ФИНАЛ. DATA SCIENCE

CHANGELLENGE CUP IT 2021

Big уточки

НАША КОМАНДА

ЮЛИЯ КОРОТКОВА

"Фундаментальная и компьютерная лингвистика"
email: koylenka15@gmail.com

АРИНА ЮСУПОВА

"Бизнес-информатика"
email: yusupovaari@yandex.ru

ВЕРОНИКА ЗЫКОВА

"Фундаментальная и компьютерная лингвистика"
email: vzykova2001@gmail.com

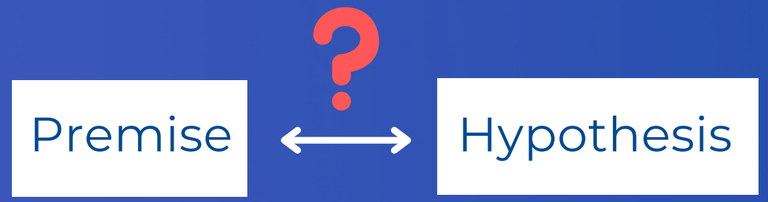
УЛЬЯНА КАЗАКОВА

" Совместный бакалавриат НИУ ВШЭ и ЦПМ (Математика)"
email: kazakovau64@gmail.com

Студентки 2-го курса бакалавриата НИУ ВШЭ

EXECUTIVE SUMMARY

Проблема



Важно:

- найти все противоречия
- работа с большими объемами данных
- официальный стиль
- юридическая лексика

Инициативы

Данные

Семантические модели

- word2vec
- tf-idf
- BERT
- RoBERTa
- ELMo

Синтаксические модели

- graph2vec
- SPINN
- CA_GCN

Логистическая регрессия

Качество модели

	precision	recall	fscore
contradiction	0.52	0.46	0.46
neutral	0.47	0.55	0.54
entailment	0.56	0.53	0.53
Weighted	0.51	0.47	0.48

Accuracy 0. 5126

ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ

МОДЕЛИ

Model	Matched	Mismatched
RoBERTa (Liu et al., 2019)	90.8	90.2
XLNet-Large (ensemble) (Yang et al., 2019)	90.2	89.8
MT-DNN-ensemble (Liu et al., 2019)	87.9	87.4
Snorkel MeTaL(ensemble) (Ratner et al., 2018)	87.6	87.2
Finetuned Transformer LM (Radford et al., 2018)	82.1	81.4

ПРОБЛЕМЫ*

Все модели учитывают семантику
НО не учитывают синтаксис

Вася бьет Петю
VS
Петя бьет Васю

- Высокая семантическая близость
- Противоречие

Premise	Hypothesis	Gold	BERT-CLS
The student saw the managers.	The managers saw the student.	**N**	E
The judge in front of the manager saw the doctors.	The doctors saw the judge.	**N**	E
The bankers admired the lawyer that the students supported.	The lawyer admired the students.	**N**	E
The secretary and the managers saw the actor.	The secretary saw the managers.	**N**	E
The manager was introduced by the professor.	The manager introduced the professor.	**N**	E

