

Ampire 360°



ПЕРСПЕКТИВНЫЙ
МОНИТОРИНГ

Ampire 360°

ПОЛИГОН КИБЕРБЕЗОПАСНОСТИ



Threat Intelligence

SOC & DFIR

Offensive Security

Киберучения

ИИ и киберзащита

Киберполигон Ampire

📅 14 октября





Тестирование ИИ моделей. Применение TextAttack для выполнения Evasion Attack

Виктор Николаевич
Лавлинских

исследователь

Перспективный мониторинг



План

1. Объект и цель
2. AITG-MOD и evasion attack
3. 2 идеи состязательных примеров в NLP
4. Стенд
5. TextAttack. Применение TextFooler и TextBugger



Объект и цель

Используемые
методы: TextFooler
и TextBugger

Цель атаки: неверно
классифицировать
новости

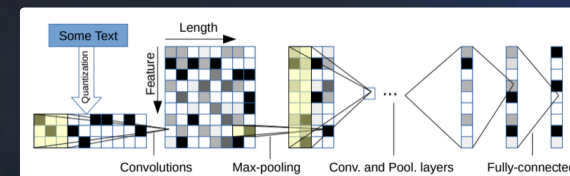
TextAttack 



API



Классификация новостей



AITG-MOD

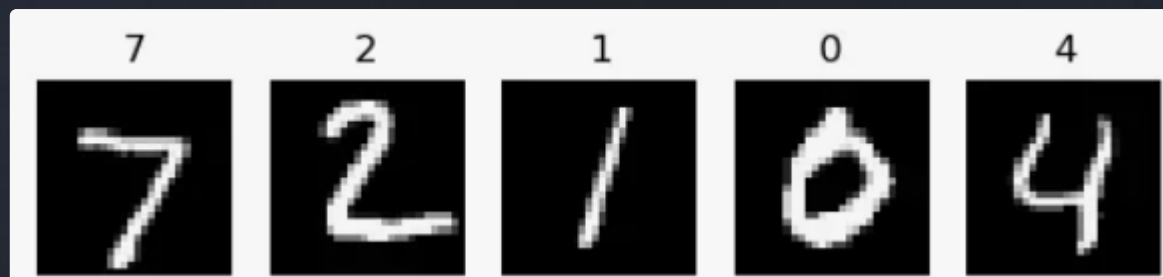
- ▶ AITG-APP — AI Application Testing (Тестирование ИИ приложений)
- ▶ AITG-MOD — AI Model Testing (Тестирование ИИ моделей)
- ▶ AITG-INF — AI Infrastructure Testing (Тестирование инфраструктуры ИИ)
- ▶ AITG-DAT — AI Data Testing (Тестирование данных ИИ)



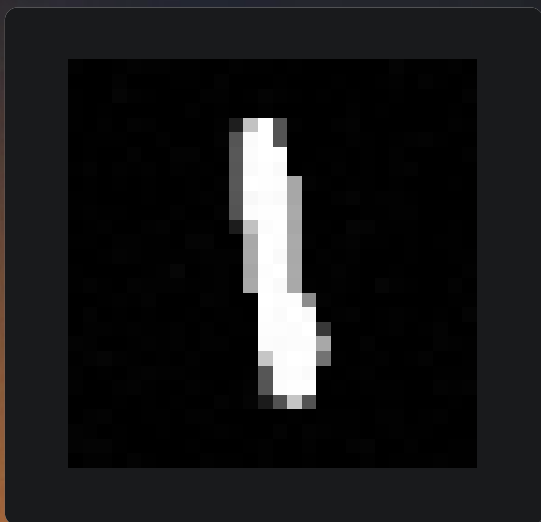
Evasion Attacks (атаки уклонения)

Изображения

Пример данных



Оригинал — цифра 1



После evasion attack — цифра 8



Evasion Attacks (атаки уклонения)

Текст

Ampire 360°



Анализ настроений

Original

Perfect performance by the actor → **Positive (99%)**

Adversarial

Spotless performance by the actor → **Negative (100%)**

Группы атаки уклонения



Используется: нецелевые, итеративные, черный ящик

Идеи состязательных примеров (adversarial example) в NLP

Original Input	Connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: <u>Positive (77%)</u>
Adversarial example [Visually similar]	<u>Aonnoisseurs</u> of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: <u>Negative (52%)</u>
Adversarial example [Semantically similar]	Connoisseurs of Chinese <u>footage</u> will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: <u>Negative (54%)</u>

