Aladdin Genies:
Arabic Adversarial Text Normalization Attacks

Presented by: Saied Alshahrani & Norah Alshahrani
Supervision: Prof. Jeanna Matthews & Dr. Soumyabrata Dey

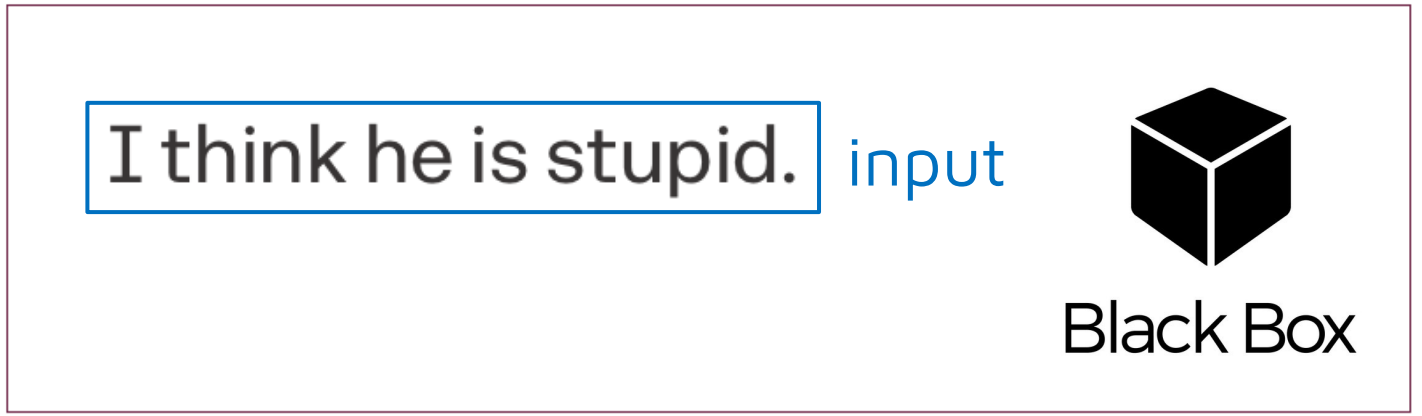
Introduction

- **White-box testing/setting:**
  - The attackers **can** access the model architecture, weights, parameters, or training datasets.
- **Black-box testing/setting:**
  - The attackers **cannot** access the model architecture, weights, parameters, or training datasets.
  - They **only** query the model and get a prediction in return.