РЕШЕНИЕ

Семантика
+
Синтаксис

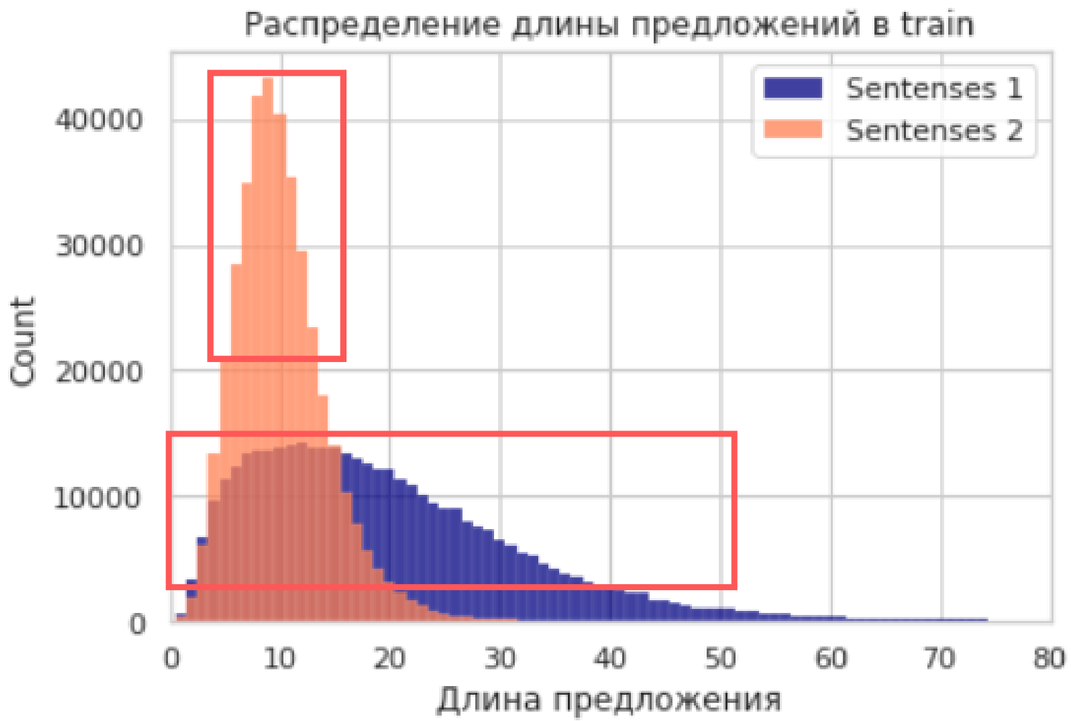
- Мало синтаксических моделей
- .Слабо изучено применение синтаксиса для решения задачи NLI

- graph2vec
- ~~SPINN~~
- ~~CA_GCN~~

*He Q., Wang H., Zhang Y. Enhancing Generalization in Natural Language Inference by Syntax Online: Association for Computational Linguistics, 2020.C. 4973–4978

