

<https://www.kaggle.com/>

[datasets/andrewmvd/](https://www.kaggle.com/datasets/andrewmvd/)

[cyberbullying-classification](https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification)

Команда:

Закирова К.
Матущак А.
Свиридов В.
Шестаков М.

Кибербуллинг

**выявление
и определение вида
с помощью моделей ML
(Twitter)**



Кибербуллинг

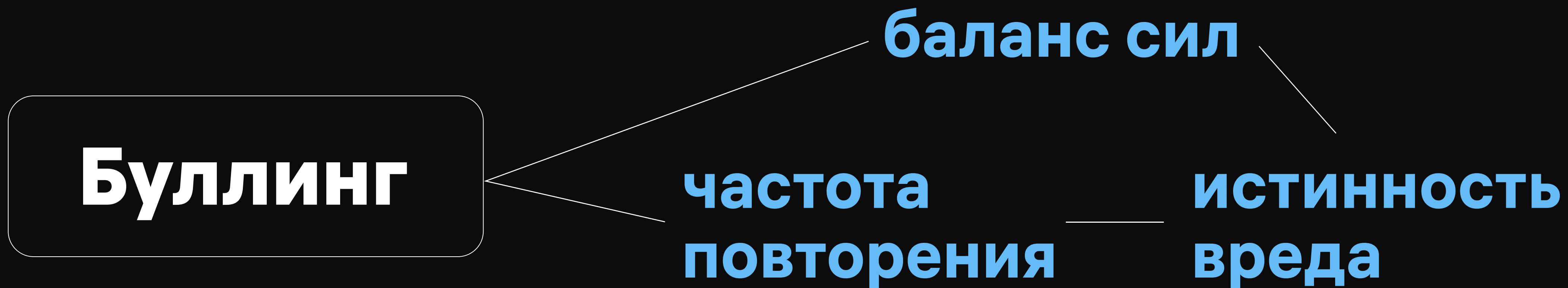
— использование цифровых технологий
для причинения вреда или для запугивания



Wang, Jason & Fu, Kaiqun & Lu, Chang-Tien. (2020). SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection. 1699-1708. 10.1109/BigData50022.2020.9378065.

Проблема феномена

— цифровой характер и относительная
анонимность



Wang, Jason & Fu, Kaiqun & Lu, Chang-Tien. (2020). SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection. 1699-1708. 10.1109/BigData50022.2020.9378065.

СОЦ.
СЕТИ



СОЦ.
СЕТИ



Актуальность

- современная социально-острая проблема, вытекающая в эмоциональный и физический дискомфорт людей IRL.
- достаточно сложная разработка, востребованная в продуктовых решениях: выявление кибербуллинга позволяет влиять на пользовательский опыт в соц. медиа.



(глобально)

36,5%

**людей чувствовали
на себе кибербуллинг**

[https://cyberbullying.org/2019-
cyberbullying-data](https://cyberbullying.org/2019-cyberbullying-data)

60%

**тинейджеров подвергались
разному виду кибербуллинга**

[https://cyberbullying.org/
summary-of-our-cyberbullying-
research](https://cyberbullying.org/summary-of-our-cyberbullying-research)

Новизна

- использование фреймворка Streamlit, предоставляющего непосредственный доступ к использованию моделей
- попытка охватить как можно больше моделей и подходов к векторизации для выявления лучшего результата определения хайта

(локально)



Обзор

Эволюция
детекции
хейта

7

- Keyword-based classifiers
- Classifiers using distributed semantics
- Deep learning classifiers with advanced linguistic features

Salminen, Joni & Hopf, Maximilian & Chowdhury, Shammur & Jung, Soon-Gyo & Almerekhi, Hind & Jansen, Jim. (2020). Developing an online hate classifier for multiple social media platforms. 10. 1. 10.1186/s13673-019-0205-6.

Keyword-based classifiers

наличие/отсутствие ключевых слов, словари

Проблемы: сложность распознавания сарказма/юмора; словари требуют постоянного обновления (неологизмы, сленг); восприятие ненависти в различных сообществах (пр. Stackoverflow); word-sense disambiguation — полисемия

Salminen, Joni & Hopf, Maximilian & Chowdhury, Shammur & Jung, Soon-Gyo & Almerekhi, Hind & Jansen, Jim. (2020). Developing an online hate classifier for multiple social media platforms. 10. 1. 10.1186/s13673-019-0205-6.