Introduction

- **Example of black—box classification systems:**

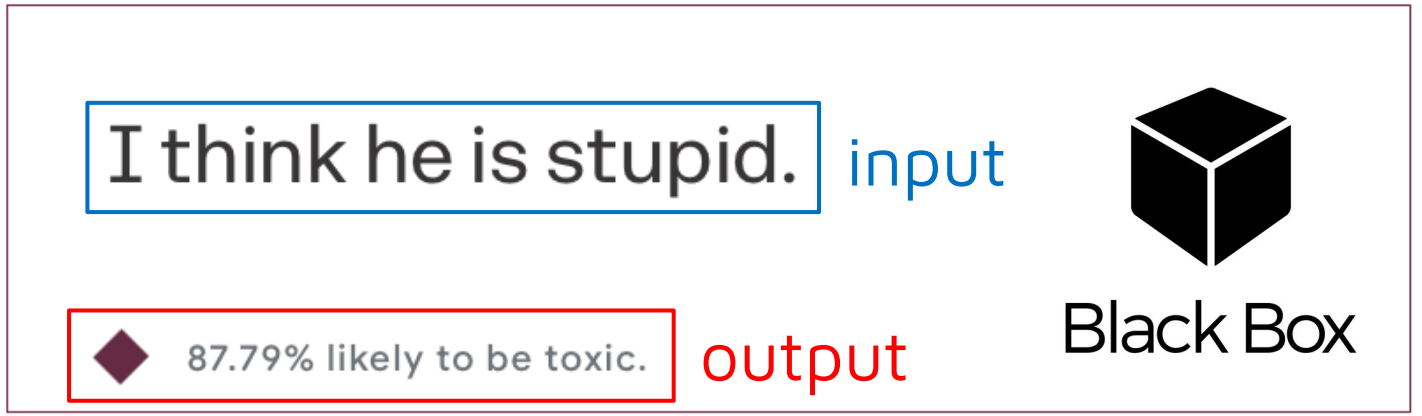
Google Perspective API



Introduction

- **Example of black—box classification systems:**

Google Perspective API



Background

- **Arabic Text Normalization:**
  - Normalization **unifies** the orthography of Arabic text.
  - Normalization **reduces** the space of word embeddings.
  - Normalization **deletes** the noise characters in Arabic.

Background

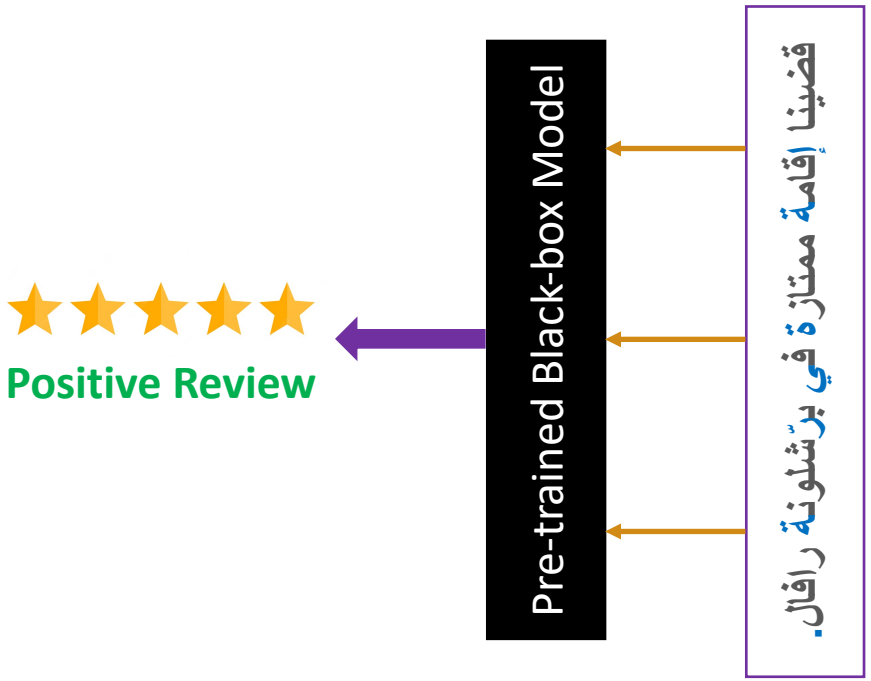
- **Arabic Text Normalization Techniques:**
  - Normalizing **Alef Variants** [ا آ أ إ] → [ا].
  - Normalizing **Alef Maksura** [ى] → [ي].
  - Normalizing **Teh Marbuta** [ة] → [ه].
  - Removing **Diacritical** marks.
  - Removing **Punctuation** marks.

Motivations

1. Study the robustness of Deep Neural Networks (DNN) pre-trained/black-box Arabic text classification models.
2. Highlight the importance of text normalization in Arabic.

