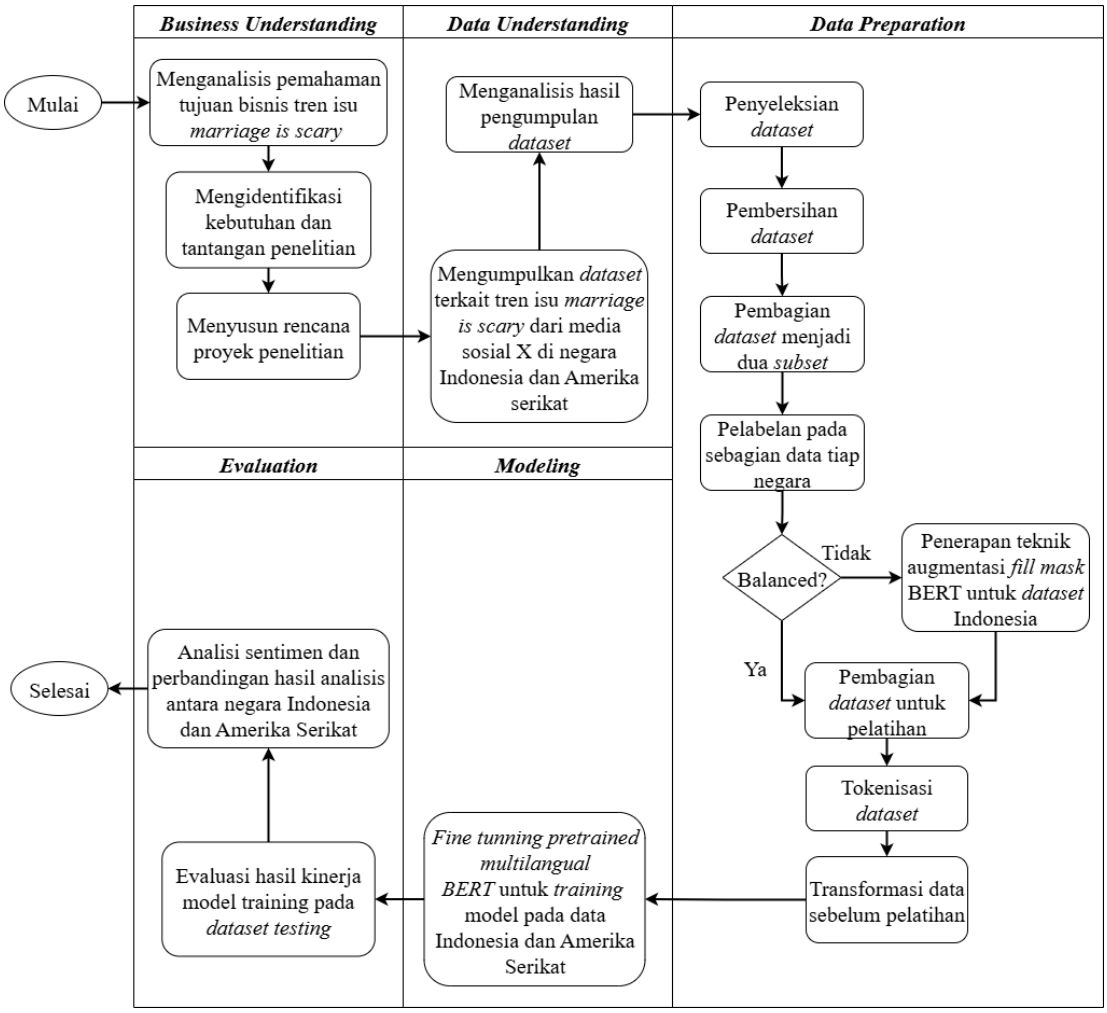
BAB III

METODOLOGI PENELITIAN

Pada bab ini merupakan penjelasan mengenai langkah-langkah penelitian yang memanfaatkan metodologi CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Tahapan-tahapan penelitian ini dilakukan dengan menggunakan 5 dari 6 tahapan metodologi CRISP-DM. Tahapan-tahapan tersebut diuraikan dalam Gambar 3.1 berikut.



Gambar 3.1 Diagram Alur Tahapan Penelitian CRISP-DM

Pada Gambar 3.1 merupakan rangkaian tahapan penelitian yang dilakukan pada tugas akhir ini. Tahapan tersebut dimulai dari pemahaman bisnis hingga evaluasi analisis sentimen. Setiap tahapan memiliki peran yang penting agar penelitian tugas akhir dapat berjalan dengan rencana serta sistematis dan menghasilkan output yang baik.

3.1. *Business Understanding*

Tahapan awal dalam metodologi penelitian dengan menggunakan CRISP-DM adalah dengan *business understanding*. Pada tahapan ini berfokus kepada identifikasi masalah dan pemahaman kebutuhan penelitian serta tujuan dari penelitian. Selain itu merancang tahapan proyek awal penelitain terkait isu *marriage is scary* di negara Indonesia dan Amerika Serikat pada media sosial X, dimana media sosial X merupakan salah satu platform yang berguna bagi pengguna untuk menyampaikan opini mereka terhadap suatu hal, sehingga menjadi tempat yang baik untuk melakukan analisis sentimen.

3.1.1. Menganalisis Pemahaman Tujuan Bisnis Penelitian

Langkah awal yang dilakukan adalah mengidentifikasi tujuan bisnis terhadap tren *marriage is scary* dari penelitian yang akan dilakukan. Dalam menentuan tujuan bisnis tersebut perlu diketahui hal-hal yang harus disiapkan terlebih dahulu. Hal-hal yang dibutuhkan dalam penelitian ini antara lain seperti berikut.

a. Sumber data yang akan didapatkan berasal dari media sosial X (Twitter) dengan menggunakan teknik *scraping* yang memanfaatkan *Tweet-Harvest* sebagai alat untuk mendapatkan data penelitian.
b. Pendefinisian label data dilakukan dengan menggunakan dua klasifikasi label sentimen yakni label positif, dan negatif pada data Indonesia dan Amerika Serikat.
c. Pelatihan model memanfaatkan *pretrained model Multilangual* BERT dengan mengaplikasikan *fine tunning* agar dapat menghasilkan model yang terbaik.
d. Mengetahui tujuan spesifik yakni untuk mengetahui demografi sentimen analisis dari media sosial X terhadap isu *marriage is scary* yang ada di negara Indonesia dan Amerika Serikat.

