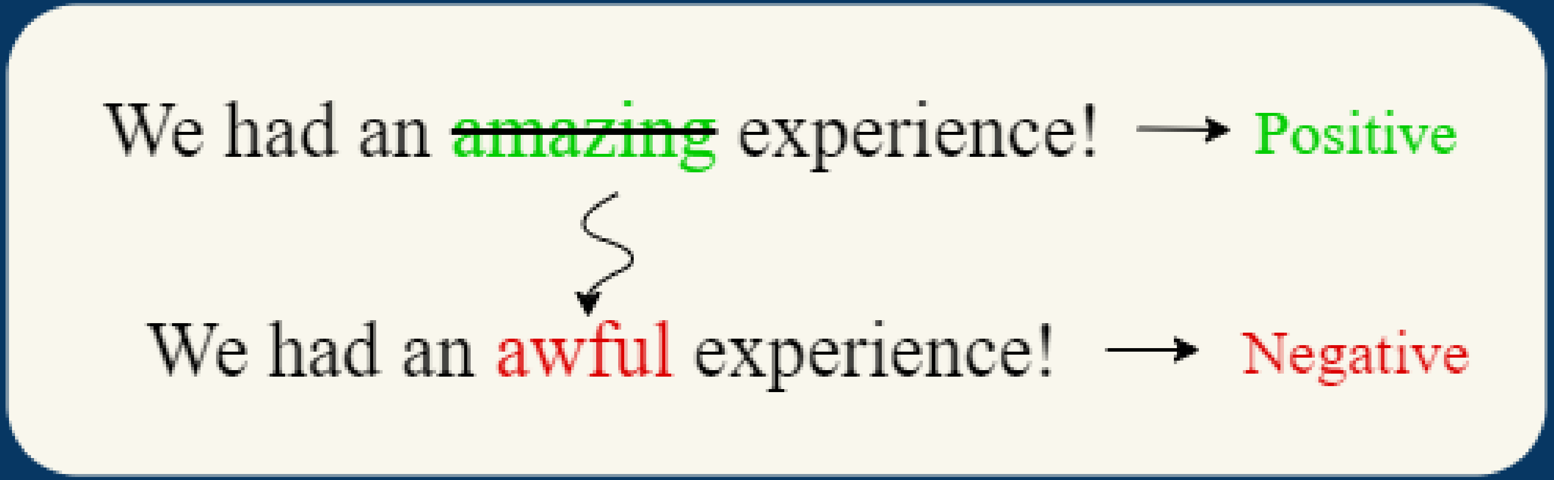
Exploring Text Counterfactual Explanations:

A Multi-Metric Evaluation Approach for Counterfactual Editors



Diploma Thesis

Karavangelis Athanasios

Table of contents

01.

Introduction

Introduction

Our Objective

The evaluation of counterfactual editors

Introduction

Our Approach

Explore multiple counterfactual generation methods and evaluate them based on various metrics

Introduction

Our Motivation

Counterfactuals of counterfactuals
Filandrianos et al.
May 2023

Introduction

Examined NLP Tasks

Text generation, Part-of-speech tagging, Sentiment analysis, Topic classification

Introduction

Our Objective

The evaluation of counterfactual editors

Explore multiple counterfactual generation methods and evaluate them using novel metrics

Our Approach

Our Motivation

Counterfactuals of counterfactuals

Text generation, Part-of-speech tagging, Sentiment analysis, Topic classification

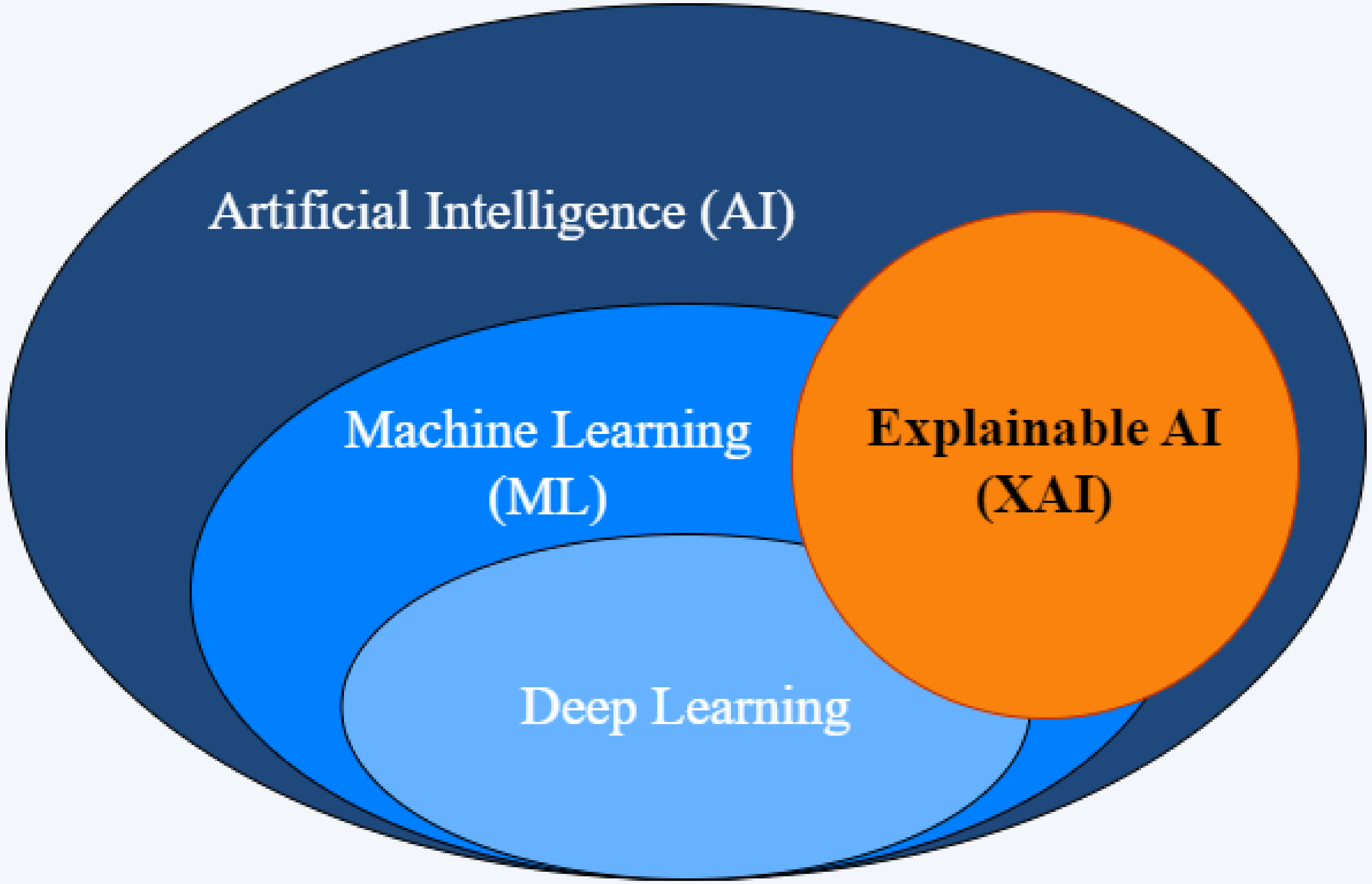
Examined NLP Tasks

02.

Theoretical background

Theoretical background

Explainable AI



Provides explanations of how models make decisions

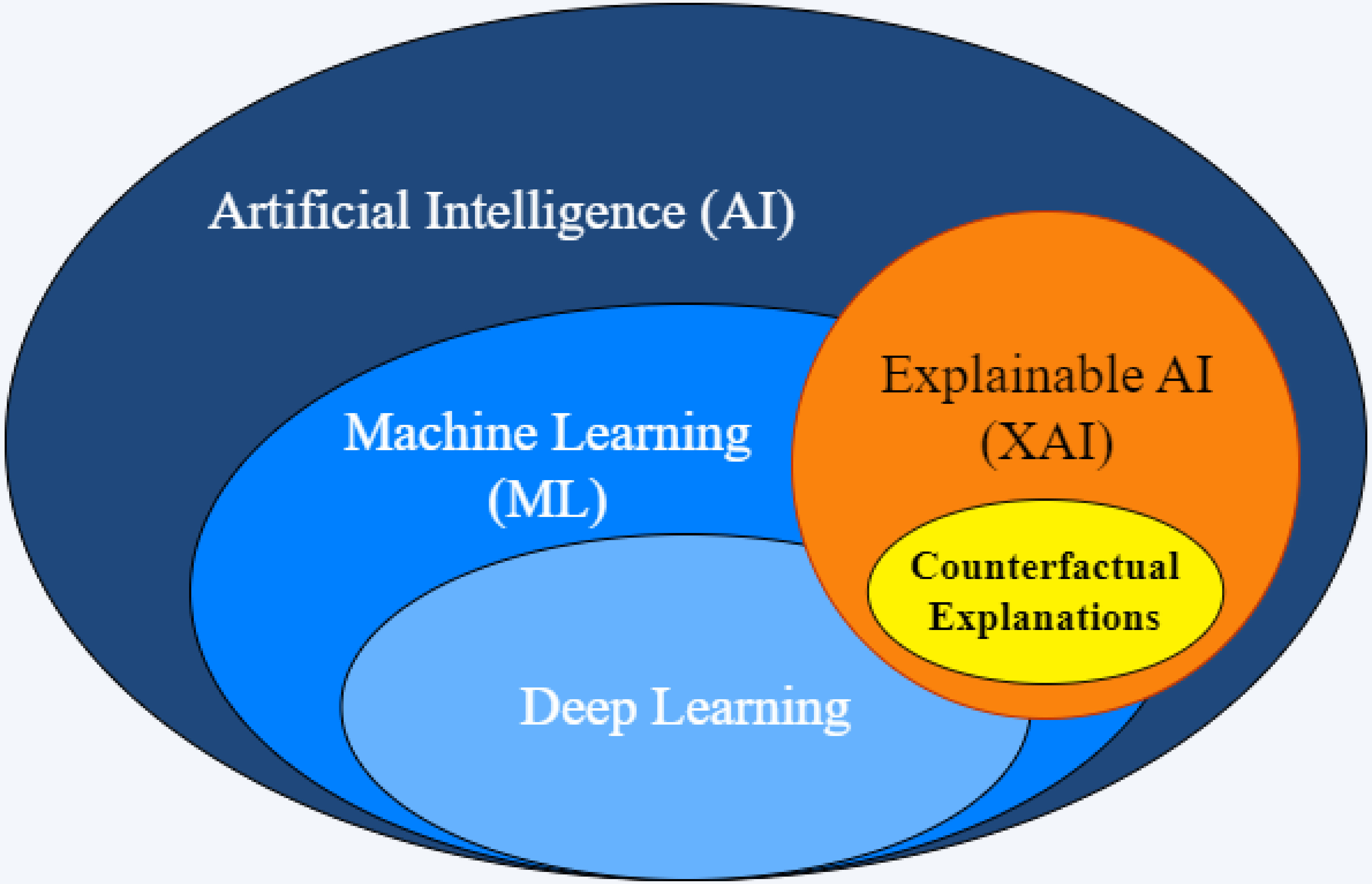
Increases

- Transparency
- Interpretability
- Trust

of ML models

Theoretical background

Counterfactual Explanations



Theoretical background

Counterfactual Explanations > Definition

Counterfactual Explanation

Definition: A feature-based explanation that identifies the minimal changes in input variables required to produce a different model prediction.