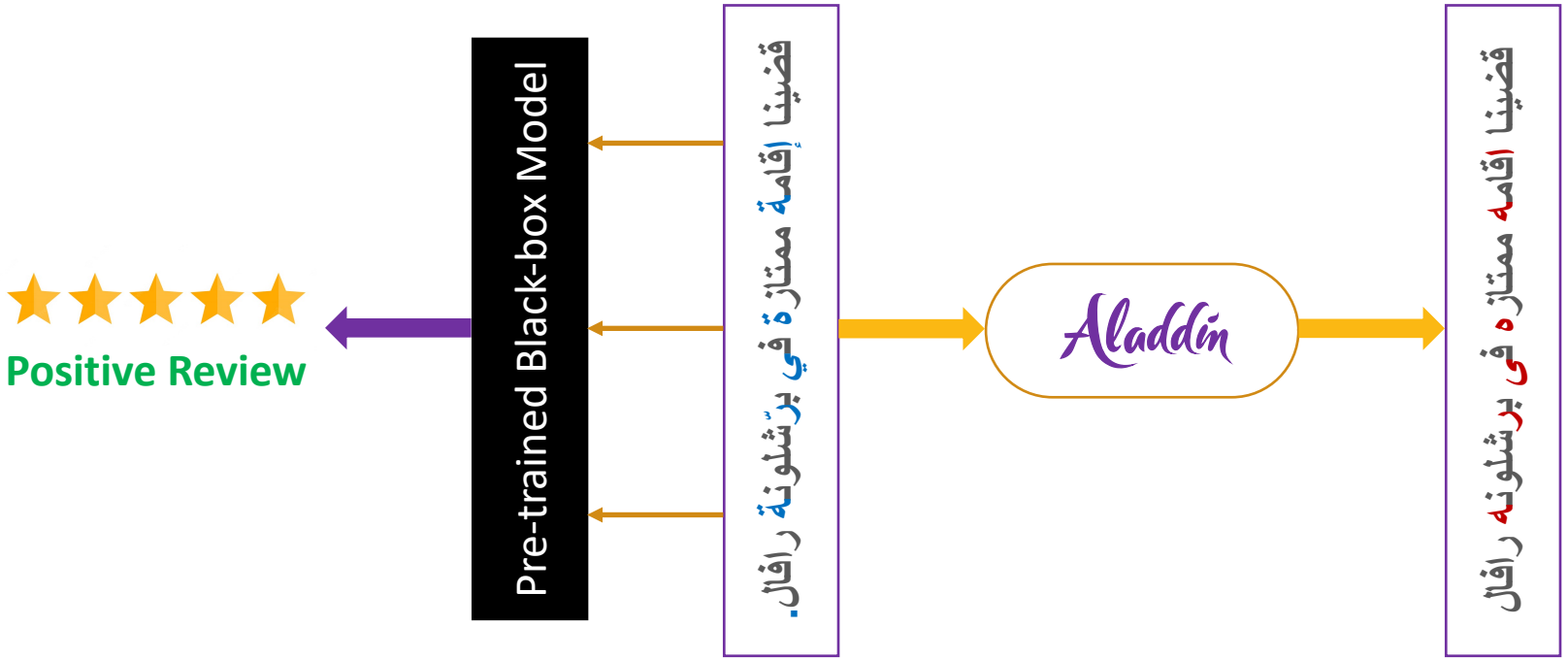
Aladdin Algorithm

- **Goal:** Flip the prediction of an Arabic text classifier.