3.1.2. Megidentifikasi Kebutuhan dan Tantangan Penelitian

Langkah pada tahapan ini adalah mengetahui kebutuhan-kebutuhan serta tantangan yang akan dihadapi dalam proses penelitan. Kebutuhan-kebutuhan serta tantangan penelitian tersebut sebagai berikut.

a. Sumber data yang didapatkan harus baik dan relevan dengan isu *marriage is scary* di media sosial X pada negara Indonesia dan Amerika Serikat.

b. Teknologi yang dibutuhkan harus berupa komputasi yang cukup dan memadai seperti CPU atau GPU dalam membantu pengolahan data.
c. Alat-alat yang dibutuhkan dalam membantu proses penelitian yakni bahasa pemrograman python, *hugging face, library transformer, matplotlib, pandas,* dan lain-lain.
d. Tantangan dalam *imbalanced* data yang dapat memengaruhi dalam optimalisasi proses pelatihan dan hasil dari pelatihan data.
e. Keketatan keamanan sosial media X yang memberikan batas pengambilan *scraping* data sehingga menghambat proses pengumpulan data, serta harga API *development Twitter* yang mahal.

3.1.3. Menyusun Rencana Proyek Penelitian

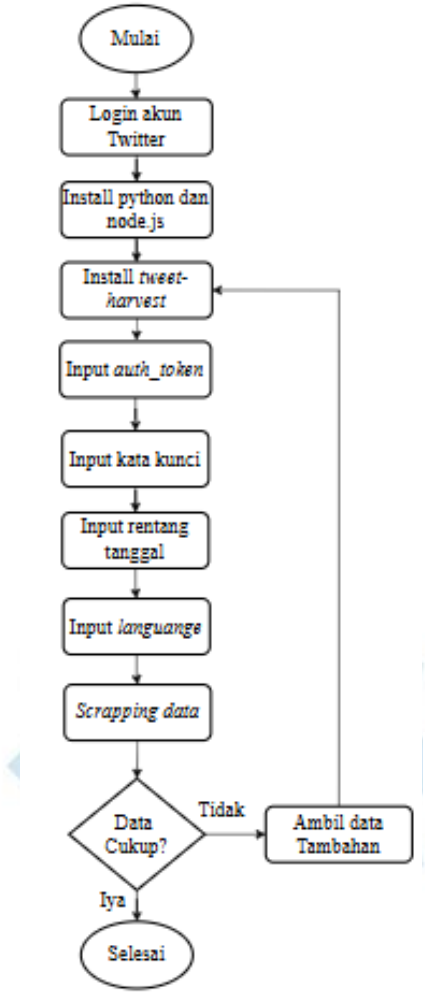
Tahapan penutup pada langkah ini adalah merancang rencana penelitian secara terperinci. Hal tersebut, seperti mencakup tehadap rincian waktu dalam setiap aktivitas penelitian, serta kebutuhan dalam sumber daya yang diperlukan sepanjang penelitian. Setiap tahapan CRISP-DM, dimulai dari pengumpulan data dari media sosial hingga evaluasi hasil model, semuanya dirancang dengan memiliki tenggat waktu yang jelas agar tetap menjaga alur kerja tetap terstruktur. Selain itu, memuat juga identifikasi terkait potensi hambatan yang akan muncul, seperti keterbatasan jumlah dan kualitas data serta daya komputasi, dan langkah-langkah untuk menghadapi tantangan tersebut.

3.2. *Data Understanding*

Tahapan ini merupakan tahapan yang bertujuan untuk mengumpulkan serta mengeksplorasi data, serta pemahaman terhadap data penelitian sebelum digunakan. Tahapan-tahapan yang dilakukan pada fase ini adalah sebagai berikut.

3.2.1. Pengumpulan Data Penelitan

Tahapan ini merupakan tahapan pengumpulan data penelitan. Pengumpulan data penelitian menggukankan teknik *scraping* data dari media sosial X dengan memanfaatkan *tweet-harvest*. Alur dari proses *scraping data* pada platform X dijelaskan seperti pada Gambar 3.2 berikut.



Gambar 3.2 Alur Pengambilan Data

Pada Gambar 3.2 merupakan alur dari pengambilan data penelitian. Dalam proses pengambilan data, diketahui bahwa dilakukan dengan pengambilan data menggunakan *filter* seperti kata kunci, rentang tanggal, dan bahasa.

Pengambilan data memanfaatkan *filter* tersebut dalam mengambil data informasi dari media sosial X agar dapat memiliki jumlah data yang cukup digunakan dalam penelitian. *Filter* data tersebut seperti.

a. Data yang akan dikumpulkan menggunakan kata kunci *marriage is scary,* takut nikah, takut cerai, trauma nikah, dan pernikahan indonesia untuk negara Indonesia. Serta kata kunci seperti *Marriage is scary, fear of marriage, boycott marriage, marriage strike, divorce,* dan *be single* untuk negara Amerika Serikat.
b. Pengambilan data juga menggunakan *filter* bahasa, seperti *lang=id* untuk Indonesia dan *lang=eng* untuk Amerika Serikat.

c. Rentang waktu yang diambil dimulai dari tanggal 1 Agustus 2024 sampai 20 Mei 2025. Rentang waktu tersebut dipilih dari bulan agustus dikarena tren *marriage is scary* mulai banyak diperbincangkan di Indonesia.

3.2.2. Menelaah Data

Kemudian tahapan berikutnya setelah data terkumpul, data tersebut dilakukan ekspolarasi data agar dapat mengetahui karakteristik dari data. Karakteristik data tersebut mencakup sumber data, format file, serta elemen-elemen karakteristik seperti *tweet*, waktu dan tanggal *tweet*, *retweet*, bahasa, serta metadata penunjang lainnya. Hal ini dilakukan agar mengetaui representasi data sudah sesuai dengan penelitian, sehingga dapat membantu dalam mengetahui tingkat representatif data yang telah terkumpul.