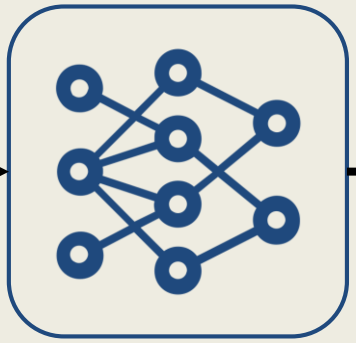
Life

cause → event

a slightly modified cause can result in a different event

Explainable AI

ML model

input →  → prediction

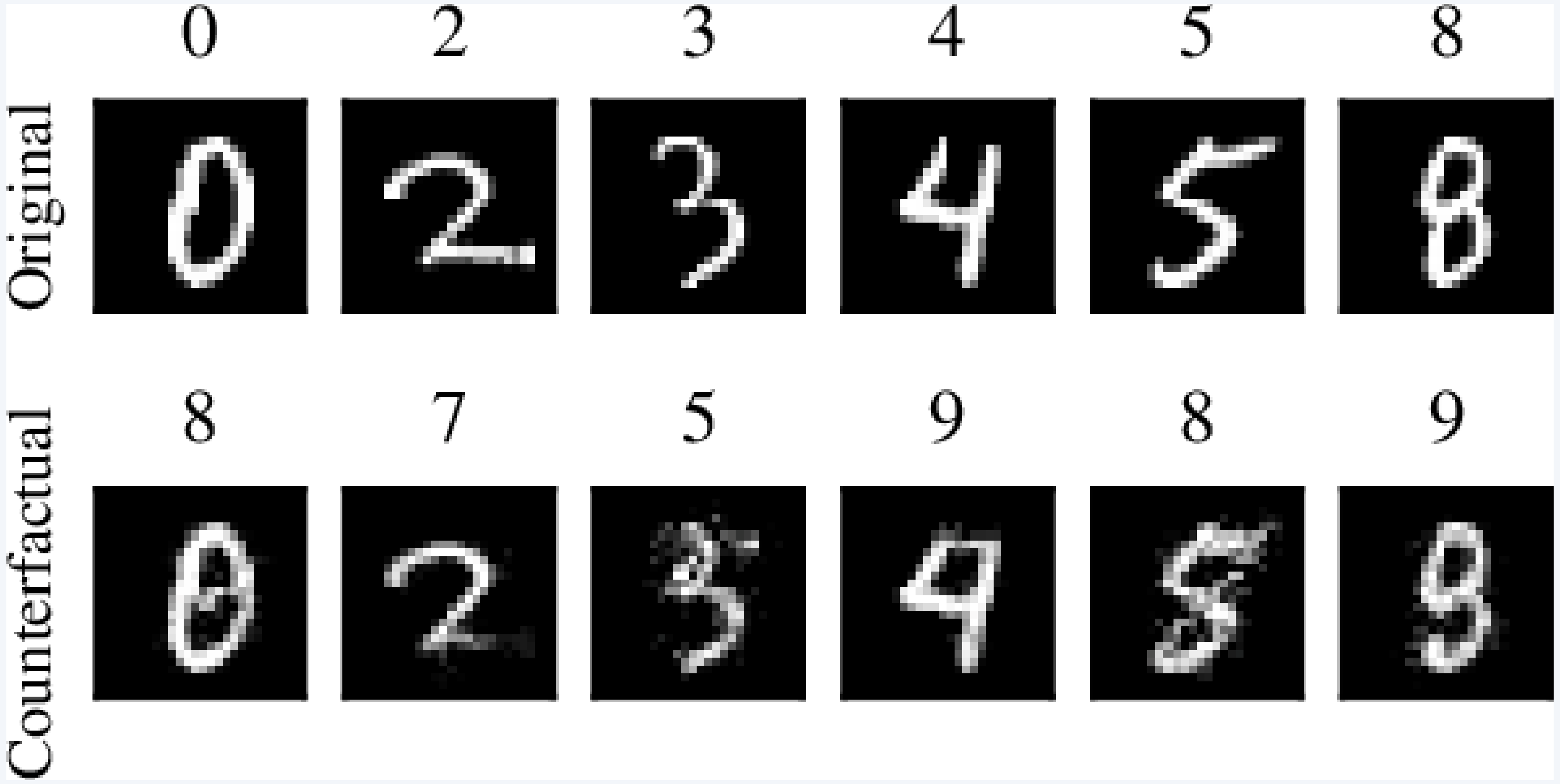
minimal changes to the input's feature values can lead to a different prediction

Theoretical background

Counterfactual Explanations > Examples

Image counterfactuals

Example on the Image Classification task with handwritten digits



Theoretical background

Counterfactual Explanations > Examples

Text counterfactuals

We had an ~~amazing~~ experience! → Positive

We had an awful experience! → Negative

Example on the Sentiment Analysis task

Example on the Topic Classification task

I think it's a Canon, but it's hardwired. Can it be used? → Miscellaneous

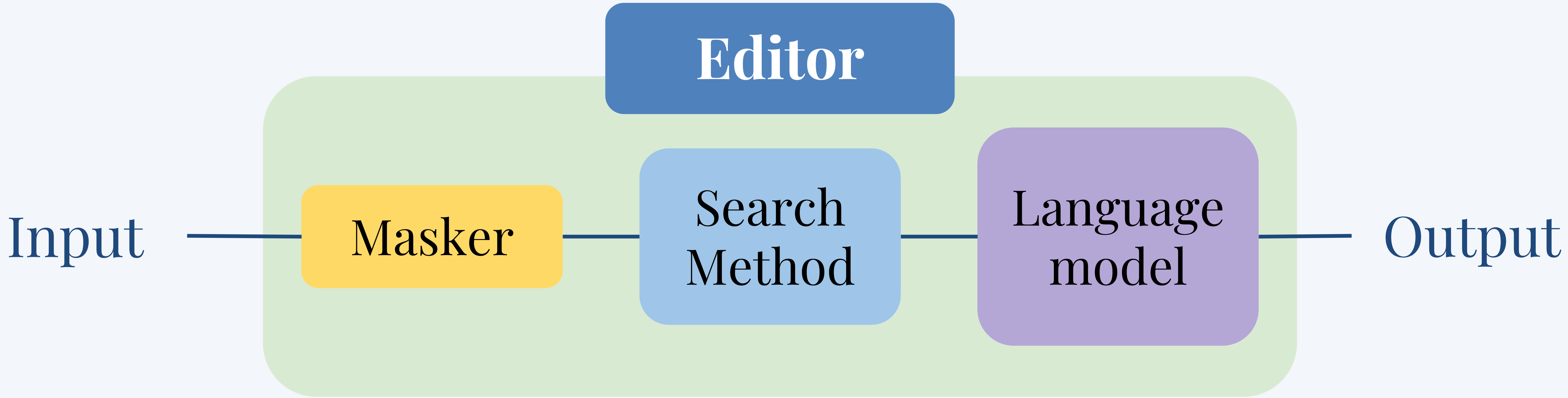
I think it's a Mac, but it's hardwired. Can it be used? → Computers

Theoretical background

Counterfactual Explanations > Counterfactual Editors

Counterfactual Editor

Definition: A framework that aims to edit a given text instance order to change the prediction of a classifier.



Theoretical background

Examined NLP Tasks

Sentiment Analysis

Uses computational methods to categorize the sentiment expressed in a piece of text

The experience so far has been fantastic!

POSITIVE

The experience has been ok.

NEUTRAL

The experince has been awful!

NEGATIVE

Topic Classification

Assigns predefined labels to text documents based on their content in order to classify them into distinct topics

I have to get my laptop fixed ASAP. → computers

NASA scientists have published some very promising findings. → science

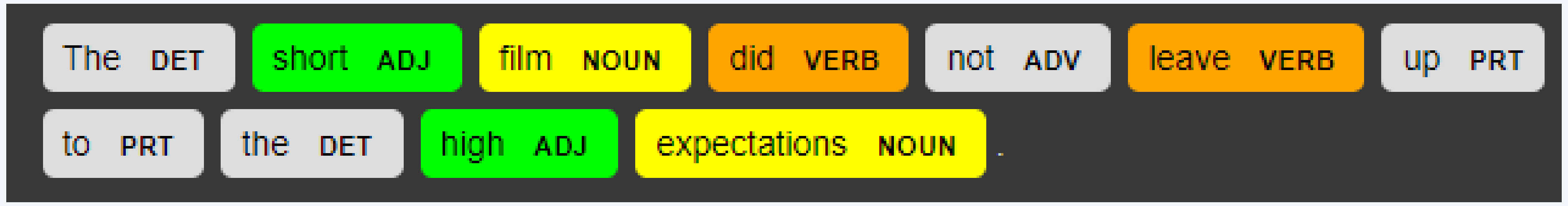
Theoretical background

Examined NLP Tasks

Part-of-speech (POS) tagging

Assigns grammatical tags to individual words in a given text that indicate their part-of-speech

The short film did not leave up to the high expectations.



Text Generation

Generates text that resembles human written text using various approaches like language models

We took <mask> for a walk in the <mask>. We had a <mask>.

Language model

We took **the dog** for a walk in the **park**. We had a **fun time**.

03.

Overview

Overview

Our motivation

Academic paper

Counterfactuals of counterfactuals
Filandrianos et al.
May 2023

1

Counterfactuals of counterfactuals
A new evaluation method for counterfactual editors.

2

Inconsistency
A novel evaluation metric for counterfactual edits.

Overview

Our work

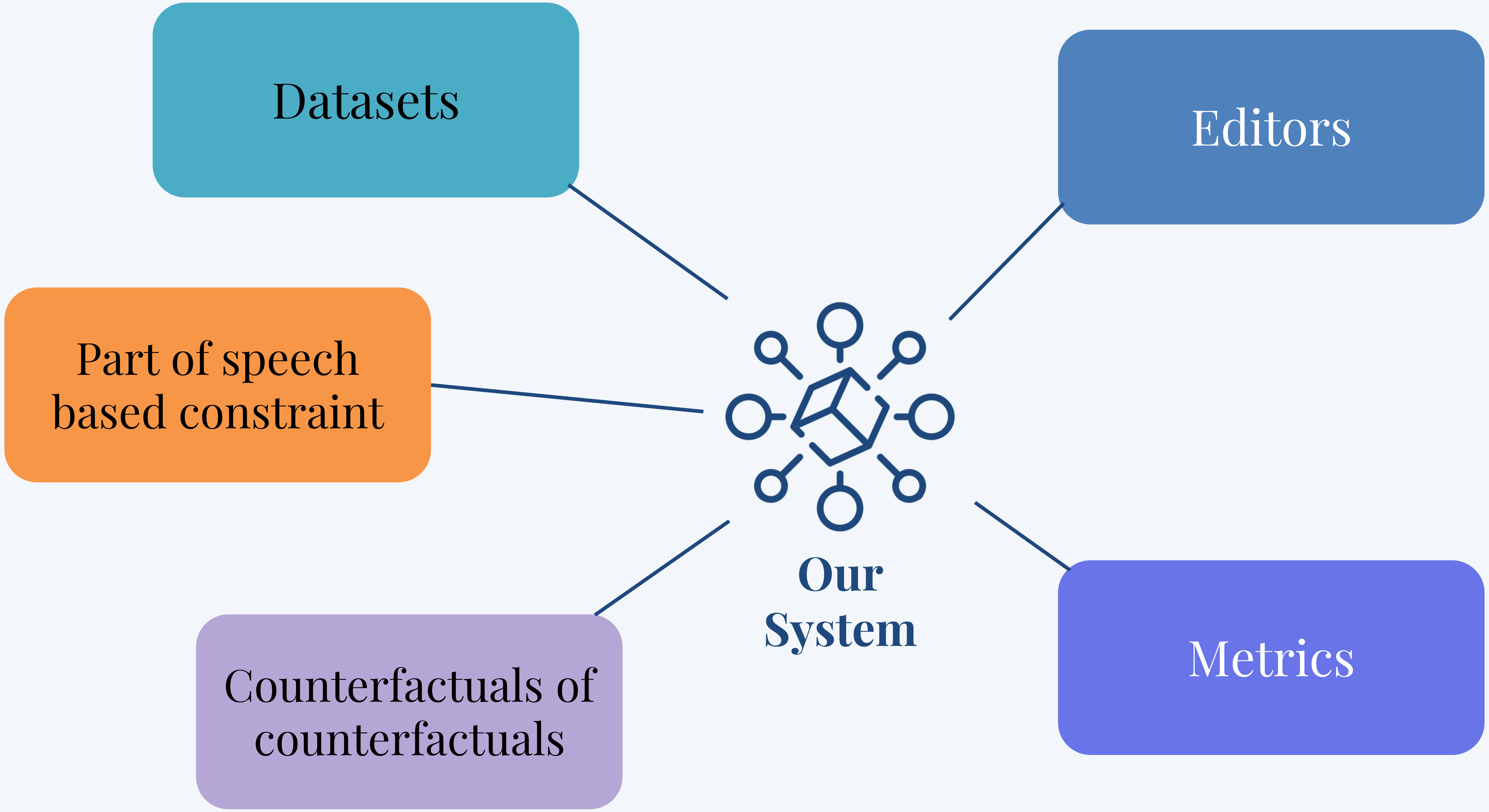
We introduce a new constraint on counterfactual generation based on part-of-speech tags