АНАЛИЗ ИСХОДНЫХ ДАННЫХ

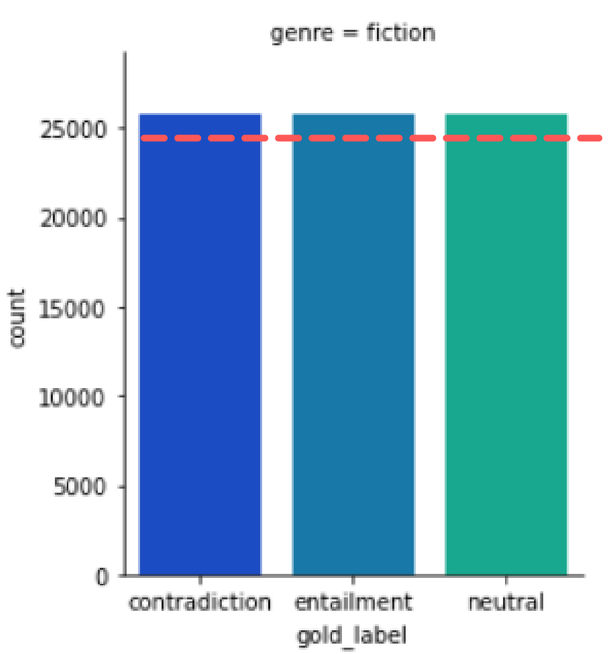
1 Количество токенов



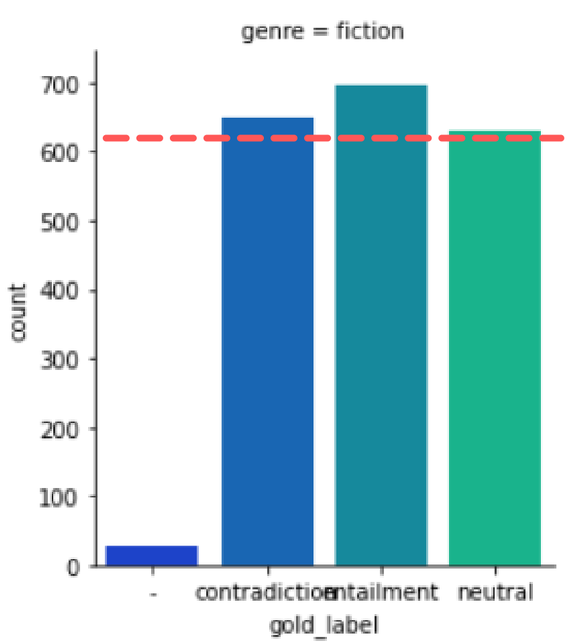
АНАЛИЗ ИСХОДНЫХ ДАННЫХ

2 Распределение лейблов

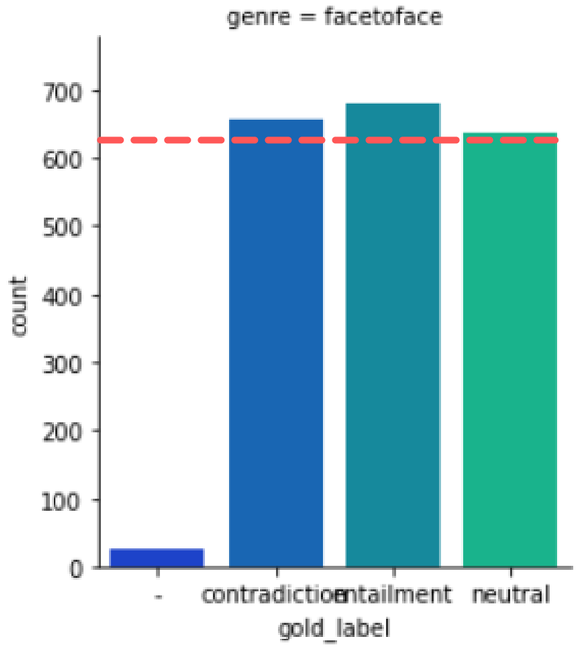
train