3.3. *Data Preparation*

Pada tahapan ini dilakukan persiapan data sebelum melanjutkan kepada tahap pelatihan. Data yang telah dikumpulkan untuk digunakan dalam proses analisis sentimen, dilakukan proses *preparation* terlebih dahulu agar data telah siap digunakan sebelum pemrosesan pelatihan data. Proses *data preparation* ini dilakukan terhadap kedua data dengan secara terpisah agar hasil dapat secara optimal.

3.3.1. Seleksi *Dataset* Penelitian

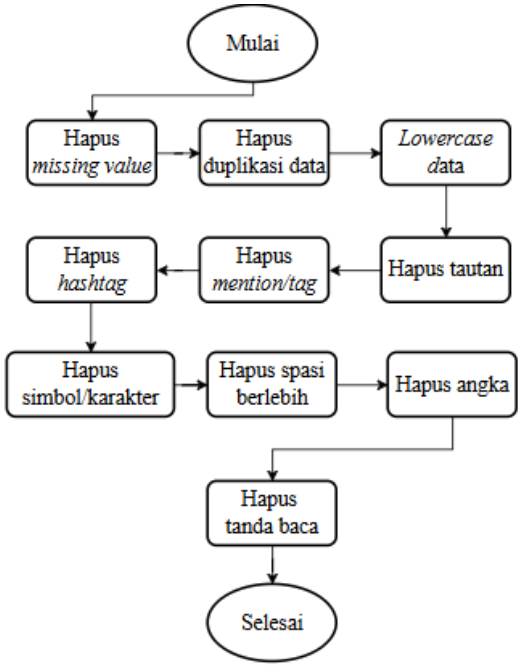
Tahapan ini merupakan tahapan untuk melakukan seleksi data agar data lebih optimal sesuai dengan tujuan penelitian. Tahapan seleksi data ini dilakukan berdasarkan beberapa hal berikut.

a. Data berdasarkan kata kunci tiap-tiap negara.
b. Bahasa yang terdapat dalam teks harus bahasa Indonesia untuk data negara Indonesia dan bahasa Inggris untuk negara Amerika Serikat.
c. Lokasi dari data yang telah terkumpul harus berkaitan dengan kedua negara tersebut.

3.3.2. Pembersihan *Dataset*

Pada tahap pembersihan data ini, proses dilakukan dengan beberapa hal agar data yang dihasilkan bersih dari *noise* yang dapat mengganggu dalam proses selanjutnya. Proses ini penting dilakukan, karena model tidak dapat mengolah data apabila data tersebut masih banyak *noise* yang mengganggu model dalam pelatihan,

sehingga oleh karena itu perlu dilakukan adanya pembersihan data dalam proses *preprocessing data.* Proses *preprocessing data* yang dilakukan pada penelitian dijelaskan pada Gambar 3.3 berikut.



Gambar 3.3 Alur Pembersihan Data

Berdasarkan pada Gambar 3.3, merupakan alur dari pemrosesan pembersihan data. Proses pembersihan data akan dilakukan dengan beberapa jenis penghapusan. Proses penghapusan data-data tersebut merupakan tahapan *preprocessing* yang sangat penting dilakukan agar data menjadi bersih dan siap untuk digunakan. Proses pembersihan data yang dilakukan tersebut antara lain sebagai berikut.

a. Penanganan *missing value data*

Proses pembersihan data yang pertama ada penanganan *missing value* pada data. Data yang kosong akan mengganggu hasil dari kinerja model. Hal ini dilakukan karena model nantinya akan menghasilkan prediksi yang tidak akurat. Proses penangan *missing value* dilakukan dengan mengecek pada setiap kolom data dan mencari baris data yang memiliki nilai kosong. Kemudian akan dihapus agar model tidak menerima input yang tidak lengkap.

b. Penangan duplikasi data

Selain proses penanganan *missing value,* tahapan berikutnya dilakukan juga pemeriksaan terhadap duplikasi data. Duplikasi merupakan baris data yang sama persis dan muncul lebih dari satu kali. Adanya duplikasi ini dapat menyebabkan

bias dalam pelatihan model. Hal ini karena model dapat belajar terlalu banyak dari pola yang berulang secara tidak wajar. Dengan menghilangkan data yang duplikat tersebut, proses pelatihan nantinya tapi lebih baik dan data dapat lebih bersih untuk dilakukan pelatihan data.

c. Mengubah data menjadi huruf kecil

Proses ini bertujuan untuk mengubah semua huruf menjadi huruf kecil. Tujuan dari proses ini berguna untuk menghindari perbedaan makna antara kata yang sama, akan tetapi berbeda huruf seperti kapital. Hal ini karena nantinya akan dianggap satu hal yang sama oleh model.

Proses ini dilakukan dengan mengecek data yang mengandung huruf kapital, secara otomatis akan diubah menjadi *lowecase.* Contoh dari hasil penguunaan *lowercase* tersebut seperti pada Tabel 3.1 berikut.

Tabel 3.1 Proses *Lowercase*

Sebelum *Lowercase*	Sesudah Lowercase
@cutieuwwu Aku lupa waktu itu dengerin kajian ustad2 siapa. *Marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita :)!!. #viral #fyp https://t.co/r86fSWp9fo	@cutieuwwu aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita :)!!. #viral #fyp https://t.co/r86fswp9fo

d. Menghapus URL atau tautan

Proses ini bertujuan untuk mengubah semua data dengan menghilangkan tautan pada data yang tidak berguna dalam proses pelatihan nantinya. Proses tahapan penghapusan dalam *preprocessing* ini dilakukan agar data yang dihasilkan bersih dari adanya *noise,* seperti adanya URL yang dapat mengganggu dalam proses pelatihan model. Sehingga sebelum melakukan pengolahan data lebih lanjut perlu dilakukannya proses penghapusan URL ini. Berikut contoh dari proses menghapus URL pada Tabel 3.2 berikut.

