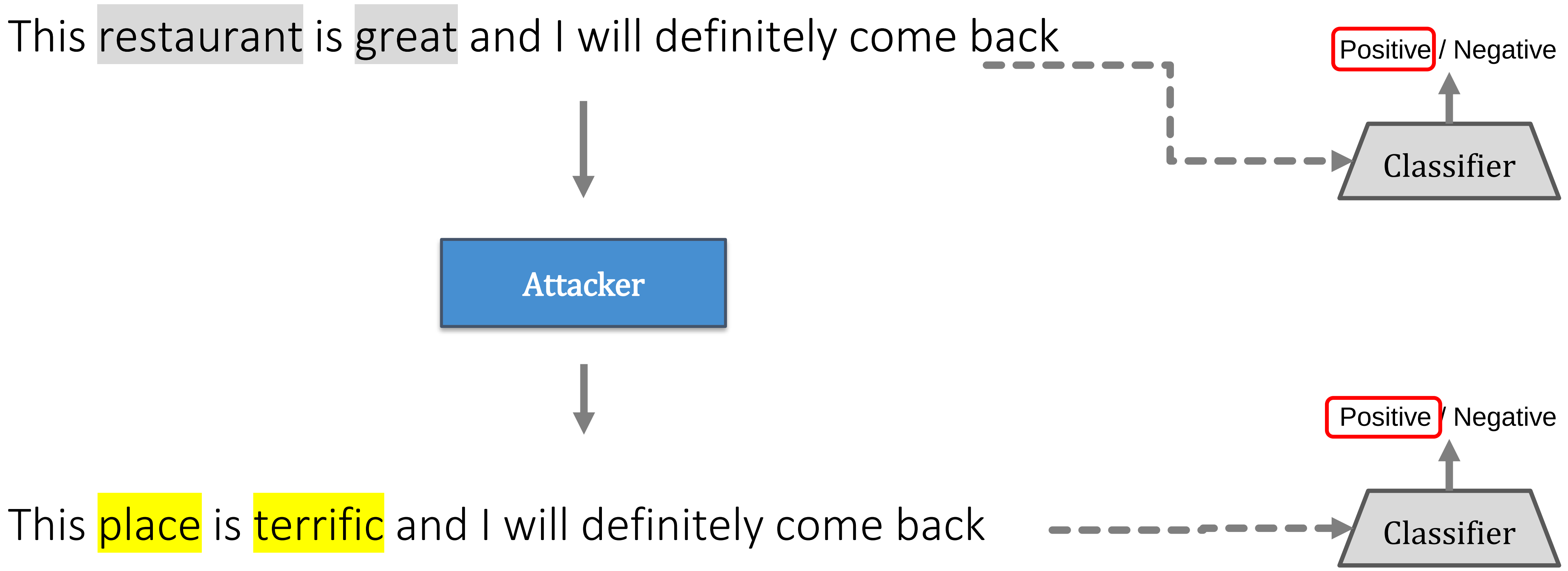
Explain2Attack: Text Adversarial Attacks via Cross-Domain Interpretability