Experiments based on multiple editors combined with various generation methods

Our evaluation helps explain various aspects of the models' decisions

Overview

Counterfactual Generation and Evaluation System



Overview

Experiments and Results

Datasets

Part of speech based constraint

Counterfactuals of counterfactuals

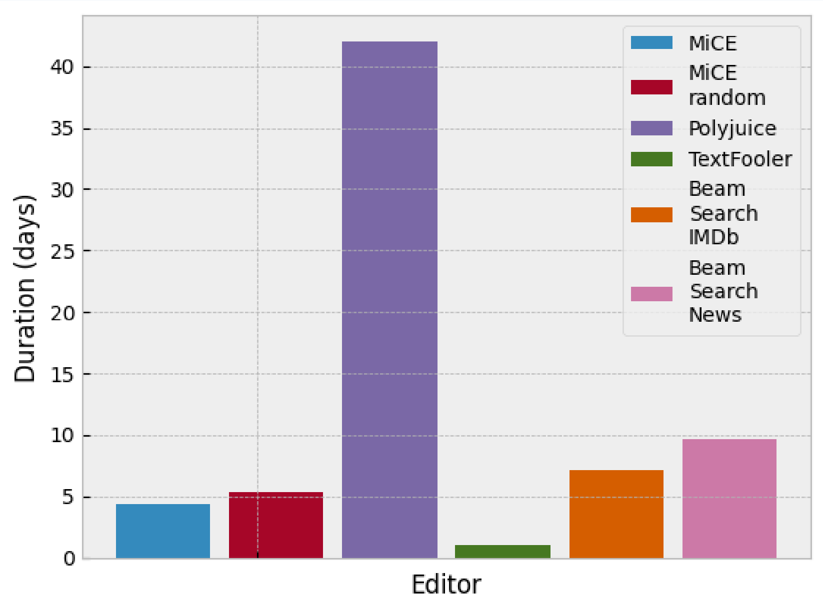
Our System

Editors

Metrics

Conduct Experiments

Extract valuable comparative results and explanations

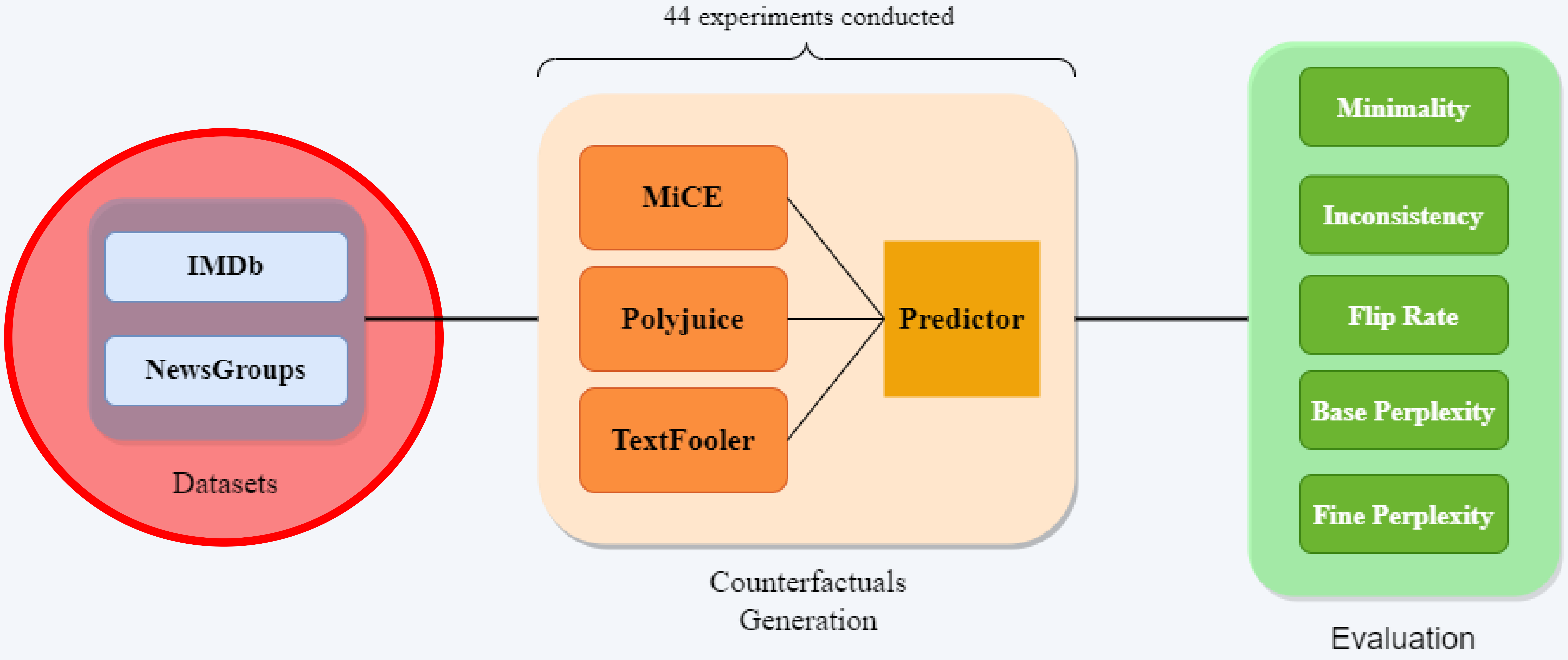


04.

Implementation

Implementation

An overview



Implementation

Datasets

IMDb

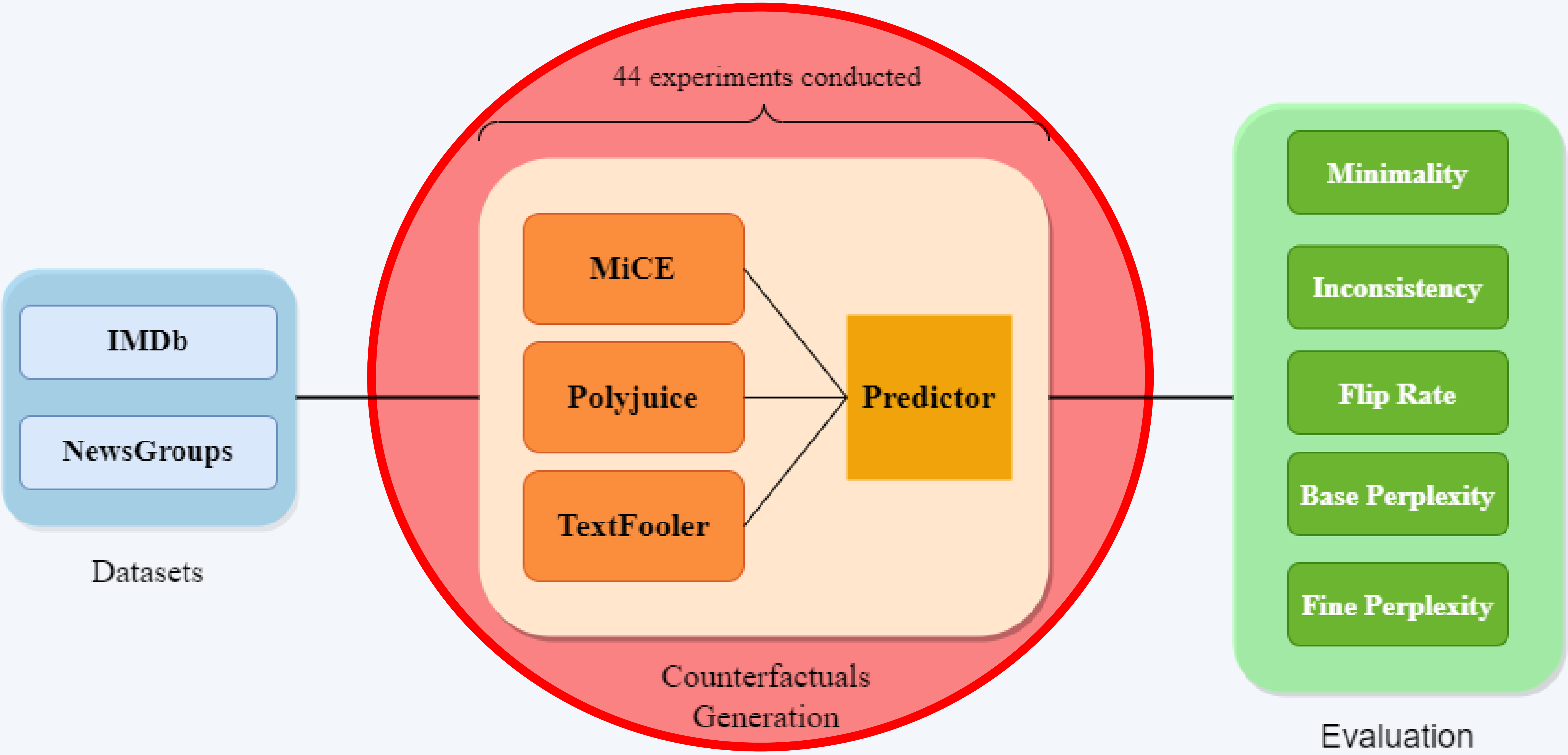
- 50.000 movie reviews
- labels:
  - 0 ⟶ negative
  - 1 ⟶ positive
- we use a subset of 500 reviews
- mean of 200 words per sentence

NewsGroups

- 20.000 documents
- 7 labels for 7 different topics
- we use a subset of 1.000 documents
- mean of 60 words per sentence

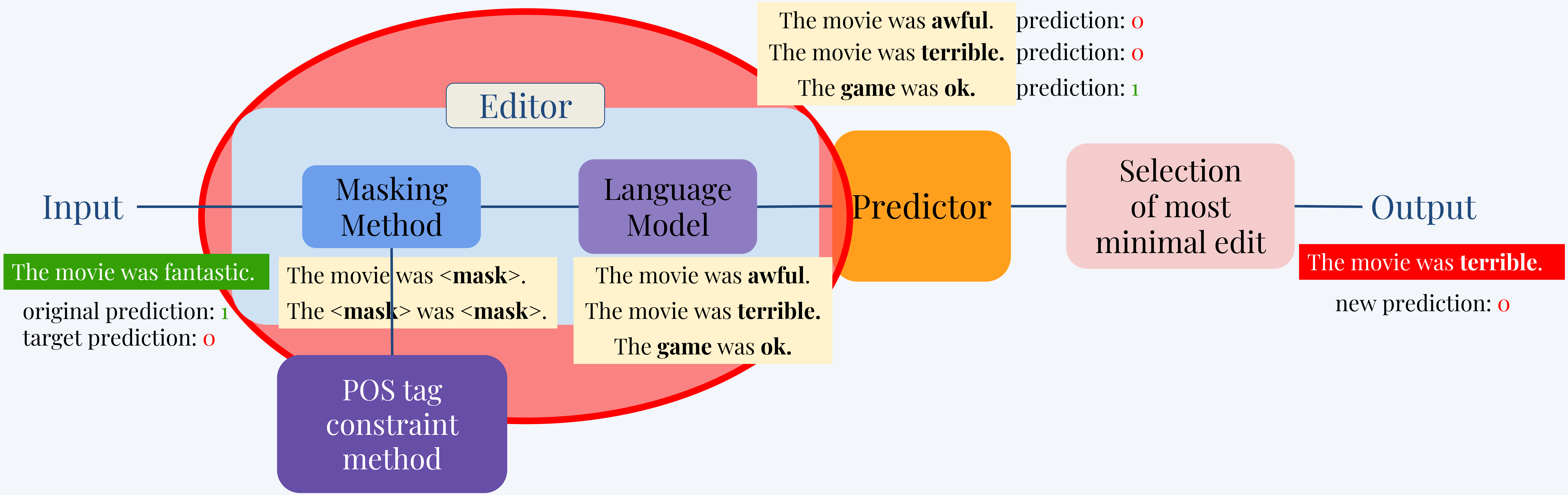
Implementation

Counterfactual Generation



Implementation

Our Counterfactual Editing System



Implementation

Counterfactual Editors

MiCE

- fine-tuned T5 Transformer
- selects edits based on minimality
- uses gradient masking and random masking

Polyjuice

- fine-tuned GPT2 model
- generates edits based on specific control codes e.g. negation, surprise
- uses random masking

TextFooler

- generates adversarial edits
- uses word embeddings to find synonyms
- employs several deterministic rules e.g. on POS tags
- uses word importance ranking for masking

Implementation

Masking methods

Random masking

Randomly selects the tokens that will be masked

MiCE **Polyjuice**

Gradient masking

Uses the predictor's self-attention to retrieve the most influential words

MiCE

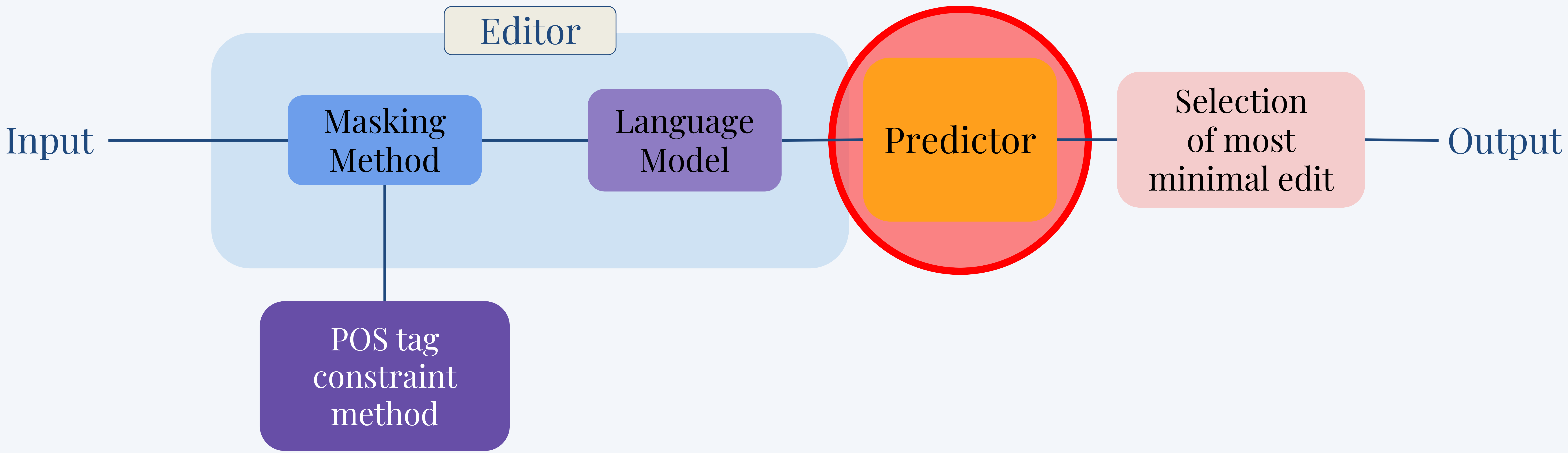
Word importance ranking

Calculates the prediction change before and after deleting each word. Then based on this difference, ranks the words from most to less important

TextFooler

Implementation

Predictors



Implementation

Predictors

- Two pre-trained predictors **fine-tuned** on IMDb and NewsGroups
- Built on RoBERTa Large
- Calculate the **probability** of the labels in the range of 0 to 1.