Tabel 3.2 Proses Menghapus URL

Sebelum Hapus URL	Sesudah Hapus URL
@cutieuwwu aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita :)!!. #viral #fyp https://t.co/r86fswp9fo	@cutieuwwu aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita :)!!. #viral #fyp

e. Menghapus *mention* atau *tag* pengguna

Pada proses ini, semua data yang mengandung *mention* nama pengguna yang disebutkan dalam atau *tag* pada data akan dihapus. Penghapusan ini berguna agar data menjadi bersih dan mempermudah model dalam mempelajari data karena tidak terdapat *noise* sehingga dapat dilakukan secara optimal.

Penghapusan *mention* atau *tag* ini menggunakan *regex* sehingga akan mencari data yang mengandung hal tersebut, kemudian pola tersebut akan dihapus dari *dataset*. Contoh dari proses menghapus *mention* atau *tag* pada Tabel 3.3 berikut.

Tabel 3.3 Proses Menghapus *Mention*

Sebelum Hapus *Mention*	Sesudah Hapus *Mention*
@cutieuwwu aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita :)!!. #viral #fyp	aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita :)!!. #viral #fyp

f. Menghapus tanda *hashtag*

Proses berikutnya merupakan proses untuk menghapus tanda *hashtag* pada data. Proses ini berguna dalam membantu dalam data untuk *training* agar hasil data yang akan digunakan menjadi bersih dan siap pakai.

Penghapusan *hashtag* ini dilakukan karena *hashtag* mengandung simbol ”#” yang sering dianggap sebagai *noise* dalam analisis. Hal ini karena nantinya akan fokus ke makna kalimat, bukan kepada tren atau metadata. Contoh penghapusan *hashtag* sebagai berikut pada Tabel 3.4.

Tabel 3.4 Proses Menghapus *Hashtag*

Sebelum Hapus *Hashtag*	Sesudah Hapus *Hashtag*
aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita :)!!. #viral #fyp	aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita :)!!.

g. Menghapus simbol atau karakter non-alfanumerik

Pada proses ini, data yang mengandung simbol-simbol atau karakter yang bukan numerik akan dihapus. Proses penghapusan ini bertujuan untuk menghapus semua karakter yang bukan huruf, angka atau spasi karena karakter tersebut tidak memiliki kontribusi penting dalam makna kalimat, serta dapat menjadi *noise* dalam

analisis model nantinya.

Proses ini dilakukan dengan mencari semua karakter yang cocok dengan *regex*, kemudian setiap karakter yang cocok akan dihapus. Proses ini berguna untuk mempermudah dalam tokenisasi nantinya dan menghindari dari token yang tidak dikenal. Contoh dari penggunaan penghapusan simbol atau karakter non-alfanumerik pada Tabel 3.5.

Tabel 3.5 Proses Hapus Simbol

Sebelum Hapus Simbol	Sesudah Hapus Simbol
aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita :)!!.	aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita !!

h. Menghapus spasi berlebih dan memotong spasi di awal/akhir teks

Proses selanjutnya merupakan penghapusan spasi berebih serta memotong spasi di awal atau akhir teks data. Selain menghapus spasi yang berlebih dari data aslinya, kemudian bertujuan untuk merapikan spasi yang sering berantakan setelah menghapus karakter-karakter seperti simbol, *hashtag*, URL dan proses lainnya.

Proses ini dilakukan dengan mencari satu atau lebih karakter spasi, kemudian diganti semua kumpulan tersebut menjadi satu spasi. Selain itu pada proses ini akan menghapus spasi di awal dan akhir dari kalimat. Contoh dari proses ini pada Tabel 3.6 sebagai berikut.

Tabel 3.6 Proses Menghapus Spasi Berlebih

Sebelum Hapus Spasi	Sesudah Hapus Spasi
aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita !!	aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita !!

i. Menghapus angka

Proses berikutnya merupakan proses untuk menghapus semua angka yang terkandung dalam teks pada data. Proses ini dilakukan bertujuan karena angka dianggap tidak relevan untuk pemrosesan selanjutnya, sehingga dapat menghindari bias pada data.

Proses penghapusan ini dilakukan dengan mencari pada data yakni satu digit angka atau lebih yang kemudian akan dihapus pada seluruh data. Contoh penggunaan penghapusan angka tersebut pada Tabel 3.7 sebagai berikut.

Tabel 3.7 Proses Menghapus Angka

Sebelum Hapus Angka	Sesudah Hapus Angka
aku lupa waktu itu dengerin kajian ustad2 siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita !!	aku lupa waktu itu dengerin kajian ustad siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita !!

j. Menghapus tanda baca (*punctuation*)

Proses berikutnya merupakan penghapusan tanda baca pada seluruh data. Hal ini dilakukan karena tanda baca sering tidak menambah makna dalam model, serta dapat membuat hasil tokenisasi tidak konsisten. Selain itu berguna untuk mengurangi *noise* pada data.

Proses yang dilakukan ketika terdapat kumpulan tanda baca pada data, kemudian proses akan membaca teks dan menghapus semua karakter yang ada di dalam data tersebut. Contoh dari penerapan menghapus tanda baca pad Tabel 3.8 sebagai berikut.

Tabel 3.8 Proses Menghapus Tanda Baca

Sebelum Hapus Tanda Baca	Setelah Hapus Tanda Baca
aku lupa waktu itu dengerin kajian ustad siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita !!	aku lupa waktu itu dengerin kajian ustad siapa. *marriage is scary* itu bisa jadi usaha setan untuk menyesatkan kita

3.3.3. Pembagian *Dataset*

Pada proses ini *dataset* keduanya antara Indonesia dan Amerika Serikat yang telah dibersihkan pada masing-masing *dataset* akan dibagi menjadi dua *subset*. Proses ini dilakukan secara terpisah untuk masing-masing negara agar bertujuan untuk menjaga proporsi data dan konteks yang terdapat dalam tiap *subset*. Pembagian ini memiliki dua tujuan yakni:

1. *Subset* untuk pelatihan pada sebagian data dari masing-masing negara yang digunakan untuk melatih model.