Mahmoud Hossam, Trung Le, He Zhao and Dinh Phung

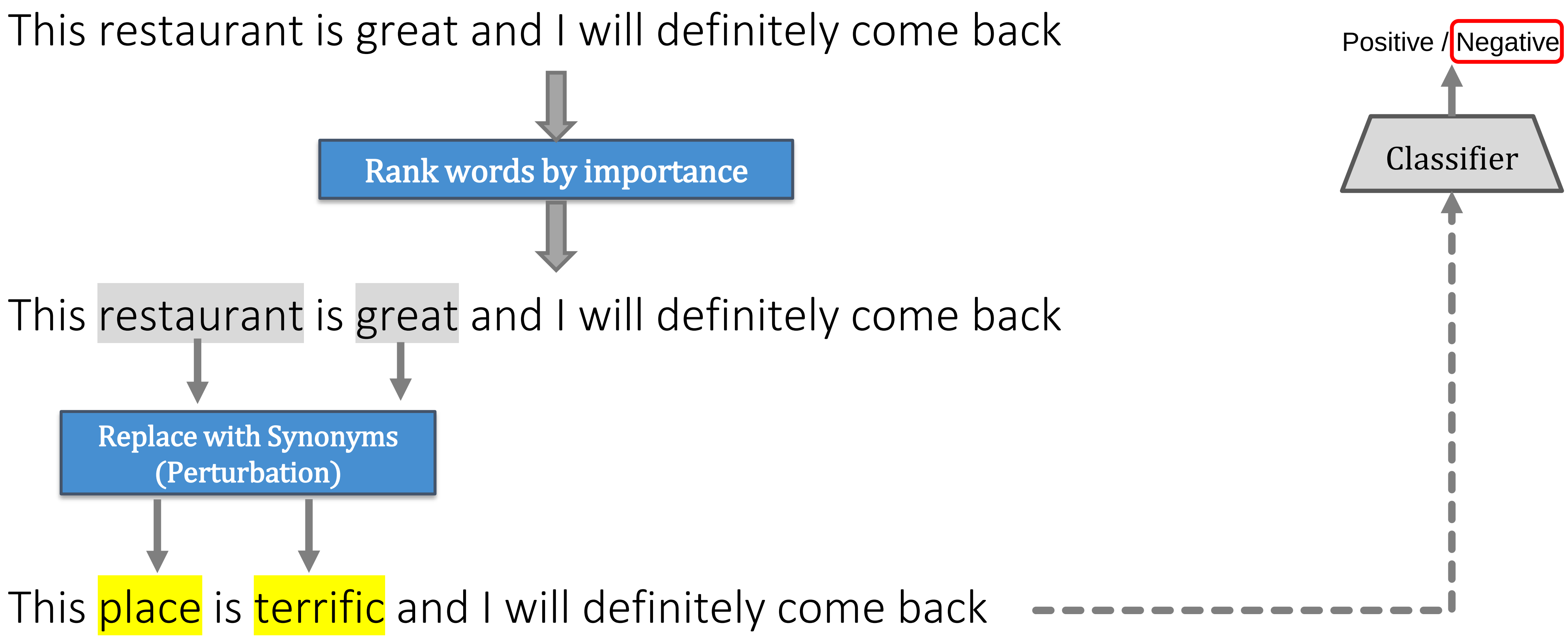
Faculty of Information Technology, Monash University
mahmoud.hossam@gmail.com, {mhossam, trunglm, ethan.zhao, dinh.phung}@monash.edu

Poster: 2463

Adversarial Examples Natural Language



Generation Steps



Generation Steps: Word Importance Ranking

- **Words importance score** I_{w_i} for word w_i is a function Φ of the target model's probability P for the whole sentence excluding w_i: $I_{w_i} = \Phi(\, P(Y \mid X_{1:T}), P(Y \mid X_{1:T \setminus \{i\}})\,)$
- Is done word by word:

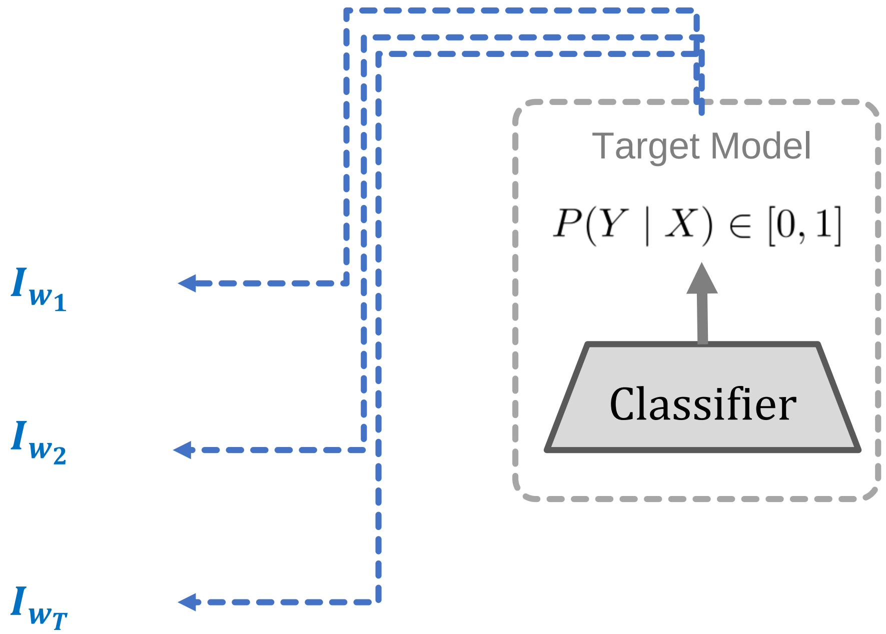
This restaurant is great and I will definitely come back

[ ] restaurant is great and I will definitely come back

This [ ] is great and I will definitely come back

⋮ ⋮

This restaurant is great and I will definitely come [ ]