Distributional semantics

9

n-grams, векторные представления слов,
word vector space models, синтаксические
особенности

Deep learning classifiers

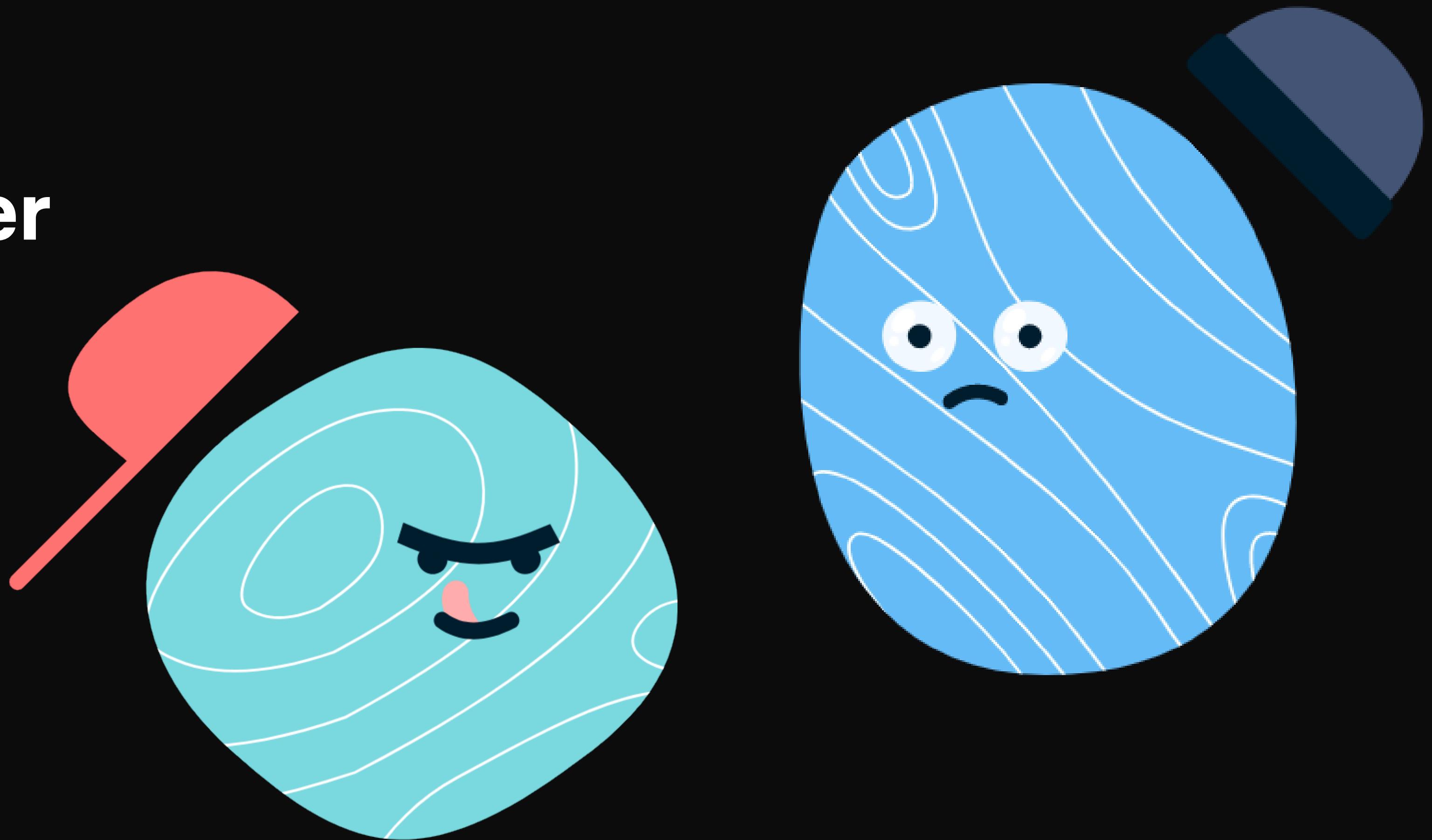
нейронные сети (CNN, RNN)

Salminen, Joni & Hopf, Maximilian & Chowdhury, Shammur & Jung, Soon-Gyo & Almerekhi, Hind & Jansen, Jim. (2020). Developing an online hate classifier for multiple social media platforms. 10. 1. 10.1186/s13673-019-0205-6.