val_mismatched



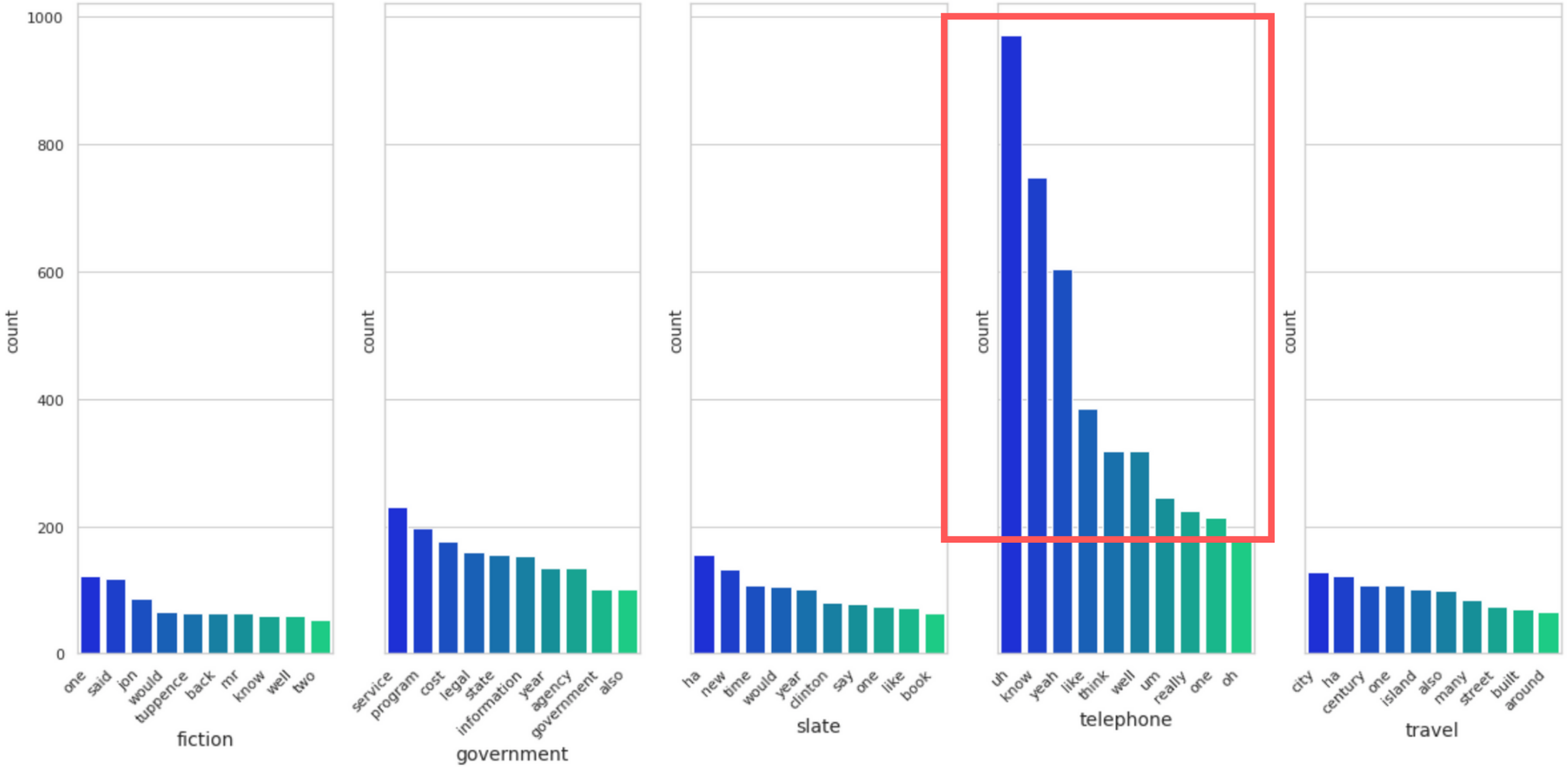
val_matched



остальное - в приложениях 1-3

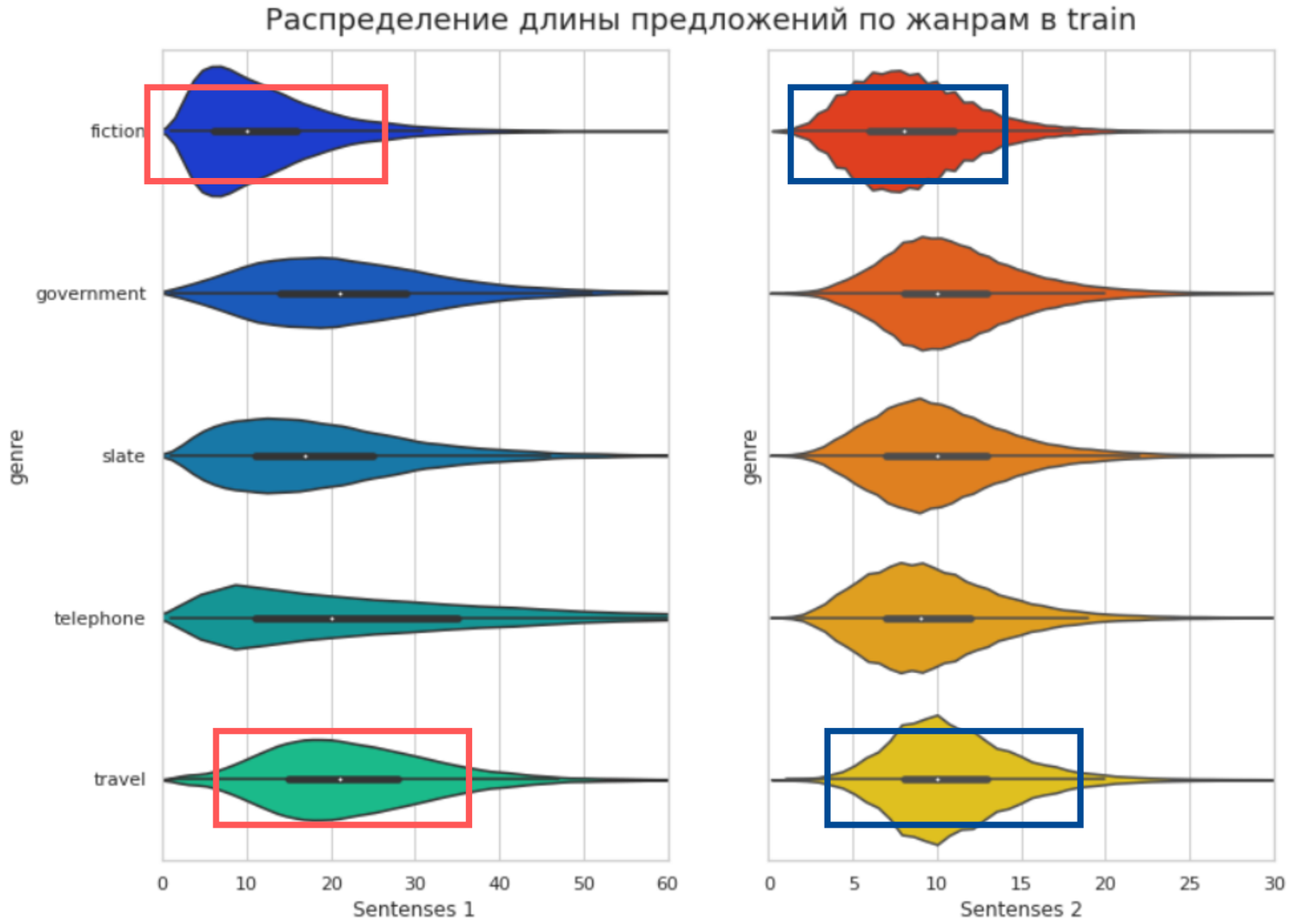
АНАЛИЗ ИСХОДНЫХ ДАННЫХ

3 Частотные слова



АНАЛИЗ ИСХОДНЫХ ДАННЫХ

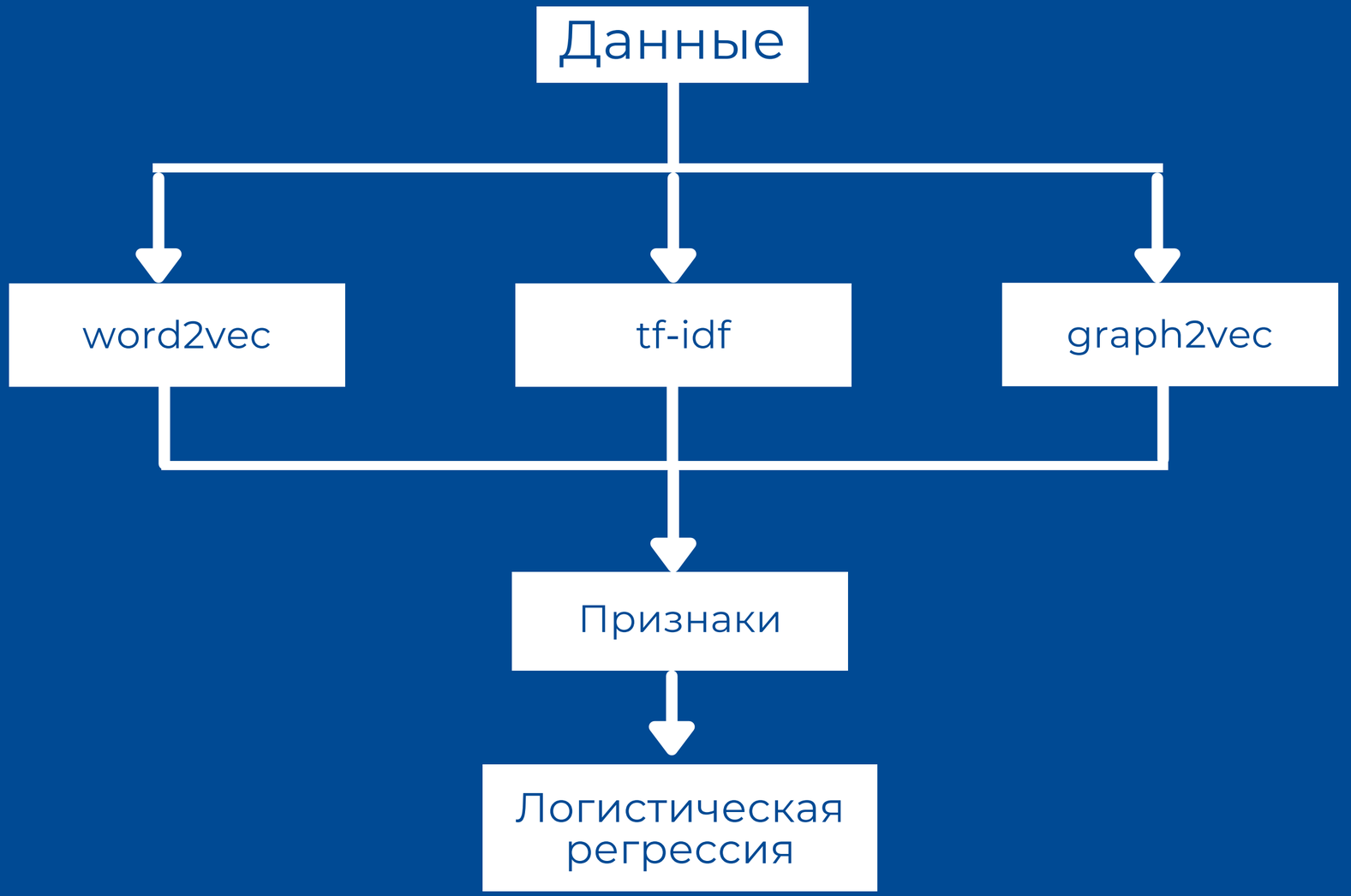
4 Длина слов



АНАЛИЗ МОДЕЛЕЙ

	Скорость	Простота реализации	Accuracy	F-score
word2vec	✓	✓	0.44	0.30
TF-IDF	✓	✓	0.51	0.47
graph2vec	✓	✓	0.32	0.18
RoBERTa	✗	✓	0.61	0.65
ELMo	✗✗	✓		
BERT		✗		
SPINN		✗		
CA_GCN		✗		

НАША МОДЕЛЬ


Данные
word2vec
tf-idf
graph2vec
Признаки
Логистическая регрессия

МЕТРИКА

$$f = \sum_{i=1}^{3} w_i * (1 + \beta^2) * \frac{precision_i * recall_i}{(\beta^2 * precision_i) + recall_i}$$, где $\beta = 3,$ $w = \{0.8, 0.1, 0.1\}$