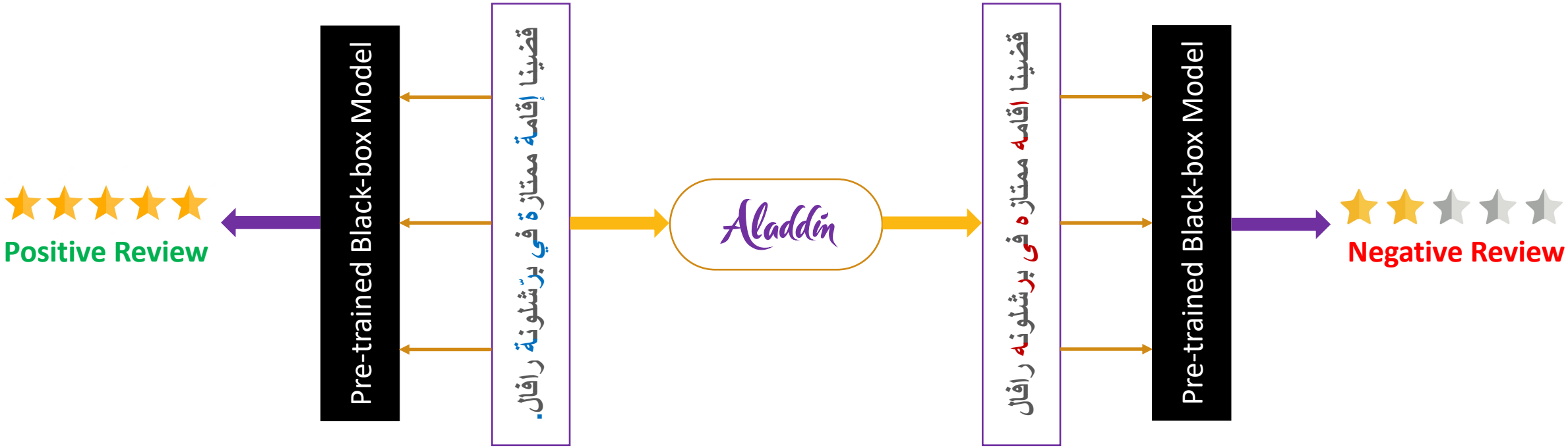
Aladdin Algorithm

- **Goal:** Flip the prediction of an Arabic text classifier.



Aladdin Algorithm

- **Goal:** Flip the prediction of an Arabic text classifier.



Aladdin Genies

- **Aladdin Genies:**

1. Genie of Teh Marbuta:

قضينا إقامة ممتازة في برّشلونة رافال.

قضينا إقامه ممتازه في برّشلونه رافال.

English Translation: "We had an excellent stay at the Barcelona Raval."
Arabic Transliteration: "qDynA AqAmo mmtAzo fy brElwno rAfAl."

Aladdin Genies

- **Aladdin Genies:**

  2. Genie of Alef Variants:

قضينا إقامة ممتازة في برّشلونة رافال.

قضينا اقامة ممتازة في برّشلونة رافال.

English Translation: "We had an excellent stay at the Barcelona Raval."
Arabic Transliteration: "qDynA AqAmo mmtAzo fy brĒlwno rAfAl."

Aladdin Genies

- **Aladdin Genies:**

  3. Genie of Diacritical Marks:

قضينا إقامة ممتازة في برّشلونة رافال.

قضينا إقامة ممتازة في برشلونة رافال.

English Translation: "We had an excellent stay at the Barcelona Raval."
Arabic Transliteration: "qDynA AqAmo mmtAzo fy brElwno rAfAl."

Aladdin Genies

- **Aladdin Genies:**

4. Genie of Punctuation Marks:

قضينا إقامة ممتازة في برّشلونة رافال.

قضينا إقامة ممتازة في برّشلونة رافال

English Translation: "We had an excellent stay at the Barcelona Raval."
Arabic Transliteration: "qDynA AqAmo mmtAzo fy brElwno rAfAl."

Aladdin Genies

- **Aladdin Genies:**

5. All Genies:

قضينا إقامة ممتازة في برّشلونة رافال.

قضينا اقامه ممتازه في برشلونه رافال

English Translation: "We had an excellent stay at the Barcelona Raval."
Arabic Transliteration: "qDynA AqAmo mmtAzo fy brElwno rAfAl."