Example on IMDb

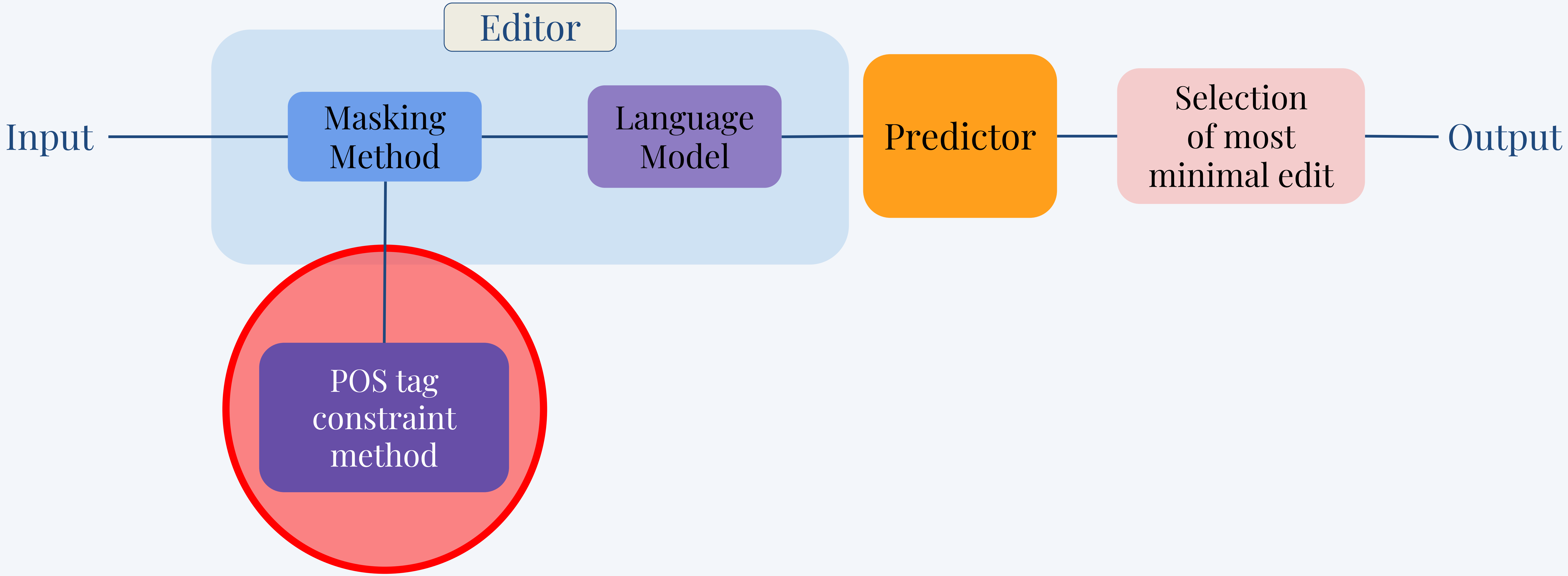
Read the book, forget the movie!

Predictor

[0.9972, 0.0028]
(0) (1)

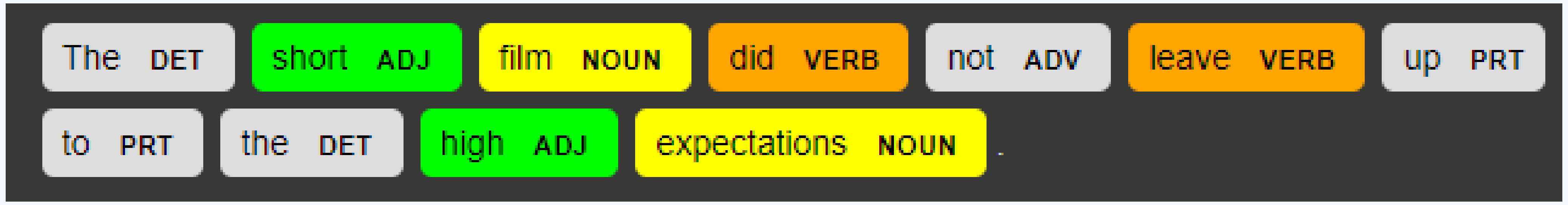
Implementation

Our POS tag constraint



Implementation

Our Part-of-speech (POS) tag constraint



Implementation

Our Part-of-speech (POS) tag constraint

What we do

- We use **part-of-speech tagging** to constrain the words that can be edited
- Aim to **minimize the needed modifications**
- Intervene in the **masking** stage of the editors to enforce the constraint

If you like good thrillers, this amazing film is just what you need!

ADJ (adjectives)

If you like good thrillers, this amazing film is just what you need!

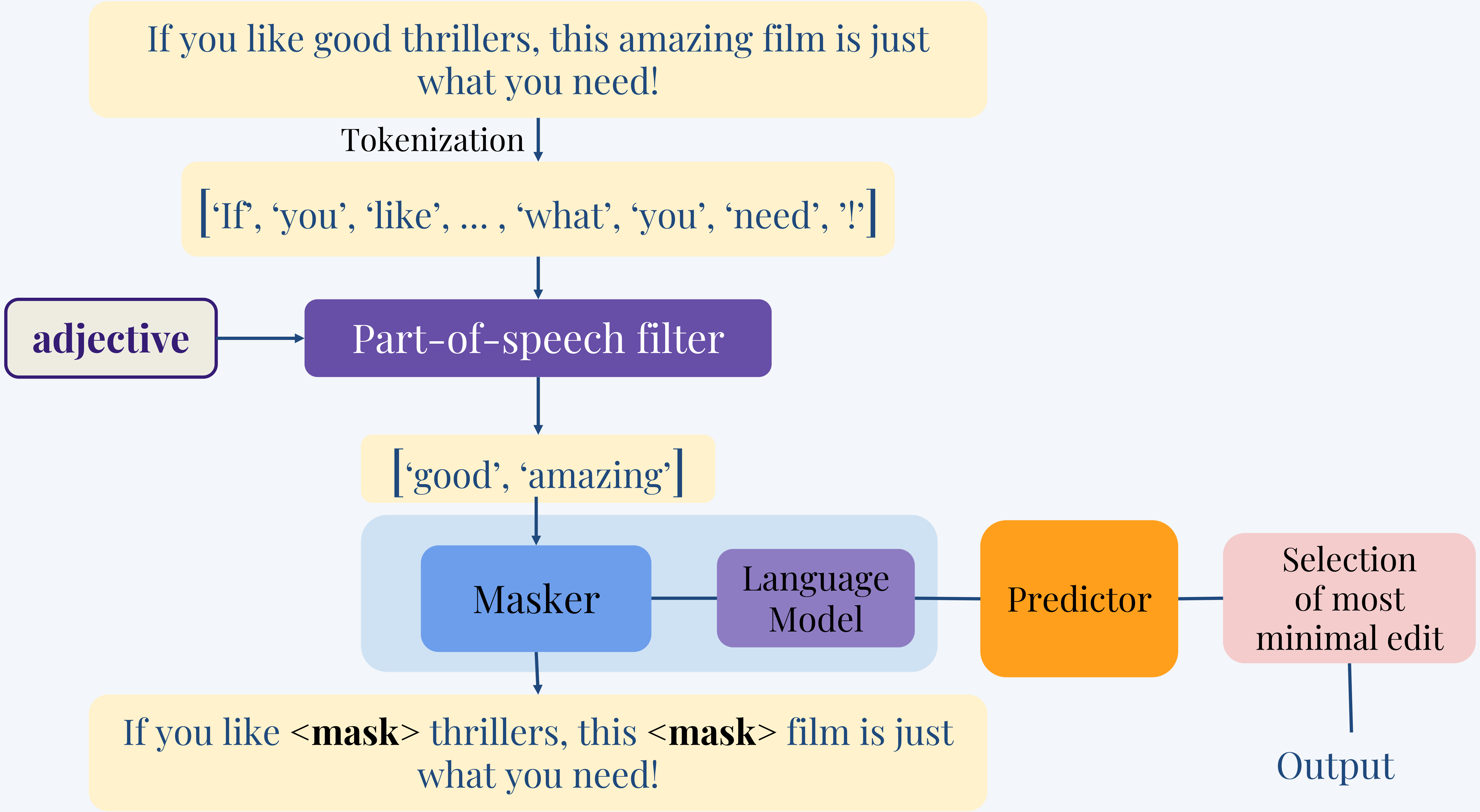
NOUN (nouns)

If you like good thrillers, this amazing film is just what you need!

VERB (verbs)