2. *Subset* untuk analisis sentimen lanjutan, yakni sisa data tersebut disimpan terpisah dan tidak digunakan dalam pelatihan. *Subset* ini digunakan untuk melakukan analisis sentimen lebih lanjut setelah model selesai dilatih.

3.3.4. Pelabelan *Dataset*

Proses berikutnya merupakan proses pelabelan data. Dalam penelitian ini, pelabelan memanfaatkan *labeling* otomatis dengan menggunakan *pretrained* BERT yang diterapkan pada tiap-tiap data dari kedua negara yang disesuaikan dengan bahasa masing-masing data. Penggunaan ini dilakukan agar dapat memberikan hasil label sentimen yang efektif dengan dikategorikan ke dalam dua kategori yakni *positive* dan *negative*.

a. Pelabelan data Indonesia

Pelabelan pada *dataset* Indonesia dilakukan dengan menggunakan *pretrained* BERT dari *hugging face.* Proses pelabelan dilakukan dengan menggunakan jenis *pretrained* BERT, yakni *Aardiiiiy/indobertweet-base-Indonesian-sentiment-analysis.* Model tersebut merupakan sebuah model BERT yang telah dilatih secara khusus untuk analisis sentimen dalam bahasa Indonesia, khususnya pada media sosial. Model ini mengklasifikasikan sentimen menjadi tiga jenis yakni negatif, netral, dan positif. Akan tetapi pada penelitian ini hanya menggunakan dua kelas saja yakni positif dan negatif.

Proses pelabelan yang dilakukan dengan menerapkan tokenisasi pada setiap teks menggunakan *tokenizer* dari model tersebut. Kemudian, model memprediksi label sentimen beserta *confidence score*-nya melalui fungsi *softmax* pada *output logits*. Hasil prediksi label dan *confidence* disimpan ke dalam dua kolom baru dalam *dataset*.

b. Pelabelan data Amerika Serikat

Proses pelabelan pada *dataset* Amerika Serikan dilakukan sama seperti *dataset* Indonesia, yakni dengan pelabelan otomatis menggunakan *pretrained* BERT dari *hugging face*. Jenis model yang digunakan adalah *finiteautomata/bertweet-base-sentiment-analysis*, yaitu merupakan BERTweet yang telah dilatih untuk analisis sentimen dalam bahasa Inggris, terutama pada teks-teks dari media sosial.

Proses pelabelan dilakukan serupa dengan *dataset* Indonesia, yakni menggunakan *tokenizer* dan model untuk menghasilkan prediksi sentimen dan

confidence. Sama seperti sebelumnya, kelas sentimen yang digunakan adalah *negative* dan *positive*.

3.3.5. Penanganan Data Tidak Seimbang

Pada proses pengolahan data setelah pelabelan terdapat hasil label kelas yang tidak seimbang secara signifikan yakni pada *dataset* Indonesia. Hal ini dimana terdapat dominasi kelas sentimen negatif terhadap sentimen positif, sehingga perlu dilakukannya proses augmentasi pada data kelas minoritas yakni positif. Untuk *dataset* Amerika Serikat tidak dilakukan augmentasi karena sebaran data masih dalam lingkup cukup.

Dalam melakukan penyeimbangan *dataset* tersebut, proses memanfaatkan teknik augmentasi dengan menggunakan *IndoBERT fill-Mask.* Langkah ini bertujuan agar meningkatkan jumlah data kelas positif agar setara dengan kelas negatif sehingga model tidak condong kepada kelas mayoritas saat pelatihan. Langkah-langkah yang dilakukan sebagi berikut.

1. Normalisasi teks

Langkah pertama yang dilakukan adalah normalisasi teks pada data terlebih dahulu. Normalisasi ini bertujuan untuk memperbaiki struktur kalimat agar lebih sesuai dengan pemahaman model bahasa IndoBERT.

2. Menerapkan teknik *fill-mask*

Langkah berikutnya merupakan penggunaan teknik *fill-mask* dari model *IndoBERT* untuk menghasilkan kalimat baru dari data asli. Teknik ini bekerja dengan menyisipkan token khusus “[MASK]” secara acak ke dalam kalimat, kemudian memprediksi kata yang paling sesuai untuk mengisi posisi tersebut. Sehingga hasilnya tersebut merupakan kalimat baru yang memiliki makna serupa namun struktur atau pilihan kata yang sedikit berbeda [45].

Proses pada tahapan ini dilakukan secara berulang hingga jumlah positif yang dihasilkan mencukupi untuk menyamai jumlah data negatif. Agar hasil augmentasi baik, kemudian dilakukan juga penghapusan duplikasi baik di dalam data augmentasi itu sendiri maupun terhadap data asli. Kemudian setelah selesai data akan digabungkan dengan data asli.

3.3.6. Pembagian Data Untuk Pelatihan

Langkah selanjutnya adalah melakukan pembagian data untuk melakukan pelatihan model. Pembagian ini dibagi menjadi tiga *subset* yakni data latih, data validasi, dan data uji, dengan pembagian dilakukan menggunakan rasio 80:10:10.

Penggunaan data latih bertujuan untuk melatih model agar dapat mempelajari pola, hubungan antar kata, serta fitur-fitur penting dalam kata. Sedangkan data validasi bertujuan untuk mengevaluasi performa model sementara, seperti membantu mencegah *overfitting*, sehingga dapat memberikan pemahaman dalam menentukan parameter terbaik. Serta data uji bertujuan untuk memberikan evaluasi akhir atas kinerja model setelah melakukan proses pelatihan.

Penggunaan pembagian data ini menggunakan *stratify,* dimana berguna untuk membagi data dengan berdasarkan jumlah sentimennya. Sehingga jumlah sentimen tiap *subset* memiliki jumlah yang ideal.

3.3.7. Tokenisasi Data

Pada tahapan ini merupakan tahapan tokenisasi pada *dataset*, baik *dataset* Indonesia maupun Amerika Serikat dengan memanfaatkan *tokenizer bert-base-multilingual-cased* dari *hugging face.* Proses tokenisasi ini bertujuan untuk mengubah teks mentah menjadi format numerik yang biasa dipahami oleh model. Tahapan dari tokenisasi tersebut terdiri dari.

