INTISARI

INDOSUM-ATTACKER: MODEL SERANGAN ADVERSARIAL UNTUK PERINGKASAN TEKS ABSTRAKTIF

Oleh

Damarjati Maulana Hilmi

23/525106/PPA/06590

Dalam era digital saat ini, peningkatan volume informasi menuntut adanya sistem yang mampu meringkas teks secara otomatis dengan akurasi tinggi. Namun, model-model peringkasan teks yang ada rentan terhadap serangan adversarial, seperti *typo* dan perbedaan gaya penulisan, yang dapat mengurangi performa dan kualitas ringkasan yang dihasilkan. Penelitian ini mengusulkan IndoSum-Attacker, sebuah model serangan adversarial berbasis penambahan data yang mencakup lima teknik modifikasi teks, yaitu: pergantian sinonim, penyisipan acak, penghapusan acak, penukaran acak, dan penerjemahan balik.

Evaluasi dilakukan terhadap dua model peringkasan teks abstraktif, yaitu BART dan mBART, dengan menggunakan dataset IndoSum serta metrik ROUGE untuk pengukuran performa model. Hasil pengujian menunjukkan bahwa semua jenis serangan mengakibatkan penurunan performa model secara signifikan. Operasi penerjemahan balik memberikan dampak penurunan paling besar, dengan skor ROUGE-L F1 pada BART turun dari 0,532 menjadi 0,491 pada tingkat modifikasi 50%, sedangkan pada mBART terjadi penurunan lebih drastis dari 0,791 menjadi 0,674.

BART terbukti lebih tangguh terhadap modifikasi input, sedangkan mBART unggul dalam menghasilkan ringkasan yang lebih terstruktur dan informatif. Perbandingan performa menunjukkan adanya pertukaran antara kualitas ringkasan dan ketangguhan terhadap serangan adversarial. Temuan ini memberikan gambaran awal mengenai potensi dampak serangan adversarial terhadap performa model peringkasan teks serta membuka peluang untuk penelitian lanjutan dalam mengeksplorasi aspek ketahanan dan kualitas ringkasan secara lebih mendalam.

Kata Kunci: *Natural Language Processing*, Peringkasan Teks Abstraktif, Serangan Adversarial, Penambahan Data, Ketangguhan Model

ABSTRACT

INDOSUM-ATTACKER: ADVERSARIAL ATTACK MODEL FOR ABSTRACTIVE TEXT SUMMARIZATION

By

Damarjati Maulana Hilmi

23/525106/PPA/06590

In today's digital era, the increasing volume of information demands a system capable of automatically summarising text with high accuracy. However, existing text summarisation models are vulnerable to adversarial attacks, such as typos and writing style differences, which can reduce the performance and quality of the resulting summaries. This research proposes IndoSum-Attacker, a data augmentation-based adversarial attack model that includes five text modification techniques, namely: synonym replacement, random insertion, random deletion, random swap, and back translation.

Evaluation was conducted on two abstractive text summarization models, BART and mBART, using IndoSum dataset and ROUGE metric for model performance measurement. The test results show that all types of attacks result in a significant decrease in model performance. The back translation operation had the greatest decreasing impact, with the ROUGE-L F1 score on BART dropping from 0.532 to 0.491 at the 50% modification level, while on mBART there was a more drastic drop from 0.791 to 0.674.

BART proved to be more resilient to input modifications, while mBART excelled in producing more structured and informative summaries. The performance comparison shows a trade-off between summarization quality and robustness against adversarial attacks. These findings provide an initial insight into the potential impact of adversarial attacks on the performance of text summarization models, and open up opportunities for further research in exploring the aspects of resilience and summarization quality in more depth.

Keywords: Natural Language Processing, Abstractive Text Summarization, Adversarial Attack, Data Augmentation, Model Robustness