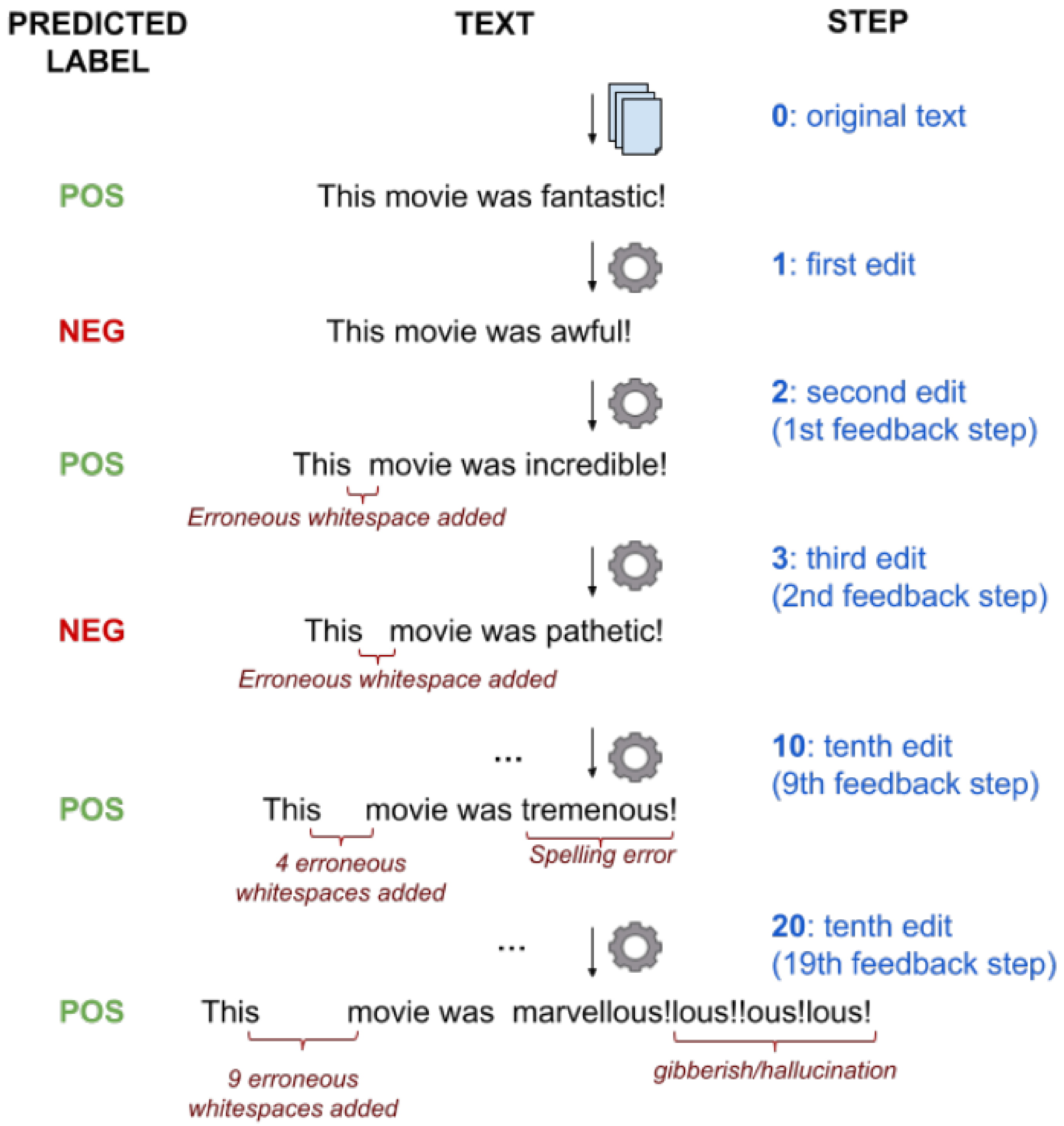
Implementation

Our Part-of-speech tag constraint



Implementation

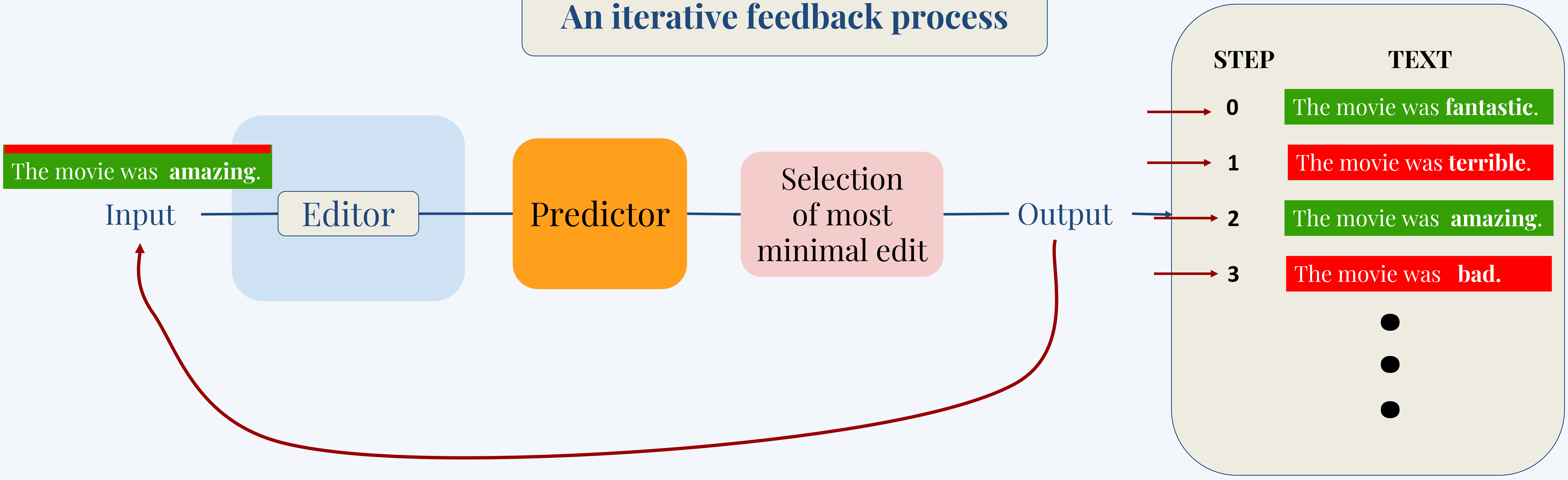
Counterfactuals of counterfactuals



Implementation

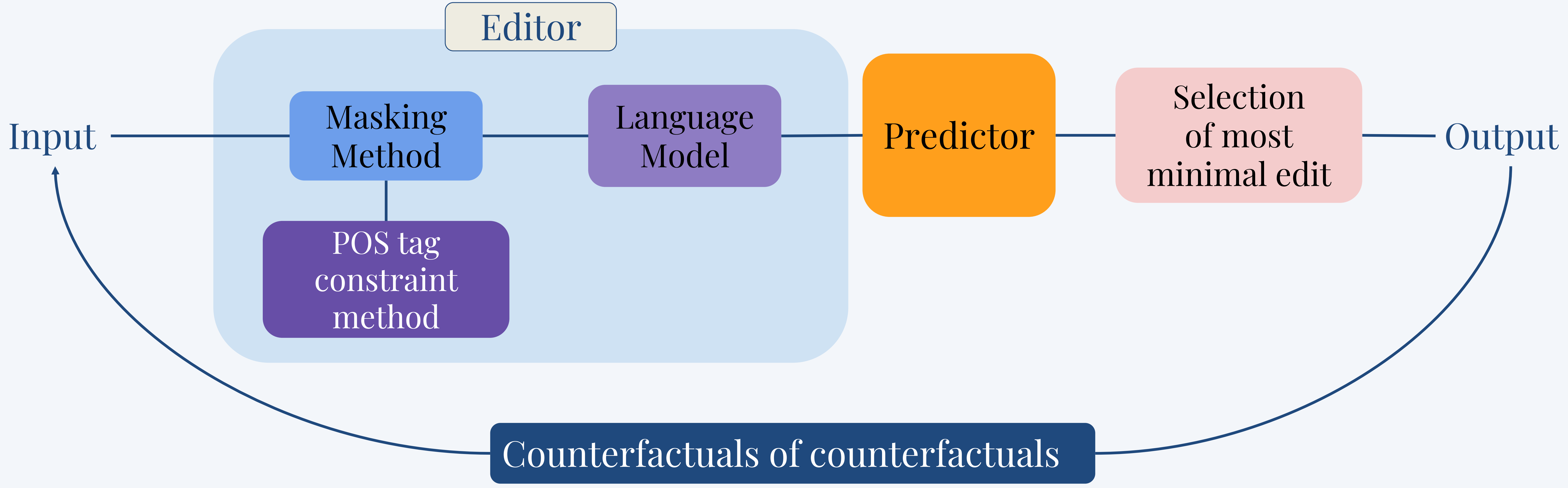
Counterfactuals of counterfactuals

An iterative feedback process



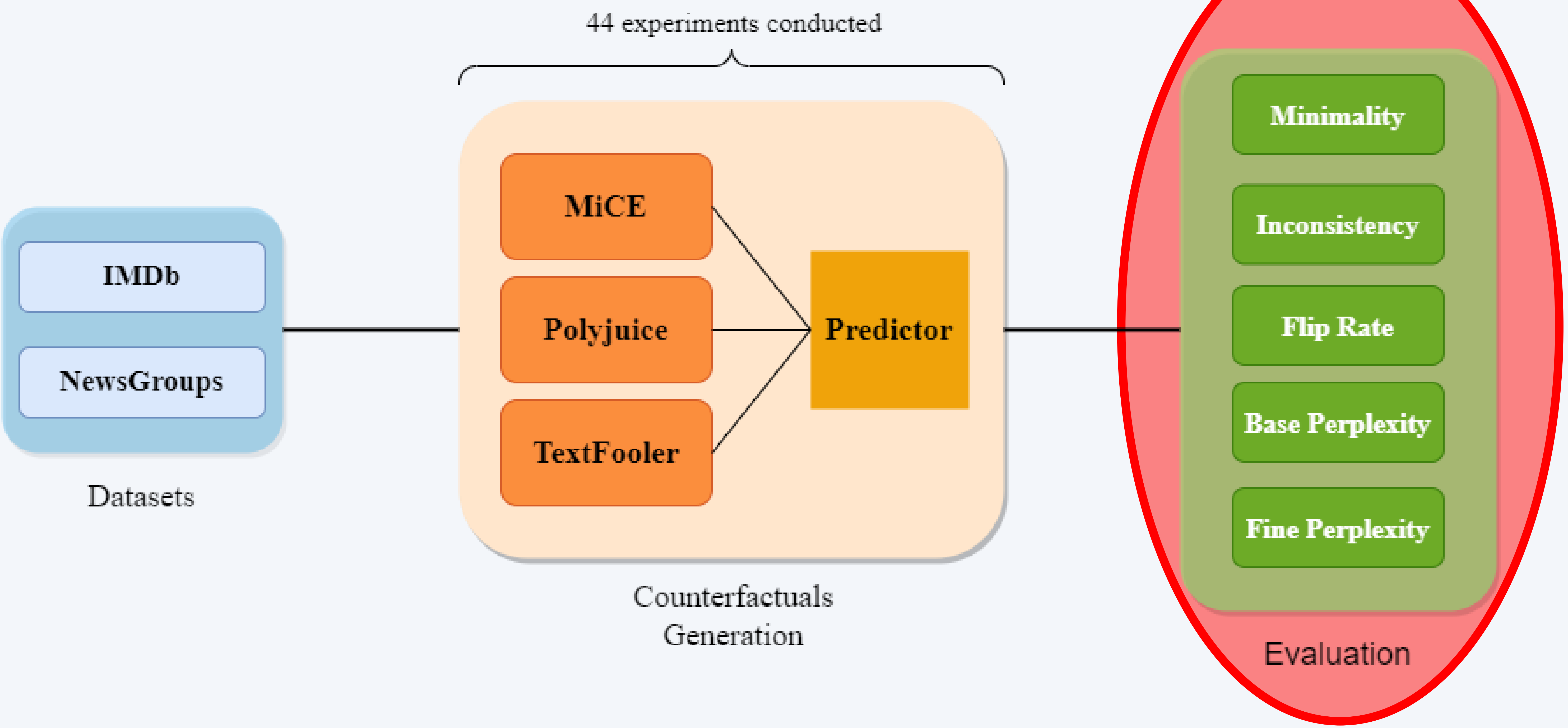
Implementation

Combining the methods



Implementation

Evaluation



Implementation

Evaluation

Minimality

- Calculates the word-level **Levenshtein edit distance**
- Shows how many words were changed in the sentence after the edit

Intuitively: Low values of the metric indicate more minimal changes by the editor

The **movie** was a **great** exhibition of **classic** cinema.

minimality: 3

The **play** was a **valid** exhibition of **bad** cinema.

Implementation

Evaluation

Inconsistency (of minimality)

- **Novel metric** introduced by Filandrianos et al.
- Measures how "**consistent**" an editor is with respect to a metric (e.g. minimality)
- Values of **0** mean that the editor generates the **most minimal edit possible**

Intuitively: Small positive values indicate almost optimal series of edits

$$inc(f,x) = relu[d(f(f(x)), f(x)) - d(f(x), x)]$$

$$inc@n(f,x) = \frac{1}{n}\sum_{i=0}^{n-1} inc(f_{i+1}(x), f_i(x))$$

Step 0: The movie was a **great** exhibition of **classic** cinema.

minimality: **2**

Step 1: The play was a **valid** exhibition of **bad** cinema.

minimality: **3**

Step 2: The **film** was a **good** exhibition of **good** cinema.

inc@2 = 3 - 2 / 2 = **0.5**

Implementation

Evaluation

Flip Rate

- Used with many counterfactual editors (MiCE, TextFooler etc)
- Shows how often the output of the predictor is flipped
- Also called: attack success rate

Intuitively: The higher the flip rate of an editor, the more edits it succeeds flipping

$$flip\_rate = \frac{edits\ with\ successful\ flip\ to\ the\ desired\ class}{number\ of\ inputs\ to\ the\ editor}$$

We had an ~~amazing~~ experience! → Positive

We had an awful experience! → Negative

Accomplished flip!

Implementation

Evaluation

Base perplexity

- A proxy for evaluating **fluency**
- Calculates the likelihood of the next token conditioned on the preceding tokens. based on some language model, e.g. we use **GPT2**

Intuitively: Lower values mean more predictable edits. Higher values mean more diverse - surprising edits.

Fine perplexity

- Same as base perplexity but the language model used is **fine-tuned** on a dataset.

Intuitively: Lower values mean that the edits converge to the dataset's distribution. Assesses how the model has adapted to the specific dataset.

Source: https://huggingface.co/docs/transformers/perplexity

Implementation

Technologies used

Python

Various ML and NLP libraries

PyTorch

Access to pretrained models

Text Generation

GPU acceleration

spaCy

Part-of-speech tagging

GPU accelerated environments

Our experiments needed **1670 GPU hours** (!) , this translates to 70 days for one GPU.

ARIS HPC

Supercomputer operated by GRNET

kaggle

colab

05.

Experiments

Experiments

Overview

Editors without making any modifications in their code

Experiments on the generation algorithm of MiCE

10 steps of edits

Experiment Types

4 editors	Out-of-the-box	ADJ	NOUN	VERB	Beam-Search
MiCE	✓	✓	✓	✓	✓
MiCERandom	✓	✓	✓	✓	–
Polyjuice	✓	✓	✓	✓	–
TextFooler	✓	✓	✓	✓	–

Interpreting the qualitative results

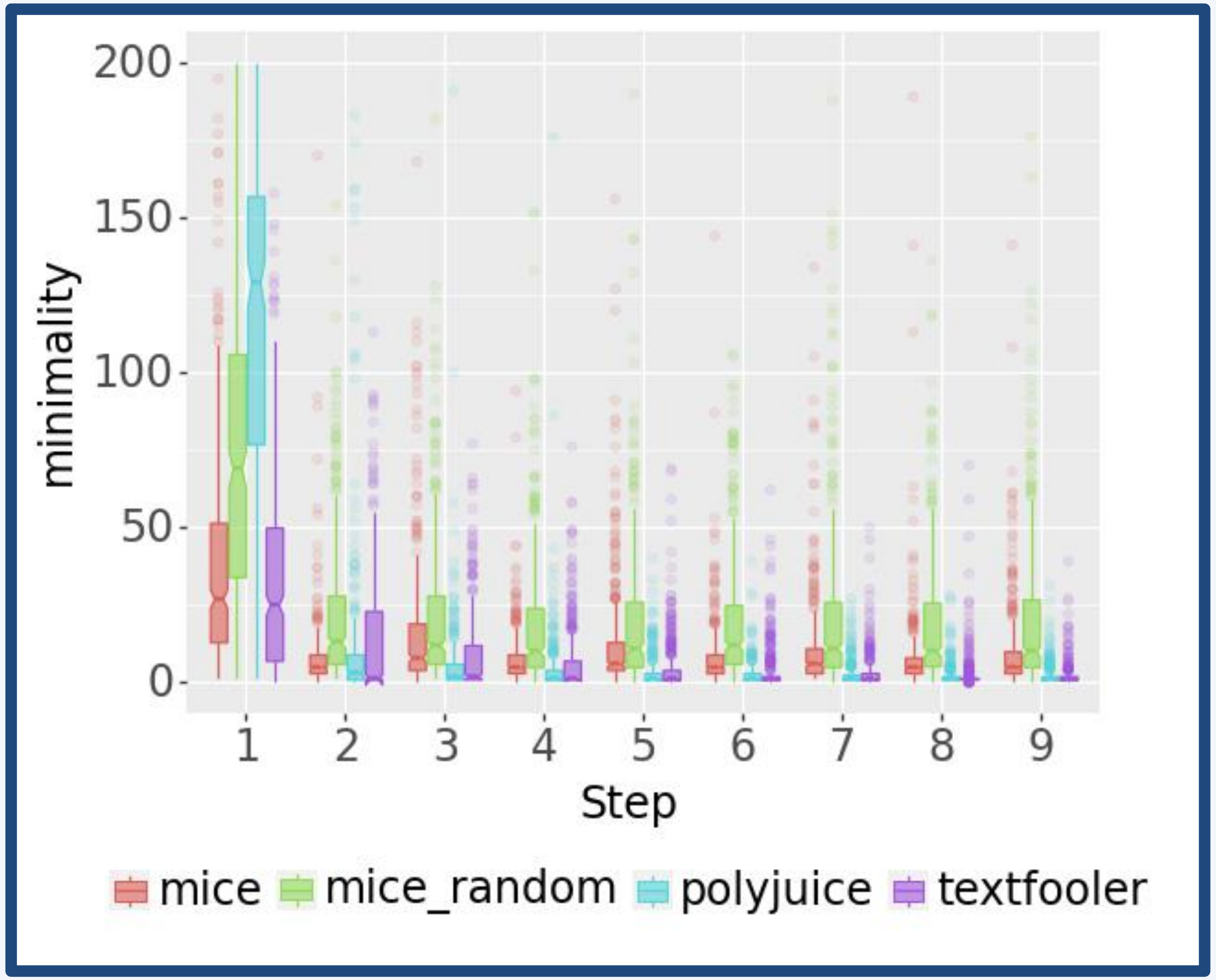
Experiments

Interpreting the qualitative results

Minimality

Out-of-the-box

Intuitively: Low values of the metric indicate more minimal changes by the editor



- **TextFooler** produces the most minimal edits. Deterministic approach with many constraints.
- **MiCE and Polyjuice** edits that use language models are less minimal
- Editors with **random masking** are less minimal than those that use **attention masking**

Experiments

Interpreting the qualitative results

Minimality

Out-of-the-box

Intuitively: Low values of the metric indicate more minimal changes by the editor

MiCE

0: You may like **Tim Burton**'s fantasies, but not in a **commercial-like show off lasting** 8 minutes. It **demonstrates** good **technical** points without real **creativity** or some established **narrative** pace.

