BAB 3
METODOLOGI PENELITIAN

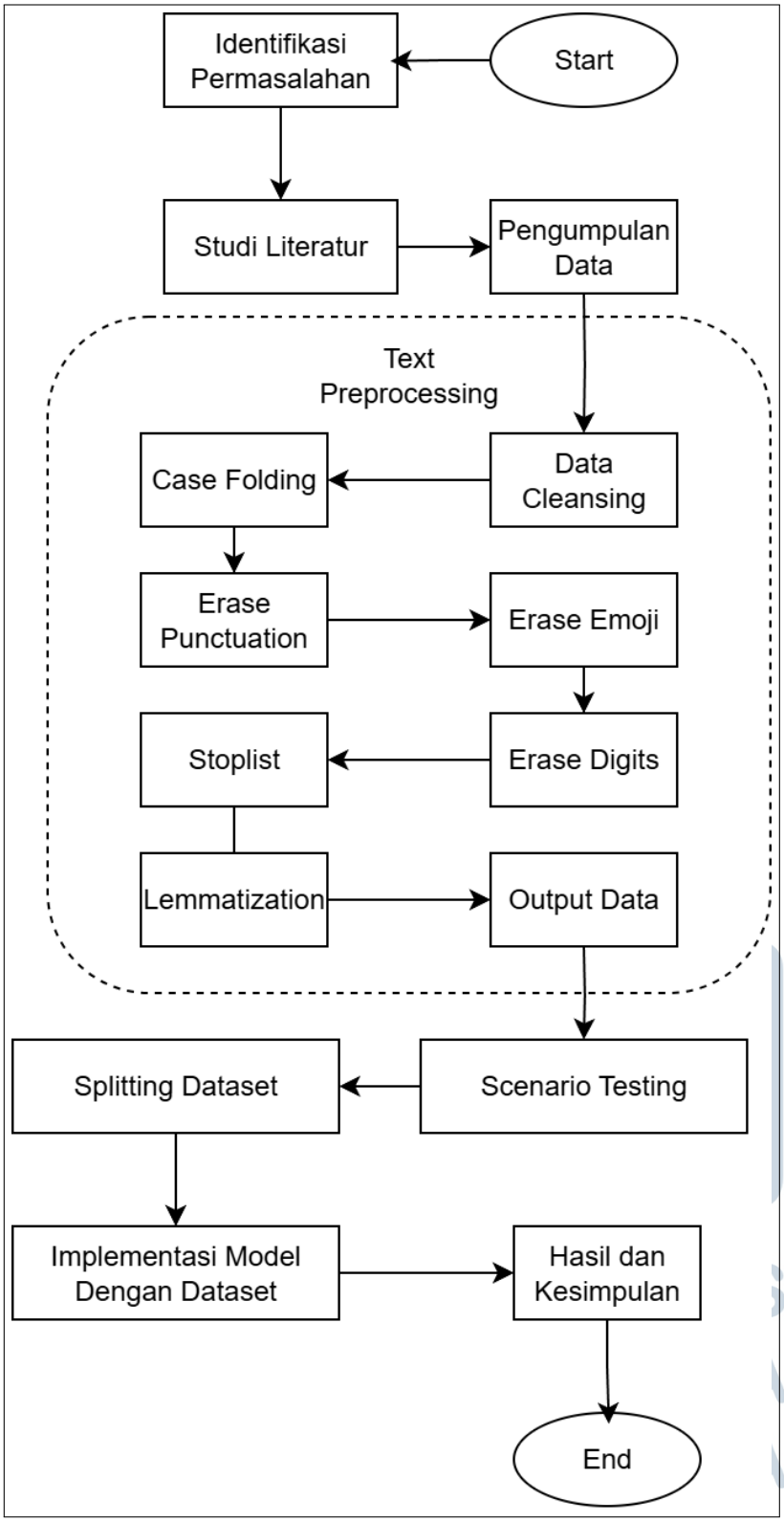
3.1 Objek Penelitian

Objek penelitian dari laporan ini merupakan analisis sentimen berbasis model *NLP (Natural Language Processing)* menggunakan *IndoNLP* untuk mengklasifikasikan sentimen mengenai penggunaan karya digital yang dibuat menggunakan kecerdasan buatan pada platform Twitter menjadi tiga label, yaitu positif, negatif, dan netral. Pengumpulan data akan dilakukan dengan menarik data melalui API media sosial *Twitter* dengan menggunakan *twscrapper*, salah satu metode *scraping data* yang dapat menarik data tertentu dari Twitter. Setelah itu, dengan menggunakan data yang sudah terkumpul, akan dilakukan proses NLP menggunakan model yang sudah terpilih untuk mengukur dan membandingkan sentimen yang bersifat positif, negatif, dan netral terhadap AI-generated artwork di *Twitter*.

3.2 Alur Penelitian

Alur penelitian yang akan dilakukan dalam melakukan analisis sentimen mengenai penggunaan *AI-generated artwork* di *Twitter* akan dimulai dengan proses identifikasi masalah untuk memahami isu yang dihadapi, diikuti dengan studi literatur mengenai topik yang relevan. Setelah itu akan dilakukan proses pengumpulan data, dan perancangan serta pembuatan model NLP menggunakan IndoNLP untuk melakukan Data Preprocessing menggunakan data yang telah dikumpulkan. Setelah melakukan testing dan evaluasi berdasarkan hasil yang telah didapatkan, akan ditarik kesimpulan dari data yang telah terproses.