1. *Tokenization*

Proses tokenisasi ini dilakukan dengan memecah kata menjadi sub-kata jika kata tersebut tidak ada dalam kosakata. Pada sub-token diawali dengan “##” jika merupakan bagian dari kata sebelumnya. Penggunaan ini dapat membantu dalam menangani kata-kata baru atau langka yang tidak ada dalam kosakata model.

Proses tokenisasi dilakukan dengan secara otomatis membaca teks dan memecahnya menjadi sub-token. Contoh dari hasil tokenization ini seperti pada Tabel 3.9 berikut.

Tabel 3.9 Proses Tokenisasi

Sebelum Tokenisasi	Sesudah Tokenisasi
lebih ke sadar kalo masih miskin sih gue juga takut nikah karena ekonomi yang gak stabil	['lebih', 'ke', 'sada', '##r', 'ka', '##lo', 'masih', 'mis', '##kin', 'si', '##h', 'gu', '##e', 'juga', 'tak', '##ut', 'ni', '##kah', 'karena',

Sebelum Tokenisasi	Sesudah Tokenisasi
	'ekonomi', 'yang', 'ga', '##k', 'stabil']

2. Penambahan Spesial Token

Tokenisasi yang dilakukan menggunakan BERT ini, dimana setiap input tidak langsung diberikan dalam bentuk teks biasa, tetapi harus disiapkan dalam format khusus menggunakan token spesial. Tujuan ini karena BERT perlu tahu dimana kalimat dimulai dan diakhiri, kemudian kalimat mana yang sedang diproses, serta input yang terdiri dari satu atau dua kalimat. Contoh dari penggunaan token ini pada Tabel 3.10 sebagi berikut.

Tabel 3.10 Proses Penambahan Spesial Token

Sebelum Penambahan Spesial Token	Sesudah Penambahan Spesial Token
['lebih', 'ke', 'sada', '##r', 'ka', '##lo', 'masih', 'mis', '##kin', 'si', '##h', 'gu', '##e', 'juga', 'tak', '##ut', 'ni', '##kah', 'karena', 'ekonomi', 'yang', 'ga', '##k', 'stabil']	['[CLS]', 'lebih', 'ke', 'sada', '##r', 'ka', '##lo', 'masih', 'mis', '##kin', 'si', '##h', 'gu', '##e', 'juga', 'tak', '##ut', 'ni', '##kah', 'karena', 'ekonomi', 'yang', 'ga', '##k', 'stabil', '[SEP]']

3. *Truncation* dan *Padding*

Pada tahapan ini dilakukan *truncation* yakni proses memotong jumlah token input jika panjangnya melebihi yang ditentukan. Tujuan dari dilakukan ini adalah untuk semua input memiliki panjang maksimal tertentu dan model tidak menerima input yang terlalu panjang. Sedangkan *padding* berguna untuk proses menambahkan token khusus ke input agar semua input memiliki panjang yang sama. Sehingga penggunaan ini membantu agar tidak ada *noise* pada data. Contoh dari penerapan tersebut yakni pada Tabel 3.11 sebagai berikut.

Tabel 3.11 Proses *Truncation* dan *Padding*

Sebelum *Truncation* dan *Padding*	Sesudah *truncation* dan *Padding*
['[CLS]', 'lebih', 'ke', 'sada', '##r', 'ka', '##lo', 'masih', 'mis', '##kin', 'si', '##h', 'gu', '##e', 'juga', 'tak', '##ut', 'ni', '##kah', 'karena', 'ekonomi', 'yang', 'ga', '##k', 'stabil', '[SEP]']	[101, 13394, 11163, 54651, 10129, 10730, 10715, 20535, 12606, 12130, 10294, 10237, 75980, 10112, 11535, 12741, 11159, 10414, 28977, 15786, 34727, 10265, 11887, 10174, 71338, 102, 0, 0, 0, 0, 0, 0]

4. *Attention Mask*

Pada proses ini mengubah menjadi *vektor* biner yang berisi angka "0" dan angka "1". Attention Mask merupakan vektor biner yang dihasilkan dari proses tokenisasi BERT, dimana dengan nilai "1" untuk token yang merupakan bagian dari teks asli dan "0" untuk token *padding*. Tujuannya adalah memberi tahu model BERT token mana yang harus diperhatikan dan token mana yang perlu diabaikan, sehingga saat proses *self-attention*, model hanya fokus pada token yang relevan dan tidak terpengaruh oleh *padding*. Berikut contoh dari penerapan tersebut pada Tabel 3.12 sebagai berikut.

Tabel 3.12 Proses *Attention Mask*

Sebelum *Attention Mask*	Sesudah *Attention Mask*
[101, 13394, 11163, 54651, 10129, 10730, 10715, 20535, 12606, 12130, 10294, 10237, 75980, 10112, 11535, 12741, 11159, 10414, 28977, 15786, 34727, 10265, 11887, 10174, 71338, 102, 0, 0, 0, 0, 0, 0]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0]

3.3.8. Transformasi Data Sebelum Pelatihan

Sebelum dilakukannya pelatihan model, *dataset* terlebih dahulu harus dilakukan perubahan agar format data sesuai dengan model. Tujuan perubahan ini adalah agar *dataset* sesuai dengan format model, sehingga model dapat melakukan pelatihan secara optimal.

1. Mengubah kolom *sentiment* menjadi *label*

Pada tahapan sebelum pelatihan, perlu adanya penyesuaian terhadap struktur *dataset* agar dapat selaras dengan format input yang dibutuhkan pembelajaran model. Proses tersebut yakni mengubah nama kolom *sentiment* menjadi *label*. Hal ini dilakukan karena konvensi umum dalam pustaka *transformers* dan *datasets* dari *hugging face*, di mana kolom target klasifikasi harus diberi nama label agar dikenali secara otomatis oleh fungsi pelatihan model.

