REPOSITOR, Vol. 7, No. 2, Mei 2025, Pp. 157-168
ISSN : 2714-7975
E-ISSN : 2716-1382

Adversarial Attack Pada Teks Berbahasa Indonesia Menggunakan Framework TextAttack

Setio Basuki*[1], Ilham Aulady Miftakhurrizqy*[2]
[1,2]Universitas Muhammadiyah Malang
setio_basuki@umm.ac.id[1], ilhamam462@webmail.umm.ac.id*[2]

Abstrak

Machine learning merupakan sebuah bidang keilmuan yang sangat cepat perkembangannya. Natural language processing atau NLP merupakan salah satu dari beberapa bagian keilmuan machine learning. Perkembangan NLP yang cepat ini juga telah diimplementasikan pada keseharian manusia, seperti chatbot, search engine, analisa sentiment, dan lain-lain. Perkembangan ini juga diikuti dengan adanya resiko seperti adversarial attack. Adversarial attack merupakan serangan yang bertujuan untuk menipu model machine learning, termasuk NLP. Adversarial attack pada NLP dilakukan dengan mengubah data yang akan diproses oleh model, sehingga model bisa terkelabuhi saat mengeluarkan output. Proses adversarial attack pada NLP tersebut bisa dipermudah dengan mengguanakan sebuah framework TextAttack. Penelitian ini bertujuan untuk melakukan perbandingan pada model NLP dengan karakteristik aristektur yang berbeda dan melihat bagaimana adversarial attack bisa mempengaruhi model NLP tersebut. Proses penelitian ini memiliki beberapa tahap, yaitu mencari dataset, preprocessing, pembuatan 3 model NLP, pengubahan resep pada framework, dan evaluasi model saat sebelum diserang serta sesudah diserang. Model yang digunakan pada penelitian ini merupakan model Multinomial Naïve Bayes dengan arsitektur klasik, model dengan arsitektur LSTM, dan model dengan arsitektur BERT. Hasil penelitian menunjukkan jika model arsitektur BERT memiliki ketahanan yang lebih baik dari pada model klasik dan model LSTM. Sedangkan model klasik memberikan hasil serangan yang paling sedikit daripada model LSTM dan model BERT. Penelitian ini diharapkan dapat memberikan kontribusi dalam bidang adversarial attack khususnya pada teks yang berbahasa Indonesia.

Kata Kunci: *Adversarial Attack, Natural Language Processing (NLP), Multinomial Naïve Bayes (MNB), Long-Short Term Memory (LSTM), BERT*

Abstract

Machine learning is a rapidly developing field of science. Natural language processing or NLP is one of several parts of machine learning science. The rapid development of NLP has also been implemented in everyday human life, such as chatbots, search engines, sentiment analysis, and others. This development is also accompanied by risks such as adversarial attacks. Adversarial attacks are attacks that aim to trick machine learning models, including NLP. Adversarial attacks on NLP are carried out by changing the data that will be processed by the model, so that the model can be fooled when producing output. The adversarial attack process on NLP can be simplified by using the TextAttack framework. This study aims to compare NLP models with different architectural characteristics and see how adversarial attacks can affect NLP models. This research process has several stages, namely dataset search, preprocessing, creation of 3 NLP models, recipe changes to the framework, and evaluation of the model before and after being attacked. The models used in this study are the Multinomial Naïve Bayes model with classical architecture, a model with LSTM architecture, and a model with BERT architecture. The results of the study show that the BERT architecture model has better robustness than the classical model and the LSTM model. While the classical model provides the fewest attack results compared to the LSTM model and the BERT model. This research is expected to contribute to the field of adversarial attacks, especially on Indonesian language texts.

Keywords: *Adversarial Attack, Natural Language Processing (NLP), Multinomial Naïve Bayes (MNB), Long-Short Term Memory (LSTM), BERT*

1. Pendahuluan

Machine Learning atau ML merupakan salah satu bidang ilmu yang cepat berkembang. ML memproses data yang ada menjadi sebuah pengetahuan. ML ini adalah keilmuan yang berada pada perbatasan antara statistika, kecerdasan buatan dan ilmu komputer [1]. ML juga dibagi menjadi beberapa bagian, salah satunya adalah Natural Language Processing atau NLP. Dengan menggunakan pendekatan ML, NLP bisa digunakan untuk memproses bahasa manusia. Contoh-contoh pemrosesan yang dilakukan menggunakan NLP yaitu, sentiment analysis, klasifikasi teks, fraud detection, text summarization, Large Language Model (LLM), dan lain-lain. Pemrosesan pada NLP tentunya juga mengalami perkembangan, pada tahun 2010 muncul sebuah teknik baru bernama Deep Learning (DL) yang merupakan bagian dari ML. DL memanfaatkan jaringan neural untuk memproses data. Lalu, dengan DL dikembangkan sebuah model Transformers [2]. Model ini menggunakan mekanisme "attention" yang cara kerjanya dengan menggunakan Query(Q), Key(K), dan Value(V) untuk memulai transformasi linear demi menghasilkan bobot dinamis untuk relasi yang berbeda [3]. Mekanisme ini lalu diimplementasikan pada model baru yaitu BERT [4]. BERT merupakan model pre-trained yang menggunakan metode bidirectional yang berarti bisa menganalisa teks dari kiri ke kanan dan/atau dari kanan ke kiri yang mengakibatkan model bisa memahami konteks dari teks. Banyak aspek dari kehidupan bergantung pada ML NLP sejak dikembangkannya model BERT, mulai dari sentiment analysis, question answering, natural language inference, dan lain-lain.

