Comparative Analysis of BERT-Based and Traditional Machine Learning Approaches for Climate Change Sentiment Classification on Twitter

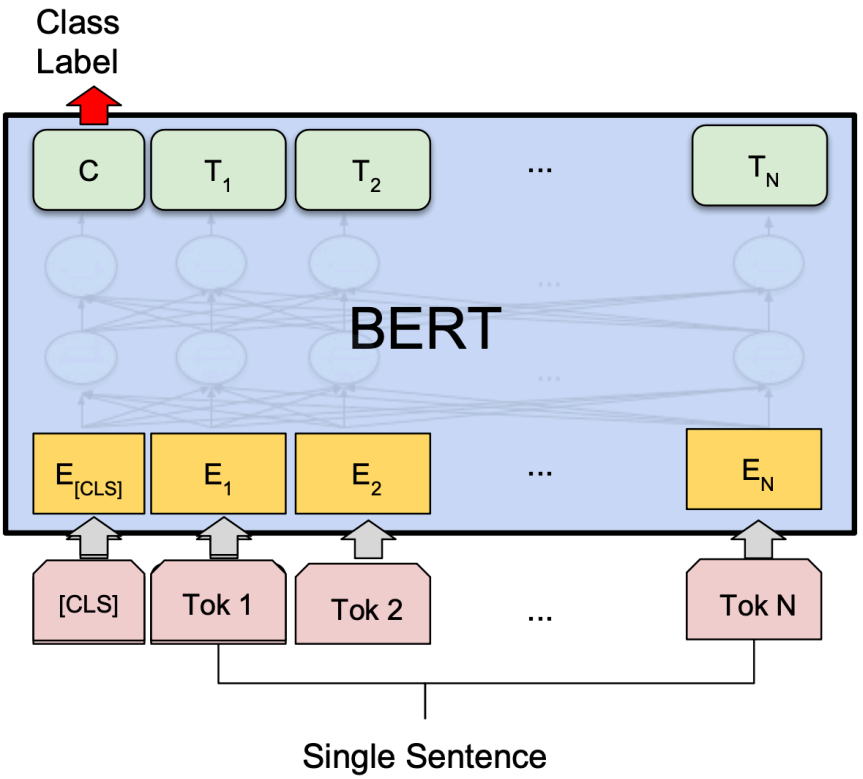
Nama Kelompok	Regu Elang
Nomor Kelompok	2
Anggota Kelompok	Nafis Naufal Rahman
	Taqiyudin Miftah Adn

1. Latar Belakang

Perubahan iklim merupakan isu global yang semakin mendesak dan menjadi perhatian utama di berbagai bidang, termasuk lingkungan, ekonomi, dan kebijakan publik. Media sosial seperti Twitter telah menjadi wadah utama bagi masyarakat untuk mengekspresikan opini, pandangan, dan emosi mereka terhadap isu ini. Analisis sentimen terhadap percakapan publik di Twitter dapat memberikan wawasan berharga mengenai persepsi masyarakat terhadap perubahan iklim, termasuk tingkat kesadaran, kekhawatiran, serta dukungan terhadap upaya mitigasi dan adaptasi yang dilakukan. Oleh karena itu, memahami sentimen publik terkait perubahan iklim memiliki relevansi yang kuat dalam domain lingkungan.

Dalam konteks tersebut, pendekatan berbasis Artificial Intelligence (AI) dan Deep Learning menjadi solusi yang menjanjikan untuk mengolah data teks dalam jumlah besar dengan tingkat kompleksitas tinggi. Model berbasis transformer seperti BERT (Bidirectional Encoder Representations from Transformers) memiliki kemampuan memahami konteks dan makna kata secara mendalam dalam kalimat, sehingga mampu mengatasi keterbatasan model tradisional yang hanya bergantung pada representasi

statistik seperti *bag-of-words* atau *CountVectorizer dan TF-IDF*. Meskipun pendekatan *traditional machine learning* (seperti Random Forest, Naive Bayes, atau Logistic Regression) masih banyak digunakan dan bergantung pada proses *preprocessing* yang ekstensif—seperti stemming, lemmatization, dan pembersihan teks—hasil penelitian menunjukkan bahwa model BERT tetap memberikan performa yang lebih baik bahkan dengan *preprocessing* yang minimal.



Gambar BERT architecture

Namun demikian, masih terdapat sejumlah kesenjangan (gap) dalam penelitian sebelumnya. Penelitian yang dilakukan oleh Anoop V. S. dkk. (2024) dalam *Climate Change Sentiment Analysis Using Domain Specific Bidirectional Encoder Representations From Transformers* menunjukkan beberapa keterbatasan yang perlu diperhatikan. Studi tersebut menerapkan teknik *oversampling* sebelum proses *train-test split*, yang dapat menyebabkan hasil evaluasi menjadi bias dan tidak merepresentasikan

performa model secara aktual. Selain itu, penelitian tersebut menggunakan jumlah data yang relatif kecil sekitar 5.500 tweet, belum mengeksplorasi variasi *n-grams* secara optimal pada pendekatan *traditional machine learning*, serta kurang memberikan penjelasan mendalam mengenai proses *fine-tuning* pada model BERT. Oleh karena itu, penelitian ini dilakukan untuk memberikan perbandingan yang lebih objektif antara pendekatan tradisional berbasis *CountVectorizer* dengan algoritma *machine learning* klasik, dan pendekatan modern berbasis BERT dalam menganalisis sentimen publik terhadap perubahan iklim menggunakan dataset Twitter Climate Change Sentiment dari Kaggle yang lebih besar dan beragam.

2. Tujuan

Tujuan utama dari penelitian ini adalah untuk melakukan analisis komprehensif terhadap sentimen publik mengenai perubahan iklim di Twitter dengan memanfaatkan pendekatan *Natural Language Processing* (NLP) berbasis *deep learning* dan *traditional machine learning*. Penelitian ini bertujuan untuk membandingkan performa model BERT yang telah di-*fine-tuning* (menggunakan *vinai/bertweet-base*) dengan empat model *machine learning* konvensional, yaitu Logistic Regression, Multinomial Naive Bayes, Ridge Classifier, dan Random Forest, yang diimplementasikan dengan representasi teks *CountVectorizer, TF-IDF, dan Word2Vec*. Perbandingan ini dilakukan untuk mengevaluasi keunggulan model berbasis transformer dalam memahami konteks linguistik yang kompleks dibandingkan pendekatan statistik tradisional yang bergantung pada proses *preprocessing* dan representasi kata yang terbatas.