Alur atau metodologi penelitian yang akan dilakukan dapat dilihat dari Gambar 3.1.



Gambar 3.1. Metodologi Penelitian

Berikut merupakan penjelasan dari metodologi penelitian yang telah dipersiapkan.

1. *Identifikasi Permasalahan* Tahap awal dalam penelitian untuk merumuskan isu atau pertanyaan utama yang ingin dijawab. Dalam konteks ini, permasalahan berfokus pada bagaimana opini publik terhadap AI-generated artwork dapat dianalisis melalui pendekatan NLP berbasis Bahasa Indonesia.

2. *Studi Literatur* Kegiatan pengumpulan referensi ilmiah dari jurnal, buku, dan publikasi terkait yang membahas analisis sentimen, model NLP, IndoBERT, serta pendekatan fine-tuning. Studi ini digunakan untuk memperkuat landasan teori dan menentukan metode yang paling sesuai.

3. *Pengumpulan Data* Tahapan ini dilakukan dengan mengambil data dari platform Twitter menggunakan tools seperti Twint atau Twscraper. Data yang dikumpulkan berupa tweet yang memuat opini pengguna terkait karya seni berbasis AI.

4. *Data Cleansing* Proses pembersihan data mentah dari atribut atau informasi yang tidak relevan, seperti ID pengguna, jumlah pengikut, URL gambar profil, dan metadata lainnya. Tujuannya adalah untuk mengurangi noise sebelum teks diproses lebih lanjut.

5. *Case Folding* Mengubah seluruh huruf dalam teks menjadi huruf kecil (lowercase) untuk menyeragamkan bentuk kata. Contohnya, kata ”Bagus”, ”BAGUS”, dan ”bagus” akan diubah menjadi satu bentuk yang sama: ”bagus”.

6. *Erase Punctuation* Penghapusan seluruh tanda baca seperti titik, koma, tanda seru, dan simbol lainnya. Tanda baca sering tidak menambah makna dalam klasifikasi sentimen, sehingga dihapus untuk menyederhanakan representasi teks.

7. *Erase Emoji* Emoji dihapus karena maknanya sangat kontekstual dan ambigu. Misalnya emoji bisa berarti sedih, terharu, atau bahkan tertawa, tergantung konteks tweet. Penghapusan emoji dilakukan agar model tidak salah menafsirkan makna sentimen.

8. *Erase Digits* Penghapusan seluruh angka atau digit numerik dari teks. Umumnya angka seperti “2024”, “100 persen”, atau “50rb” tidak

memberikan kontribusi bermakna terhadap konteks sentimen dan hanya menambah kerumitan tokenisasi.

9. *Stoplist* Menghapus kata-kata umum yang sering muncul tetapi tidak mengandung makna penting untuk klasifikasi, seperti “yang”, “dan”, “karena”, “itu”. Ini bertujuan untuk mengurangi noise dan membantu model fokus pada kata kunci penting.

10. *Lemmatization* Mengubah kata menjadi bentuk dasarnya (lemma). Misalnya, “bermain”, “dimainkan”, dan “pemainnya” akan diubah menjadi “main”. Proses ini dilakukan dengan library Sastrawi khusus Bahasa Indonesia.

11. *Output Data* Teks yang telah melalui seluruh tahapan preprocessing disimpan dalam format yang siap digunakan untuk proses pelatihan model, bebas dari karakter asing, noise, dan bentuk tidak seragam.

12. *Skenario Testing*
    Pada tahap ini, model IndoBERT diujicobakan menggunakan dataset benchmark IndoNLU untuk menentukan konfigurasi parameter terbaik, seperti jumlah epoch dan rasio split dataset. Tujuannya adalah menemukan kombinasi hyperparameter dan skenario pelatihan yang paling optimal sebagai dasar untuk melabel data analisis utama.

12. *Pembagian Dataset (Data Splitting)*
    Setelah tahap pengujian awal pada skenario testing, data kemudian dibagi menjadi tiga bagian, yaitu *training*, *validation*, dan *testing*. Pembagian ini dilakukan untuk mengevaluasi performa model secara objektif. Rasio pembagian bervariasi sesuai dengan skenario yang telah dirancang, antara lain 40:30:30, 60:20:20, 70:15:15, dan 80:10:10. Masing-masing skenario bertujuan untuk mengetahui bagaimana performa model berubah tergantung proporsi data pelatihan dan pengujian yang digunakan.

14. *Implementasi Model dengan Dataset*
    Setelah model terbaik diperoleh dari skenario testing, model tersebut digunakan untuk melabel dataset utama yang sebelumnya belum berlabel. Data hasil labeling ini kemudian dibagi ke dalam berbagai skenario pembagian dataset (40:30:30, 60:20:20, 70:15:15, dan 80:10:10) untuk pelatihan ulang dan evaluasi guna mencari performa terbaik berdasarkan metrik evaluasi seperti akurasi, F1 Score, precision, dan recall.

15. *Hasil dan Kesimpulan*

    Setelah dilakukan pelatihan dan evaluasi pada seluruh skenario, diperoleh hasil bahwa rasio 40:30:30 memberikan performa paling stabil dan akurat. Model dengan performa terbaik digunakan kembali untuk memproses keseluruhan data yang belum terlabel, sehingga menghasilkan klasifikasi sentimen yang menyeluruh dan representatif terhadap opini publik. Temuan menunjukkan dominasi sentimen negatif, yang memberikan wawasan penting dalam memahami persepsi masyarakat terhadap AI-generated artwork.