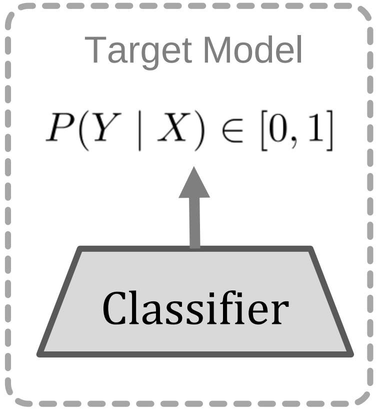
Problem: Number of queries needed for word ranking = Length (Sentence)

Generation Steps: Word Importance Ranking

- Challenges in black-box setting
  - Number of required queries

This restaurant is great and I will definitely come back

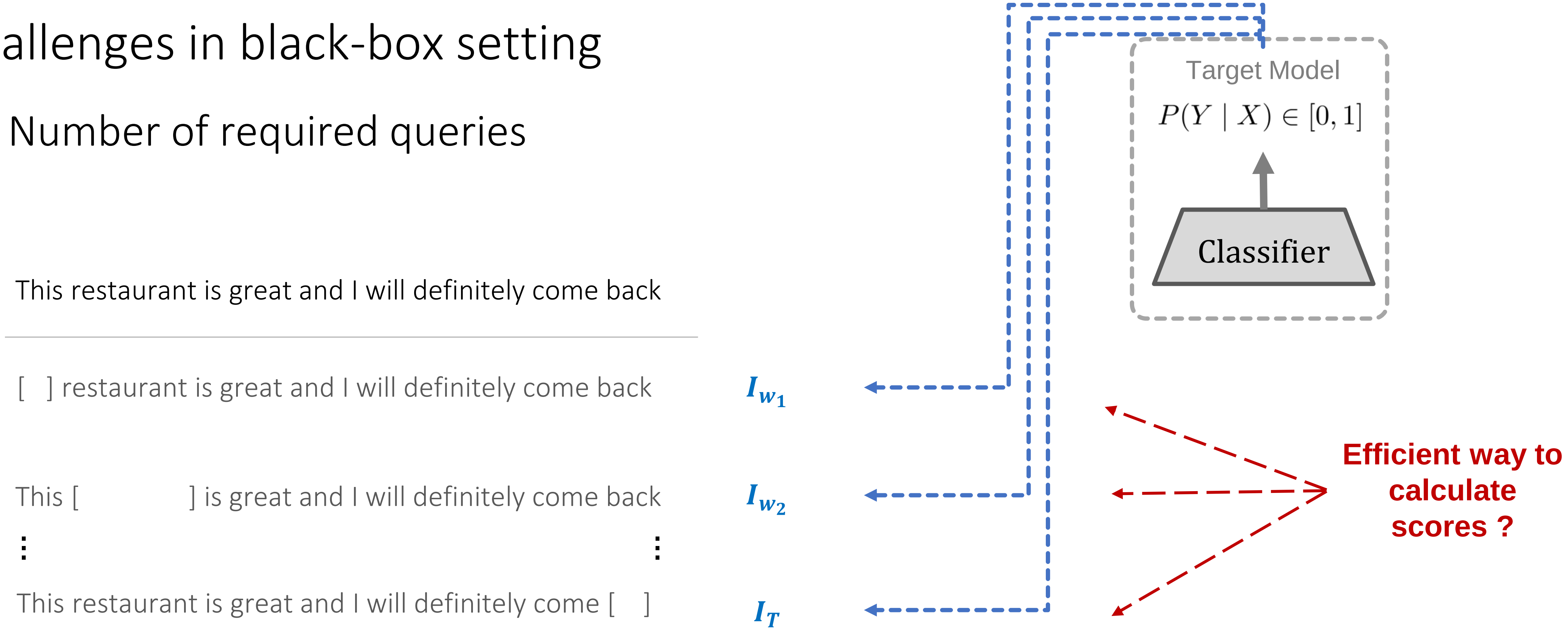
I wonder how I didn't know about this before, but this place is the best ! **Needs a query for each word in a sentence**