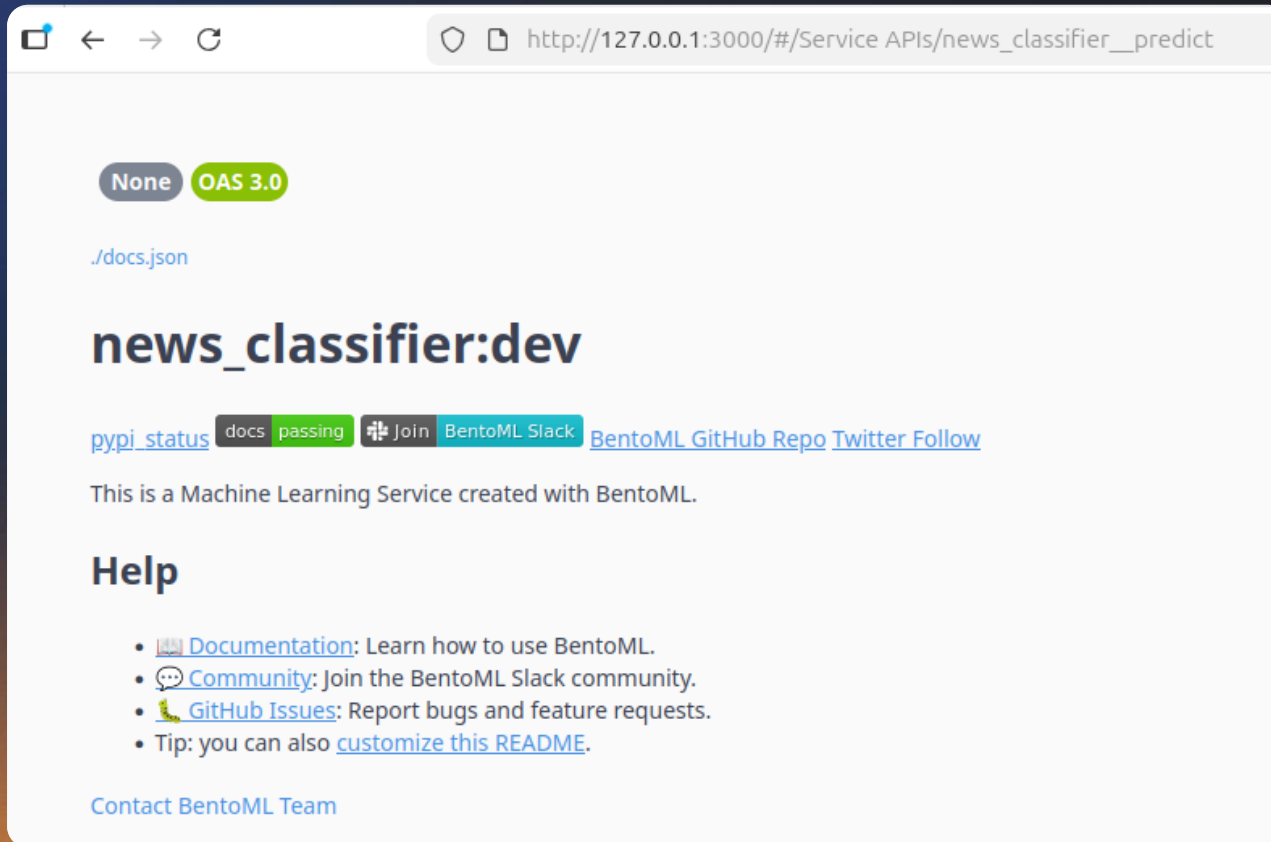
Описание модели

Сверточная нейронная сеть классифицирует новости на 4 класса:

1. Мир
2. Спорт
3. Бизнес
4. Наука и технологии

Предсказанный класс: Sports

```
"ATHENS (Reuters) - A weary Michael Phelps targeted his  
"fourth Olympic gold medal in Athens, turning his attention on  
"Wednesday to the 200 meters individual medley and settling for  
"the second-fastest overall time in the heats"
```



Заныск: `bentoml serve service.py:NewsClassifier --reload`

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:3000/predict' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
    "text": "USATODAY.com - The Securities and Exchange Commission is expected to vote Wednesday to prohibit mutual fund companies from funneling stock trades to brokerage firms that agree to
```

Request URL

http://127.0.0.1:3000/predict

Server response

Code	Details
------	---------

200

Response body

```
{
  "predicted_class": 3,
  "probabilities": [
    0.0013986504636704922,
    0.0000011119382179458626,
    0.9979859590530396,
    0.0006142202764749527
  ]
}
```



Download

TextAttack

1. Использование готовых методов атак
2. Разработка собственных методов атак
3. Аугментация обучающих данных
4. Обучение NLP моделей

Почему TextFooler и TextBugger



Вывод готовых
методов: `textattack`
`list attack-recipes`