3.3 Identifikasi Permasalahan

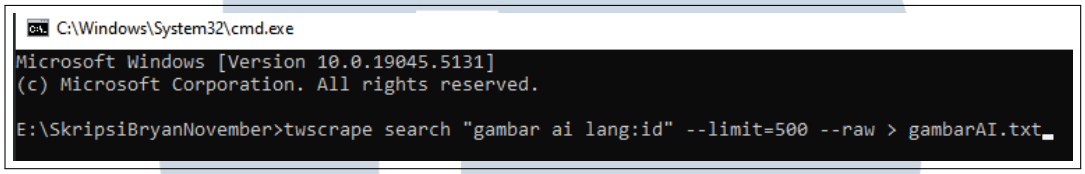
Sebelum melakukan penelitian, langkah awal yang dilakukan adalah identifikasi permasalahan untuk memahami konteks dan urgensi dari topik yang dibahas. Fenomena meningkatnya penggunaan AI-generated artwork di media sosial, khususnya Twitter, telah menimbulkan perdebatan publik terkait isu plagiarisme, pelanggaran etika, dan dampaknya terhadap industri kreatif. Banyak pengguna mengutarakan opini yang beragam mengenai keabsahan karya seni yang dihasilkan oleh kecerdasan buatan, yang sebagian besar dilatih menggunakan dataset tanpa persetujuan seniman. Permasalahan ini menjadi dasar dilakukannya penelitian analisis sentimen untuk memahami sikap masyarakat terhadap teknologi tersebut.

3.4 Studi Literatur

Selanjutnya, dilakukan studi literatur terhadap berbagai sumber yang relevan guna memperkuat landasan teoritis dan metodologis penelitian. Literatur yang dikaji mencakup konsep dasar Artificial Intelligence, perkembangan AI dalam bidang seni, serta penerapan Natural Language Processing (NLP) dan Natural Language Understanding (NLU) untuk tugas analisis sentimen. Fokus utama studi diarahkan pada model IndoBERT yang merupakan bagian dari benchmark IndoNLU, hasil kolaborasi berbagai institusi di Indonesia dan internasional. Model ini dipilih karena memiliki performa tinggi dan telah disesuaikan dengan struktur dan keragaman bahasa Indonesia formal maupun informal. Berdasarkan studi ini, ditentukan bahwa model SmSA dari IndoNLU paling sesuai untuk kebutuhan klasifikasi sentimen terhadap opini masyarakat di Twitter mengenai AI-generated artwork.

3.5 Pengumpulan Data

Proses pengumpulan data dilakukan dengan metode *scraping data* menggunakan *Twscrapper* melalui *Command Prompt (CMD) dan VSCode*. Data yang diambil dari *Twitter* melalui *Twscrapper* berupa isi teks dalam bentuk *post* atau '*tweet*' yang telah dibuat oleh pengguna Twitter dalam kurun waktu empat tahun terakhir, yaitu 2020-2024.

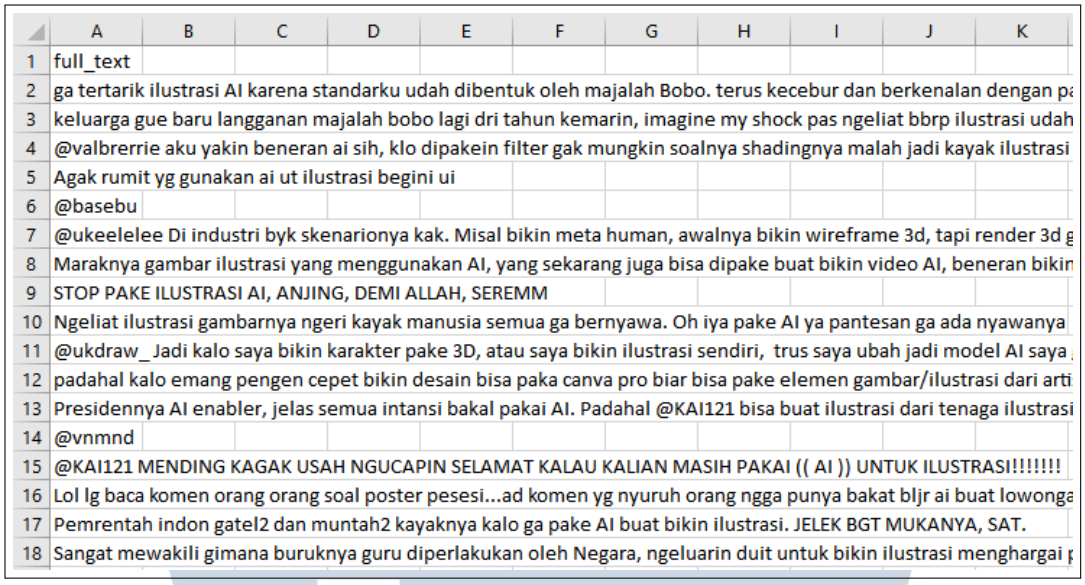


Gambar 3.2. Penarikan *post/tweet* menggunakan kata kunci 'gambar ai' dalam bahasa Indonesia.

Kata kunci yang akan digunakan dalam proses pengambilan data atau scraping data dari sosial media Twitter adalah 'ilustrasi ai', dan 'gambar ai'. Alasan pemilihan kedua kata kunci ini dikarenakan pemilihan model NLP yang akan digunakan dalam topik ini terlatih dalam bahasa Indonesia. Oleh sebab itu data yang akan ditarik akan difokuskan terhadap *post*-ingan dengan bahasa Indonesia. Kata tambahan seperti 'ilustrasi buatan ai' tidak akan digunakan karena 'ilustrasi ai' sudah mencakup teks atau kata dari data/ yang akan diambil.