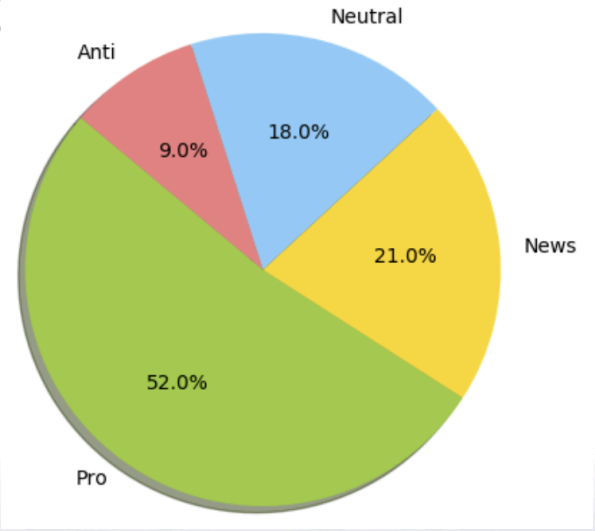
Selain aspek performa model, penelitian ini juga bertujuan untuk melakukan eksplorasi data (EDA) guna memahami distribusi sentimen, pola kata yang dominan, serta potensi bias dalam dataset *Twitter Climate Change Sentiment* dari Kaggle. Dengan demikian, hasil penelitian ini diharapkan dapat memberikan perbandingan objektif, meningkatkan akurasi prediksi sentimen, serta memberikan wawasan mengenai

faktor-faktor linguistik dan metodologis yang mempengaruhi hasil analisis sentimen terkait isu perubahan iklim di media sosial.

3. Dataset

Penelitian ini menggunakan dataset Twitter Climate Change Sentiment Dataset yang bersumber dari Kaggle. Dataset ini berisi kumpulan tweet yang mengekspresikan opini publik terhadap isu perubahan iklim, dengan total 43.943 entri data. Setiap entri terdiri atas dua atribut utama, yaitu *message* (teks tweet) dan *sentiment* (label sentimen), serta satu kolom tambahan *tweet_id* sebagai identifikasi unik.

Dataset ini dikategorikan ke dalam empat label sentimen, yaitu -1 (negative) sebanyak 3.990 data, 0 (neutral) sebanyak 7.715 data, 1 (positive) sebanyak 22.962 data, dan 2 (news) sebanyak 9.276 data. Distribusi label menunjukkan adanya ketidakseimbangan yang cukup signifikan (*class imbalance*), di mana kelas *positive* mendominasi sekitar setengah dari total data, sementara kelas *negative* memiliki jumlah data paling sedikit. Tweet yang terdapat dalam dataset bersifat multilingual, namun pada tahap pemrosesan awal, hanya tweet berbahasa Inggris yang dipertahankan agar model dapat dilatih secara konsisten menggunakan representasi bahasa yang sama.



Visualisasi distribusi kelas

Tahapan Pra-pemrosesan (Preprocessing)

1. Penanganan Duplikasi

   Menghapus tweet duplikat berdasarkan isi pesan menggunakan fungsi `df.drop_duplicates()`.

2. Pembersihan Teks

   Melakukan pembersihan simbol dan karakter tidak lazim (misalnya Ã¢â‚¬â„¢, Ã¢â‚¬Å“, Ã¢â‚¬Â¦), menghapus karakter non-ASCII, awalan "RT @", URL, string aneh seperti $q$$q$, serta menghapus spasi berlebih.

3. Deteksi Bahasa

   Menggunakan pustaka *langdetect* untuk mengidentifikasi bahasa, kemudian mempertahankan hanya tweet dengan bahasa Inggris (*detected_language == 'en'*).

4. Penghapusan Label "News"

   Tweet dengan label sentimen 2 (news) dihapus agar fokus analisis tertuju pada opini publik (positif, netral, negatif).

5. Pembuatan Subset

   Membuat *DataFrame* baru berisi kolom *message* dan *sentiment*, serta menghapus duplikasi berdasarkan kombinasi kedua kolom tersebut.

6. Penyimpanan Data Bersih

   Menyimpan hasil pembersihan ke dalam file `cleaned_tweets.csv` sebagai dataset akhir yang siap digunakan untuk pelatihan model.

Baik pendekatan traditional machine learning maupun deep learning (BERTweet) menggunakan dataset yang sama setelah tahap pra-pemrosesan tersebut. Untuk model tradisional, dilakukan tahapan

tambahan berupa *feature extraction* menggunakan TF-IDF, CountVectorizer, dan Word2Vec, serta transformasi teks melalui *tokenization*, *stemming*, *lemmatization*, dan penghapusan *stop words*. Sementara itu, pendekatan BERTweet menerapkan *tokenization* dan *embedding* berbasis transformer dengan proses *fine-tuning* pada model vinai/bertweet-base.

4. Metodologi & Eksperimen

Jelaskan informasi terkait metodologi dan eksperimen yang dilakukan. Misalnya:

- Arsitektur Model:

Model	Basis Arsitektur	Param (≈)	Layers	Hidden	Heads	Kelebihan Utama
Roberta-large -mnli	RoBERTa - Large	355 M	24	1024	16	Semantik kuat, zero shot, via NLI
distilroberta-base-climate-detector	DistillRoberta	82M	6	768	12	Khusus domain iklim
bertweet-base	RoBERTa-base	135M	12	768	12	Dioptimalkan untuk data Twitter

1. FacebookAI/roberta-large-mnli
   RoBERTa-Large yang sudah di-fine-tune pada tugas NLI (entailment/contradiction/neutral), model ini juga kuat dalam memahami relasi semantik antar kalimat dikarenakan parameternya yang besar.
2. climatebert/distilroberta-base-climate-detector

Model ini adalah DistilRoBERTa yang lebih ringan dan sudah diadaptasi pada domain iklim.

3. vinai/bertweet-base
   Model ini dilatih secara khusus pada 850 juta tweet berbahasa Inggris. Model ini juga dilatih dengan Penuh dengan *Slang* dan Singkatan Kata-kata seperti "LOL", "IDK", "IMO", menggunakan Emoji, dan mengandung @mentions dan #hashtags, Struktur kalimat menjadi tidak standar. Gaya Bahasa Informal: Tata bahasa dan ejaan seringkali tidak sempurna.

- Baseline:

Untuk memberikan perbandingan yang objektif terhadap pendekatan *deep learning* berbasis BERT, penelitian ini menggunakan beberapa model konvensional (baseline) berbasis *traditional machine learning*. Empat algoritma yang digunakan adalah Logistic Regression, Ridge Classifier, Multinomial Naive Bayes, dan Random Forest. Keempat model tersebut dipilih karena mewakili beragam paradigma pembelajaran, yaitu model linear, probabilistik, dan *ensemble*, yang umum digunakan dalam tugas klasifikasi teks.

Setiap model dilatih menggunakan tiga jenis teknik representasi teks, yaitu TF-IDF, CountVectorizer, dan Word2Vec, sehingga total dilakukan 9 kombinasi pelatihan (3×3). Selain itu, dilakukan satu eksperimen tambahan dengan data yang telah dibersihkan lebih lanjut, menggunakan CountVectorizer, teknik SMOTE Tomek untuk penyeimbangan kelas, serta optimasi hiperparameter melalui GridSearchCV.. Model-model tersebut dievaluasi menggunakan metrik yang dihasilkan dari *classification report*, meliputi akurasi, presisi, recall, dan F1-score (weighted).