```
a2t (textattack.attack_recipes.A2TYoo2021)
alzantot (textattack.attack_recipes.GeneticAlgorithmAlzantot2018)
bae (textattack.attack_recipes.BAEGarg2019)
bert-attack (textattack.attack_recipes.BERTAttackLi2020)
checklist (textattack.attack_recipes.CheckList2020)
clare (textattack.attack_recipes.CLARE2020)
deepwordbug (textattack.attack_recipes.DeepWordBugGao2018)
faster-alzantot (textattack.attack_recipes.FasterGeneticAlgorithmJia2019)
hotflip (textattack.attack_recipes.HotFlipEbrahimi2017)
iga (textattack.attack_recipes.IGAWang2019)
input-reduction (textattack.attack_recipes.InputReductionFeng2018)
kuleshov (textattack.attack_recipes.Kuleshov2017)
morpheus (textattack.attack_recipes.MorpheusTan2020)
pruthi (textattack.attack_recipes.Pruthi2019)
pso (textattack.attack_recipes.PSOZang2020)
pwws (textattack.attack_recipes.PWWSRen2019)
seq2sick (textattack.attack_recipes.Seq2SickCheng2018BlackBox)
textbugger (textattack.attack_recipes.TextBuggerLi2018)
textfooler (textattack.attack_recipes.TextFoolerJin2019)
```

Статьи, в которых использовали TextFooler и TextBugger:

2025 "Tougher Text, Smarter Models: Raising the Bar for Adversarial Defence Benchmarks"

2024 "Accuracy of TextFooler black box adversarial attacks on 01 loss sign activation neural network ensemble"

2022 "Text Adversarial Purification as Defense against Adversarial Attacks"

Greedy — Word Importance Ranking

Оригинал:

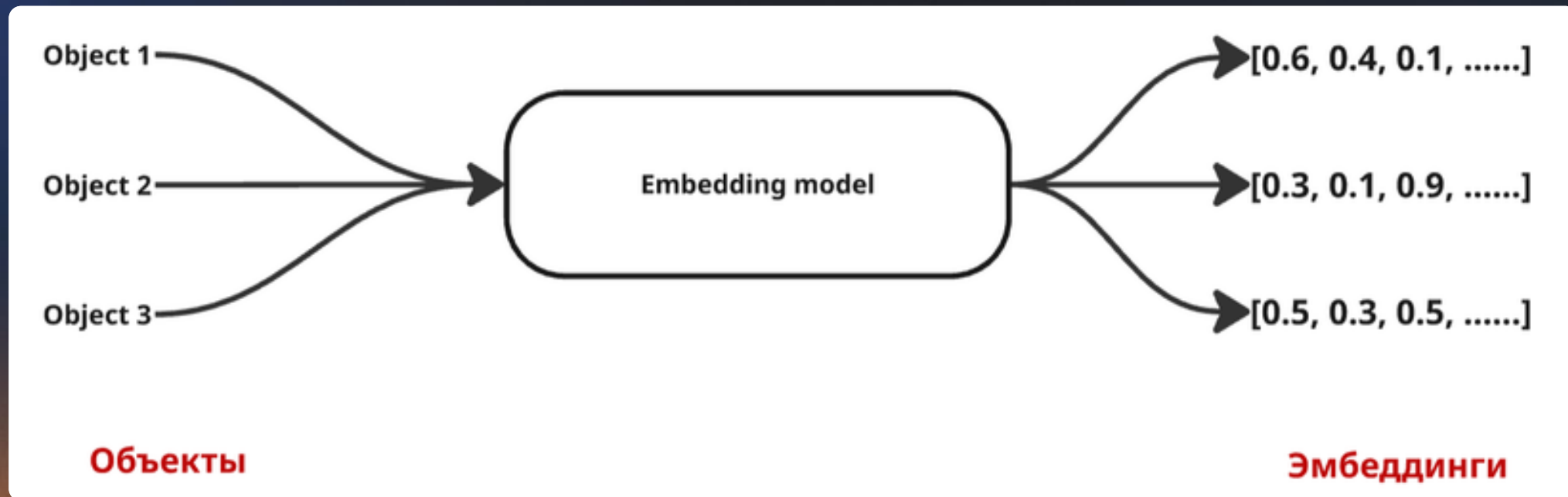
The top two U.S. automakers continue to deal with declining sales while the little guy grows.

Вероятности: [0.0053 0.0004 0.9737 0.0206] - 3 класс (Бизнес)

Значимость	Слово
1	sales
2	declining
3	guy
4	automakers
5	grows
6	deal

Значимость	Слово
7	The
8	top
9	two
10	S
11	little
12	U
13	continue

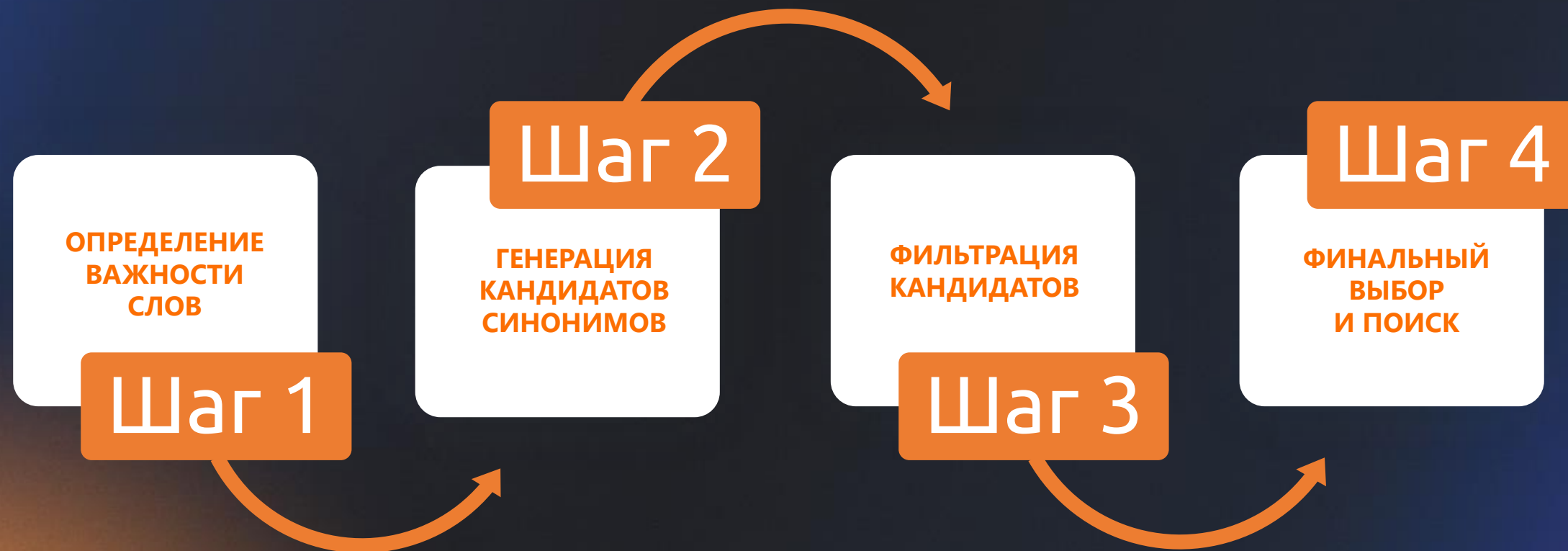
Поиск кандидата и семантическое сходство



Поиск синонима — counterfitted_GLOVE