Gambar 3.3. Hasil data yang telah ditarik menggunakan *Twscrapper*.

Pada gambar 3.3, bisa dilihat contoh data yang telah berhasil diambil dengan estimasi jumlah 500 data untuk kata kunci ’gambar ai’ dalam bentuk file .txt. Data yang telah ditarik memiliki informasi berlebih seperti identitas lengkap dari setiap akun yang telah membuat *tweet*, seperti jumlah *followers* (pengikut), link gambar profil mereka, dan lainnya. Dikarenakan kondisi data yang sangat berantakan, dan karena proses *data preprocessing* hanya membutuhkan isi teks dari *post* atau *tweet* yang telah ditarik melalui *Twscraper*, data perlu diolah kembali agar bisa digunakan dengan efisien melalui proses Data Preprocessing.



	A	B	C	D	E	F	G	H	I	J	K
1	full_text										
2	ga tertarik ilustrasi AI karena standarku udah dibentuk oleh majalah Bobo. terus kecebur dan berkenalan dengan pa										
3	keluarga gue baru langganan majalah bobo lagi dri tahun kemarin, imagine my shock pas ngeliat bbrp ilustrasi udah										
4	@valbrerrie aku yakin beneran ai sih, klo dipakein filter gak mungkin soalnya shadingnya malah jadi kayak ilustrasi										
5	Agak rumit yg gunakan ai ut ilustrasi begini ui										
6	@basebu										
7	@ukeelelee Di industri byk skenarionya kak. Misal bikin meta human, awalnya bikin wireframe 3d, tapi render 3d g										
8	Maraknya gambar ilustrasi yang menggunakan AI, yang sekarang juga bisa dipake buat bikin video AI, beneran bikin										
9	STOP PAKE ILUSTRASI AI, ANJING, DEMI ALLAH, SEREMM										
10	Ngeliat ilustrasi gambarnya ngeri kayak manusia semua ga bernyawa. Oh iya pake AI ya pantesan ga ada nyawanya										
11	@ukdraw_ Jadi kalo saya bikin karakter pake 3D, atau saya bikin ilustrasi sendiri, trus saya ubah jadi model AI saya										
12	padahal kalo emang pengen cepet bikin desain bisa paka canva pro biar bisa pake elemen gambar/ilustrasi dari arti										
13	Presidennya AI enabler, jelas semua intansi bakal pakai AI. Padahal @KAI121 bisa buat ilustrasi dari tenaga ilustrasi										
14	@vnmnd										
15	@KAI121 MENDING KAGAK USAH NGUCAPIN SELAMAT KALAU KALIAN MASIH PAKAI ((AI)) UNTUK ILUSTRASI!!!!!!!										
16	Lol lg baca komen orang orang soal poster pesesi...ad komen yg nyuruh orang ngga punya bakat bljr ai buat lowonga										
17	Pemrentah indon gatel2 dan muntah2 kayaknya kalo ga pake AI buat bikin ilustrasi. JELEK BGT MUKANYA, SAT.										
18	Sangat mewakili gimana buruknya guru diperlakukan oleh Negara, ngeluarin duit untuk bikin ilustrasi menghargai p										

Gambar 3.4. Data yang sudah dirapikan dalam bentuk .csv.

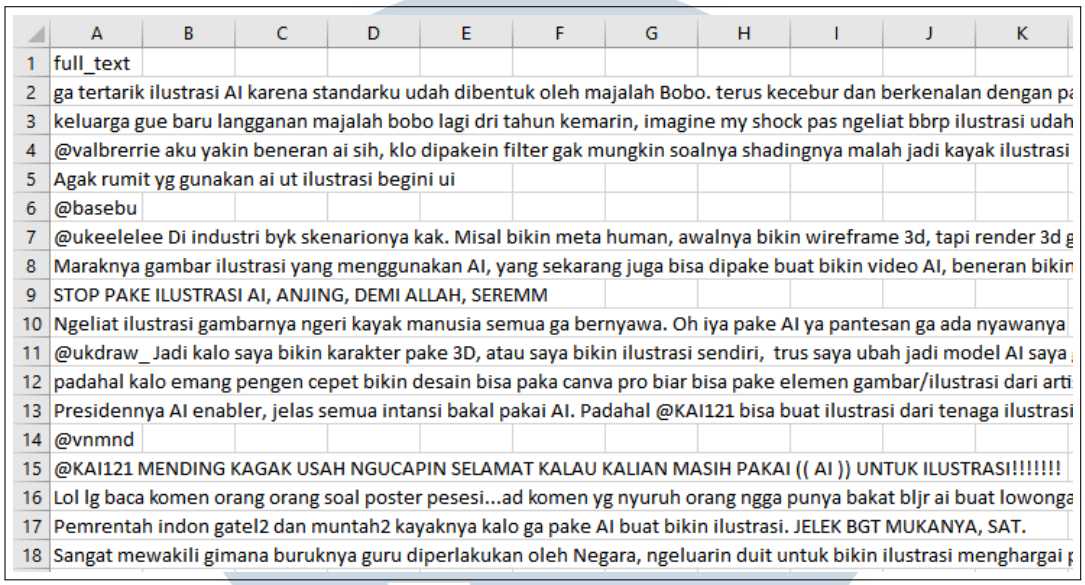
Bisa dilihat pada gambar 3.5, isi teks dari file .txt yang ditarik melalui Twscraper sudah berhasil dirapikan dalam bentuk file .csv. File yang sudah diolah akan digunakan untuk proses selanjutnya, yaitu *data preprocessing*.

