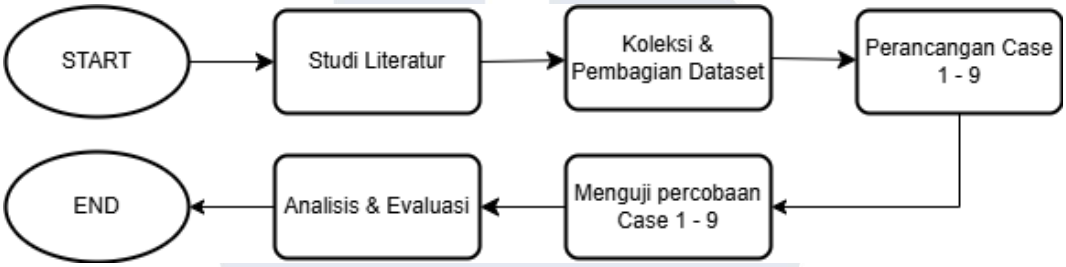
BAB III

ANALISIS DAN PERANCANGAN SISTEM

3.1 Metode Penelitian

Dalam penelitian ini terdapat 6 tahap utama pada gambar 3.1, yaitu studi literatur, koleksi dan pembagian dataset, perancangan skenario atau case 1 – 9, menguji percobaan case 1 – 9, dan tahap terakhir melakukan analisis dan evaluasi tiap case.



Gambar 3.1 Flowchart metode penelitian.

3.2 Studi Literatur

Penelitian dimulai dengan melakukan pencarian, pengumpulan, dan pembuatan sintesis penelitian terdahulu yang berkaitan dengan deteksi teks menggunakan machine learning, judi online, dan algoritma yang sering digunakan dalam deteksi teks seperti BERT dan *Fuzzy Matching*. Hal ini dilakukan untuk mengetahui pada kondisi dan metode seperti apa model mampu memberikan performa terbaik. Studi literatur dilakukan dengan sumber utama jurnal ilmiah dan juga review artikel pada situs web untuk mencari pemahaman lebih rinci mengenai topik penelitian dari berbagai penggiat machine learning.

3.3 Koleksi & Pembagian Dataset

Penelitian ini menggunakan 3 tipe dataset yang bersumber dari 2 dataset yang diambil dari situs kaggle.com. Dataset pertama berjudul "youtube_chat_jogja_clean.csv" dan dataset kedua berjudul "judol_tiktok.csv". Keduanya berisi komentar dari platform media sosial YouTube dan TikTok yang banyak mengandung unsur terkait judi online [17][18]. Kedua dataset tersebut kemudian diklasifikasikan ke dalam 3 tipe untuk keperluan analisis.

3.3.1 Dataset Tipe 1

Dataset Tipe 1 adalah dataset “youtube_chat_jogja_clean.csv” yang merupakan bentuk original dari dataset dengan jumlah baris 6350. Ciri dari dataset ini adalah penulisan normal dengan sedikit gabungan angka dan huruf, contohnya teks yang mengandung unsur judi online seperti “depo” menjadi “d3p0”. Gambaran dataset tipe 1 dapat dilihat pada gambar 3.2. Kolom yang akan digunakan adalah “cleaned_message”

	datetime	author_name	message	cleaned_message
0	2024-10-07 08:32:03	KUSUMA	assalamu'alaikum..	assalamualaikum
1	2024-10-07 08:58:42	Tata PaNda	wa'alaikumussalam	waalaikumussalam
2	2024-10-07 09:20:29	Nimas putri Paranata	udah lewat 22 menit ni	udah lewat 22 menit ni
3	2024-10-07 09:30:23	proboyuwono djogdja	16:30 wib	1630 wib
4	2024-10-07 09:34:59	Vian Noorcha Putra	Tribun Tv Mana Ini Kenapa Acaranya Belum Dimul...	tribun tv mana ini kenapa acaranya belum dimul...
...	...	...	...	...
6345	2024-10-07 14:22:01	may	Met ultah kotakuu,love you	met ultah kotakuulove you
6346	2024-10-07 14:22:01	Johnnie Boyles	**d3P0 100 jd 2jt** buruan gas **GARANSI 100**% di :fi...	d3p0 100 jd 2jt buruan gas garansi 100% di met...
6347	2024-10-07 14:22:02	JURAGAN99	:hand-pink-waving: NATUNATOTO:hand-pink-waving...	handpinkwaving natunatotohandpinkwavingpasti g...
6348	2024-10-07 14:22:06	Nunung Kusmawati	menyala jogjaku:fire::fire:	menyala jogjaku
6349	2024-10-07 14:22:15	Johnnie Boyles	**d3P0 100 jd 2jt** buruan gas **GARANSI 100**% di :fi...	d3p0 100 jd 2jt buruan gas garansi 100% di met...