Mathematical Definitions

- **Adversarial Examples:**
  - Suppose a deep learning Arabic text classifier $\mathcal{F}(\cdot)\colon \mathbb{X} \longrightarrow \mathbb{Y}$
  - Let original example be $x \in \mathbb{X}$ and adversarial example be $\acute{x}$
  - Let true label of x be $y \in \mathbb{Y}$ and predicted label of $\acute{x}$ be $\hat{y}$
  - Let prediction score of x be $p \in \mathbb{P}$ and predicted score of $\acute{x}$ be $\acute{p}$
  - An adversarial example $\acute{x}$ follows:

$$\acute{x} = x + \Delta x \text{ , where } \| \Delta x \|_p < \epsilon \text{ and } \acute{x} \in \mathbb{X}$$

  - The successful attack forces: $\mathcal{F}(x) \neq \mathcal{F}(\acute{x})$ and $y \neq \hat{y}$

Mathematical Definitions

- **Success Rate:**
  - Suppose a deep learning Arabic text classifier $\mathcal{F}(\cdot)\colon \mathbb{X} \longrightarrow \mathbb{Y}$
  - Let original example be $x \in \mathbb{X}$ and adversarial example be $\acute{x}$
  - Let true label of x be $y \in \mathbb{Y}$ and predicted label of $\acute{x}$ be $\hat{y}$
  - Let prediction score of x be $p \in \mathbb{P}$ and predicted score of $\acute{x}$ be $\acute{p}$
  - A success rate $s \in \mathbb{S}$ follows:

$$s = \frac{\#successful\ \acute{x}}{\#original\ x} * 100$$

where successful $\acute{x}$ forces:

$\mathcal{F}(x) \neq \mathcal{F}(\acute{x})$ and $y \neq \hat{y}$

Mathematical Definitions

- **Decrease Rate:**
  - Suppose a deep learning Arabic text classifier $\mathcal{F}(\cdot)$: $\mathbb{X} \longrightarrow \mathbb{Y}$
  - Let original example be $x \in \mathbb{X}$ and adversarial example be $\acute{x}$
  - Let true label of x be $y \in \mathbb{Y}$ and predicted label of $\acute{x}$ be $\hat{y}$
  - Let prediction score of x be $p \in \mathbb{P}$ and predicted score of $\acute{x}$ be $\acute{p}$
  - A decrease rate $d \in \mathbb{D}$ follows:

$$d = \sum_{i=1}^{n} \frac{p_{xi}}{n} - \sum_{i=1}^{\acute{n}} \frac{\acute{p}_{x'i}}{\acute{n}}$$

where $\acute{n}$ is #failed $\acute{x}$ and n is #x and $\mathcal{F}(x) = \mathcal{F}(\acute{x})$ and $y = \hat{y}$

Datasets

Dataset	Task	# Classes	# Samples	# Representative Samples
MADAR Parallel Corpus	Arabic Dialect Identification	26	112,000	253
Arabic Dialects Corpus		5	53,403	627
Arabic Sarcasm V2 Corpus		5	15,548	293
Arabic Books Reviews Corpus	Arabic Sentiment Analysis	5	63,257	6,939
Arabic Sentiment Corpus		3	15,572	3,814
Multi-domain Arabic Sentiment Corpus		3	6,733	346

Sampling Datasets

Step 1
- Calculate the total samples.

Step 2
- Compile the representative samples.

Step 3
- Select only 250 representative samples.