  - Raise suspicion towards attacking agent

Generation Steps: Word Importance Ranking

- Challenges in black-box setting
  - Number of required queries

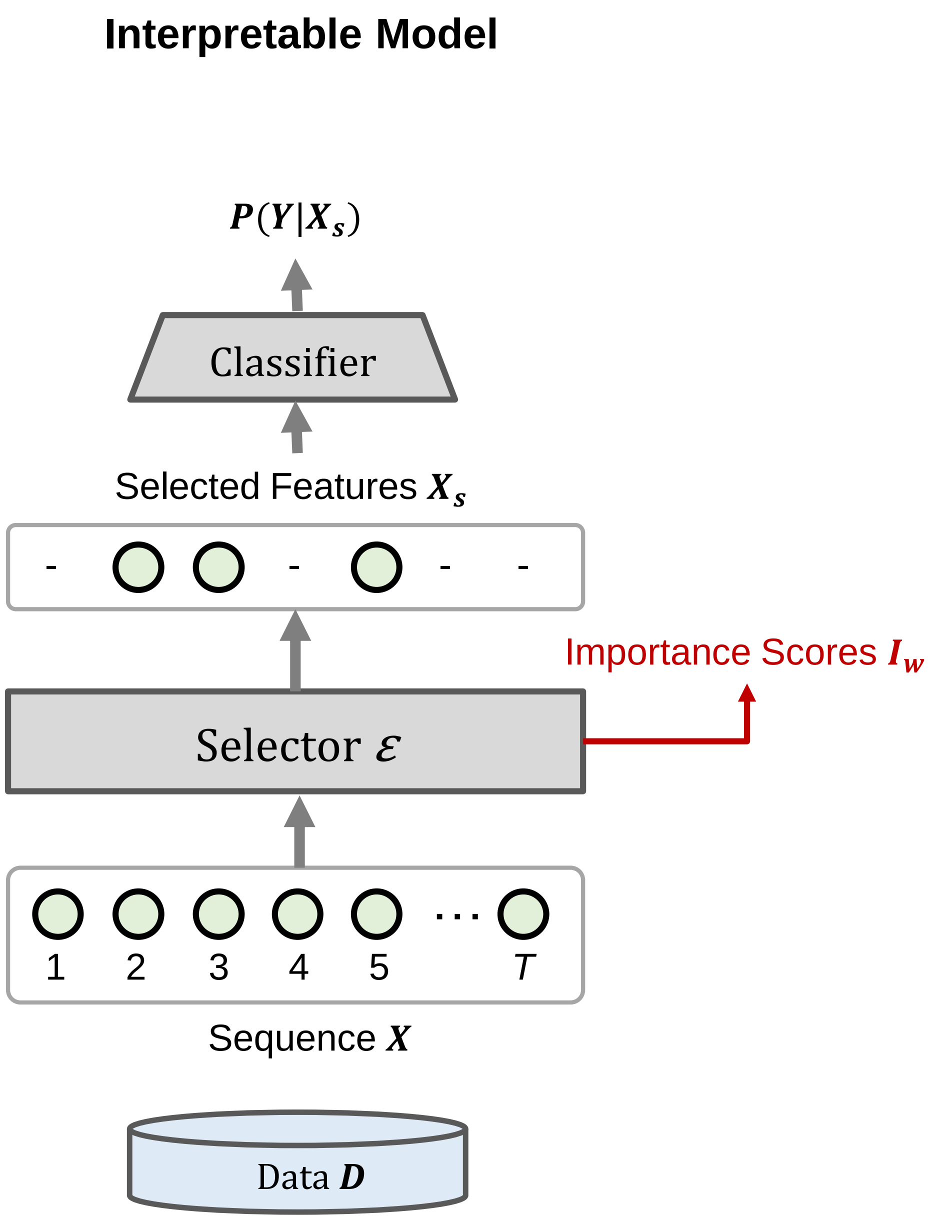


Interpretability

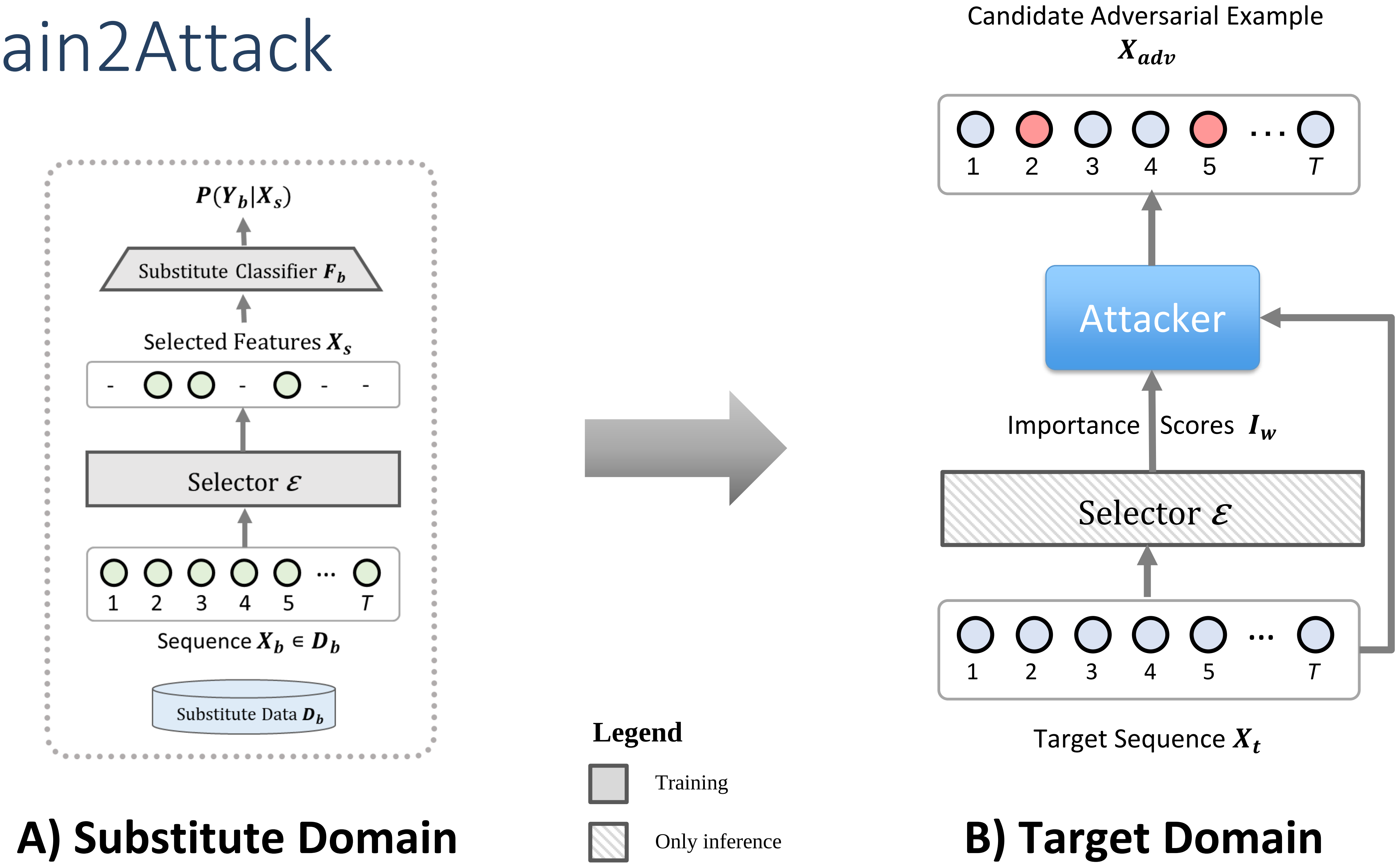
- Employ Interpretability
  - Can learn important features from X
  - Objective: Maximize Mutual Information

$$\max_{\mathcal{E}} I\left(X_S; Y\right)$$

  - Logits can be used as importance scores I_w



Explain2Attack



Results

- **Explain2Attack** reduced the average number of queries compared the baseline **TextFooler.**
- And achieves same or better attack rate, with higher **Query Efficiency (QE)**

Statistic of Used Datasets

Dataset	Train	Test	Avg. Length
IMDB	25K	25K	215
MR	9K	1K	20
Amazon MR	25K	25K	100
Yelp	560K	38K	152