3.6 Data Preprocessing

Data preprocessing merupakan tahap krusial dalam proses Natural Language Processing (NLP), yang bertujuan untuk meningkatkan akurasi model serta meminimalkan potensi kesalahan dalam interpretasi data oleh sistem. Pada tahapan awal, proses pembersihan data dilakukan untuk menghilangkan elemen-elemen yang tidak relevan dan tidak memberikan kontribusi signifikan terhadap analisis teks. Informasi seperti followers count, user mentions, profile banner URL, profile image URL, dan atribut metadata lainnya akan dihapus, karena keberadaannya justru dapat memperkenalkan noise dan mengganggu kinerja model dalam memahami konteks linguistik yang sebenarnya.

Perlu ditegaskan bahwa proses pelabelan pada penelitian ini tidak dilakukan secara manual ataupun oleh *annotator* manusia *(Ground Truth)*. Label sentimen (positif, negatif, netral) pada data tweet dihasilkan secara otomatis menggunakan model IndoBERT yang telah dilatih dengan model *Sentiment Analysis (SMSA)*. Proses ini dikenal sebagai *pseudolabeling*, di mana label diperoleh dari hasil prediksi model tanpa ground truth manusia. Oleh karena itu, tidak terdapat proses validasi antar-*annotator* dalam penelitian ini.

3.6.1 Data Cleansing



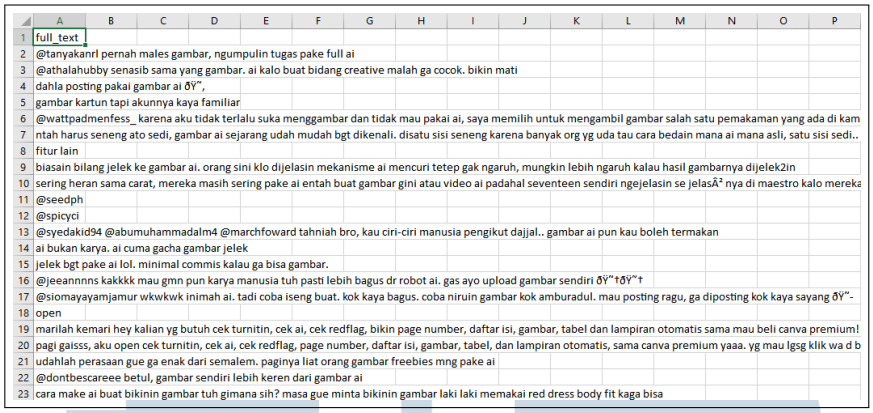
	A	B	C	D	E	F	G	H	I	J	K
1	full_text										
2	ga tertarik ilustrasi AI karena standarku udah dibentuk oleh majalah Bobo. terus kecebur dan berkenalan dengan pa										
3	keluarga gue baru langganan majalah bobo lagi dri tahun kemarin, imagine my shock pas ngeliat bbrp ilustrasi udah										
4	@valbrerrie aku yakin beneran ai sih, klo dipakein filter gak mungkin soalnya shadingnya malah jadi kayak ilustrasi										
5	Agak rumit yg gunakan ai ut ilustrasi begini ui										
6	@basebu										
7	@ukeelelee Di industri byk skenarionya kak. Misal bikin meta human, awalnya bikin wireframe 3d, tapi render 3d g										
8	Maraknya gambar ilustrasi yang menggunakan AI, yang sekarang juga bisa dipake buat bikin video AI, beneran bikin										
9	STOP PAKE ILUSTRASI AI, ANJING, DEMI ALLAH, SEREMM										
10	Ngeliat ilustrasi gambarnya ngeri kayak manusia semua ga bernyawa. Oh iya pake AI ya pantesan ga ada nyawanya										
11	@ukdraw_ Jadi kalo saya bikin karakter pake 3D, atau saya bikin ilustrasi sendiri, trus saya ubah jadi model AI saya										
12	padahal kalo emang pengen cepet bikin desain bisa paka canva pro biar bisa pake elemen gambar/ilustrasi dari arti										
13	Presidennya AI enabler, jelas semua intansi bakal pakai AI. Padahal @KAI121 bisa buat ilustrasi dari tenaga ilustrasi										
14	@vnmnd										
15	@KAI121 MENDING KAGAK USAH NGUCAPIN SELAMAT KALAU KALIAN MASIH PAKAI ((AI)) UNTUK ILUSTRASI!!!!!!!										
16	Lol lg baca komen orang orang soal poster pesesi...ad komen yg nyuruh orang ngga punya bakat bljr ai buat lowonga										
17	Pemrentah indon gatel2 dan muntah2 kayaknya kalo ga pake AI buat bikin ilustrasi. JELEK BGT MUKANYA, SAT.										
18	Sangat mewakili gimana buruknya guru diperlakukan oleh Negara, ngeluarin duit untuk bikin ilustrasi menghargai p										

Gambar 3.5. Data siap diproses.

Seperti yang ditunjukkan pada Gambar 3.5, isi teks dari file berformat .txt yang diperoleh melalui proses pengambilan data menggunakan Twscraper telah berhasil diubah dan dirapikan ke dalam format .csv. Proses konversi ini memudahkan pengelolaan data. File hasil konversi ini selanjutnya akan digunakan pada tahap berikutnya, yaitu proses case folding.