Hasil evaluasi menunjukkan bahwa model terbaik diperoleh dari model Logistic Regression berbasis Countvectorizer yang diadaptasi dari kode referensi pada bagian *Code* dari dataset yang sama di

Kaggle. Model ini dimodifikasi dengan penghapusan label *news* (sentimen 2) agar hanya memprediksi tiga kelas utama (*negative*, *neutral*, dan *positive*). Selain itu, dilakukan variasi *n-grams* yang lebih luas untuk meningkatkan performa model. Jika studi sebelumnya tidak menggunakan n grams, penelitian ini menambahkan kombinasi unigram–bigram yang terbukti menghasilkan performa tertinggi. Model ini mencapai akurasi 0.78, recall 0.78, precision 0.77, dan F1-score tertimbang 0.77, sehingga dijadikan baseline utama untuk dibandingkan dengan hasil *fine-tuning* model BERTweet.

Model ML Accuracy tertinggi	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Ridge Classifier (Unigram & Bi-gram)	**78**	77	78	77

Secara keseluruhan, eksperimen ini menunjukkan bahwa meskipun model *traditional machine learning* dapat mencapai performa yang kompetitif melalui proses *preprocessing* dan *feature engineering* yang ekstensif, model berbasis transformer seperti BERTweet tetap unggul dalam memahami konteks linguistik yang kompleks tanpa memerlukan tahap pembersihan dan representasi fitur yang terlalu rumit.

- Langkah Eksperimen:
    1. Langkah pertama adalah memuat dataset mentah `twitter-climate-change-sentiment-dataset` dari Kaggle. Tanpa melalui tahap pra-pemrosesan, data ini langsung digunakan untuk melatih (fine-tuning) model-model Transformer. Tujuannya adalah untuk mendapatkan baseline performa dan mengukur seberapa baik model dapat menangani data yang kotor dan penuh *noise*.

2. Proses pelatihan menggunakan konfigurasi *hyperparameter* yang seragam, yaitu Optimizer AdamW, Learning Rate 2e-5, 3 Epoch.

Model	**Accuracy**	**F1-Score**
FacebookAI/roberta-large-mnli	**84**	**84**
climatebert/distilroberta-base-climate-detector	81	80
vinai/bertweet-base	**84**	83

Hasil akurasi yang tinggi pada data mentah memicu investigasi lebih lanjut terhadap kualitas dataset. Analisis mendalam (EDA) dilakukan untuk memeriksa potensi anomali dalam data. Ditemukan bahwa dataset mentah mengandung masalah integritas data yang signifikan:

- Terdapat 3.446 baris data duplikat.
- Teks mengandung banyak simbol aneh, karakter non-ASCII, URL, dan awalan "RT @".
- Ditemukan 659 tweet dalam berbagai bahasa selain Inggris, seperti Italia, Tagalog, dan Belanda.

Temuan adanya data duplikat dalam jumlah besar mengindikasikan adanya potensi kebocoran data (*data leakage*). Artinya, data yang sama kemungkinan besar hadir di kedua set, baik data latih maupun data uji.

- Langkah Eksperimen Lanjutan:
    1. Pemuatan dan Eksplorasi Data Awal (EDA)
        a. Dataset dimuat dari KaggleHub.
        b. Sebagai bagian dari EDA, analisis sentimen awal dilakukan menggunakan VADER (Valence Aware

Dictionary and sEntiment Reasoner) untuk mendapatkan skor sentimen leksikal sebagai perbandingan dan pemahaman data.

sentiment	message	tweetid	vader_sentiment_score
-1	@tiniiebeany climate change is an interesting hustle as it was global warming but the planet stopped warming for 15 yes while the suv boom	792927353886371840	0.6428
1	RT @NatGeoChannel: Watch #BeforeTheFlood right here, as @LeoDiCaprio travels the world to tackle climate change https://t.co/LkDehj3tNn htt…	793124121511883264 1	0.0000
1	Fabulous! Leonardo #DiCaprio's film on #climate change is brilliant!!! Do watch. https://t.co/7rV6BrmxjW via @youtube	793124402388832256	0.8544
1	RT @Mick_Fanning: Just watched this amazing documentary by leonardodicaprio on	793124635873275904	0.6705

	climate change. We all think this… https://t.co/kNSTE8K8im		
2	RT @cnalive: Pranita Biswasi, a Lutheran from Odisha, gives testimony on effects of climate change & natural disasters on the po…	793125156185137153	-0.2732

c. Dilakukan juga analisis terhadap elemen-elemen spesifik Twitter seperti jumlah retweet (RT), mention (@), dan hashtag (#) untuk memahami karakteristik dataset.

2. Penghapusan Data Duplikat
   a. Pesan (tweet) yang identik atau duplikat diidentifikasi dalam dataset. Ditemukan sebanyak 3.446 baris data yang merupakan duplikat.
3. Pembersihan dan Normalisasi Teks

Kode Pembersihan Teks dari URL, Simbol, dan Karakter Non-ASCII

```
pd.set_option('display.max_colwidth', None)
import re

def remove_urls_and_rt(text):
    text = text.replace('ÃƒÂ¢Ã¢â€šÂ¬Ã¢â€žÂ¢',
"'").replace('ÃƒÂ¢Ã¢â€šÂ¬Ã‚Â ',
'').replace('ÃƒÂ¢Ã¢â€šÂ¬Ã‚','')
    text = text.replace('ÃƒÂ¢Ã¢â€šÂ¬Ã¦', '...')
    text = re.sub(r'[^\x00-\x7F]+', ' ', text)
    text = re.sub(r'RT @\w+:', '', text)
    text = re.sub(r'http\S+|https\S+', '', text)
    text = text.replace('$q$', '')
    text = text.strip()
    return text

def contains_non_ascii(text):
```

```
    return any(ord(char) > 127 for char in text)

messages_with_strange_symbols =
df[df['message'].apply(contains_non_ascii)]
sample_messages_before =
messages_with_strange_symbols['message'].head(10).to
list()
sample_messages_after = [remove_urls_and_rt(msg) for
msg in sample_messages_before]
```

a. Normalisasi Karakter Khusus, mengganti simbol-simbol Unicode yang salah encoding (misalnya, Ã¢â‚¬Â¦ menjadi …).
b. Penghapusan Karakter Non-ASCII, menghilangkan karakter yang tidak termasuk dalam set ASCII standar.
c. Pembersihan Elemen Twitter, menghapus prefiks retweet (RT @username:), URL (http atau https), dan simbol spesifik seperti q.
d. Penghapusan spasi yang tidak perlu di awal dan akhir teks.