Семантическое сходство предложений — Universal Sentence Encoder

TextFooler — алгоритм



TextFooler применение

Оригинал:

UAW said its members at Caterpillar Inc on Sunday voted to reject the construction equipment maker's contract proposal, the second time this year workers have voted against a Caterpillar contract offer

Вероятности: [0.0116 0.0015 0.9023 0.0847] - 3 класс (Бизнес)

После evasion attack:

UAW said its members at Caterpillar Lnc on Sunday voted to reject the construction equipment maker's contract proposal, the second time this year agents have voted against a Caterpillar contract offer

Вероятности: [0.0527 0.5735 0.262 0.1118] - 2 класс (Спорт)

Значимость	Слово
1	Inc
2	workers
3	Sunday
4	year
5	second
6	contract

TextBugger гибрид



1. Insert (Вставка): awesome → awe some
2. Delete (Удаление): awesome → awesme
3. Swap (Перестановка): awesome → awesmoe
4. Substitute-Character (Замена символа на похожий): awesome → awes0me
5. Substitute-Word (Замена слова на синоним): awesome → amazing

TextBugger алгоритм



TextBugger применение

Оригинал:

UAW said its members at Caterpillar Inc on Sunday voted to reject the construction equipment maker's contract proposal, the second time this year workers have voted against a Caterpillar contract offer

Вероятности: [0.0116 0.0015 0.9023 0.0847] - 3 класс (Бизнес)

После evasion attack:

UAW said its members at Caterpillar Inc on Sunday voted to reject the construction equipment maker's contract proposal, the second time this year laborers have voted against a Caterpillar contract offer

Вероятности: [0.0513 0.7269 0.157 0.0648] - 2 класс (Спорт)

Значимость	Слово
1	Inc
2	workers
3	Sunday
4	year
5	second
6	contract



- ▶ AITG-MOD и evasion attack.
- ▶ Стенд для применения TextAttack
- ▶ Разобраны идеи состязательных атак на примере TextFooler и TextBugger

Благодарю за внимание

Виктор Николаевич
Лавлинских
исследователь



Ampire 360°