1: You may like **Cary Grant**'s **play**, but not in a **full- length** 8 minutes. It **contains** good **plot** points without real **surprises** or some established **frantic** pace.

2: You may like Cary Grant's play, but not in a **mere** 8 minutes. It contains **good** plot points without real **interest** or some established **stable** pace.

Experiments

Interpreting the qualitative results

Minimality

Out-of-the-box

Intuitively: Low values of the metric indicate more minimal changes by the editor

TextFooler

0: You may **like** Tim Burton's fantasies, but not in a commercial-like show off lasting 8 **minutes**. It demonstrates good technical points without real **creativity** or some established narrative pace.

1: You may **such** Tim Burton's fantasies, but not in a commercial-like show off lasting 8 **mn**. It demonstrates good technical points without real **groundbreaking** or some established narrative pace.

2: You may such Tim Burton's fantasies, but not in a commercial-like show off **longstanding** 8 mn. It demonstrates good technical points without real groundbreaking or some established narrative pace.

Experiments

Interpreting the qualitative results

Minimality

POS tag constraint

Intuitively: Low values of the metric indicate more minimal changes by the editor

- **Constraining** the editors to a specific POS tag reduces the candidate words for modification
- **More minimal edits** generated
- Most efficient POS:

IMDb	NewsGroups
ADJ	NOUN

MiCE ADJ

0: You may like Tim Burton's fantasies, but not in a **commercial-like** show off lasting 8 minutes. It demonstrates good **technical** points without real creativity or some established **narrative** pace.

1: You may like Tim Burton's fantasies, but not in a **light- hearted** show off lasting 8 minutes. It demonstrates good **plot** points without real creativity or some established **predictable** pace.

2: You may like Tim Burton's fantasies, but not in a **boring** show off lasting 8 minutes. It demonstrates **basic** plot points without real creativity or some established predictable pace.

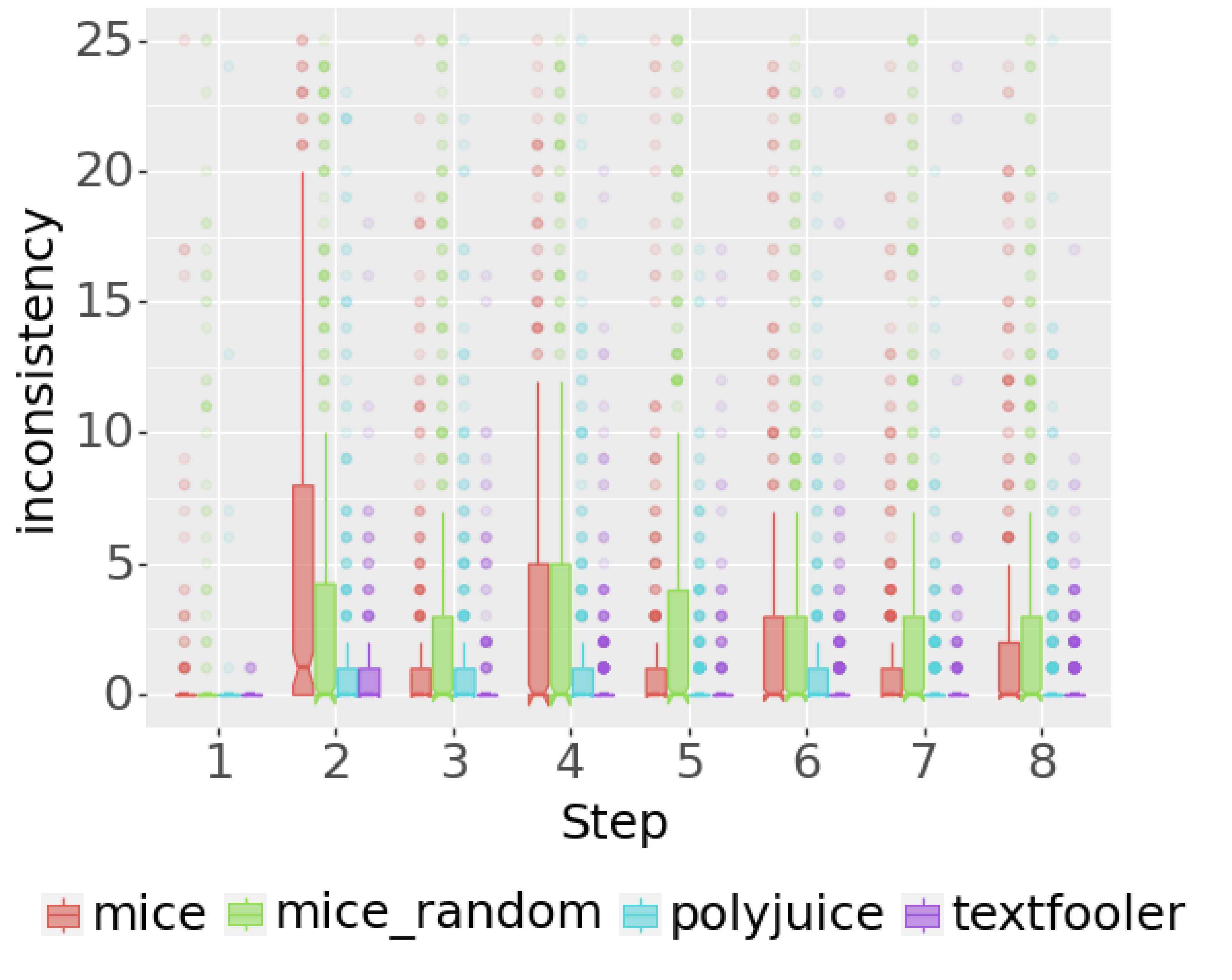
Experiments

Interpreting the qualitative results

Inconsistency

Out-of-the-box

Intuitively: Small positive values indicate almost optimal series of edits



- **TextFooler** produces the most consistent edits. *>Inconsistency values, nearly 0.*
- Language models are more sensitive to input modifications. *> MiCE and Polyjuice are less consistent.*
- Editors turn more consistent in later feedback steps.

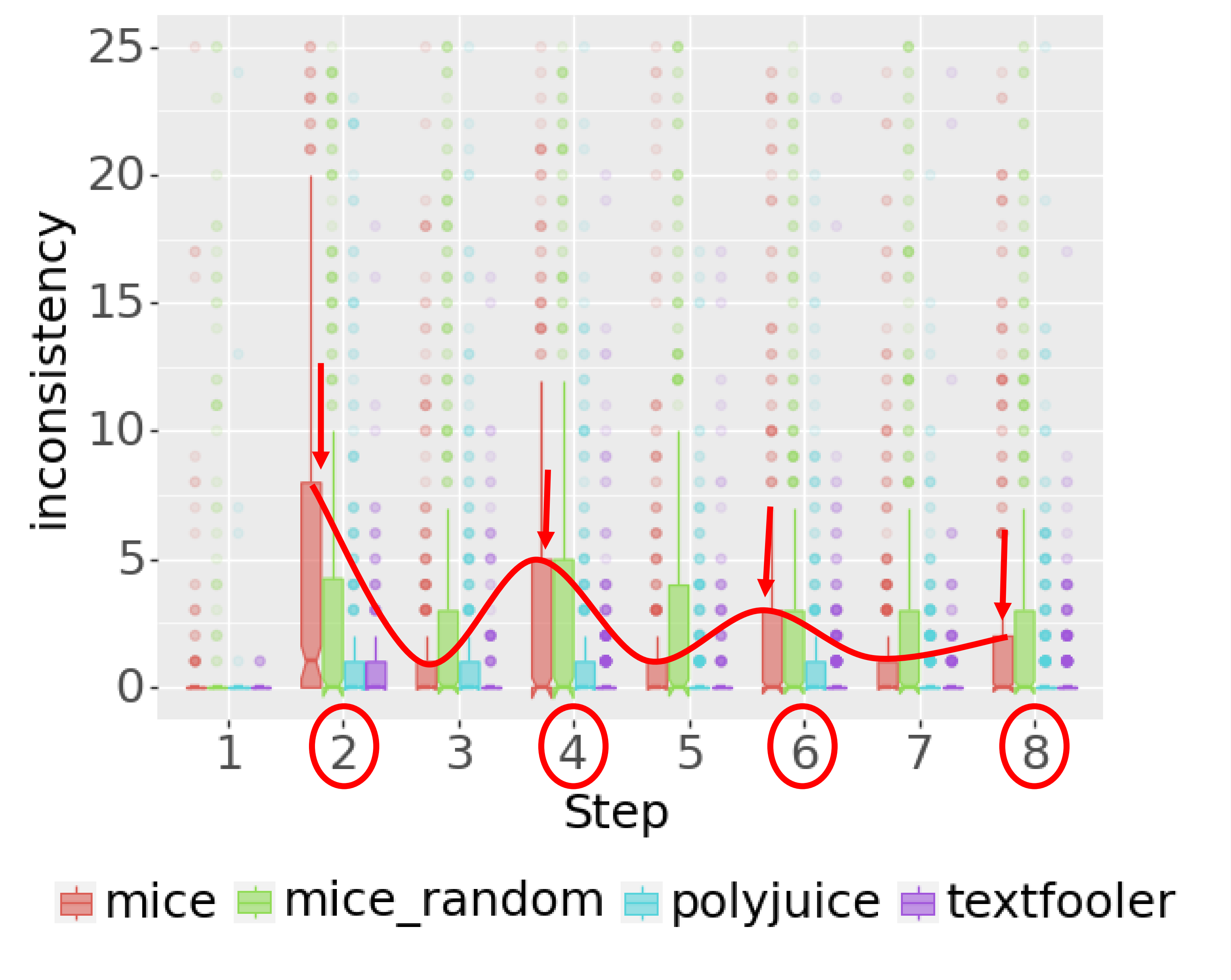
Experiments

Interpreting the qualitative results

Inconsistency

Out-of-the-box

Intuitively: Small positive values indicate almost optimal series of edits



- MiCE editors present **higher inconsistency values in even steps**
- Even steps represent flips from the original prediction to the contrast one
- MiCE moves easier back to the original than the contrast one
  > *One cause is text residue from the original sentence*

Experiments

Interpreting the qualitative results

Inconsistency

POS tag constraint

Intuitively: Small positive values indicate almost optimal series of edits

- Constraining the editors to a specific POS tag makes the editors **far more consistent**
- All editors present mean inconsistency **values around 0 and 1**, hinting at almost optimal edits in terms of minimality
- Most consistent POS: **IMDb & NewsGroups** ADJ

Experiments

Interpreting the qualitative results

Inconsistency

POS tag constraint

IMDb | **MiCERandom ADJ**

0: The **biggest** heroes, is one of the **greatest** movies ever. A good story, great actors and a brilliant ending is what makes this film the jumping start of the director Thomas Vinterberg's great carrier.

1: The **great carrier,** heroes, is one of the **worst** horror movies ever. A good story, great actors and a **brilliant** ending is what makes this film the jumping start of the director Thomas Vinterberg's great carrier.

2: The great carrier, heroes, is one of the **best** horror movies ever. A good story, great actors and a **surprisingly satisfying** ending is what makes this film the jumping start of the director Thomas Vinterberg's great carrier.

Experiments

Interpreting the qualitative results

Inconsistency

Out-of-the-box

IMDb | MiCERandom

0: The biggest heroes, is one of the **greatest** movies ever. **A good** story, great actors and a brilliant ending is what **makes this film the jumping start of** the director **Thomas** Vinterberg's great carrier.

1: The **biggest** heroes, **is not** one of the **best** movies ever. **Stupid story, great** actors and **a brilliant script is what saved the** director **Wolfgang** Vinterberg's great carrier.

2: The **original** heroes, **heroes of slash** movies ever **created** - great actors and **actresses ! This is legendary** director **Üne** Vinterberg's great carrier **-opera duo**.

Experiments

Interpreting the qualitative results

Inconsistency

POS tag constraint

NewsGroups | TextFooler NOUN

0: Wayne: Look for these advertised in sailboat supplies catalogs.