Model ML dibuat dengan melatih (training) sistem menggunakan sebuah algoritma. Algoritma tersebut lalu diberi data yang telah diproses agar sistem dapat memahami pola dari data tersebut menggunakan algoritma yang telah dipilih. Setelah sistem dapat memahami data yang telah diberikan, maka selanjutnya dilakukan pengujian model dengan data yang berbeda (testing). Pengujian atau testing dilakukan untuk melihat bagaimana kinerja dari model ML tersebut. Tolak ukur pengujian berbeda-beda tergantung dengan use case model ML. Model klasifikasi, sentiment analisis, dan sebagainya memiliki tolak ukur precision, recall, F1-score, dan accuracy. Untuk model regresi yang menjadi tolak ukur merupakan MSE, RMSE, MAE, R-squared, dan sebagainya. Output dari pengujian merupakan persentase yang menandakan seberapa bagus kinerja model, semakin besar persentasenya maka semakin bagus model tersebut dalam melaksanakan use case-nya .Meskipun model yang dihasilkan itu sudah dinilai baik, masih ada resiko yang dipertimbangkan, seperti Adversarial Attack. Secara istilah adversarial berarti bermusuhan, diambil dari kata adversary yang berarti berhadap-hadapan dan attack memiliki makna serangan. Serangan adversarial attack adalah teknik serangan yang digunakan untuk menipu model ML. Contoh pada text classification, serangan dilakukan dengan merubah struktur kalimat dengan menggunakan sinonim, menghapus kata, dan sebagainya agar model ML salah dalam mengklasifikasi teks. Serangan adversarial attack dapat menimbulkan ketidakstabilan pada model ML. Hal ini menunjukkan bahwa model yang dibuat rentan terhadap manipulasi dengan penambahan gangguan kecil pada input. Teknik penambahan ini dapat dilakukan dengan menggunakan framework yang bernama TextAttack [5]. Karena teknik yang digunakan untuk menambah gangguan ini mudah diterapkan yaitu dengan sebuah library pada bahasa pemrograman, maka perlu adanya perhatian untuk serangan adversarial pada teks ini.

Serangan adversarial pada teks merupakan serangan yang bertujuan untuk menipu model NLP dengan membuat sebuah adversarial examples berbentuk teks yang telah diperturbasi. Contoh teks perturbasi yang digunakan antara lain, penggantian kata, penggantian sinonim, menyisipkan atau menghapus kata, kesalahan pada grammar, penggantian tanda baca, dan penggantian konteks. Pertrubasian pada teks ini dapat mengakibatkan kesalahan mengklasifikasi teks oleh model NLP meskipun manusia dapat mengklasifikasi teks tersebut dengan baik.

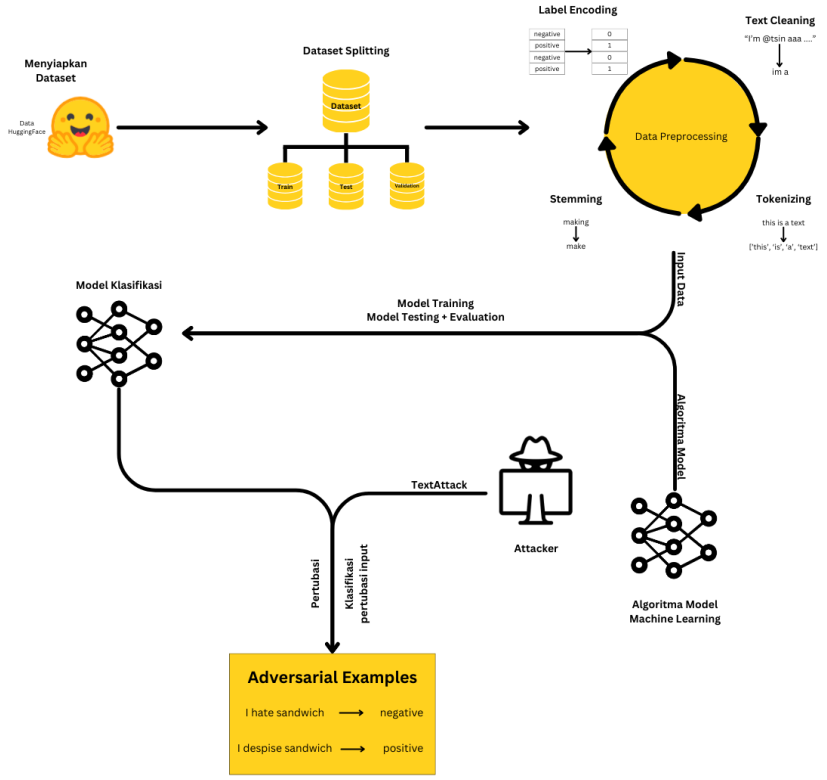
Penelitian serangan adversarial ini bertujuan untuk membuat sebuah metode yang dapat menyerang model sentiment analysis dengan Bahasa Indonesia dan menghasilkan sebuah adversarial example. Untuk metode akan digunakan penggantian kata dan word swap pada tiga model NLP. Tiga model tersebut adalah Multinomial Naïve Bayes, LSTM, dan pre-trained BERT. Penggunaan model yang berbeda ini karena adanya beda karakteristik dari setiap model, dimana model Multinomial Naïve Bayes merupakan metode klasik dari ML, LSTM menggunakan metode jaringan neural sedangkan BERT menggunakan metode attention secara Bi-directional. Dengan ketiga model itu lalu digunakan algoritma pencarian dengan yang ada pada paper "Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency" [6] dimana

menggunakan algoritma untuk mencari kemungkinan bobot dari kata yang penting atau Probability Weighted Word Saliency (PWWS).

2. Metode Penelitian

2.1 Alur Penelitian

Tahapan penelitian menggambarkan proses penelitian yang akan dilakukan dan menggambarkan alur dari penelitian. Tahapan penelitian dapat dilihat pada Gambar 1 berikut.



Gambar 1. Alur Penelitian

2.2 Dataset

Pemilihan data dari sekumpulan data yang perlu dilakukan sebelum memproses data dimulai. Data yang dipilih akan digunakan untuk proses natural language processing dan disimpan pada suatu berkas. Data yang digunakan pada penelitian ini merupakan data Hugging Face dari repository tyqiangz/multilingual-sentiments dengan total 12.760 instance teks. Data dari Hugging Face ini sudah terpisah menjadi data latih, data uji dan data validasi. Berikut merupakan contoh dari instance teks pada data Hugging Face (Gambar 2, Gambar 3, Gambar 4).

	text	label	source
0	warung ini dimiliki oleh pengusaha pabrik tahu...	positive	indonlue/smsa
1	mohon ulama lurus dan k212 mmbri hujjah partai...	neutral	indonlue/smsa
2	lokasi strategis di jalan sumatera bandung . t...	positive	indonlue/smsa
3	betapa bahagia nya diri ini saat unboxing pake...	positive	indonlue/smsa
4	duh . jadi mahasiswa jangan sombong dong . kas...	negative	indonlue/smsa
...	...	...	...
10995	tidak kecewa	positive	indonlue/smsa
10996	enak rasa masakan nya apalagi kepiting yang me...	positive	indonlue/smsa
10997	hormati partai-partai yang telah berkoalisi	neutral	indonlue/smsa
10998	pagi pagi di tol pasteur sudah macet parah , b...	negative	indonlue/smsa
10999	meskipun sering belanja ke yogya di riau junct...	positive	indonlue/smsa

Gambar 2. Gambar Data Train dari Hugging Face

	text	label	source
0	kemarin gue datang ke tempat makan baru yang a...	negative	indonlue/smsa
1	kayak nya sih gue tidak akan mau balik lagi ke...	negative	indonlue/smsa
2	kalau dipikir-pikir , sebenarnya tidak ada yan...	negative	indonlue/smsa
3	ini pertama kalinya gua ke bank buat ngurusin ...	negative	indonlue/smsa
4	waktu sampai dengan gue pernah disuruh ibu lat...	negative	indonlue/smsa
...	...	...	...
495	kata nya tidur yang baik itu minimal enam jam ...	neutral	indonlue/smsa
496	indonesia itu ada di benua asia .	neutral	indonlue/smsa
497	salah satu kegemaran anak remaja indonesia sek...	neutral	indonlue/smsa
498	melihat warna hijau bisa bikin mata jadi lebih...	positive	indonlue/smsa
499	bondan winarno yang suka bilang maknyus sekara...	neutral	indonlue/smsa

Gambar 3. Gambar Data Test dari Hugging Face

	text	label	source
0	meski masa kampanye sudah selesai , bukan bera...	neutral	indonlue/smsa
1	tidak enak	negative	indonlue/smsa
2	restoran ini menawarkan makanan sunda . kami m...	positive	indonlue/smsa
3	lokasi di alun alun masakan padang ini cukup t...	positive	indonlue/smsa
4	betapa bejad kader gerindra yang anggota dprd ...	negative	indonlue/smsa
...	...	...	...
1255	film tncfu , tidak cocok untuk penonton yang t...	negative	indonlue/smsa
1256	indihome ini mahal loh bayar nya . hanya , pen...	negative	indonlue/smsa
1257	be de gea , cowok cupu yang takut dengan pacar...	negative	indonlue/smsa
1258	valen yang sangat tidak berkualitas . konentat...	negative	indonlue/smsa
1259	restoran ini menjadi tempat pilihan saya berbu...	positive	indonlue/smsa

Gambar 4. Gambar Data Validation dari Hugging Face

2.3 Data Preprocessing

Tahap ini merupakan tahap dimana data yang ada harus sesuai dengan use case model yang akan dibuat. Pada penelitian ini dilakukan dengan cara berikut :

a. Text Cleaning

Text cleaning dilakukan dengan membersihkan data seperti membuang baris atau kolom data dengan missing value, membuang atribut yang tidak sesuai, melakukan uppercase atau lowercase, menghilangkan kata atau frasa yang tidak ada kaitannya dengan analisa sentiment. Proses ini bisa dilakukan dengan menggunakan regular expression (regex). Regex merupakan sebuah bahasa yang digunakan untuk menentukan pencarian teks dalam sebuah string. Regex menggunakan notasi algebra untuk mencari karakteristik dalam sebuah string lalu memberikan teks yang cocok dengan pola dalam notasi tersebut [7].

b. Label Encoding

Proses label encoding dalam preprocessing penelitian ini adalah sebagai berikut, label “positif”, “netral” dan “negatif” menjadi 2, 1 dan 0. Proses label encoding ini dilakukan dengan menggunakan library Scikit-Learn [8].

c. Tokenizing

Tokenizing dalam penelitian ini dilakukan dengan menggunakan library NLTK [9]. NLTK merupakan open source program yang bisa melakukan mencakup pemrosesan bahasa secara simbolik dan statistik, seperti tokenizing. Proses ini merubah teks menjadi token dengan cara memisah setiap kata pada teks. Setiap token direpresentasikan dengan angka atau index.

d. Stemming

Stemming dalam penelitian ini dilakukan dengan menggunakan library Sastrawi. Sastrawi merupakan library stemming. Library ini bisa diakses pada link https://github.com/sastrawi/sastrawi [10]. Sastrawi memberikan fungsi untuk melakukan stemming pada teks berbahasa Indonesia.

e. Model Machine Learning

Pada penelitian ini, model yang kami gunakan ada tiga macam, yaitu model klasik, model deep learning dan model pre-trained.

1) Model Klasik

Model klasik yang digunakan di penelitian ini adalah model dengan algoritma Multinomial Naïve Bayes (MNB). Algoritma ini melakukan perhitungan dengan menghitung probabilitas suatu elemen pada suatu kelas. Sebelum melatih algoritma ini dengan data latih, dilakukan embedding terlebih dahulu pada data. Embedding dilakukan dengan menggunakan TF-IDF (Term Frequency-Inverse Document Frequency). Embedding ini menghitung seberapa banyak/relevan sebuah kata berada pada sebuah teks dalam dokumen.

2) Model Deep Learning

Model deep learning yang digunakan di penelitian ini adalah model dengan algoritma LSTM (Long Short-Term Memory). Arsitektur yang dibuat pada penelitian ini terdiri dari beberapa layer: input, embedding layer, LSTM layer, fully connected layer, dan output. Illustrasi dari model bisa dilihat pada Gambar 5 berikut.

Pada Gambar 5 menunjukkan layer model yang digunakan dalam penelitian ini. Layer pertama yaitu input merupakan layer inputtan dari teks yang telah di-preprocessing. Layer kedua merupakan layer embedding. Layer ini merupakan proses untuk merubah angka yang pada kasus ini jumlah kata pada data menjadi sebuah vector. Layer ketiga merupakan layer LSTM. Pada layer ini diberikan dropout dan recurrent dropout sebesar 0.2 agar tidak terjadi overfitting yang dapat menyebabkan bias pada model. Layer keempat merupakan layer penyambung antara layer LSTM dan output. Pada layer ini diberikan kernel regulizer l2 dengan factor regulasi sebesar 0.01. Kernel regulaizer diperlukan untuk mencegah overfitting pada model. Layer terakhir merupakan layer output dengan 3 neurons untuk memberikan output antara 0, 1, dan 2.

3) Model Pre-Trained

Penelitian ini menggunakan IndoBERTweet sebagai model pre-trained. Model ini telah dilatih untuk melakukan tugas NLP dengan teks yang layaknya seperti tweet pada aplikasi X. Penelitian ini menggunakan optimizer AdamW untuk mengoptimalkan training pada model. Pada parameter AdamW diberikan learning rate sebesar 0.00001 dan epsilon sebesar 0.00000001.

2.5 Attacker (TextAttack)

TextAttack merupakan framework yang digunakan untuk mempermudah dalam pembuatan adversarial examples teks pada model NLP. Adapun dalam penelitian ini akan menggunakan framework TextAttack untuk menyerang dan membuat adversarial examples pada model NLP. Framework ini memiliki formula yang sudah terintegrasi untuk mempermudah penyerangan pada model NLP. Formula-formula tersebut merupakan algoritma yang sudah dibuat oleh penelitian sebelumnya, seperti PWWSRen2019, A2T (Attack for Adversarial Training Recipe), BERT-Attack, dan sebagainya. Selain dengan formula TextAttack juga menyediakan metode untuk membuat formula sendiri. Pada penelitian ini formula yang digunakan merupakan formula sendiri yang hampir sama dengan PWWSRen2019.

2.6 Adversarial Examples

Adversarial examples merupakan output dari adversarial attack. Output ini berisi teks awal, teks yang telah dipertubasi, label awal, label dengan teks yang telah dipertubasi, label yang sebenarnya, dan hasil serangan. Output dari adversarial attack juga memberikan persentase akurasi model sebelum diserang, persentase akurasi model setelah diserang, persentase keberhasilan serangan, dan banyaknya setiap hasil serangan yang berhasil, gagal, serta yang dilewati.

3. Hasil Penelitian dan Pembahasan

Pada bab ini merupakan penjelasan penelitian pada bab sebelumnya. Bab hasil dan pembahasan mencakup penjelasan model yang akan diserang, implementasi TextAttack, hasil adversarial examples, dan dampak serangan pada model.

3.1 Model yang Dihasilkan

Model machine learning yang digunakan merupakan model dari hasil latih dengan data HuggingFace. Dataset ini berasal dari repository tyqiangz/multilingual-sentiment pada platform HuggingFace. Dataset ini memiliki total teks sebanyak 12.760 yang sudah terbagi menjadi 3, data latih dengan total 11.000 teks, data test dengan total 500 teks, dan data validasi dengan total 1260 teks. Data tersebut menjadi input pada model NLP dengan penjelasan hasil model sebagai berikut.

a) Model Multinomial Naïve Bayes

Pada algoritma multinomial naïve bayes, data dilakukan embedding dengan word embedding tf-idf sebelum dijadikan input sebagai data latih model. Setelah itu dilakukan testing dan berikut merupakan Tabel 1 berupa hasil evaluasi model tersebut.

Tabel 1. Tabel Evaluasi Model MNB

	Precision	Recall	F1-Score	Support
0	0.49	0.78	0.60	204
1	0.53	0.20	0.30	88
2	0.77	0.41	0.53	208
Akurasi			0.54	500

Label 0 merupakan label negatif, label 1 merupakan label netral, dan label 2 merupakan label positif.

b) Model LSTM

Model LSTM menggunakan beberapa layer. Layer yang digunakan antara lain, layer embedding, layer LSTM, layer dense dengan aktivasi relu, dan layer dense sebagai output dengan aktivasi sigmoid. Pada layer embedding. Lalu data latih diinputkan sebanyak 10 epoch dengan batch sebesar 256. Setiap epoch juga dilakukan validasi menggunakan data validasi. Berikut merupakan Tabel 2 berupa hasil dari model setelah testing.

Tabel 2. Tabel Evaluasi Model LSTM

	Precision	Recall	F1-Score	Support
0	0.89	0.72	0.79	254
1	0.43	0.83	0.57	46
2	0.79	0.82	0.80	200
Akurasi			0.77	500

Label 0 merupakan label negatif, label 1 merupakan label netral, dan label 2 merupakan label positif. Berikut adalah gambar dari loss dan accuracy serta confussion matrix dari model.

c) Model IndoBERTweet

Implementasi model IndoBERTweet dilakukan dengan menggunakan API dari library Transformers dari database HuggingFace. Optimizer yang digunakan pada model ini merupakan AdamW dengan learning rate sebesar 0.00001 dan epsilon sebesar 0.00000001. Lalu batch data latih sebesar 4 dan batch data test sebesar 1. Hasil yang didapat adalah pada Tabel 3 sebagai berikut.

Tabel 3. Tabel Evaluasi Model IndoBERTweet

Training Loss	0.1269
Validation Loss	0.4538
F1-Score (Weighted)	0.9123

Tabel 4. Hasil Akurasi Per-label

Label	Akurasi
0	204/204
1	60/88
2	194/208

3.2 Implementasi TextAttack

Implementasi TextAttack dilakukan dengan menggunakan resep yang telah disediakan oleh frameworknya. Pada penelitian ini, resep yang dipilih adalah PWWSRen2019. Resep serangan memiliki beberapa parameter, antara lain transformation, constraint, goal_function, dan search_method. Pada parameter transformation digunakan metode Word Swap dengan class WordSwapWordNet. Pada parameter constrain menggunakan class RepeatModification dan StopwordModification. Goal_function pada resep ini menggunakan class UntargettedClassification. Lalu parameter search_method menggunakan class GreedyWordSwapWIR dengan parameter "weighted-saliency". Secara default, resep ini hanya bisa digunakan pada dataset teks berbahasa inggris. Maka dari itu penelitian ini mengubah/membuat class yang kompatibel dengan framework agar bisa digunakan pada teks berbahasa Indonesia.

Class yang diubah merupakan kelas resep PWWSRen2019 dan class WordSwapWordNet. Pengubahan class WordSwapWordNet dilakukan dengan mengganti inisialisasi bahasa pada NLTK wordnet dari inggris menjadi Indonesia. Pada class PWWSRen2019 diperlukan untuk menyesuaikan pada parameter transformation dengan class WordSwapWordNet yang telah diubah. Parameter yang juga diubah pada kelas ini merupakan parameter constraint pada class StopwordModification yang secara default parameter pada class ini merupakan inggris, maka parameter diubah menjadi Indonesia.

Selain mengubah resep, juga dilakukan pembuatan/pengubahan pada model wrapper. Hal ini bertujuan agar TextAttack dapat berinteraksi dengan model untuk mendapatkan query, prediksi model, dan logit. Ketiga model yang telah dipersiapkan memiliki model wrapper yang berbeda. Pada model MNB dan LSTM dibuat model wrapper kustom untuk model masing-masing. Tetapi model IndoBERTweet menggunakan model wrapper dari framework API karena model tersebut kompatibel dengan model wrapper dari framework.

Setelah mempersiapkan resep dan model wrapper selanjutnya menginisialisasi dataset yang akan digunakan untuk melakukan serangan. Pada penelitian ini, data test sebanyak 500 digunakan dalam serangan.

3.3 Dampak Serangan pada Ketiga Model

Dalam penelitian ini, dampak yang disebabkan oleh adversarial attack adalah adanya penurunan dari akurasi model dalam melaksanakan tugas analisis sentimen. Setiap model juga mengalami perbedaan dalam persentase keberhasilan serangan. Perbedaan ini bisa disebabkan oleh beda arsitektur, besar akurasi orisinil model, atau parameter resep. Berikut tabel summary hasil adversarial attack.

Tabel 5. Summary Hasil Serangan

Model	Successful	Failed	Skipped
MNB	249	18	233
LSTM	309	68	123
IndoBERTweet	368	101	31

Pada Tabel 5, kolom successful adalah total serangan yang berhasil mengelabuhi model NLP. Kolom failed adalah total serangan yang gagal karena resep tidak bisa menemukan pertubasi yang bisa mengelabuhi model. Kolom skipped adalah total serangan yang tidak terjadi karena model yang tidak bisa memprediksi klasifikasi secara tepat.

Dari hasil pada tabel 5, model IndoBERTweet memiliki total serangan yang berhasil dan gagal terbanyak, tetapi serangan yang terlewati yang paling sedikit. Sedangkan model MNB memilki total serangan yang berhasil dan yang gagal paling sedikit, namun banyak serangan yang terlewati. Model LSTM berada di antar kedua model tersebut. Secara angka terlihat jika model IndoBERTweet paling banyak terserangnya, tetapi itu tidak membuktikan jika model IndoBERTweet memiliki ketahanan yang paling rendah daripada model MNB dan LSTM, karena juga harus dipertimbangkan serangan yang gagal dan serangan yang terlewati. Maka, agar dapat menentukan model yang memiliki ketahanan yang paling bagus, diperulkan beda persentase antara akurasi orisinil dan akurasi setelah di serang serta persentase keberhasilan serangan pada setiap model. Berikut merupakan tabel persentase dari summary hasil serangan.

Tabel 6. Summary Persentase Serangan

Model	Akurasi Orisinil	Akurasi Setelah Diserang	ΔAkurasi	Persentase Keberhasilan Serangan
MNB	0.534	0.036	0.498	0.932
LSTM	0.754	0.136	0.618	0.819
IndoBERTweet	0.938	0.202	0.736	0.784

Akurasi orisinil pada serangan dan akurasi pada hasil model berbeda karena TextAttack melakukan testing lagi menggunakan data test. Setelah itu dilakukan query serangan.

Pada Tabel 6 dapat terlihat bahwa model MNB dengan akurasi orisinil terendah memiliki akurasi yang paling rendah setelah terkena adversarial attack. Dan IndoBERTweet dengan akurasi tertinggi memilki akurasi yang paling tinggi meskipun terkena serangan. Model LSTM tetap berada diantara kedua model. ΔAkurasi (Akurasi orisinil – Akurasi setelah diserang) pada model IndoBERTweet menunjukkan angka yang paling tinggi dari kedua model. Sedangkan model MNB memiliki ΔAkurasi terendah. Hal ini bisa terjadi karena tingginya akurasi awal dari model IndoBERTweet yang dapat menyebabkan ΔAkurasi memiliki angka yang tertinggi. Namun, secara persentase keberhasilan IndoBERTweet memiliki angka terendah dari kedua model. Sedangkan, model MNB dengan angka tertinggi pada persentase keberhasilan serangan. Dari angka tersebut bisa disimpulkan model dengan ketahanan dari yang terbaik hingga terburuk merupakan model IndoBERTweet, model LSTM, dan model MNB.

3.4 Adversarial Examples

Adversarial examples merupakan output setelah serangan adversarial attack dilakukan. Berikut merupakan tabel contoh adversarial examples dari setiap model yang diserang.

Tabel 7. Adversarial Examples Model MNB

Original Text	***Pertubed Text***	***Output***	***Pertubed Output***	***True Output***	***Result***
melihat komen nya negatif jadi pikir pikir buat mencari tiket kereta di sini mending aplikasi yang lain saja yang sudah terbukti bagus bertahun bertahun	melihat komen nya negatif jadi pikir pikir buat mencari tiket kereta di sini mending aplikasi yang lain saja yang sudah terbukti bagus bertahun bertahun	2	2	0	*Skipped*
ada apa dengan young lex kenapa dia harus mengirim hal hal tidak [[berfaedah]] kayak gitu merusak moral banget sumpah	ada apa dengan young lex kenapa dia harus mengirim hal hal tidak [[bagus]] kayak gitu merusak moral banget sumpah	0	2	0	*Successful*
samsung galaxy s saya sudah diluncurkan semalam saya kecewa desain nya standar [[banget]] ekspektasi saya	samsung galaxy s saya sudah diluncurkan semalam saya kecewa desain nya standar [[sangat]] ekspektasi saya [[memang]]	0	0	0	*Failed*

[[cuma]] terwujud percuma terbawa sampai mimpi	terwujud percuma terbawa sampai mimpi

Tabel 8. Adversarial Examples Model LSTM

Original Text	***Pertubed Text***	***Output***	***Pertubed Output***	***True Output***	***Result***
biasanya pesan tiket pesawat lancar lancar saja terakhir beli tiket kereta eksekutif untuk orang sudah ditransfer sebelum waktu berakhir tetapi cek order pemesanan kadaluarsa menyebalkan	biasanya pesan tiket pesawat lancar lancar saja terakhir beli tiket kereta eksekutif untuk orang sudah ditransfer sebelum waktu berakhir tetapi cek order pemesanan kadaluarsa menyebalkan	2	2	0	*Skipped*
aku lagi makan mangga potong aku [[suka]] banget buah mangga besok kalau kepasar mau beli banyak banyak lagi ah	aku lagi makan mangga potong aku [[harap]] banget buah mangga besok kalau kepasar mau beli banyak banyak lagi ah	2	0	2	*Successful*
buat gue kece [[banget]] lenovo itu saking kece nya layar laptop lenovo yoga bisa ditekuk derajat [[keren]] [[banget]] memang benar sih kadang different adalah lebih baik	buat gue kece [[sangat]] lenovo itu saking kece nya layar laptop lenovo yoga bisa ditekuk derajat [[cergas]] [[terlalu]] memang benar sih kadang different adalah lebih baik	2	2	2	*Failed*

Tabel 9. Adversarial Examples Model IndoBERTweet

Original Text	***Pertubed Text***	***Output***	***Pertubed Output***	***True Output***	***Result***
iya enak banget soal nya iya dong indomie mah teh terbaik iya lalu aku juga tidak suka orang yang padahal jelas jelas dia salah dan harus nya minta maaf tapi adem ayem saja sedangkan aku nya sudah menderita	iya enak banget soal nya iya dong indomie mah teh terbaik iya lalu aku juga tidak suka orang yang padahal jelas jelas dia salah dan harus nya minta maaf tapi adem ayem saja sedangkan aku nya sudah menderita	0	0	2	*Skipped*

saya [[kecewa]] sama aplikasi ini padahal [[cek]] in nya [[jam]] malahan saya ketinggal pesawat kembalikan [[uang]] saya	saya [[menjelajah]] sama aplikasi ini padahal [[audit]] in nya [[ketika]] malahan saya ketinggal pesawat kembalikan [[penukar]] saya	0	1	0	*Successful*
saya [[kecewa]] karena saya sudah transfer dan belum dikonfirmasi saya [[telepon]] malah [[pelanggan]] [[sibuk]] terus aduh [[kecewa]] nih gue	saya [[gagal]] karena saya sudah transfer dan belum dikonfirmasi saya [[telefon]] malah [[pengguna]] [[kegiatan]] terus aduh [[urung]] nih gue	0	0	0	*Failed*

Tabel 7, Tabel 8, dan Tabel 9 tersebut diambil dari 500 total adversarial examples. Kolom Original Text merupakan kolom teks orisinil dari dataset. Pertubed Text merupakan kolom teks yang telah dipertubasi. Jika hasil dari serangan adalah skipped maka tidak akan terjadi pertubasi teks. Output merupakan kolom output label saat TextAttack melakukan klasifikasi menggunakan model. Pertubed Output merupakan output label setelah diserang dengan adversarial attack. True Output merupakan output sebenarnya yang berada pada dataset. Result merupakan hasil serangan.

4. Kesimpulan

Penelitian ini bertujuan untuk menyerang dan membandingkan tiga model machine learning dengan arsitektur yang berbeda. Ketiga model tersebut mendapatkan dampak yang berbeda terhadap serangan adversarial attack. Model yang awalnya memiliki tingkat akurasi yang cukup tinggi akan menurun secara drastis setelah mendapat serangan dari adversarial attack.

Dari hasil penelitian, model MNB dengan akurasi sebesar 53% menurun menjadi 3% setelah diserang dengan adversarial attack dengan tingkat keberhasilan serangan sebesar 93%. Model LSTM dengan akurasi sebesar 75% menurun menjadi 13% setelah diserang dengan adversarial attack dengan Tingkat keberhasilan serangan sebesar 81%. Model pre-trained IndoBERTweet dengan akurasi sebesar 93% menurun menjadi 20% setelah terkena serangan adversarial attack dengan tingkat keberhasilan serangan sebesar 78%. Dari semua hasil tersebut, model MNB memiliki beda akurasi yang paling kecil dan memiliki tingkat keberhasilan serangan yang paling tinggi. Beda akurasi yang kecil pada model MNB dikarenakan akurasi model awal yang paling kecil, akan tetapi model tersebut paling rentan terhadap serangan adversarial attack. model pre-trained memiliki ketahanan yang paling besar diantara dua model lainnya.

Penelitian ini memiliki keterbatasan pada data teks yang hanya diambil pada situs HuggingFace. Keterbatasan yang lain terdapat pada keterbatasan waktu, karena untuk melakukan training dan implementasi TextAttack pada model memerlukan waktu yang cukup lama, terutama pada model IndoBERTweet.

Untuk penelitian selanjutnya disarankan untuk melakukan preprocessing yang lebih mendalam seperti penggunaan data balancing dan membuat arsitektur model yang lebih optimal. Hal tersebut bisa berpengaruh pada akurasi model yang mana dapat memberikan hasil perbandingan serangan yang lebih bagus.

Referensi

[1] A. C. Müller dan S. Guido, *Introduction to Machine Learning with Python A GUIDE FOR DATA SCIENTISTS Introduction to Machine Learning with Python*, 1 ed. Sebastopol: O'Reilly Media, Inc., 2016.

[2] A. Vaswani *dkk.*, "Attention Is All You Need," dalam *31st Conference on Neural Information Processing Systems*, Long Beach, Jun 2017. [Daring]. Tersedia pada: http://arxiv.org/abs/1706.03762

[3] X. Luo, H. Ding, M. Tang, P. Gandhi, Z. Zhang, dan Z. He, "Attention Mechanism with BERT for Content Annotation and Categorization of Pregnancy-Related Questions on a Community QA Site," dalam *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, Institute of Electrical and Electronics Engineers Inc., Des 2020, hlm. 1077–1081. doi: 10.1109/BIBM49941.2020.9313379.

[4] J. Devlin, M.-W. Chang, K. Lee, dan K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," dalam *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, dan T. Solorio, Ed., Minneapolis, Minnesota: Association for Computational Linguistics, Jun 2019, hlm. 4171–4186. doi: 10.18653/v1/N19-1423.

[5] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, dan Y. Qi, "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP," 2020. [Daring]. Tersedia pada: https://arxiv.org/abs/2005.05909

[6] S. Ren, Y. Deng, K. He, dan W. Che, "Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency," dalam *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence: Association for Computational Linguistics, Jul 2019, hlm. 1085–1097. [Daring]. Tersedia pada: https://wordnet.princeton.edu/

[7] L. Li, R. Ma, Q. Guo, X. Xue, dan X. Qiu, "BERT-ATTACK: Adversarial Attack Against BERT Using BERT," dalam *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, dan Y. Liu, Ed., Online: Association for Computational Linguistics, Nov 2020, hlm. 6193–6202. doi: 10.18653/v1/2020.emnlp-main.500.

[8] D. Jin, Z. Jin, J. T. Zhou, dan P. Szolovits, "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment," dalam *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, Jul 2019, hlm. 8018–8025. Diakses: 21 Januari 2025. [Daring]. Tersedia pada: https://ojs.aaai.org/index.php/AAAI/article/view/6311/6167

[9] J. Li, S. Ji, T. Du, B. Li, dan T. Wang, "TEXTBUGGER: Generating Adversarial Text Against Real-world Applications," dalam *26th Annual Network and Distributed System Security Symposium, NDSS 2019*, San Diego: The Internet Society, Feb 2019. doi: 10.14722/ndss.2019.23138.

[10] D. Jurafsky dan J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. 2025. [Daring]. Tersedia pada: https://web.stanford.edu/~jurafsky/slp3/