Selain itu, kolom *text* dihapus dari setiap *subset dataset*, yaitu data latih (*train_dataset*), data validasi (*val_dataset*), dan data uji (*test_dataset*). Penghapusan ini dilakukan karena data teks mentah sudah tidak ada lagi dibutuhkan pada tahap pemodelan, karena seluruh teks telah ditokenisasi ke dalam bentuk numerik.

2. Mengubah label menjadi numerik

Setelah kolom *sentiment* di ubah namanya menjadi label. Proses kemudian berlanjut pada tahapan berikutnya. Pada langkah ini dilakukan pemrosesan untuk mengubah label menjadi bentuk numerik. Kategori positif dan negatif pada kolom label diubah menjadi bentuk numerik. Hal ini dilakukan karena sebagian besar *machine learning*, termasuk model BERT hanya dapat memproses data dalam bentuk numerik bukan dalam bentuk string atau teks.

Dalam *dataset,* isi data pada kolom label berupa *positive* dan *negative*. Oleh karena itu, dibuat sebuah *mapping* atau pemetaan label ke dalam format angka, di mana *negative* dikodekan sebagai 0 dan *positive* sebagai 1. Proses ini dilakukan secara otomatis pada seluruh data latih (*train_dataset*), data validasi (*val_dataset*), dan data uji (*test_dataset*) dengan menggunakan fungsi *map()* dari pustaka *Hugging face Dataset*s.

Dengan mengubah label menjadi angka tersebut, model dapat lebih mudah memproses dan mempelajari pola dari data selama proses pelatihan. Selain itu, representasi numerik ini juga mempermudah dalam perhitungan metrik evaluasi, seperti akurasi, *presicion*, dan *recall* yang digunakan untuk menilai kinerja model.

3.4. *Modeling*

Setelah data telah dipersiapkan, tahap selanjutnya adalah membangun model untuk pelatihan pada data yang sudah siap digunakan sebelumnya. Model yang telah dibangun berguna untuk menganalisis data *tweet* untuk mengetahui pola-pola atau tren dari sentimen terkait isu *marriage is scary.*

3.4.1. *Fine-tuning* Model Pada Pelatihan

Pada tahap penelitian ini merupakan tahapan pelatihan data. Proses pelatihan data dilakukan dengan menggunakan *fine-tuning* model, dengan memanfaatkan arsitektur model *pretrained* BERT yang telah dilatih sebelumnya. Proses dari *fine-tuning* ini adalah untuk mengklasifikasikan sentimen teks negara Indonesia serta Amerika Serikat.

Pada penelitian ini, pelatihan memanfaatkan model *multilangual* BERT yang telah dilatih sebelumnya, yaitu *bert-base-multilingual-cased* dari *hugging face*. Model ini dipilih karena mendukung berbagai bahasa termasuk bahasa Indonesia dan Inggris, selain itu telah melalui *pretraining* dengan data dalam banyak bahasa,

sehingga cocok untuk *training* data berbeda bahasa. *Fine*-tunnning yang dilakukan pada penelitian ini terdiri atas beberapa hal berikut.

1. Pemilihan Model Serta Tokenizer

Model *bert-base-multilingual-cased* diunduh melalui *library Transformers* dari *hugging face*. Bersama dengan modelnya, kemudian digunakan juga *tokenizer* resmi BERT *multilingual* untuk melakukan tokenisasi teks menjadi format yang bisa dipahami oleh model. *Tokenizer* ini secara otomatis menambahkan token khusus seperti [CLS] dan [SEP], melakukan *padding* dan *truncation*, serta mengeluarkan ID numerik untuk setiap token.

2. Pelatihan Model Untuk Data Indonesia Dan Amerika Serikat

Pelatihan model dilakukan dengan *corpus* yang berbeda yakni antara data Indonesia dengan Amerika Serikat. Eksperimen pembangunan model pada penelitian ini terdiri dari 3 skenario untuk data Indonesia, serta 4 skenario untuk data Amerika Serikat. Eksperimen-eksperimen ini dilakukan dengan menggunakan jenis-jenis *fine-tuning* yang berbeda. Tujuan dilakukan eksperimen *fine-tuning* pada kedua data tersebut dapat menemukan model terbaik pada pelatihan di tiap data, antara data Indonesia dan Amerika Serikat. Skema eksperimen dilakukan dengan disesuaikan pada karakteristik masing-masing data. Pada Indonesia, fokus utama adalah memulihkan representasi kelas minoritas melalui augmentasi sehingga eksperimen difokuskan pada pengaturan *epoch* untuk memantau stabilitas dan risiko *overfitting* setelah dilakukan augmentasi. Sedangkan pada Amerika Serikat, fokusnya adalah mengelola jumlah data antar kelas positif dan negatif yang sedikit berbeda dengan *gap* yang masih wajar. Sehingga dari pelatihan pada kedua data tersebut dapat memiliki model pelatihan terbaiknya sendiri. Maka dari itu disusunlah skema eksperimen pada penelitian ini.

a) Pada Data Indonesia

Pada data indonesia, eksperimen dilakukan dengan 3 jenis eksperimen yakni dilakukan dengan menggunakan jenis *hyperparameter* yang berbeda. Eksperimen-eksperimen yang dilakukan pada data Indonesia dengan dilakukan menggunakan

hyperparameter berikut pada Tabel 3.13.

Tabel 3.13 Eksperimen Pelatihan Data Indonesia

Eksp ke-	*Epoch*	*Learning rate*	*Batch Size*	*Weight decay*	*Eval strategy*	*Save strategy*	*Logging strategy*
1	3	1e-5	16	0.01	50	50	50
2	5	1e-5	16	0.01	50	50	50
3	8	1e-5	16	0.01	50	50	50

b) Pada Data Amerika Serikat

Berbeda dengan data Indonesia, data Amerika Serikat dilakukan pelatihan dengan menggunakan *hyperparameter* yang berbeda pada setiap eksperimennya. Pelatihan dilakukan dengan 4 kali eksperimen. *Hyperparameter* pada eksperimen ini dapat dilihat pada Tabel 3.14 berikut.