After-Attack Accuracies, Queries and Query Efficiency

Classifier		BERT		WordCNN			WordLSTM		
Target Model		IMDB	MR	IMDB	MR	Amazon MR	IMDB	MR	Amazon MR
	Clean_Acc.	92.18	89.97	87.32	79.85	90.14	88.78	81.82	91.30
Adv_Acc. ↓	TextFooler (Jin et al., 2019)	11.88	13.59	**0.60**	1.50	**3.92**	**0.04**	**2.06**	**2.15**
	(Substitute Data)	*(Yelp)*	*(Amazon MR)*	*(Yelp)*	*(IMDB)*		*(Amazon MR)*		*(IMDB)*
	Explain2Attack (ours)	**11.32**	**13.34**	0.61	**1.31**	3.97	0.06	2.27	2.38
Avg_Queries ↓	TextFooler	980.5	**181.6**	444	112.8	378.7	500.2	117.5	392.7
	Explain2Attack	**873.5**	184.07	**404.5**	**108.7**	**349.4**	**440.5**	**114.2**	**369.3**
Query Efficiency (QE) ↑	TextFooler	0.082	**0.421**	0.195	0.695	0.228	0.177	0.679	0.227
	Explain2Attack	**0.093**	0.416	**0.214**	**0.723**	**0.247**	**0.201**	**0.697**	**0.241**