1: Wayne: Look for these advertised in sailboat wares catalogs.

2: Wayne: Look for these advertised in sailboat foodstuffs catalogs.

3: Wayne: Look for these advertised in sailboat wares catalogs.

4: Wayne: Look for these advertised in sailboat foodstuffs catalogs.

5: Wayne: Look for these advertised in sailboat wares catalogs.

Experiments

Interpreting the qualitative results

Inconsistency

Out-of-the-box

NewsGroups | TextFooler

0: Wayne: Look for these advertised in sailboat supplies catalogs.

1: Thomas: Look for these shown in sailboat wares catalogs.

2: Thomas: Look for these shown in sailboat supplies catalogs.

3: Thomas: Look on these shown in spacecraft foodstuff catalogs.

4: Thomas: Look on these shown in boat foodstuffs catalogs.

5: Thomas: Observe on these shown in spacecraft foodstuffs catalogs.

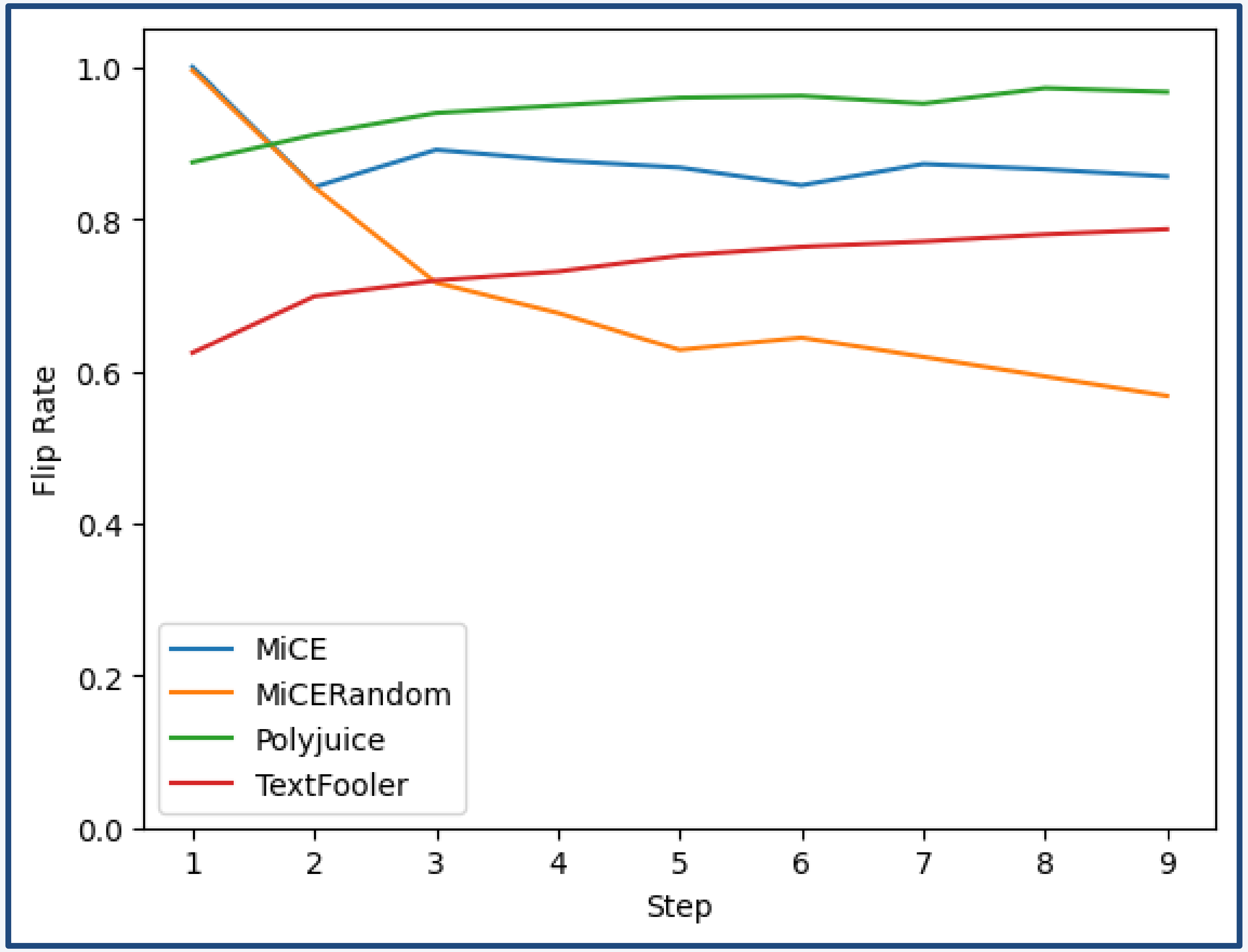
Experiments

Interpreting the qualitative results

Flip Rate

Out-of-the-box

Intuitively: The higher the flip rate of an editor, the more edits it succeeds flipping



- Combined with counterfactuals of counterfactuals, it reveals editors imperfections or strengths
  > *e.g. for **MiCERandom***
- At Step 1 **MiCE** flipped 100% of the input , at Step 9: 85%
- **Polyjuice** and **TextFooler** become more effective at later steps
- Flip rate reveals that the editors present **different behavior** when they are **not dataset-dependent**

Experiments

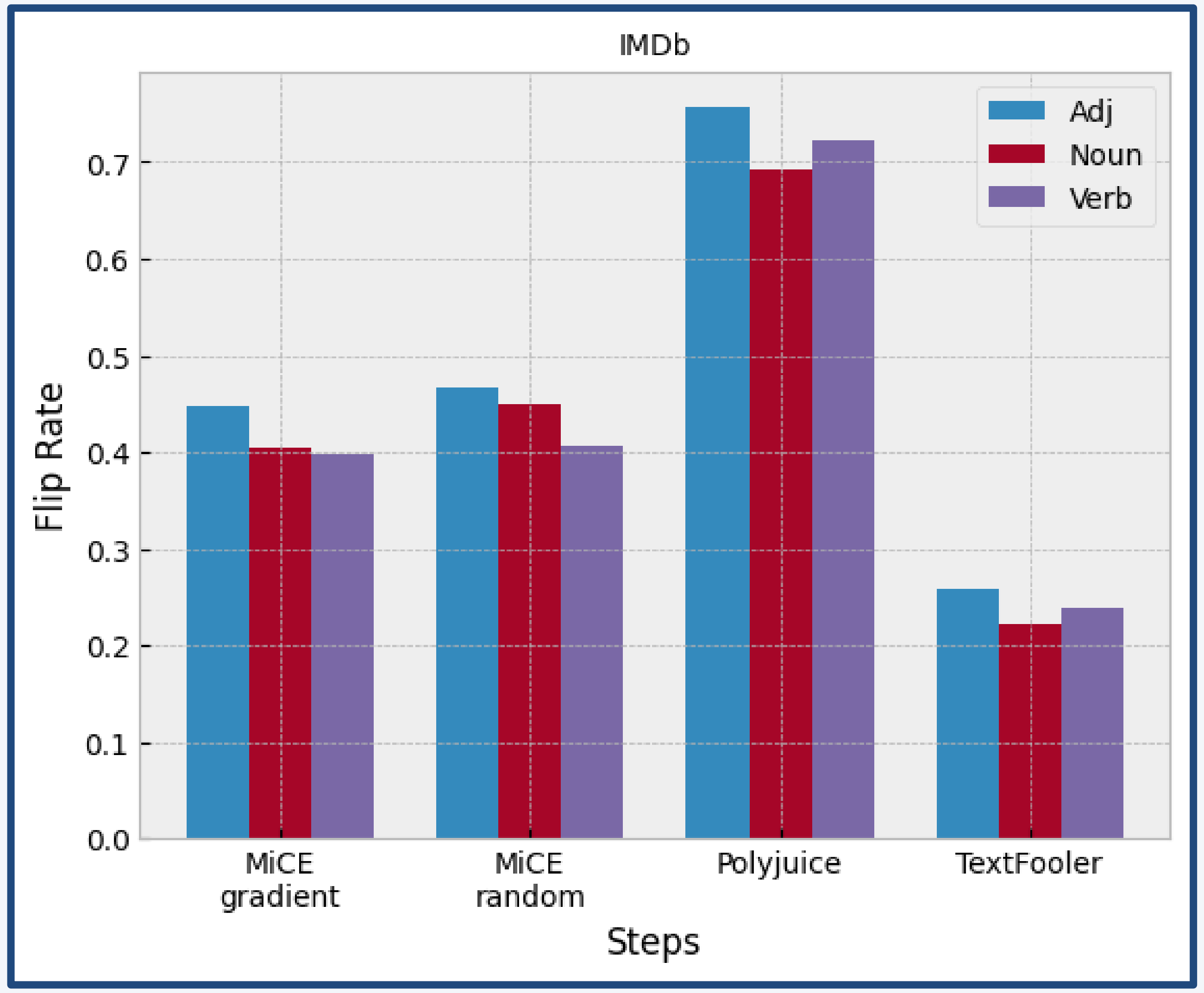
Interpreting the qualitative results

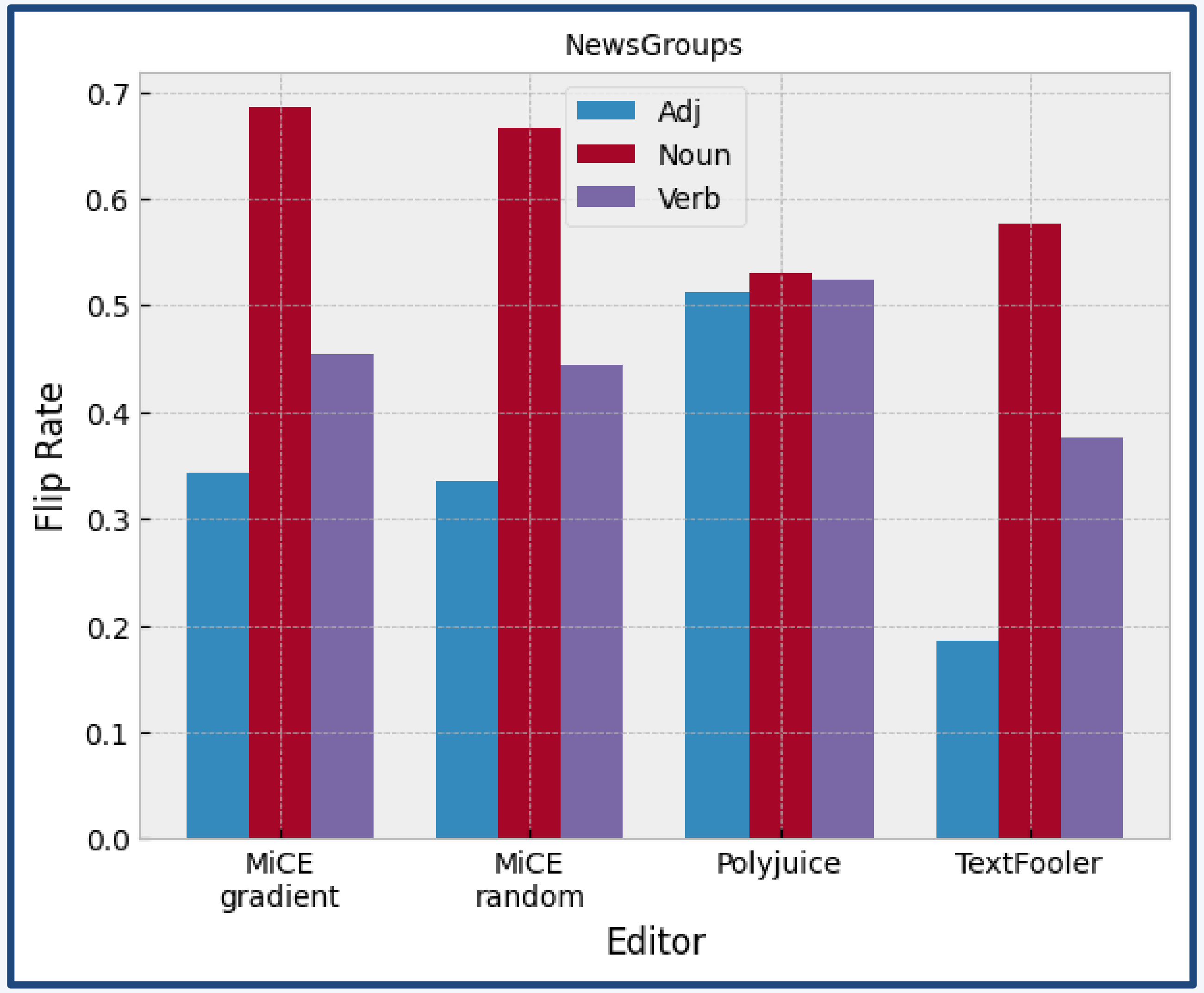
Flip Rate

POS tag constraint

Intuitively: The higher the flip rate of an editor, the more edits it succeeds flipping

- Generally, much lower flip rates
- In IMDb, adjectives perform better
- In NewsGroups, nouns perform better





