technologia*, vol. 13, no. 1, hlm. 65–71, Jan 2022, doi: http://dx.doi.org/10.31602/tji.v13i1.6268.

[12] A. R. Pangestu, A. A.-G. Shabir, F. Susanto, dan T. Setiadi, "PENERAPAN FINITE STATE AUTOMATA PADA PROSES PENDAFTARAN PRAKTIKUM JURUSAN INFORMATIKA UNIVERSITAS AHMAD DAHLAN," *Jurnal Ilmiah Ilmu Komputer*, vol. 10, no. 1, hlm. 14–17, Apr 2024, doi: https://doi.org/10.35329/jiik.v10i1.293.

[13] A. W. Mahastama, "Model Berbasis Aturan Untuk Transliterasi Bahasa Jawa Dengan Aksara Latin Ke Aksara Jawa," *Jurnal Buana Informatika*, vol. 13, no. 2, hlm. 146–154, Okt 2022, doi: 10.24002/jbi.v13i02.6526.

[14] P. A. Gagniuc, I. B. Păvăloiu, dan M. I. Dascălu, "The Aho-Corasick Paradigm in Modern Antivirus Engines: A Cornerstone of Signature-Based Malware Detection," *Algorithms*, vol. 18, no. 12, hlm. 1–18, Des 2025, doi: 10.3390/a18120742.

[15] M. Fayez, A. Abbas, H. Khaled, dan S. Ghoniemy, "Enhanced Aho-Corasick Algorithm for Network Intrusion Detection Systems," *International Journal of Intelligent Computing and Information Sciences*, vol. 24, no. 3, hlm. 83–92, Sep 2024, doi: 10.21608/ijicis.2024.315432.1349.

[16] L. Wikarsa, T. Suwanto, dan C. H. Loha, "Development Of A Bilingual Dictionary Of Sahu Tala'i – Indonesia Using The Aho-Corasick Algorithm," *Jurnal Pekommas*, vol. 9, no. 1, hlm. 81–91, Jun 2024, doi: 10.56873/jpkm.v9i1.5281.