Pretrained Models

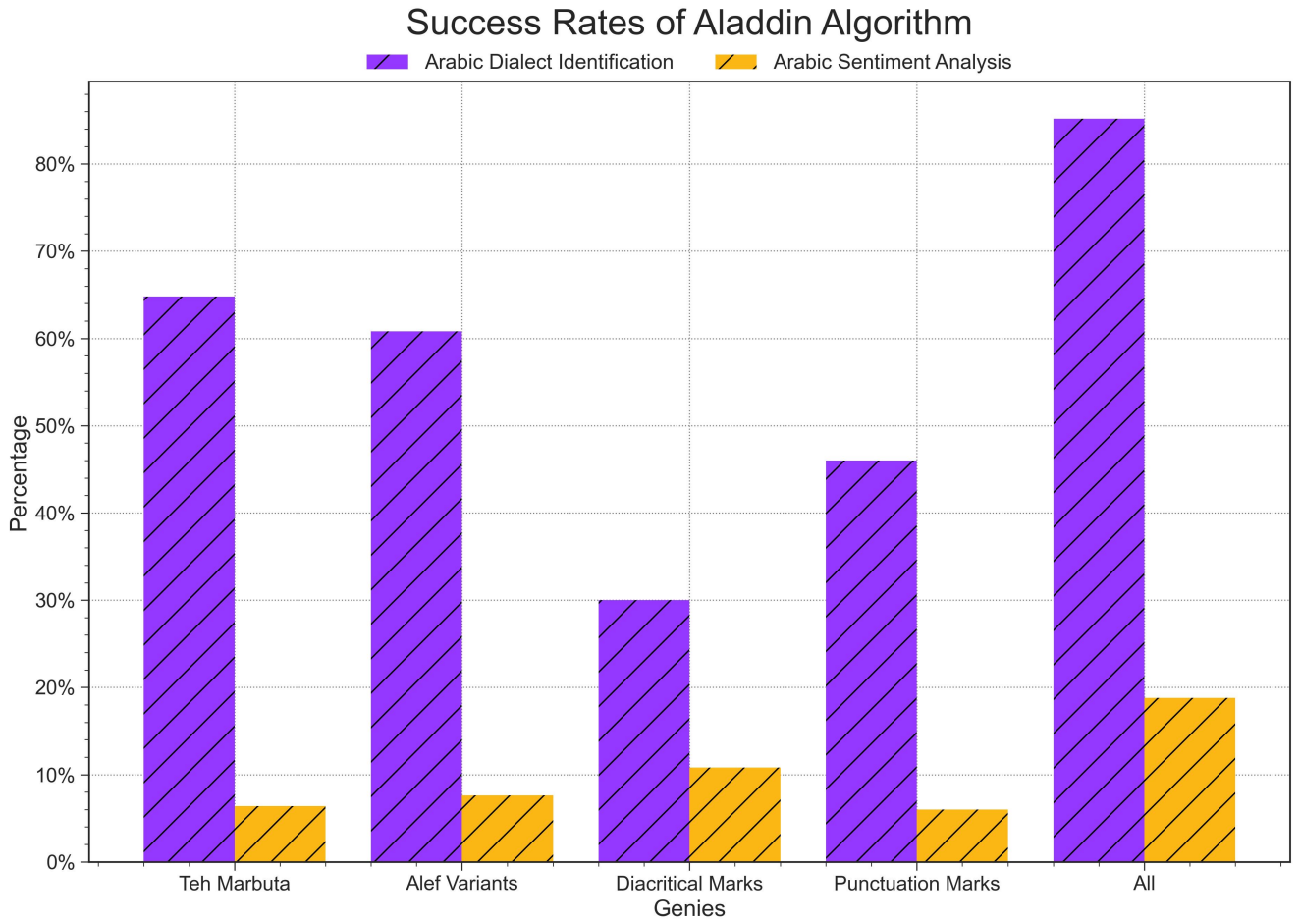
Pretrained Model	Task
ADI 1: Arabic MARBERT Dialect Identification City	Arabic Dialect Identification (ADI)
ADI 2: CAMeLBERT Mix DID Madar Corpus-26	
ADI 3: CAMeLBERT MSA DID MADAR Twitter-5	
ASA 1: Arabic MARBERT Sentiment	Arabic Sentiment Analysis (ASA)
ASA 2: CAMeLBERT Dialectal Arabic Sentiment	
ASA 3: CAMeLBERT Mix Sentiment	

Results: Success Rates

Success Rates of Arabic Dialect Identification Task					
	Teh Marbuta	**Alef Variants**	**Diacritical Marks**	**Punctuation Marks**	**All**
ADI 1	47.60%	0	0	7.60%	52.00%
ADI 2	**64.80%**	**60.80%**	**30.00%**	13.60%	**85.20%**
ADI 3	42.80%	28.00%	29.60%	**46.00%**	60.40%