3.6.2 Case Folding

Tahapan kedua dalam proses data preprocessing adalah penerapan teknik *case folding*, yaitu proses mengubah seluruh huruf dalam teks menjadi huruf kecil (lowercase). Tujuan utama dari case folding adalah untuk menyamakan representasi kata yang semestinya memiliki makna sama namun dituliskan dengan perbedaan kapitalisasi. Sebagai contoh, kata “AI”, “Ai”, dan “ai” akan dianggap sebagai entitas yang berbeda oleh model jika tidak dilakukan normalisasi bentuk huruf terlebih dahulu. Dengan demikian, proses ini dapat mengurangi redundansi kata serta menyederhanakan struktur data masukan.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	full_text															
2	@tanyakanrl pernah males gambar, ngumpulin tugas pake full ai															
3	@athalahubby senasib sama yang gambar. ai kalo buat bidang creative malah ga cocok. bikin mati															
4	dahla posting pakai gambar ai ðŸ˜,															
5	gambar kartun tapi akunnya kaya familiar															
6	@wattpadmenfess_ karena aku tidak terlalu suka menggambar dan tidak mau pakai ai, saya memilih untuk mengambil gambar salah satu pemakaman yang ada di kam															
7	ntah harus seneng ato sedi, gambar ai sejarang udah mudah bgt dikenali. disatu sisi seneng karena banyak org yg uda tau cara bedain mana ai mana asli, satu sisi sedi..															
8	fitur lain															
9	biasain bilang jelek ke gambar ai. orang sini klo dijelasin mekanisme ai mencuri tetep gak ngaruh, mungkin lebih ngaruh kalau hasil gambarnya dijelek2in															
10	sering heran sama carat, mereka masih sering pake ai entah buat gambar gini atau video ai padahal seventeen sendiri ngejelasin se jelasÃ² nya di maestro kalo mereka															
11	@seedph															
12	@spicyci															
13	@syedakid94 @abumuhammadalm4 @marchfoward tahniah bro, kau ciri-ciri manusia pengikut dajjal.. gambar ai pun kau boleh termakan															
14	ai bukan karya. ai cuma gacha gambar jelek															
15	jelek bgt pake ai lol. minimal commis kalau ga bisa gambar.															
16	@jeeannnns kakkkk mau gmn pun karya manusia tuh pasti lebih bagus dr robot ai. gas ayo upload gambar sendiri ðŸ˜¨ðŸ˜¨															
17	@siomayayamjamur wkwkwk inimah ai. tadi coba iseng buat. kok kaya bagus. coba niruin gambar kok amburadul. mau posting ragu, ga diposting kok kaya sayang ðŸ˜															
18	open															
19	marilah kemari hey kalian yg butuh cek turnitin, cek ai, cek redflag, bikin page number, daftar isi, gambar, tabel dan lampiran otomatis sama mau beli canva premium!															
20	pagi gaisss, aku open cek turnitin, cek ai, cek redflag, page number, daftar isi, gambar, tabel, dan lampiran otomatis, sama canva premium yaaa. yg mau lgsg klik wa d b															
21	udahlah perasaan gue ga enak dari semalem. paginya liat orang gambar freebies mng pake ai															
22	@dontbescareee betul, gambar sendiri lebih keren dari gambar ai															
23	cara make ai buat bikinin gambar tuh gimana sih? masa gue minta bikinin gambar laki laki memakai red dress body fit kaga bisa															

Gambar 3.6. Semua huruf kapital diubah menjadi huruf kecil.

Seperti yang ditampilkan pada Tabel 3.1, terlihat perbandingan isi teks sebelum dan sesudah dilakukan case folding pada sejumlah tweet. Perubahan ini tampak sederhana, namun berperan penting dalam standarisasi data sebelum memasuki tahap pemrosesan linguistik lanjutan.

Tabel 3.1. Pengubahan Kapitalisasi Teks pada Data Twitter

No	Sebelum case folding	Sesudah Case folding
1	Guys saran aja dari aku, kalau semisal gak bisa gambar dan gak ada uang buat commis, mending gak usah bagi-bagi freebies aja daripada pake gambar buatan AI. Woozi bikin Maestro buat ngasih pesan kalau pada akhirnya the real maestro ya manusia, bukan robot.	guys saran aja dari aku, kalau semisal gak bisa gambar dan gak ada uang buat commis, mending gak usah bagi-bagi freebies aja daripada pake gambar buatan ai. woozi bikin maestro buat ngasih pesan kalau pada akhirnya the real maestro ya manusia, bukan robot.
2	Ga mau suudzon tapi itu gambar AI ga, sih?	ga mau suudzon tapi itu gambar ai ga, sih?