Gambar 3.2 Gambaran dataset tipe 1.

3.3.2 Dataset Tipe 2

Dataset tipe 2 dibangun dari dataset "youtube_chat_jogja_clean.csv". Perbedaannya terletak pada karakteristik teks dimana ada gabungan angka dan huruf lebih banyak. Sebagai contoh, teks yang awalnya "depo 100 jd 2jt buruan gas garansi 100%" diubah menjadi "d3epo 100 j/d3 2jt buru_a0n gas gara_nsi 100%". Modifikasi ini dilakukan melalui teknik augmentasi, yaitu proses menghasilkan data baru secara artifisial dengan cara mengubah struktur atau bentuk teks. Ilustrasi dari dataset tipe 2 dapat dilihat pada Gambar 3.3. Dalam dataset ini, kolom "cleaned_message" merepresentasikan bentuk teks asli yang telah dibersihkan, sedangkan kolom "augmented_message" menunjukkan versi teks yang telah dimodifikasi. Dataset tipe 2 dikembangkan dari tipe 1 untuk menguji performa model dalam menghadapi data dengan bentuk dan variasi yang lebih kompleks.

	author_name	cleaned_message	augmented_message
0	KUSUMA	assalamualaikum	assalamualai!kum
1	Tata PaNda	waalaikumussalam	wa@la5ikumuss4l@m
2	Nimas putri Paranata	udah lewat 22 menit ni	ud@h le;wat 22 menit <ni
3	proboyuwono djogdja	1630 wib	1630 wib
4	Vian Noorcha Putra	tribun tv mana ini kenapa acaranya belum dimul...	tribun 78v mana in1 ken'4pa ac@ra,_nya[belum ...
...	...	...	...
6345	may	met ultah kotakuulove you	m3t you ko7akuulove- u@ltah
6346	Johnnie Boyles	depo 100 jd 2jt buruan gas garansi 100% di met...	jd met~eorw!n b%uruan 91+00 gas topup d1 2j7 j...
6347	JURAGAN99	handpinkwaving natunatotohandpinkwavingpasti g...	h@ndp!-nkwav1ng /natunatotohand=pinkwa.vin9p@s...
6348	Nunung Kusmawati	menyala jogjaku	meny{ala j0gjaku
6349	Johnnie Boyles	depo 100 jd 2jt buruan gas garansi 100% di met...	d3epo 100 j/d3 2jt buru_a0n gas gara_nsi 100% ...

Gambar 3.3 Gambaran dataset tipe 2.

3.3.3 Dataset Tipe 3

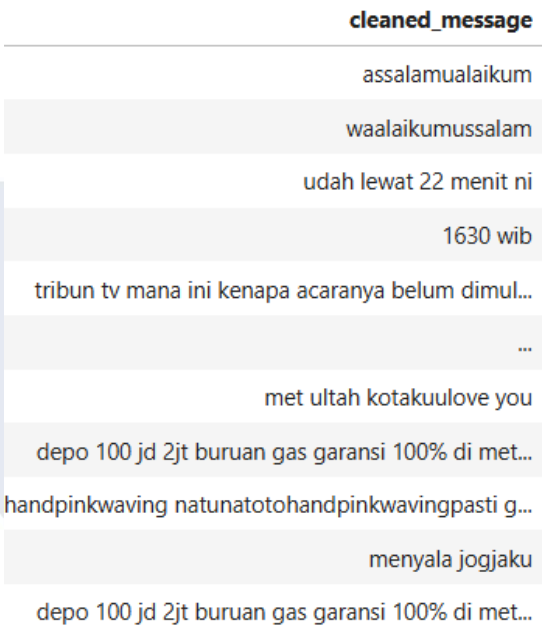
Dataset tipe 3 merupakan dataset "judol_tiktok.csv" yang berbeda dari 2 tipe sebelumnya. Dataset ini terdiri dari 99.065 baris dan berisi komentar yang masih mengandung unsur judi online, namun tidak banyak mengandung teks gabungan antara huruf dan angka seperti tipe 1 dan 2. Gambaran dataset tipe 3 dapat dilihat pada Gambar 3.4. Kolom yang digunakan dalam dataset ini adalah "text", dan tidak dilakukan modifikasi apa pun terhadap isi teksnya. Sama seperti dataset tipe 2, dataset tipe 3 digunakan untuk menguji performa model dalam kondisi data yang bervariasi.