Results

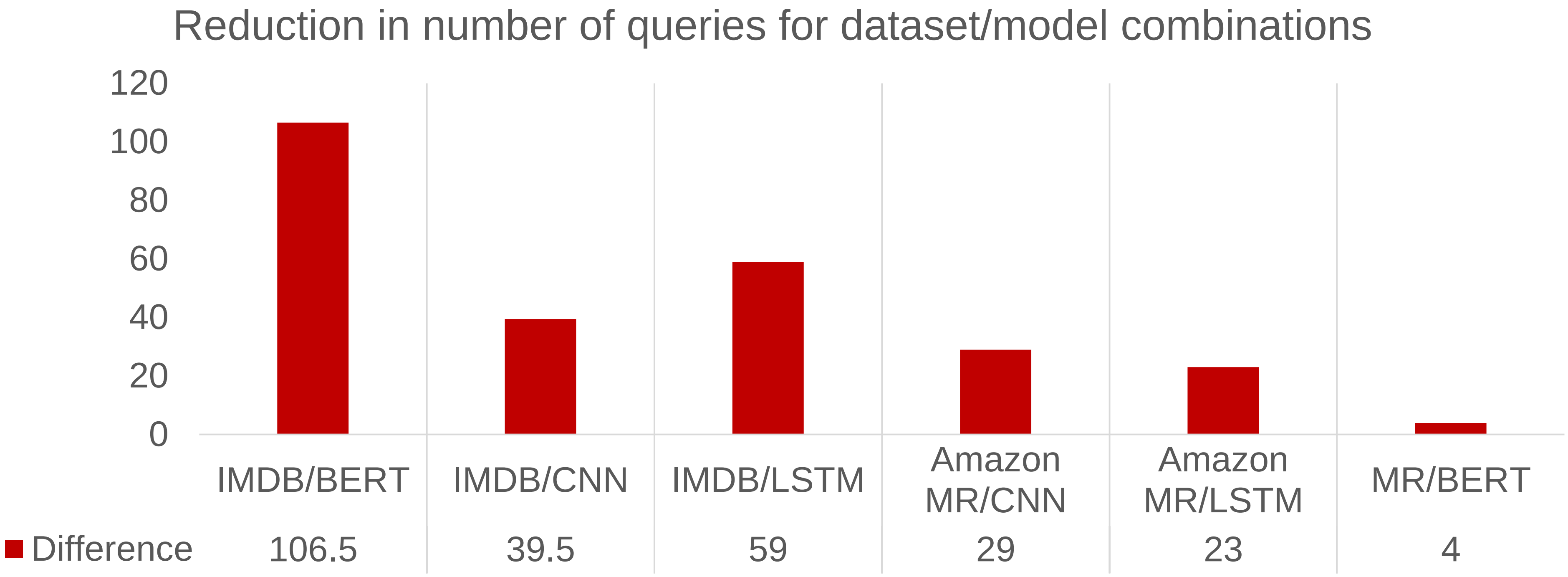


Table 5.3: Effect of Sentence Length on Number of Queries

Target Dataset		IMDB			Amazon MR		MR		
Classifier		BERT	CNN	LSTM	CNN	LSTM	BERT	CNN	LSTM
Average Sentence Length		215			100		20		
Avg_Queries $\downarrow$	TextFooler	980.5	444	500.2	378.7	392.7	112.8	117.5	**181.6**
	Explain2Attack	**873.5**	**404.5**	**440.5**	**349.4**	**369.3**	**108.7**	**114.2**	184.07
Difference		106.5	39.5	59.7	29.3	23.4	4.1	3.3	-3.0

Conclusion

- First framework to learn word importance in black-box setting.
- Reduces query cost and computational complexity.
- Achieves similar or better attack rates than state-of-the-art.
- Not affected by input length
  - Very efficient for long input sentences

Thank You

Poster: 2463

mahmoud.hossam@gmail.com

@mahossam