Experiments

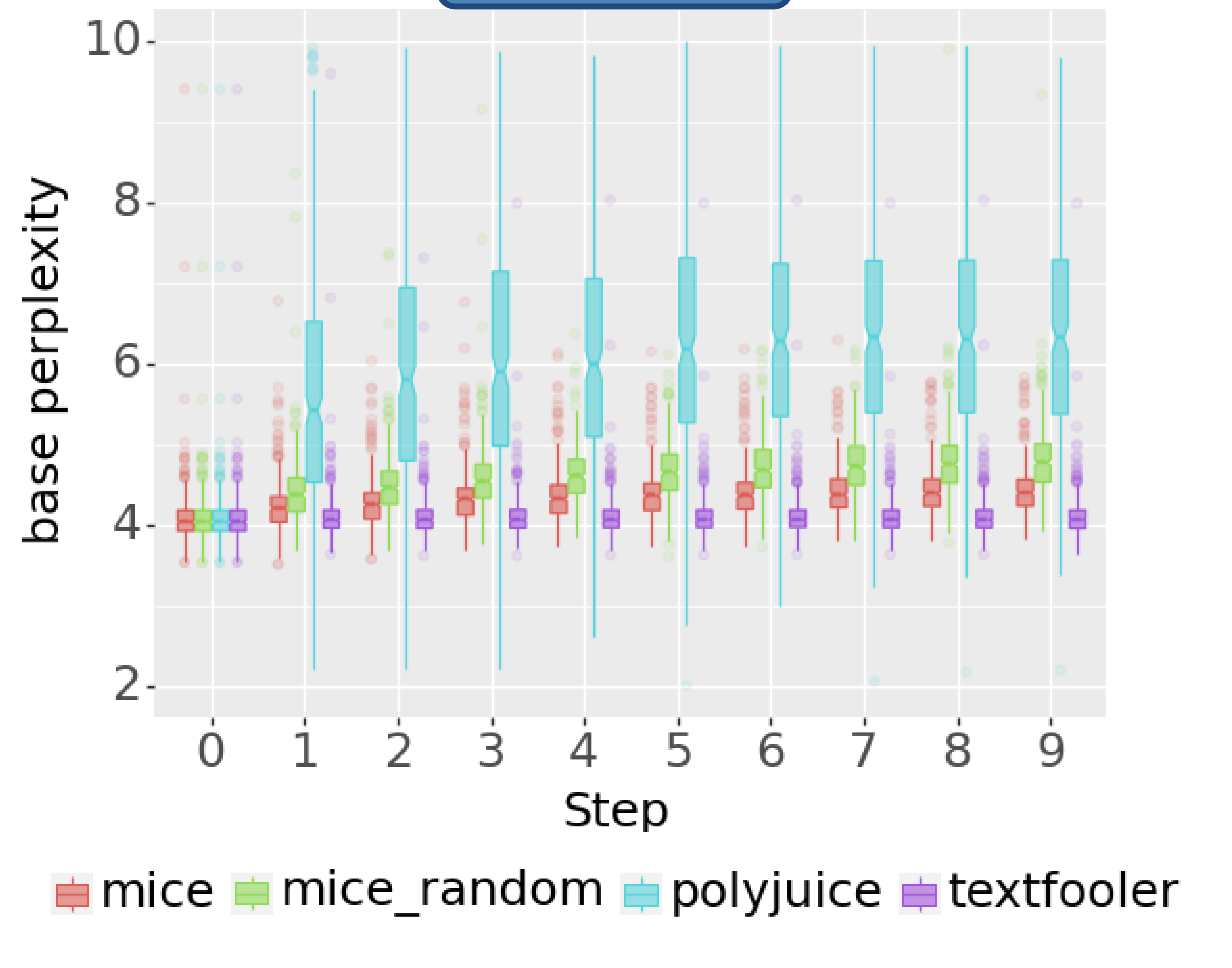
Interpreting the qualitative results

Base Perplexity

Out-of-the-box

Intuitively: Lower values mean more predictable edits. Higher values mean more diverse - surprising edits.

IMDb



Polyjuice creates more diverse text->has increasing ppl values
>model trained on many datasets

TextFooler's perplexity does not deteriorate at later steps
> maintains sentence's structure

Experiments

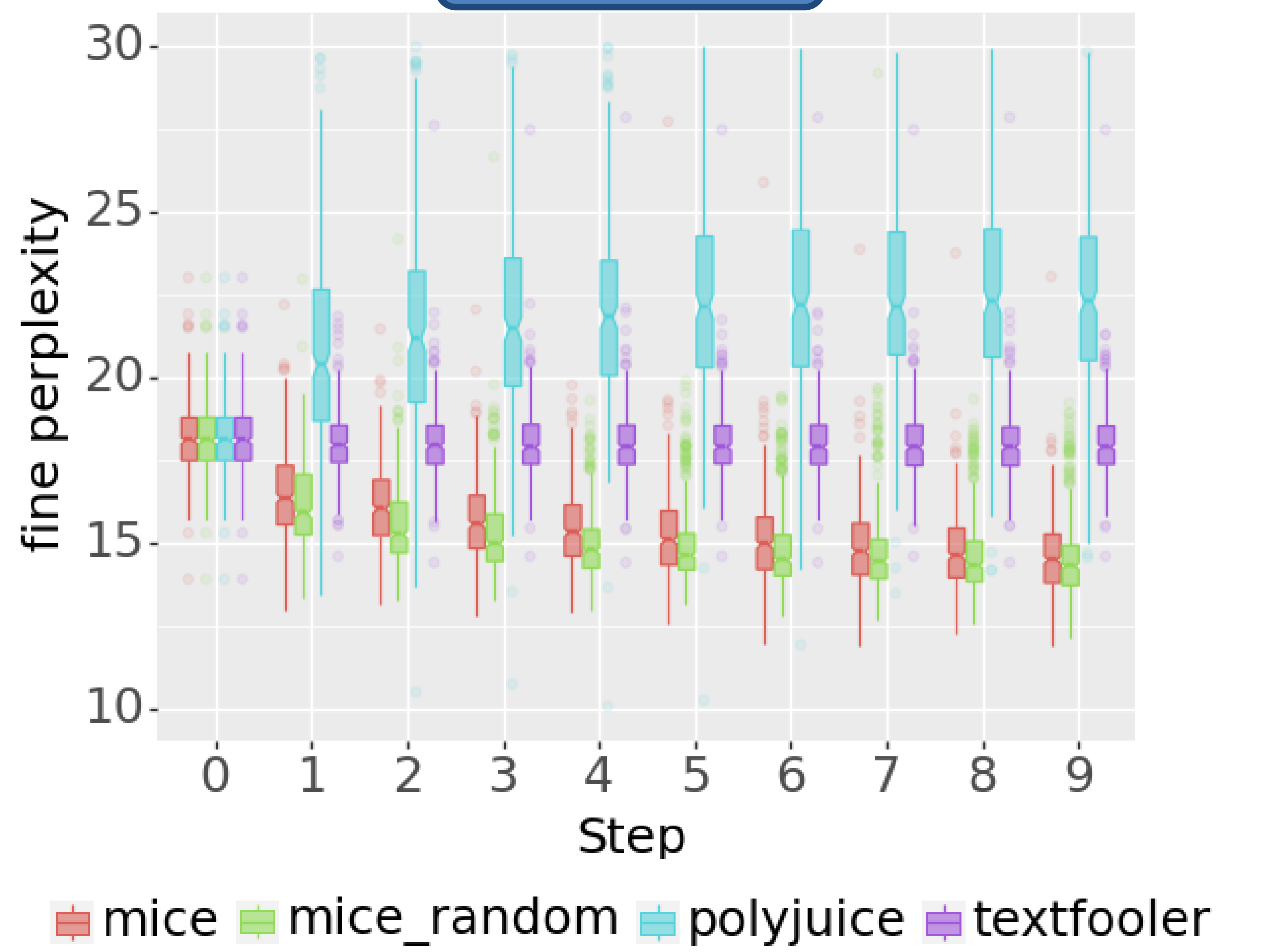
Interpreting the qualitative results

Fine Perplexity

Out-of-the-box

Intuitively: Lower values mean that the edits converge to the dataset's distribution. Assesses how the model has adapted to the specific dataset.

IMDb



- MiCE and MiCERandom present decrease in fine-ppl (!)
  > Overfitting behavior. They are pre-trained on the IMDb dataset , the same dataset the model of fine-ppl is fine-tuned on!

- **TextFooler** is stable, and **Polyjuice** generates more diverse text

Experiments

Interpreting the qualitative results

Base & Fine Perplexity

POS tag constraint

- All editors present **lower base perplexity values.**
  > Cause: Editing less tokens favors the maintenance of the sentence's structure.
  However, we have less diversity.

- The POS constraint helps to limit the overfitting behavior
  > the text converges slower to the dataset's distribution

Experiments

Interpreting the qualitative results

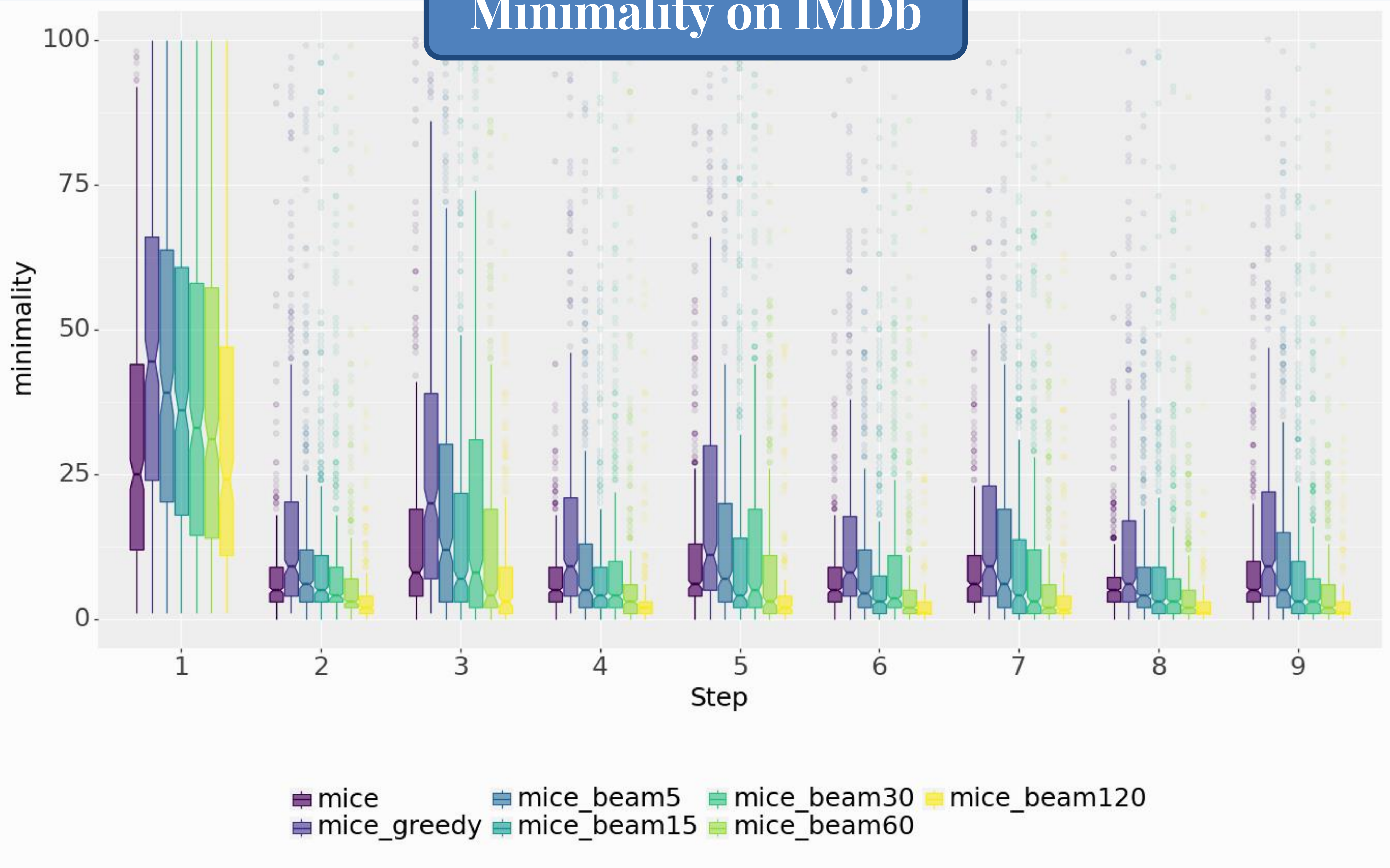
Beam-search on MiCE

- Beam search with 120 beams **outperforms MiCE's** generation method (multinomial sampling) on IMDb in terms of **minimality and inconsistency**

- Beam search with a high number of beams enables the model to explore more substitutions, increasing the possibility of minimal edits

- Slightly more diverse text is generated as beams increase

Minimality on IMDb



06.

Conclusion & Future work

Conclusion & Future Work

Conclusion

In this diploma thesis, we:

- Implement a counterfactual generation system
- Conduct experiments with multiple counterfactual editors and methods and generate thousands of counterfactuals
- Introduce a novel method for counterfactuals generation which leverages part-of-speech tagging
- Effectively utilized and expanded methods in the recent bibliography and proved their efficiency
- Explain the decisions of counterfactual editors and explore potential vulnerabilities

Conclusion & Future Work

Future work

- Experiment with more NLP tasks, e.g. Named Entity Recognition
- Study the insights inconsistency provides when combined with other metrics too
- Different predictors in order to investigate potential bias
- Create a dataset with the generated counterfactual explanations for other tasks, e.g. data augmentation
- Explore the inverse problem:
  Develop a new editor from scratch based on this work's evaluation and explanations, e.g. optimized on inconsistency for more consistent edits

Thank you for your attention!

Questions?