	aweme_id	create_time	text
0	7376590288212479237	1717846551	Uda 9 bulanan main,,,\n-+10jt,,\nbelum bisa st...
1	7376590288212479237	1717546502	bner saya juga ngerasain
2	7376590288212479237	1717649591	untan
3	7376590288212479237	1717842605	aku sudah mengiklaskan 😊
4	7376590288212479237	1717786502	entah sampai kapan..😭
...	...	...	...
99060	7282359034567625990	1696592050	judi sepak bola pling berbahaya bro!!!!masyara...
99061	7282359034567625990	1695666989	pemerintahnya lg fokus batasin tiktokshop pak,...
99062	7282359034567625990	1695755353	Uang usaha saya yang harusnya diputar rekan bi...
99063	7282359034567625990	1696589832	saya deposit 50rb...dua kli sceter hsilnya tai...
99064	7282359034567625990	1695731301	konon ada artis lg ngurus legalkan judi online.

Gambar 3.4 Gambaran dataset tipe 3.

3.3.4 Dataset Training

Dataset tipe 1 – 3 adalah dataset yang digunakan untuk testing model. Sementara untuk training adalah dataset tipe 1 yang telah melewati tahap pre-processing menjadi lebih rapi dan bersih tanpa ada gabungan angka dan huruf atau kesalahan penulisan, seperti pada gambar 3.5.



Gambar 3.5 Dataset Training.

3.4 Perancangan Skenario Pengujian

Dalam penelitian ini terdapat 9 case atau skenario percobaan. Masing - masing skenario memiliki perbedaan pada tujuan, metode, dan dataset yang digunakan untuk training dan testing. Tabel 3.1 menunjukan perbedaan dari masing-masing skenario.

Tabel 3.1. Perbedaan Metode, Labelling, dan Dataset Skenario 1 – 9.

Skenario	Metode	Labelling Dataset	Dataset	Tujuan
1	IndoBERT Training	Manual Else If	Tipe 1	Menguji Performa IndoBERT Secara Individu
2			Tipe 2	
3			Tipe 3	

4	Fuzzy Matching	Fuzzy Matching Percentage	Tipe 1	Menguji Performa Fuzzy Matching Secara Individu
5			Tipe 2	
6			Tipe 3	
7	IndoBERT + Fuzzy Matching	Fuzzy Matching Percentage	Tipe 1	Menguji Performa Gabungan IndoBERT & Fuzzy Matching
8			Tipe 2	
9			Tipe 3	

Skenario 1 - 3 berfokus pada penggunaan metode IndoBERT melalui proses training dan pelabelan dataset secara manual. Pelabelan manual dilakukan menggunakan logika else-if, yaitu dengan mencocokkan setiap baris data langsung pada daftar kata kunci dengan kemiripan 100%. Kata kunci diperoleh dari dataset tipe 1 melalui frequency analysis untuk mengidentifikasi kata-kata yang sering muncul. Hal ini didasarkan pada ciri khas promosi judi online, di mana pelaku biasanya melakukan spam secara berulang dengan komentar yang serupa. Setelah model dilatih, proses testing dilakukan pada dataset tipe 1 hingga 3 guna mengevaluasi kemampuan model dalam mendeteksi konten judi online dalam berbagai kondisi data yang bervariasi.