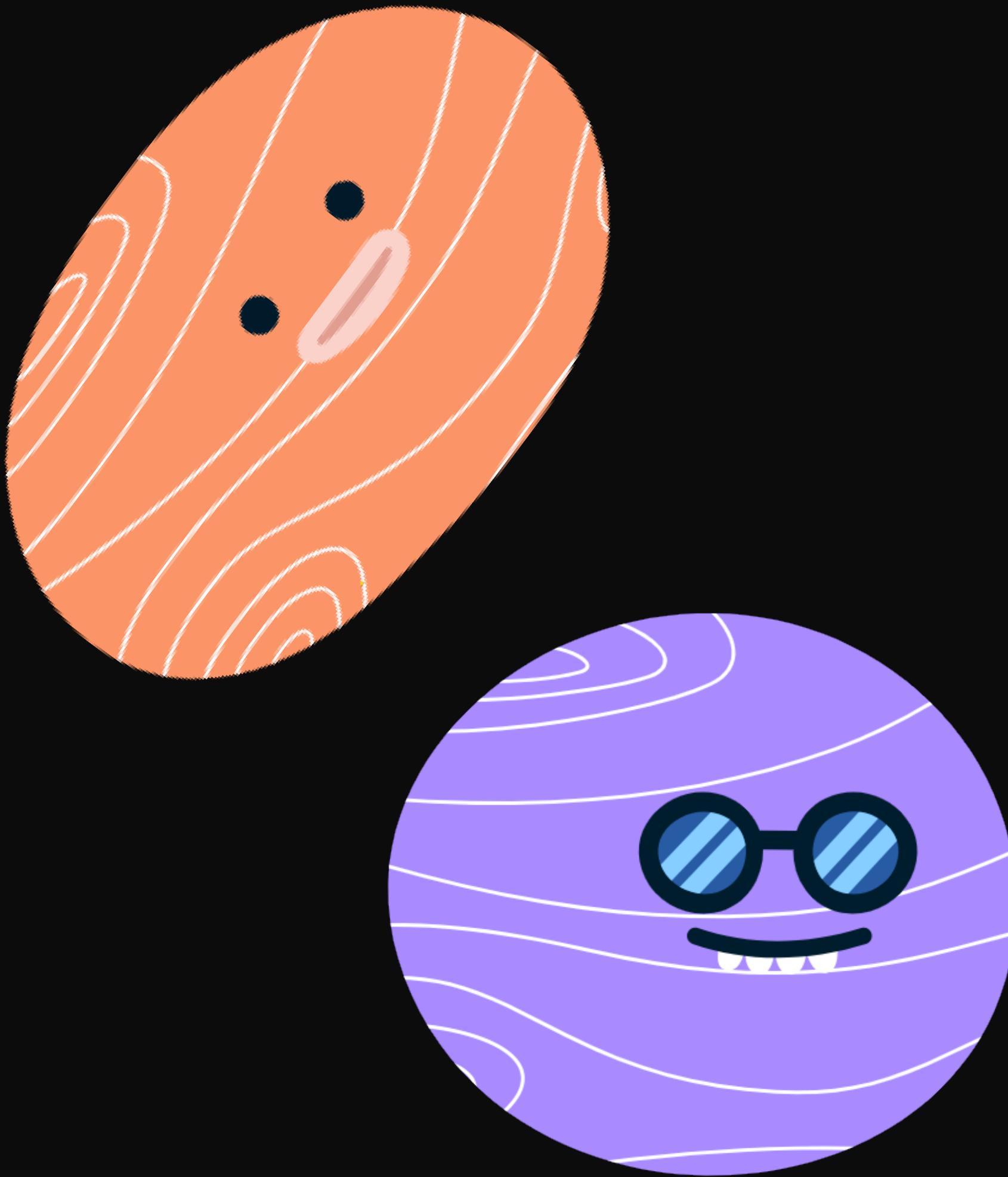
Векторизация

- TF-IDF
- Count Vectorizer



Модели

- Logistic regression
- KNN
- Decision Tree
- Random Forest
- Multinomial Naive Bayes
- SVM (SVC, Linear SVC)



Ограничения

— возможное падение точности при работе с текстами, не являющимися твитами; спецификация работы только с одной соц. сетью

there is a lack of development and testing of models using data from multiple social media platforms

Salminen, Joni & Hopf, Maximilian & Chowdhury, Shammur & Jung, Soon-Gyo & Almerekhi, Hind & Jansen, Jim. (2020). Developing an online hate classifier for multiple social media platforms. 10. 1. 10.1186/s13673-019-0205-6.

Успехи

- создание приложения
- применение разных подходов
- определяем кибербуллинг в **заданном** тексте



Я чувствую себя KNN-
классификатором,
медленный, плохо
работаю, но всё
равно стараюсь



[https://acipenserstudio-
iad.streamlit.app/
diagrams](https://acipenserstudio-iad.streamlit.app/)

Лучшие параметры¹⁵

LogReg {'solver': 'saga', 'penalty': 'l2',
'class_weight': None, 'C': 1.0}

82%

L SVC {'multi_class': 'ovr', 'loss':
'squared_hinge', 'fit_intercept': True,
'class_weight': None, 'C': 0.5}

Время обучения

- SVC обучалась дольше всего (2 мин)
- KNN и MNB обучились быстрее всего (7 сек)



Время предсказания

- LogReg — лучшее время (0.5 с)
- Остальные модели — также быстро < 1 с
(кроме KNN и SVC)



Качество

- KNN — худший результат (~50%)
- SVC, Linear SVC, LogReg — лучшие результаты (82%)
- Random Forest — незначительно хуже (81%)

Векторайзеры

- Равные настройки, ограничение на словарь (1000d)
- Count Vectorizer чуть-чуть быстрее и точнее

Результат нашей
работы в том, что мы
изучили, как люди
кибербуллят, и теперь
можем кибербулить
их ещё эффективнее



Литература

1. Wang, Jason & Fu, Kaiqun & Lu, Chang-Tien. (2020). **SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection.** 1699-1708. [10.1109/BigData50022.2020.9378065](https://doi.org/10.1109/BigData50022.2020.9378065).
2. Salminen, Joni & Hopf, Maximilian & Chowdhury, Shammur & Jung, Soon-Gyo & Almerekhi, Hind & Jansen, Jim. (2020). **Developing an online hate classifier for multiple social media platforms.** 10. 1. [10.1186/s13673-019-0205-6](https://doi.org/10.1186/s13673-019-0205-6).
3. <https://cyberbullying.org/2019-cyberbullying-data>