Tabel 3.14 Eksperimen Pelatihan Data Amerika Serikat

Eksp ke-	*Epoch*	*Learning rate*	*Batch Size*	*Class Weight*	*Threshold*	*Early Stopping*
1	3	1e-5	16	Tidak Ada	Default (*argmax*)	Tidak Ada
2	5	1e-5	16	Otomatis (*compute class weight)*	Default (*argmax*)	Ya
3	5	1e-5	16	Manual ([1.5, 1.0])	0.6	Ya
4	5	1e-5	16	Manual ([1.5,1.0])	0.5	Ya

Dari semua percobaan eksperimen tersebut untuk melihat hasil *tunning* pelatihan akan dievaluasi dengan *Weights & Biases* (WandB). Tujuan menggunakan tersebut adalah agar eksperimen dapat terlacak secara otomatis.

3.5. *Evaluation*

Setelah model selesai dibangun, kemudian dilakukan evaluasi terhadap hasil dari kinerja model tersebut agar memastikan bahwa hasil dari analisis sentimen memiliki akurasi baik dan relevan. Evaluasi merupakan tahapan penting dalam proses pemodelan untuk mengetahui seberapa baik kinerja model dalam melakukan prediksi. Proses evaluasi dilakukan pada data uji, sehingga dapat mencerminkan kemampuan generalisasi model terhadap data baru.

3.5.1. Evaluasi Hasil Kinerja Model

Evaluasi dilakukan dengan menggunakan beberapa metrik umum dalam klasifikasi dalam menampilkan hasil evaluasi dari model. Jenis evaluasi yang digunakan dalam penelitian ini menggukan beberapa metrik sebagai berikut.

a. Metrik evaluasi selama *training*: Penggunaan metrik *accuracy, precison, recall,* dan *F1-score.*
b. Metrik evaluasi setelah *training: precision* per kelas, *recall* per kelas, *F1-score* per kelas, *accuracy* dan *support* (jumlah sampel masing-masing kelas).
c. *Confusion Matrix:* menghitung prediksi besar dan salah.

3.5.2. Analisis Sentimen dan Perbandingan Hasil Analisis

Pada tahap ini, dilakukan analisis sentimen terhadap opini publik terkait isu "*marriage is scary*" yang berasal dari dua negara, yaitu Indonesia dan Amerika Serikat. Analisis ini bertujuan untuk memahami bagaimana persepsi masyarakat yang berada pada usia produktif untuk menikah, yaitu generasi muda. Secara konseptual, isu ini dikaitkan dengan Generasi Z dan Milenial, dimana menurut Badan Pusat Statistik (BPS) pada tahun 2024 menyatakan bahwa Generasi Z merupakan pemuda yang lahir pada tahun 1997 sampai 2012, dengan perkiraan usia sekarang 13 sampai 28 tahun. Sedangkan milenial merupakan generasi yang lahir pada tahun 1981 sampai 1996, sehingga sekarang berusia 29 sampai 44 tahun [46]. Dimana generasi tersebut yang dikenal aktif menggunakan media sosial dalam menyampaikan opini pribadi maupun pandangan sosial di masing-masing negara terhadap isu tersebut di media sosial X. Seperti yang disampaikan Statista bahwa jumlah pengguna media sosial X berdasarkan kelompok umur pada bulan Februari 2025 di seluruh dunia, kelompok usia 25 sampai 34 tahun merupakan yang paling banyak menggunakan sosial media X sebesar 37,5 persen, kemudian usia 18 sampai 24 tahun sebesar 32,1 persen dan usia 35 sampai 49 tahun sebesar 21,1 persen [47].

Selain itu, Badan Pusat Statistik (BPS) pada tahun 2024 juga menyatakan bahwa sebanyak 69,75 persen pemuda di Indonesia berstatus belum kawin, 29,10 persen berstatus kawin, serta 1,15 persen berstatus cerai hidup atau cerai mati. Kelompok umur 16 sampai 18 tahun terdiri atas 98,67 persen yang berstatus belum kawin, 1,24 persen berstatus kawin, dan 0,09 persen berstatus cerai hidup/cerai

mati; pada kelompok umur 19 sampai 24 tahun terdapat 83,68 persen berstatus belum kawin, 15,59 persen berstatus kawin, serta 0,73 persen berstatus cerai hidup/cerai mati; sedangkan pada kelompok umur 25 sampai 30 tahun sebanyak 40,03 persen berstatus belum kawin, 57,82 persen berstatus kawin, dan 2,15 persen berstatus cerai hidup/cerai mati [48]. Selanjutnya hasil dari survei yang dilakukan Generasi Berencana (GenRe) pada tahun 2024 yang terdiri dari 1825 responden berusia dari 21 sampai 24 tahun menyatakan mayoritas takut menikan, sedangkan sekitar 26 persen saja yang tidak takut menikah, sisanya memiliki sikap ragu-ragu sampai sangat takut menikah [49].

Kemudian *Pew Research Center* menyatakan bahwa di Amerika Serikat terdapat 69 persen berusia 18 hingga 34 tahun belum menikah dan mengatakan ingin menikah, kemudian 23 persen mengatakan tidak yakin, dan 8 persen mengatakan tidak ingin menikah. Selain itu, 20 persen mengatakan bahwa menikah sangat atau penting bagi seseorang dalam menjalani hidup yang memuaskan, dan 22 persen mengatakan memiliki anak sangat penting [50]. Maka dari itu, analisis sentimen ini untuk mengetahui sentimen opini publik pada pengguna media sosial X di kedua negara terhadap respon "*marriage is scary"*.

Data dari kedua negara, kemudian diproses melalui tahapan tokenisasi, pelabelan, dan inferensi menggunakan model BERT yang telah dilatih. Hasil prediksi sentimen kemudian dikategorikan ke dalam dua kelas, yaitu positif dan negatif. Selanjutnya, hasil analisis dari kedua negara dibandingkan untuk melihat perbedaan distribusi sentimen. Perbandingan ini dilakukan dengan menghitung persentase sentimen positif dan negatif dari masing-masing negara, serta memvisualisasikan hasil tersebut.

1. Diagram batang perbandingan jumlah kategori di kedua negara.
2. Visualisasi demografi berdasarkan sentimen pada tiap negara