Success Rates of Arabic Sentiment Analysis Task					
	Teh Marbuta	**Alef Variants**	**Diacritical Marks**	**Punctuation Marks**	**All**
ASA 1	**6.40%**	0	0	5.60%	8.00%
ASA 2	6.00%	**7.60%**	**10.80%**	**6.00%**	**18.80%**
ASA 3	6.00%	5.60%	8.80%	5.60%	14.80%

Results: Success Rates

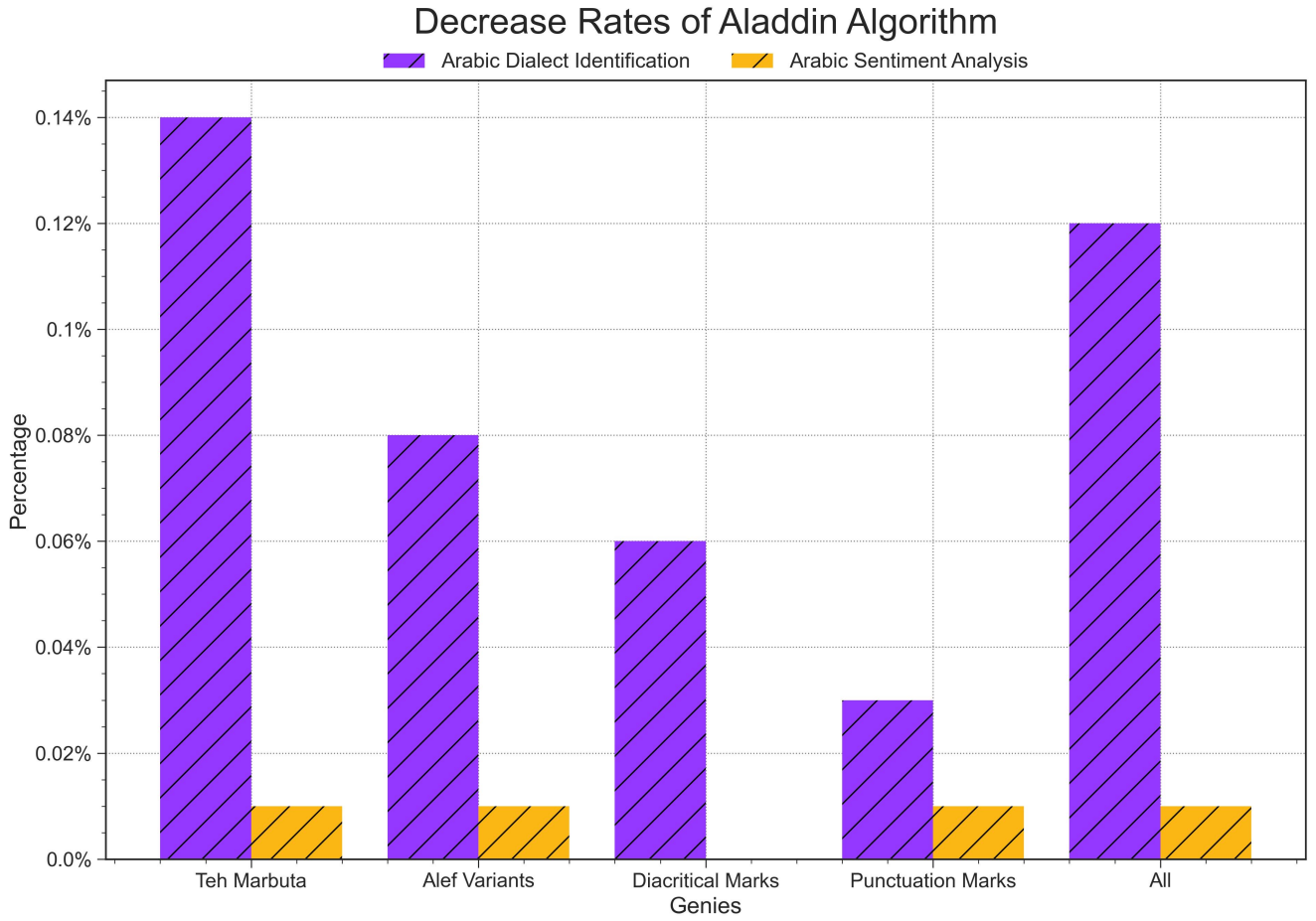


Results: Decrease Rates

Decrease Rates of Arabic Dialect Identification Task					
	Teh Marbuta	Alef Variants	Diacritical Marks	Punctuation Marks	All
ADI 1	**0.14%**	0	0	0.01%	0.11%
ADI 2	0.08%	**0.08%**	-0.05%	**0.03%**	**0.12%**
ADI 3	0.02%	-0.04%	**0.06%**	-0.06%	-0.08%

Decrease Rates of Arabic Sentiment Analysis Task					
	Teh Marbuta	Alef Variants	Diacritical Marks	Punctuation Marks	All
ASA 1	0	0	**0**	0	**0.01%**
ASA 2	**0.01%**	0	**0**	**0.01%**	**0.01%**
ASA 3	0	**0.01%**	-0.01%	**0.01%**	**0.01%**

Results: Decrease Rates



Conclusions

- Aladdin algorithm effectively generates Arabic text adversarial examples up to 94% of total original examples in a pure black-box manner.
- Aladdin algorithm reduces the performance of state-of-the-art deep learning models by up to 85.20%.

Conclusions

- Aladdin algorithm reduces the accuracy of state-of-the-art deep learning models even when the algorithm's adversarial examples failed by up to 0.14%.

- Arabic text normalization improves the accuracy of state-of-the-art deep learning models in specific cases.

Future Works

- We plan to train new DNN models for Arabic text classification, such as RNNs (LSTM or Bi-LSTM) and CNNs, then test their robustness against our algorithm and its generated adversarial examples in a black-box manner.