4. Deteksi dan Filtrasi Bahasa

Kode Deteksi Bahasa pada Teks Menggunakan langdetect

```
from langdetect import detect
from langdetect import LangDetectException

def detect_language(text):
    try:
        return detect(text)
    except LangDetectException:
        return 'unknown'

df_cleaned['detected_language'] =
df_cleaned['message'].apply(detect_language)
```

a. Library langdetect digunakan untuk mendeteksi bahasa dari setiap tweet yang telah dibersihkan.
b. Ditemukan bahwa sebagian kecil data bukan dalam bahasa Inggris

sentiment	message	tweetid	vader_sentiment_score	detected_language	
5	0	Unamshow awache kujinga na iko global warming	793125429418815489	0.1531	sw
318	0	Numinipis na ang yelo sa Arctic Circle, pati si Santa nangangayayat na. Nakakatakot talaga ang global warming	793323071796940800	0.1531	tl
356	0	jiyong: girls are so hot\njiyong: guys are hot too oh damn\njiyong: why is everyone so hot\ndaesung: global warming.	793357444738981888	-0.2732	tl
796	0	#AlGore & climate change	793646936456777728	0.0000	it
985	1	global warming	793891418402062336	0.1531	so

sebanyak 659 tweet dalam bahasa lain seperti Italia (it), Tagalog (tl), dan Belanda (nl).

c. Menjaga Konsistensi

Drop non-english tweets
`df_english = df_cleaned[df_cleaned['detected_language'] == 'en']` `df_english.shape`

Untuk menjaga konsistensi dan karena model yang akan digunakan dilatih pada korpus bahasa Inggris semua tweet yang terdeteksi bukan dalam bahasa Inggris dihapus dari dataset.

5. Filtrasi Label
   a. Dataset awal memiliki 4 kategori sentimen: -1 (Anti), 0 (Netral), 1 (Pro), dan 2 (News).
   b. Karena fokus analisis adalah pada sentimen personal (Anti, Netral, Pro), semua tweet dengan label 2 (Berita) dihapus dari dataset.

6. Persiapan Dataset Final
   a. Dataset akhir yang telah bersih hanya berisi kolom message dan sentiment, yang siap digunakan untuk tahap pelatihan model.
   b. Dataset ini kemudian disimpan ke dalam file cleaned_tweets.csv.

7. Pelatihan & Evaluasi
   a. Pembagian Data
      Dataset yang telah bersih dibagi menjadi data latih dan data uji dengan rasio 80:20.
   b. Tokenisasi
      Setiap set data ditokenisasi menggunakan *tokenizer* spesifik dari masing-masing model. Teks diubah menjadi representasi numerik yang dapat dipahami oleh model.

c. Fine-tuning
   Setiap model di-fine-tune pada data latih.
d. Evaluasi
   Setelah pelatihan selesai, performa final dari setiap model diukur pada set data uji. Metrik evaluasi yang digunakan adalah Akurasi dan F1-Score.

8. Implementasi Model Deep Learning

   Pada tahap ini, tiga model *deep learning* berbasis arsitektur Transformer yang telah di-*pre-trained* diimplementasikan untuk tugas klasifikasi sentimen dengan parameter:

   - Epoch: 3
   - Optimizer: AdamW
   - Learning Rate: 2 x 105
   - Weight Decay: 0.01
   - Learning Rate Scheduler: Linear Schedule with Warmup

9. Implementasi Model ML Tradisional

   Pada tahap ini, dilakukan implementasi berbagai model *machine learning* konvensional untuk membangun model klasifikasi sentimen sebagai pembanding terhadap model berbasis *deep learning*. Empat algoritma yang digunakan meliputi Logistic Regression, Ridge Classifier, Multinomial Naive Bayes (MNB), dan Random Forest Classifier. Seluruh model dilatih dan diuji menggunakan empat metrik evaluasi utama: *accuracy*, *precision*, *recall*, dan *F1-score (weighted)*.

Pelatihan model dilakukan dalam tiga kategori *preprocessing* yang berbeda, masing-masing disesuaikan dengan kebutuhan dan karakteristik model serta *feature extraction* yang digunakan.

1. Preprocessing Umum (Kategori 1)

Tahap ini digunakan untuk seluruh kombinasi model dengan tiga metode ekstraksi fitur: TF-IDF, CountVectorizer, dan Word2Vec, menghasilkan total 12 eksperimen (3 model × 3 metode). Proses *preprocessing* dilakukan melalui tahapan berikut:

- Hanya mengambil karakter alfabet.
- Mengubah seluruh teks menjadi huruf kecil (*lowercasing*).
- Menghapus *stopwords* berbahasa Inggris.
- Melakukan *stemming* dan *lemmatization*.
- Melakukan *random oversampling* untuk menyeimbangkan distribusi kelas.

Untuk representasi teks:

- TF-IDF dan CountVectorizer digunakan untuk membangun representasi berbasis frekuensi kata.
- Word2Vec dilatih sendiri menggunakan library *gensim* dengan parameter `vector_size=100`, `window=5`, `min_count=1`, dan `workers=4`. Karena hasil vektorisasi *Word2Vec* dapat mengandung nilai negatif, khusus untuk model Multinomial Naive Bayes, diterapkan MinMaxScaler agar data berada dalam rentang non-negatif.

2. Preprocessing Kaggle-based (Kategori 2)

Kategori kedua didasarkan pada eksperimen yang diadaptasi dari *notebook* Kaggle dengan performa terbaik, kemudian dimodifikasi untuk peningkatan hasil.
Tahapan *preprocessing* meliputi:

- *Tokenization*
- *Lowercasing*
- Penghapusan *stopwords*
- *Stemming*
- *Random Oversampling*

Representasi teks dilakukan dengan CountVectorizer yang disesuaikan menggunakan beberapa variasi *n-grams* untuk mengoptimalkan fitur kontekstual: unigram, bigram, trigram, bigram & trigram, serta kombinasi *unigram & bigram*. Pendekatan ini menghasilkan akurasi tertinggi sebesar 0.787, dengan *recall* 0.777, *precision* 0.774, dan *F1-score weighted* 0.774.

3. Preprocessing Cleaner + SMOTE Tomek (Kategori 3)

Kategori ketiga menggunakan *preprocessing pipeline* yang diselaraskan dengan tahap *deep learning*, namun ditambahkan langkah pembersihan lanjutan.
Tahapannya meliputi:

- *Tokenization*
- *Lowercasing*
- Penghapusan *stopwords*
- *Lemmatization*
- Penyeimbangan kelas menggunakan SMOTE Tomek untuk mengatasi *class imbalance*.

Ekstraksi fitur menggunakan CountVectorizer, dan model dilatih dengan GridSearchCV untuk melakukan pencarian parameter terbaik berdasarkan *cross-validation*. Eksperimen ini memberikan hasil yang lebih buruk dikarenakan perbedaan preprocessing data.