3	Commis fanart emg berapaan sih? cuma 50an lebih.. 60k lah, klo gk mampu bayar commis BELAJAR GAMBAR!! bikin chibi kek gk susah itu kok KALO MAU BELAJAR, drpd bikin freebies tp pake AI kocak mana jelek bgt lagi	commis fanart emg berapaan sih? cuma 50an lebih.. 60k lah, klo gk mampu bayar commis belajar gambar!! bikin chibi kek gk susah itu kok kalo mau belajar, drpd bikin freebies tp pake ai kocak mana jelek bgt lagi
4	Kak jangan yaa, jangan pake ai. Uji udah ngingetin jangan pake ai. Terus itu gambar kotak susu yang dipegang wonu kenapa jadi es jeruk	kak jangan yaa, jangan pake ai. uji udah ngingetin jangan pake ai. terus itu gambar kotak susu yang dipegang wonu kenapa jadi es jeruk

3.6.3 Erase Punctuation

Tahapan ketiga, yang dilakukan setelah seluruh data diubah menjadi huruf kecil, adalah penghapusan tanda baca (punctuation removal). Tanda baca seperti titik (.), koma (,), titik dua (:), tanda seru (!), tanda tanya (?), serta karakter simbol lainnya dihapus karena dianggap tidak memberikan kontribusi yang signifikan terhadap pemahaman makna dalam analisis sentimen. Selain itu, keberadaan tanda baca dapat mengganggu proses tokenisasi dan menyebabkan duplikasi kata dalam bentuk berbeda. Dengan menghapus seluruh tanda baca, data menjadi lebih bersih dan seragam. Perbandingan antara data sebelum dan sesudah penghapusan tanda baca ditampilkan pada Tabel 3.2.

Tabel 3.2. Penghapusan Tanda Baca pada Data Twitter

No	Sebelum Tanda Baca Dihapus	Sesudah Tanda Baca Dihapus
1	ayo gais, yg butuh cek turnitin, cek ai, cek redflag, page number, daftar isi, gambar, tabel, dan lampiran otomatis, sama canva premium ke aku aja ya, wa d bio zonauang	ayo gais yg butuh cek turnitin cek ai cek redflag page number daftar isi gambar tabel dan lampiran otomatis sama canva premium ke aku aja ya wa d bio zonauang

3.6.4 Erase Emoji/Emoticon

Setelah tahapan penghapusan tanda baca, **tahapan keempat** yang dilakukan adalah penghapusan emotikon atau emoji dari data. Emotikon dan emoji merupakan simbol visual yang digunakan untuk mengekspresikan emosi atau reaksi tertentu, namun maknanya sangat bervariasi tergantung pada konteks, budaya, serta preferensi pengguna. Karena sifatnya yang ambigu dan kontekstual—yang sering kali tidak dapat ditafsirkan secara akurat hanya berdasarkan teks dalam satu tweet—emoji dalam penelitian ini dianggap sebagai elemen non-linguistik yang berpotensi mengganggu interpretasi model NLP. Oleh karena itu, seluruh karakter emoji dihapus untuk menjaga konsistensi data dan meningkatkan akurasi klasifikasi. Perbandingan antara data sebelum dan sesudah proses penghapusan emoji dapat dilihat pada Tabel 3.3.

Tabel 3.3. Penghapusan Emoji pada Data Twitter

No	Sebelum Emoji Dihapus	Sesudah Emoji Dihapus
1	apa ini hari kiamat gw black box ai udah gabisa send gambar Ÿ˜Ÿ˜Ÿ˜Ÿ˜Ÿ˜Ÿ˜Ÿ˜ terus gw upchar gimana coy Ÿ˜Ÿ˜Ÿ˜Ÿ˜Ÿ˜Ÿ˜Ÿ˜Ÿ˜	apa ini hari kiamat gw black box ai udah gabisa send gambar terus gw upchar gimana coy

3.6.5 Erase Digits

Tahapan kelima adalah penghapusan angka atau digit numerik dari teks. Dalam konteks analisis sentimen, angka-angka seperti tahun, jumlah, kode, atau nilai numerik lainnya umumnya tidak memberikan kontribusi berarti terhadap makna emosional dari teks yang dianalisis. Keberadaan angka juga dapat menciptakan variasi tokenisasi yang tidak relevan serta menambah kompleksitas data secara keseluruhan. Oleh karena itu, angka yang terdapat dalam data dihapus agar model NLP dapat lebih fokus pada kata-kata bermakna yang benar-benar relevan terhadap tugas klasifikasi sentimen.

Tabel 3.4. Penghapusan Digit pada Data Twitter

No	Sebelum Digit Dihapus	Sesudah Digit Dihapus
1	litereli lu balik badan liat langsung hasil generate ke 2 baliknya lagi generate ke 3 ada juga 1 gambar dipajang di 2 tempat temanya kartun malah jadi gambar poster bjir kecewa dikit banget	litereli lu balik badan liat langsung hasil generate ke baliknya lagi generate ke ada juga gambar dipajang di tempat temanya kartun malah jadi gambar poster bjir kecewa dikit banget

3.6.6 Stopword

Setelah penghapusan digit, proses dilanjutkan ke **tahapan keenam**, yaitu penghapusan kata-kata umum yang disebut dengan stopword atau stoplist. Stopword merupakan kata-kata yang sering muncul dalam teks namun tidak memiliki kontribusi bermakna terhadap konteks atau isi informasi yang dianalisis. Contoh dari kata-kata tersebut antara lain: “karena”, “oleh”, “terus”, “dan”, “itu”, dan sebagainya. Dalam konteks analisis sentimen atau klasifikasi teks, kata-kata ini biasanya tidak membantu model dalam membedakan antara satu kategori dengan kategori lainnya. Oleh karena itu, kata-kata tersebut dihapus dari keseluruhan isi teks untuk menyederhanakan data, mengurangi noise, serta meningkatkan akurasi dan efisiensi dalam proses pemodelan NLP.