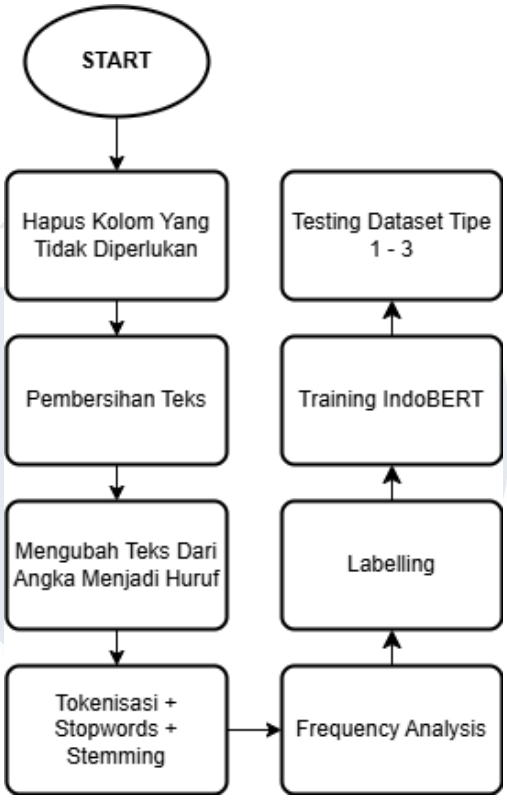
Skenario 4 - 6 berfokus pada penggunaan metode Fuzzy Matching. Pada skenario-skenario ini tidak dilakukan proses training, karena setelah melakukan frequency analysis, proses langsung dilanjutkan dengan testing melalui pencocokan kata kunci pada dataset tipe 1 hingga 3. Secara garis besar, skenario 4 - 6 dianggap selesai ketika proses labelling terhadap dataset telah selesai. Proses labelling dilakukan dengan metode fuzzy matching, yaitu mencocokkan kata-kata yang memiliki tingkat kemiripan minimal 80%.

Skenario 7 - 9 menggabungkan metode Fuzzy Matching dan metode training IndoBERT. Setelah dilakukan frequency analysis, proses pelabelan diubah menggunakan Fuzzy Matching dengan tetap berdasarkan kata kunci. Hasil pelabelan kemudian digunakan untuk melatih model IndoBERT, yang selanjutnya

diuji pada dataset tipe 1 hingga 3 guna mengevaluasi performa deteksi dalam berbagai kondisi data.

3.5 Tahapan Pengujian Pada Skenario 1 – 9

3.5.1 Skenario 1 – 3



Gambar 3.6 Flowchart Skenario 1 - 3.

cleaned_message
assalamualaikum
waalaikumussalam
udah lewat 22 menit ni
1630 wib
tribun tv mana ini kenapa acaranya belum dimul...
...
met ultah kotakuulove you
depo 100 jd 2jt buruan gas garansi 100% di met...
ıandpinkwaving natunatotohandpinkwavingpasti g...
menyala jogjaku
depo 100 jd 2jt buruan gas garansi 100% di met...

Gambar 3.7 Kolom cleaned_message yang dibutuhkan.

Gambar 3.6 menggambarkan alur kerja Skenario 1 hingga 3. Proses diawali dengan penghapusan kolom-kolom yang tidak relevan dari dataset, dan hanya menggunakan kolom "cleaned_message" sebagaimana ditampilkan pada Gambar 3.7. Setelah itu, dilakukan tahap pre-processing yang terdiri dari beberapa langkah berikut:

- Cleaning Text: Teks dibersihkan dengan memastikan encoding dalam format UTF-8, mengonversi emoji menjadi teks, menghapus format khusus seperti bold, menghilangkan spasi berlebih, dan mengubah seluruh teks menjadi huruf kecil (lowercase).
- Leetspeak: Mengonversi angka atau simbol menjadi huruf agar kata-kata tidak baku tetap dikenali. Misalnya, kata “g4c0r” diubah menjadi “gacor”.
- Tokenisasi: Teks dipecah menjadi unit-unit kecil (token) untuk memudahkan proses oleh model. Contohnya, kalimat "Promo deposit 100k langsung cair!" akan diubah menjadi token: [“Promo”, “deposit”, “100k”, “langsung”, “cair”].
- Stopwords: Menghapus kata-kata yang tidak memiliki makna penting dalam konteks analisis, seperti “dan”, “di”, “ke”, dan “yang”, agar model dapat lebih fokus pada kata-kata utama.
- Stemming: Mengubah kata ke bentuk dasarnya dengan cara menghapus imbuhan seperti awalan, akhiran, dan sisipan. Contohnya kata "mendepositkan" menjadi “deposit”.