5. Hasil & Evaluasi

Pada awalnya, model-model Transformer dilatih langsung pada data mentah dengan hipotesis bahwa arsitektur canggih mereka mampu menangani *noise*. Hasilnya menunjukkan performa yang **sangat tinggi namun menyesatkan**, dengan `roberta-large-mnli` dan `vinai/bertweet-base` mencapai akurasi hingga **84%**.

Setelah dataset dibersihkan dari duplikat, bahasa asing, dan label yang tidak relevan, semua model dilatih ulang. Hasil evaluasi pada data bersih ini memberikan gambaran performa yang jauh lebih realistis dan valid.

Model	Accuracy	Precision	Recall	F1-Score
FacebookAI/roberta-large-mnli	**84**	**84**	**85**	**84**
climatebert/distilroberta-base-climate-detector	78	79	78	77
vinai/bertweet-base	81	81	81	80

Analisis Perbandingan::

- FacebookAI/roberta-large-mnli
  Model ini menunjukkan performa yang stabil, tetap berada di 84% akurasi bahkan setelah data dibersihkan. Hal ini menandakan dua kemungkinan. Pertama, model ini sangat robust dan mampu menangani noise dengan baik. Kedua,

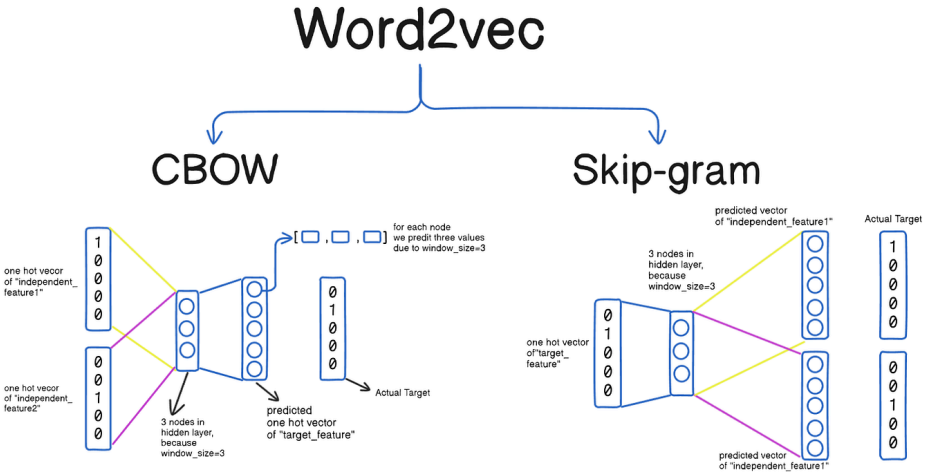
Pengaruh positif dari data bersih mengimbangi hilangnya "bocoran" dari data duplikat.

- vinai/bertweet-base & climatebert/distilroberta-base-climate-detector: Kedua model ini mengalami penurunan performa setelah data dibersihkan. Penurunan ini adalah bukti kuat bahwa skor awal mereka memang terinflasi oleh data leakage. Skor yang lebih rendah (81% dan 78%) adalah representasi yang lebih akurat dari kemampuannya yang sebenarnya.
- Durasi Training:
  - roberta-large-mnli (2x GPU T4) : 55 menit
  - distilroberta-base-climate-detector (GPU T4): 14 menit
  - Bertweet-base (GPU T4) : 29 menit

Model	Epoch 1		Epoch 2		Epoch 3	
	Acc Val	F1 Val	Acc Val	F1 Val	Acc Val	F1 Val
FacebookAI/roberta-large-mnli	**81.51**	**79.31**	**84.73**	**83.92**	**84.11**	**83.84**
climatebert/distilroberta-base-climate-detector	77.00	75.59	78.70	77.58	78.64	77.82
vinai/bertweet-base	79.05	77.28	81.06	80.60	81.26	80.56

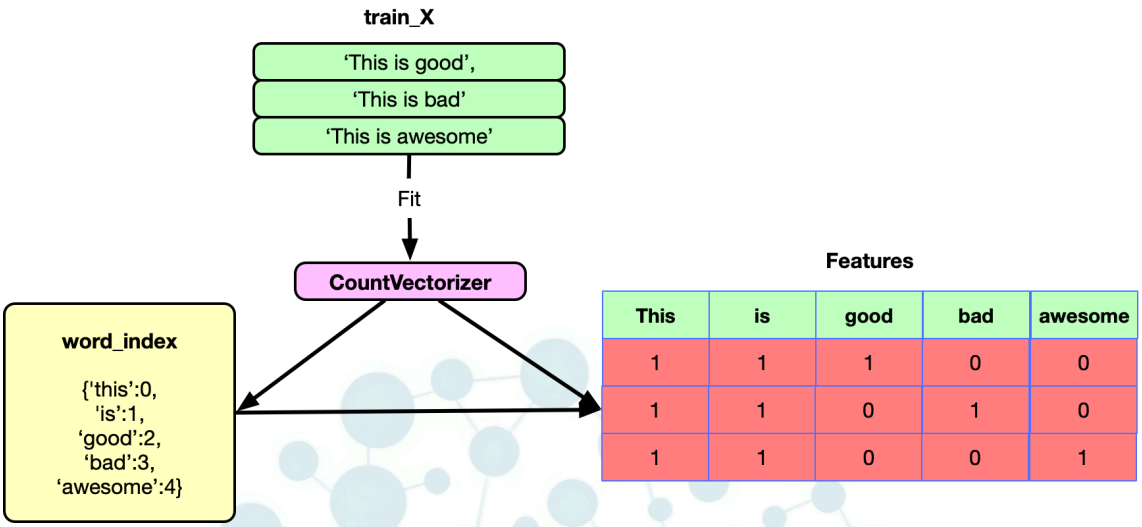
- Model ML Tradisional Feature Extraction

1. Word2Vec



Word2Vec adalah sebuah teknik dalam *natural language processing* (NLP) yang digunakan untuk mempelajari representasi kata dari sebuah korpus teks yang besar. Alih-alih memperlakukan kata sebagai simbol unik, Word2Vec mengubah setiap kata menjadi sebuah vektor numerik (daftar angka) dalam ruang multidimensi. Proses ini dilakukan dengan menganalisis konteks, yaitu kata-kata lain yang sering muncul di sekitarnya. Model ini memiliki dua arsitektur utama: CBOW (Continuous Bag-of-Words), yang memprediksi sebuah kata berdasarkan kata-kata di sekitarnya, dan Skip-gram, yang melakukan hal sebaliknya, yaitu memprediksi kata-kata di sekitar berdasarkan satu kata. Hasil akhirnya adalah sebuah ruang vektor di mana kata-kata dengan makna serupa, seperti "raja" dan "ratu," akan memiliki vektor yang posisinya berdekatan. Keunggulan utamanya adalah kemampuan untuk menangkap hubungan semantik dan sintaktik, yang bahkan memungkinkan operasi matematis pada kata, seperti vektor('raja') - vektor('pria') + vektor('wanita') yang hasilnya akan sangat dekat dengan vektor('ratu').