Future Works

- We plan to study the effectiveness of different defense mechanisms, like adversarial training and spell-checking.
- We plan to examine the transferability of our adversarial examples on other DNN Arabic text classification models.

References

- Gao, Ji, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers. In 2018 IEEE Security and Privacy Workshops (SPW), pp. 50-56. IEEE, 2018.
- Bouamor, Houda, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid et al. The MADAR Arabic Dialect Corpus and Lexicon. In LREC. 2018.
- Boujou, ElMehdi, Hamza Chataoui, Abdellah El Mekki, Saad Benjelloun, Ikram Chairi, and Ismail Berrada. An Open Access NLP Dataset for Arabic Dialects: Data Collection, Labeling, and Model Construction. arXiv preprint arXiv:2102.11000 (2021).
- Farha, Ibrahim Abu, Wajdi Zaghouani, and Walid Magdy. Overview of the WANLP 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic. In Proceedings of the sixth Arabic natural language processing workshop, pp. 296-305. 2021.
- Aly, Mohamed, and Amir Atiya. LABR: A Large-Scale Arabic Book Reviews Dataset. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 494-498. 2013.
- ElSahar, Hady, and Samhaa R. El-Beltagy. Building Large Arabic Multi-domain Resources for Sentiment Analysis. In International conference on intelligent text processing and computational linguistics, pp. 23-34. Springer, Cham, 2015.
- Al-Moslmi, Tareq, Mohammed Albared, Adel Al-Shabi, Nazlia Omar, and Salwani Abdullah. Arabic senti-lexicon: Constructing Publicly Available Language Resources for Arabic Sentiment Analysis. Journal of information science 44, no. 3 (2018): 345-362.