Tabel 3.5. Implementasi Stopword pada Data Twitter

No	Sebelum stopword	Sesudah stopword
1	ga tertarik ilustrasi ai karena standarku udah dibentuk oleh majalah bobo terus kecebur dan berkenalan dengan para pekerja seni, tau gimana perspektif dan perjuangan masing2, jadi sangat menghargai seni yg manusiawi	ga tertarik ilustrasi ai standarku udah dibentuk majalah bobo kecebur berkenalan pekerja seni tau gimana perspektif perjuangan menghargai seni yg manusiawi

3.6.7 Lemmatization

Setelah proses penghapusan stopword, **tahapan ketujuh** sekaligus terakhir dalam rangkaian data preprocessing adalah proses lemmatization. Lemmatization bertujuan untuk mengubah setiap kata dalam teks menjadi bentuk dasarnya (lemma), dengan cara menghilangkan imbuhan atau afiks seperti "-nya", "di-", "ter-", "ber-", dan sejenisnya. Dengan mengembalikan kata ke bentuk dasarnya, struktur kalimat menjadi lebih sederhana dan seragam, sehingga dapat membantu meningkatkan akurasi dan efisiensi model Natural Language Processing (NLP) yang akan digunakan dalam tahap analisis selanjutnya.

Untuk melaksanakan proses lemmatization dalam Bahasa Indonesia, digunakan pustaka (library) khusus yaitu *Sastrawi* yang ditulis dalam bahasa pemrograman PHP. Pustaka ini dipilih karena pustaka standar NLP seperti *nltk* secara default hanya mendukung Bahasa Inggris dan tidak memiliki fungsi yang cukup relevan untuk Bahasa Indonesia. Proses lemmatization menggunakan Sastrawi memungkinkan model untuk mengenali kata dalam bentuk yang lebih netral dan representatif terhadap makna aslinya.

Tabel 3.6. Implementasi Lemmatization pada Data Twitter

No	Sebelum Lemmatization	Sesudah Lemmatization
1	beruntung tumbuh generasi yg menikmati gambar gambar kosong yg ai	untung tumbuh generasi yg nikmat gambar gambar kosong yg ai
2	susah nih disuruh milih lukisan gambar maknanya gak pilih	susah nih suruh milih lukis gambar makna gak pilih

Dengan selesainya seluruh tahapan preprocessing, data teks kini telah berada dalam kondisi optimal untuk dianalisis lebih lanjut menggunakan model berbasis *IndoNLU*. Model ini akan memanfaatkan hasil dari preprocessing untuk melakukan klasifikasi sentimen secara lebih akurat dan kontekstual terhadap isi tweet yang telah dibersihkan dan disederhanakan.

3.6.8 Model IndoBERT

Secara keseluruhan, terdapat dua belas model IndoBERT yang dikembangkan oleh tim IndoNLP, masing-masing dengan fungsi dan karakteristik yang berbeda. Setiap model dirancang untuk menyelesaikan tugas tertentu dalam ranah pemrosesan bahasa alami, mulai dari klasifikasi sederhana hingga pemahaman hubungan antar kalimat. Beberapa di antaranya ditujukan untuk analisis sentimen, sementara yang lain difokuskan untuk tugas seperti *named entity recognition* (NER), *document classification*, dan *natural language inference* (NLI). Model-model ini juga memiliki arsitektur dan parameter pelatihan yang beragam, sehingga pemilihan model yang tepat harus mempertimbangkan konteks tugas serta karakteristik data yang digunakan.

Penjelasan lebih lanjut mengenai beberapa model yang tersedia dalam ekosistem *IndoBERT*, beserta fungsinya masing-masing, dapat dilihat pada Tabel 3.7.

Tabel 3.7. Model IndoBERT

No	Nama model	Fungsi model
1	fine tune casa	*Context-Aware Self-Attentive*, model ini berfungsi untuk menganalisis dan memahami konteks dari suatu teks atau kalimat. CASA akan memberikan label positif, negatif atau netral terhadap setiap inti kata yang telah dipecahkan.
2	fine tune Emot	Model ini berfungsi untuk menganalisis dan memahami emosi yang tersimpan di dalam suatu teks atau kalimat. Terdapat lima label untuk model ini, yaitu: *sadness*, *anger*, *love*, *fear*, dan *happy*.
3	fine tune SmSA	Model ini berfungsi untuk melakukan analisis sentimen di dalam suatu teks atau kalimat yang bersifat positif, negatif, dan netral.
4	fine tune NERGrit	*Named Entity Recognition Dataset*, model ini berfungsi untuk membedakan entitas dari suatu teks atau kalimat dengan label *PERSON*, *PLACE*, dan *ORGANIZATION*.
		Lanjut pada halaman berikutnya

5	fine tune WRete	Model ini merupakan dataset yang berfungsi untuk menemukan hubungan logis antar dua teks. Model ini membedakan hasil analisis mereka menjadi dua: *Entail or Paraphrase*, dan *NotEntail*.

Untuk melakukan analisis sentimen terhadap opini publik mengenai penggunaan AI-generated artwork, yang dikategorikan ke dalam tiga label sentimen yaitu positif, negatif, dan netral, digunakan **model SmSA** (Sentiment Analysis with IndoBERT). Model ini merupakan bagian dari ekosistem *IndoNLU* dan telah di-fine-tune secara khusus untuk tugas klasifikasi sentimen dalam Bahasa Indonesia. Dengan memanfaatkan model SmSA, data teks yang telah melalui tahapan preprocessing dapat dianalisis secara lebih akurat dan dibedakan secara otomatis ke dalam tiga kategori sentimen tersebut berdasarkan pola linguistik dan semantik yang terkandung dalam teks.