2. CountVectorizer



3. TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) adalah sebuah metode statistik yang digunakan dalam pemrosesan bahasa alami dan temu kembali informasi untuk mengevaluasi seberapa penting suatu kata dalam sebuah dokumen yang merupakan bagian dari suatu koleksi dokumen yang lebih besar. TF-IDF menggabungkan dua komponen:

1. **Term Frequency (TF) atau Frekuensi Term**: Mengukur seberapa sering sebuah kata muncul dalam sebuah dokumen. Frekuensi yang lebih tinggi menunjukkan tingkat kepentingan yang lebih besar. Jika suatu istilah sering muncul dalam sebuah dokumen, kemungkinan besar istilah tersebut relevan dengan konten dokumen tersebut.

$$TF(t, d) = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}}$$

2. **Inverse Document Frequency (IDF) atau Invers Frekuensi Dokumen**: Mengurangi bobot kata-kata yang umum ditemukan di banyak dokumen, sambil meningkatkan bobot kata-kata yang jarang muncul. Jika suatu istilah terdapat pada lebih sedikit dokumen, maka istilah tersebut cenderung lebih bermakna dan spesifik.

$$\text{IDF}(t, D) = \log \frac{\text{Total number of documents in corpus D}}{\text{Number of documents containing term t}}$$

Keseimbangan ini memungkinkan TF-IDF untuk menyoroti istilah-istilah yang sering muncul dalam dokumen tertentu namun juga unik di seluruh koleksi dokumen. Hal ini menjadikan TF-IDF sebagai alat yang berguna untuk berbagai tugas seperti pemeringkatan pencarian, klasifikasi teks, dan ekstraksi kata kunci.

Model ML CountVectorizer	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Unigram & Bi-gram Ridge Classifier(kaggle preprocessing)	**78**	77	78	77
Random Forest(preprocessing umum)	76	75	76	75
Ridge Classifier(clean preprocessing)	74	73	74	72

Model ML TF-IDF	Accuracy	Precision	Recall	F1-Score
Random Forest (preprocessing umum)	**76**	75	76	75
Random Forest(preprocessing umum)	75	76	76	76
Multinomial Naive Bayes(preprocessing umum)	72	74	72	72

Model ML Word2vec	Accuracy	Precision	Recall	F1-Score
Random Forest (preprocessing umum)	**71**	68	71	68
Logistic Regression(preprocessing umum)	62	65	62	64
Multinomial Naive Bayes(preprocessing umum)	59	59	59	58

6. Diskusi

- Analisis hasil model.
    - FacebookAI/roberta-large-mnli
      Model ini adalah yang terbesar, dengan 24 lapisan Transformer dan 355 juta parameter. Ukuran masif ini memberinya kapasitas tertinggi untuk mempelajari pola linguistik yang kompleks dan abstrak. Pelatihannya pada tugas NLI (*Natural*

Language Inference) membuatnya sangat kuat dalam memahami relasi semantik antar kalimat. Performanya yang tetap stabil di akurasi 84% bahkan setelah data dibersihkan menunjukkan bahwa model ini sangat robust. Setelah simbol-simbol aneh dan duplikasi data dihilangkan, arsitekturnya yang besar mampu unggul dalam menangkap makna semantik dari teks yang lebih bersih.

- vinai/bertweet-base

  Model ini secara khusus dilatih pada 850 juta data Tweet membuatnya sangat baik untuk menangani bahasa informal, slang, emoji, dan simbol khas Twitter.

  - Sebelum *pre-processing*:

    Akurasinya yang tinggi dan hampir setara dengan roberta-large-mnli dapat dijelaskan oleh kemampuannya menangkap pola dari "noise" data Twitter. Banyaknya simbol aneh dan format non-standar pada data mentah justru menjadi fitur yang dapat dikenali dengan baik oleh bertweet-base.
  - Setelah pre-processing:

    Setelah data dibersihkan dari simbol-simbol aneh tersebut, keunggulan spesialisasi bertweet-base sedikit berkurang. Tugas klasifikasi menjadi lebih bergantung pada pemahaman semantik murni dari teks yang bersih. Pada kondisi ini, model dengan parameter yang lebih besar (roberta-large-mnli) cenderung lebih unggul, sehingga performa bertweet-base mengalami sedikit penurunan.

- climatebert/distilroberta-base-climate-detector

  Model ini memiliki parameter paling sedikit di antara ketiganya. Meskipun sudah disesuaikan untuk domain iklim, performanya paling rendah karena kapasitas belajarnya yang terbatas akibat arsitekturnya yang lebih ringan. Hasilnya menunjukkan bahwa untuk tugas ini, ukuran arsitektur menjadi faktor yang

lebih dominan daripada spesialisasi domain jika model tersebut merupakan versi "distilasi".

- Mengapa model tertentu lebih baik/lebih buruk?
    1. RoBERTa-large-MNLI lebih baik karena kapasitas besar (24 layer, 355M parameter).
    2. BERTweet-base kompetitif pada data Twitter karena di *pre-training* khusus di korpus tweet (slang, emoji, ALLCAPS, hashtag, url).
    3. ClimateBERT (distilroberta) unggul dari sisi efisiensi karena hanya memiliki 82M parameter, namun sedikit di bawah dua model di atas untuk sentimen Twitter Hal tersebut wajar karena ukurannya lebih kecil.
- Kendala teknis yang ditemui.
    1. Limitasi GPU dan Waktu Eksekusi, pelatihan model berbasis Transformer khususnya FacebookAI/roberta-large-mnli yang memiliki 355 juta parameter, memerlukan VRAM GPU yang besar dan waktu pelatihan yang signifikan.
    2. Limitasi CPU dan Waktu Eksekusi, pelatihan model berbasis SVM khususnya ketika melakukan gridsearch cv untuk mencari parameter yang optimal.

- Insight dari eksperimen

```
def remove_urls_and_rt(text):
    text = text.replace('Ã¢â‚¬â„¢', "'").replace('Ã¢â‚¬Å“',
'"').replace('Ã¢â‚¬Å', '"')
    text = text.replace('Ã¢â‚¬Â¦', '...')
    text = re.sub(r'[^\x00-\x7F]+', '', text)
    text = re.sub(r'^RT @\w+:', '', text)
    text = re.sub(r'http\S+|https\S+', '', text)
    text = text.replace('$q$', '')
    text = text.strip()
    return text
```

Untuk tahap preprocessing kita melihat banyak sekali simbol aneh yang kemungkinan besar hasil dari scraping yang bisa mengganggu prediksi dari

model, maka kami perlu menghapus simbol simbol tersebut, diantaranya Penghilangan simbol aneh seperti ‘Ã¢â‚¬â„¢’ kemudian menghilangkan ‘RT @\w+:’ dan juga simbol http yang tidak memiliki makna secara langsung pada teks.

Before Cleaning	After Cleaning
RT @NatGeoChannel: Watch #BeforeTheFlood right here, as @LeoDiCaprio travels the world to tackle climate change https://t.co/LkDehj3tNn httÃ¢â‚¬Â¦	Watch #BeforeTheFlood right here, as @LeoDiCaprio travels the world to tackle climate change htt...
RT @Mick_Fanning: Just watched this amazing documentary by leonardodicaprio on climate change. We all think thisÃ¢â‚¬Â¦ https://t.co/kNSTE8K8im	Just watched this amazing documentary by leonardodicaprio on climate change. We all think this...
RT @cnalive: Pranita Biswasi, a Lutheran from Odisha, gives testimony on effects of climate change & natural disasters on the poÃ¢â‚¬Â¦	Pranita Biswasi, a Lutheran from Odisha, gives testimony on effects of climate change & natural disasters on the po...
RT @cnalive: Pranita Biswasi, a Lutheran from Odisha, gives testimony on effects of climate change & natural disasters on the poÃ¢â‚¬Â¦	Pranita Biswasi, a Lutheran from Odisha, gives testimony on effects of climate change & natural disasters on the po...
RT @CCIRiviera: Presidential Candidate #DonaldTrump is Ã¢â‚¬Å“dangerousÃ¢â‚¬Â¦ on climate change, says #monaco Ã¢â‚¬â„¢s Prince AlbertÃ¢â‚¬Â¦	Presidential Candidate #DonaldTrump is "dangerous on climate change, says #monaco 's Prince Albert...

Hasil dari penghapusan simbol aneh tersebut maka didapat data yang lebih bersih dan lebih dapat dibaca dibandingkan sebelumnya. Tetapi tidak hanya berhenti disitu, terdapat juga banyak sekali bahasa asing selain bahasa inggris pada dataset tersebut. Sehingga diperlukan preprocessing lebih lanjut untuk menghilangkan bahasa asing tersebut.

message	detected_language
Unamshow awache kujinga na iko global warming	sw

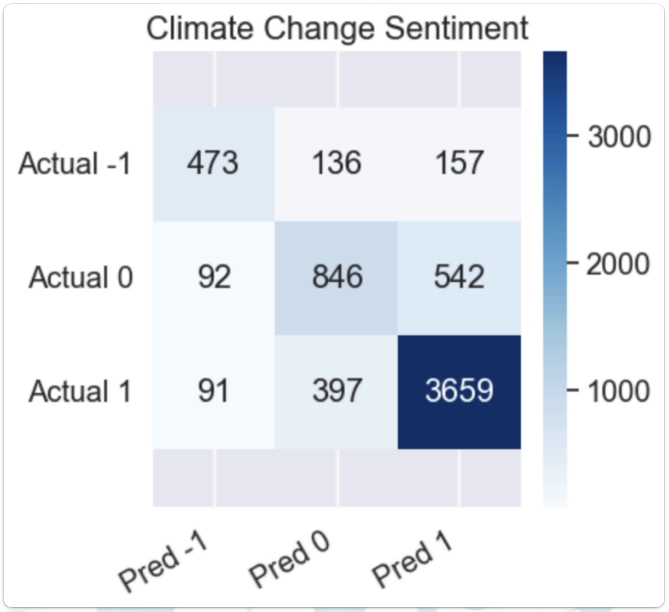
Numinipis na ang yelo sa Arctic Circle, pati si Santa nangangayayat na. Nakakatakot talaga ang global warming	tl
jiyong: girls are so hot\njiyong: guys are hot too oh damn\njiyong: why is everyone so hot\ndaesung: global warming.	tl
#AlGore & climate change	it
global warming	so

Pada dataset tersebut terdapat 659 bahasa asing yang terdeteksi. Walaupun tidak semua yang terdeteksi benar-benar menggunakan bahasa selain inggris, tetapi mayoritas hasil dari deteksi tersebut menggunakan asing

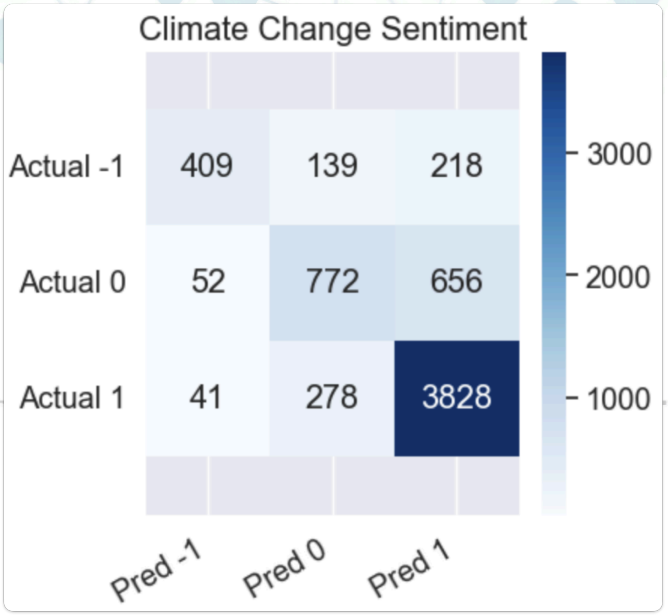
message	detected_language
#AlGore & climate change	it
Cop22, da Parigi a Marrakesh: la sfida contro il global warming	it
@i_artaza @COP22 @UNFCCC @ConversationUS regarding climate change.	it
#AccordodiParigi Stop sussidi a fonti fossili contro climate change. @LeonaroDicaprio	it
@beppe_grillo Quindi non volevate le trivelle ma esultate per uno il cui vice presidente afferma che il global warming u...	it

Contohnya pada visualisasi tersebut ada banyak bahasa Italia yang masuk ke dataset tersebut, sehingga perlu di drop agar tidak mengganggu prediksi dari model yang hanya di pre-trained menggunakan bahasa inggris.

Kemudian kami melakukan analisis perbandingan antara *confusion matrix* sebelum dan sesudah penerapan teknik *oversampling* menunjukkan dampak signifikan terhadap kemampuan model dalam menangani masalah ketidakseimbangan kelas (*class imbalance*). Pada model awal tanpa *oversampling* (gambar bawah), terlihat adanya bias yang kuat terhadap kelas mayoritas (sentimen '1'), di mana model mampu memprediksi dengan benar sebanyak 3.828 kali. Namun, performanya jauh lebih rendah pada kelas minoritas, dengan hanya 409 prediksi benar untuk kelas '-1' dan 772 untuk kelas '0'. Setelah teknik *oversampling* diterapkan (gambar atas), terjadi peningkatan performa yang substansial pada kelas-kelas minoritas tersebut. Jumlah prediksi yang benar untuk kelas '-1' meningkat menjadi 473, dan untuk kelas '0' meningkat menjadi 846. Meskipun peningkatan ini disertai dengan sedikit penurunan performa pada kelas mayoritas menjadi 3.659, *trade-off* ini dapat diterima karena hasilnya adalah model yang jauh lebih seimbang dan andal. Dengan demikian, dapat disimpulkan bahwa *oversampling* secara efektif mengurangi bias model dan meningkatkan kemampuannya untuk menggeneralisasi, sehingga lebih kapabel dalam mengidentifikasi seluruh kelas sentimen secara lebih merata dalam aplikasi dunia nyata.