Setelah melalui tahapan pre-processing, dilakukan frequency analysis untuk mengidentifikasi kata kunci yang relevan dengan indikasi konten judi online. Mengingat karakteristik utama promosi judi online yang cenderung melakukan spam, kata kunci dikelompokkan ke dalam tiga kategori: satu kata (unigram), dua kata (bigram), dan tiga kata (trigram), guna meningkatkan akurasi model dalam memahami konteks

kalimat secara menyeluruh. Hasil kata kunci kemudian disaring secara manual untuk memastikan kesesuaian dengan konteks judi online.

Selanjutnya, dilakukan proses labelling terhadap dataset training menggunakan logika else-if, dengan mencocokkan setiap baris teks terhadap daftar kata kunci yang telah diperoleh. Setelah proses pelabelan selesai, model IndoBERT dilatih selama 2 epoch. Model yang telah dilatih kemudian diuji pada dataset tipe 1 hingga 3 untuk mengevaluasi kemampuannya dalam mendeteksi konten judi online secara kontekstual dan membedakannya dari teks yang tidak terkait.

3.5.2 Skenario 4 - 6

Gambar 3.8 Flowchart Skenario 4 - 6.

Gambar 3.8 menggambarkan alur kerja Skenario 4 hingga 6. Secara umum, alurnya serupa dengan Skenario 1 hingga 3 hingga tahap frequency analysis. Perbedaan terletak saat setelah kata kunci diperoleh, di mana kata kunci tersebut langsung dicocokkan dengan dataset tipe 1 hingga 3 dengan metode Fuzzy Matching. Dalam skenario ini, teks dianggap terindikasi sebagai judi online apabila memiliki tingkat

kemiripan atau match score minimal 80%, sehingga threshold ditetapkan pada nilai 80. Tidak terdapat proses pelatihan model, melainkan langsung dilakukan pengujian untuk menilai kemampuan sistem dalam membedakan teks yang mengandung unsur judi online. Skenario ini juga didasarkan pada hipotesis awal bahwa Fuzzy Matching tidak mampu memahami konteks, tetapi dapat menjadi mekanisme awal untuk menangani variasi ejaan kata.

3.5.3 Skenario 7 - 9

Gambar 3.9 Flowchart Skenario 7 - 9.

Gambar 3.9 menggambarkan alur kerja Skenario 7 hingga 9, yang merupakan pengembangan dari Skenario 1 hingga 6. Pada skenario ini, metode IndoBERT dan Fuzzy Matching digabungkan untuk mengoptimalkan deteksi konten judi online, memahami konteks data, serta membedakan secara akurat antara teks yang mengandung unsur judi dan yang tidak.

Proses awal mengikuti tahapan yang serupa dengan skenario sebelumnya, dimulai dari pre-processing hingga frequency analysis. Selanjutnya, pelabelan pada data latih dilakukan menggunakan Fuzzy

Matching dengan kemiripan sebesar 80% berdasarkan kata kunci yang telah diidentifikasi. Hasil pelabelan ini digunakan untuk melatih model IndoBERT, yang kemudian diuji pada dataset tipe 1 hingga 3 untuk mengevaluasi performa dalam berbagai kondisi data.

3.6 Metriks Evaluasi

Setelah seluruh Skenario 1 hingga 9 dijalankan, dilakukan analisis dan evaluasi untuk menilai bagaimana dan dalam kondisi apa model memiliki performa baik dalam mendeteksi judi online pada berbagai jenis dataset. Untuk skenario yang melibatkan pelatihan IndoBERT, evaluasi dilakukan dengan melihat metrik akurasi, training loss, dan validation loss. Ketiga metrik tersebut dianalisis melalui learning curve untuk mendeteksi indikasi underfitting atau overfitting, dengan ketentuan sebagai berikut:

- Underfitting apabila akurasi berada di bawah 0,3 dan loss (baik training maupun validation) melebihi 0,7 dalam skala 1.
- Overfitting apabila akurasi mendekati 1, sementara training loss dan validation loss sangat rendah, yaitu di bawah 0,1 dalam skala 1.

Untuk skenario yang hanya menggunakan metode Fuzzy Matching, evaluasi dilakukan berdasarkan rata-rata nilai match score, dengan ambang batas minimum kemiripan sebesar 80%.

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA