Adversarial Attack Pada Teks Berbahasa Indonesia Menggunakan Framework TextAttack

Ilham Aulady Miftahurrizqy
202010370311462

Dosen Pembimbing
Setio Basuki, MT., Ph.D.
NIP. 10809070477PNS.

Introduction

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. ML juga dibagi menjadi beberapa bagian, salah satunya adalah Natural Language Processing atau NLP. 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. Meskipun model yang dihasilkan itu sudah dinilai baik, masih ada resiko yang dipertimbangkan, seperti Adversarial Attack. Serangan adversarial pada teks merupakan serangan yang bertujuan untuk menipu model NLP dengan membuat sebuah adversarial examples berbentuk teks yang telah diperturbasi. Penelitian serangan adversarial ini bertujuan untuk membuat sebuah metode yang dapat menyerang model sentiment analysis dengan Bahasa Indonesia dan menghasilkan sebuah adversarial example dengan membandingkan performa dari tiga model NLP dengan karakteristik yang berbeda.

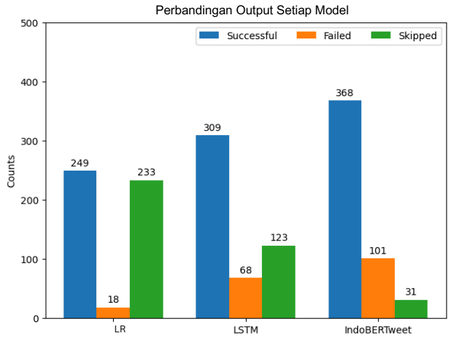
Objective

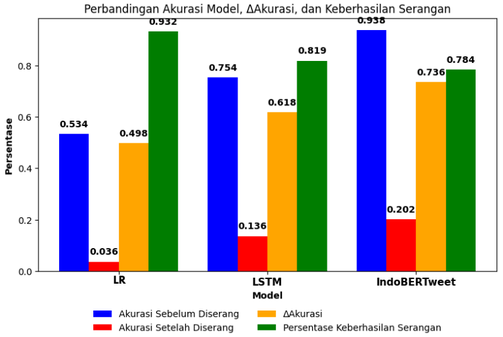
Tujuan dari penelitian ini yaitu, untuk mengetahui dampak yang ditimbulkan dan beda akurasi dari model yang telah diserang oleh adversarial attack. Mengetahui persentase serangan adversarial yang berhasil menggunakan TextAttack pada ketiga model. Mengetahui model NLP yang lebih tahan terhadap serangan adversarial dari keefektivitasan serangan adversarial.

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

Result





Feature Selection

Penelitian ini hanya menggunakan teks dan label dari dataset. Proses pelatihan didasarkan pada data dengan label teks yang akan digunakan model untuk melakukan klasifikasi. Model yang telah jadi akan diserang menggunakan resep pada framework TextAttack, yaitu PWWSRen2019. Penyerangan ini akan menghasilkan adversarial example pada setiap model NLP.

Model

Penelitian ini menggunakan tiga model NLP antara lain, Logistic Regression, LSTM, dan pre-trained IndoBERTweet. Penggunaan model yang berbeda ini karena adanya beda karakteristik dari setiap model, dimana model Logistic Regression merupakan metode klasik dari ML, LSTM menggunakan metode jaringan neural sedangkan IndoBERTweet menggunakan metode attention secara Bi-directional. Perbedaan karakteristik tersebut meliputi tingkat kompleksitas, akurasi, dan resources.

Conclusion

LR memiliki akurasi awal terendah, tetapi paling rentan terhadap serangan.
IndoBERTweet memiliki daya tahan terbaik, dengan persentase keberhasilan paling kecil.

Discussion

Dari hasil tersebut bisa disimpulkan Model dengan ketahanan dari yang terbaik adalah IndoBERTweet karena persentase keberhasilan serangan yang paling rendah. Model IndoBERTweet memiliki ketahan yang terbaik karena kompleksitas model yang tinggi dan mekanisme attention dari model yang menimbang pentingnya setiap kata dalam teks. Jika sebuah kata berubah pada teks maka model IndoBERTweet bisa mempertimbangkan kata yang lain untuk memahami konteks dari teks.