Gambar confusion matrix menggunakan oversampling



Gambar confusion matrix tidak menggunakan oversampling

7. Kesimpulan

Ringkasan hasil riset:

- Membandingkan pendekatan transformer (RoBERTa-large-MNLI, BERTweet, ClimateBERT-distil) dengan baseline tradisional (LogReg, NB, Ridge, RF) berbasis

CountVectorizer/TF-IDF/Word2Vec pada dataset Twitter Climate Change

- Melakukan EDA dan preprocessing yang relevan untuk data Twitter (hapus duplikat, normalisasi simbol, filtrasi bahasa Inggris, penghapusan label "news"), diikuti evaluasi terukur (Accuracy, Precision, Recall, Macro-F1).

- Model mana yang performanya terbaik?
    - RoBERTa-large-MNLI (fine-tuned) menjadi yang terbaik secara agregat pada set evaluasi, terutama ditinjau dari Macro-F1, berkat kapasitas besar dan resep prapelatihan yang kuat.
    - BERTweet-base sangat kompetitif pada teks Twitter "asli"/Raw (emoji, ALLCAPS, hashtag). Pada konfigurasi ini, BERTweet sering mendekati atau bahkan pada subset tertentu menyamai RoBERTa-large dengan biaya komputasi lebih rendah.
    - Di pihak baseline, Logistic Regression + CountVectorizer (unigram-bigram) adalah pembanding terkuat, namun tetap berada di bawah model transformer ketika data menuntut pemahaman konteks, sarkas, dan sinyal non-leksikal.

- Implikasi hasil proyek untuk domain terkait.

    Hasil penelitian ini memiliki dua implikasi utama, yaitu pada ranah akademik dan lingkungan. Dari sisi akademik, penelitian ini memperkuat bukti bahwa model transformer seperti RoBERTa-large-MNLI, BERTweet, dan ClimateBERT-distil mampu memberikan pemahaman konteks yang jauh lebih dalam dibandingkan pendekatan tradisional berbasis *CountVectorizer* atau *TF-IDF*. Temuan ini menunjukkan bahwa penggunaan representasi bahasa yang kontekstual dapat secara signifikan meningkatkan akurasi dan ketahanan model

dalam menghadapi variasi ekspresi khas Twitter seperti sarkasme, emoji, dan penulisan tidak baku. Selain itu, hasil ini memberikan kontribusi metodologis bagi peneliti AI, khususnya dalam studi analisis sentimen bertema lingkungan, dengan menekankan pentingnya *fine-tuning* yang tepat serta penanganan ketidakseimbangan data secara hati-hati.

Dari sisi praktis dan lingkungan, temuan bahwa sentimen publik terhadap isu perubahan iklim cenderung positif memberikan indikasi bahwa kesadaran masyarakat terhadap urgensi krisis iklim semakin meningkat. Model transformer yang terbukti unggul ini juga berpotensi digunakan oleh organisasi lingkungan, pembuat kebijakan, maupun peneliti sosial untuk memantau persepsi publik secara real time, mendeteksi disinformasi terkait iklim, serta mengevaluasi efektivitas kampanye kesadaran lingkungan. Namun demikian, keterbatasan dalam biaya komputasi dan kebutuhan sumber daya tinggi menjadi tantangan yang perlu diperhatikan dalam penerapan skala besar di masa depan.

8. *Future Works*

Penelitian ini dapat dikembangkan lebih lanjut melalui beberapa arah kerja lanjutan. Pertama, perlu dilakukan **perluasan dataset** dengan melakukan *scraping* data terbaru dari platform X (Twitter) untuk memperoleh variasi konteks dan ekspresi yang lebih luas. Proses pelabelan secara **manual dan terverifikasi** penting dilakukan guna memastikan kualitas anotasi sentimen yang lebih konsisten, serta mengatasi potensi bias label yang terdapat pada dataset awal dari Kaggle. Selain itu, eksplorasi terhadap **strategi penyeimbangan data** yang lebih adaptif, seperti *contextual data augmentation* atau *prompt-based augmentation*, dapat membantu

mengurangi dampak ketidakseimbangan kelas yang ditemukan pada dataset ini.

Kedua, dari sisi model, disarankan untuk melakukan eksperimen lanjutan menggunakan **Large Language Models (LLM)** dengan pendekatan efisien seperti **Parameter-Efficient Fine-Tuning (PEFT)** atau **Low-Rank Adaptation (LoRA)**. Pendekatan ini memungkinkan pembandingan yang lebih komprehensif antara performa model besar dan efisiensi komputasinya dibandingkan dengan model BERT konvensional.

Ketiga, untuk meningkatkan nilai interpretatif penelitian, dapat dilakukan **analisis *explainability*** menggunakan metode seperti SHAP atau LIME untuk menjelaskan faktor-faktor yang paling memengaruhi keputusan klasifikasi model. Hal ini tidak hanya meningkatkan transparansi model, tetapi juga memberikan wawasan tambahan bagi peneliti maupun pengambil kebijakan dalam memahami persepsi publik terhadap isu perubahan iklim.

9. Output

Link Akhir :

- Notebook: https://github.com/NafisNaufal/climate-change-sentiment-analysis
- Model terbaik: https://drive.google.com/drive/folders/1g-Nc4t4phYRmayv1c8EDlubfLMtJAQXL?usp=sharing
- Slide presentasi: https://www.canva.com/design/DAG17Du-TmI/4q2sO4jiMbimShcWhgNPvw/edit?utm_content=DAG17Du-TmI&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

References:

Anoop, V. S., dkk. (2024). *Climate Change Sentiment Analysis Using Domaic Specific Bidirectional Encoder Representations From Transformers*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.

Fir, A. (2020). *Mengenal Word2Vec*. Medium. Diambil dari https://medium.com/@afrizalfir/mengenal-word2vec-af4758da6b5.

Li, S. (2019). *Understanding Count Vectorizer*. Medium. Diambil dari https://medium.com/swlh/understanding-count-vectorizer-5dd71530c1b.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. In Advances in Neural Information Processing Systems (NeurIPS).

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020). *BERTweet: A pre-trained language model for English Tweets*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP).

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108.

Varshavsky, P. (2023). *Transformers*. Medium. Diambil dari https://medium.com/data-science/transformers-89034557de14.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need*. In Advances in Neural Information Processing Systems (NeurIPS).

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). *ClimateBERT: A Pretrained Language Model for Climate-Related Text*. arXiv preprint arXiv:2110.12010.