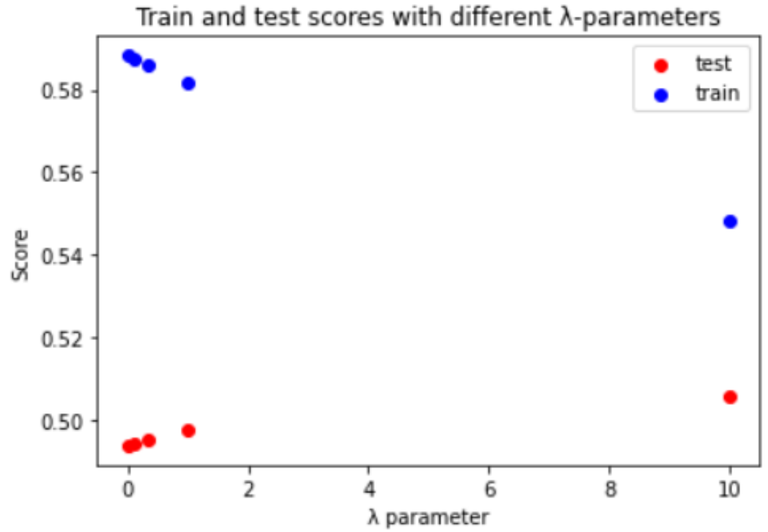
ПОЧЕМУ НЕ ACCURACY? → баланс и приоритетность классов

ПОЧЕМУ НЕ RECALL? → нужны и recall, и precision, поэтому f-мера
!но recall важнее, поэтому β=3

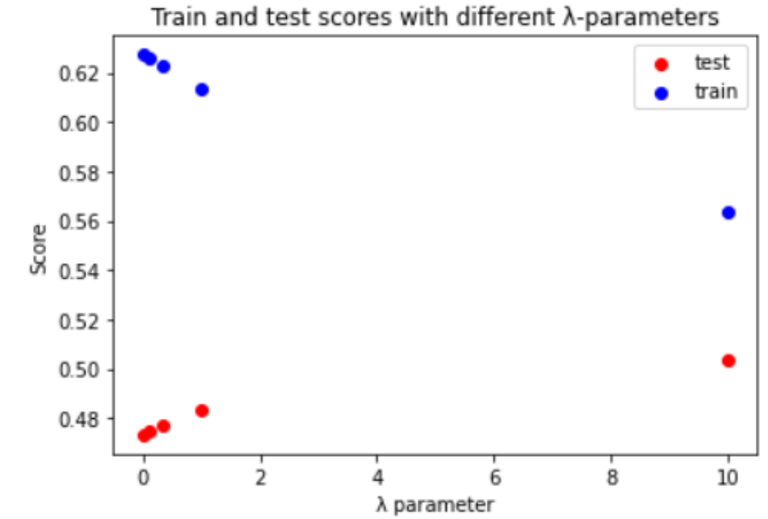
ПОЧЕМУ ВЕСА НЕ ОДИНАКОВЫЕ? → приоритетным является
определение класса "противоречий"

ТЬЮНИНГ МОДЕЛИ И РЕЗУЛЬТАТЫ

saga



sag



solver: saga

C: 0.1

penalty: L2

	acc	f-score
train	0.53	0.48
val	0.51	0.48

НАША КОМАНДА

ЮЛИЯ КОРОТКОВА

"Фундаментальная и компьютерная лингвистика"

- Победитель кейс-чемпионата FutureTech 2020
- Разработчик чат-бота «Эйнштейн» для Geek Picnic
- Разработчик разметки китайских текстов в НКРЯ
- Английский – Advanced

Email: koylenka15@gmail.com
Телефон: +79824429642

АРИНА ЮСУПОВА

"Бизнес-информатика"

- Agro Data Science Cup 2020 (21 место)
- Мисс непосредственность 2008
- Английский – Advanced

Email: yusupovaari@yandex.ru
Телефон: +79196838343

ВЕРОНИКА ЗЫКОВА

"Фундаментальная и компьютерная лингвистика"

- Телеграмм-бот, анализирующий водность и лексическое разнообразие текста
- Поддержка информационного сайта Русско-китайского параллельного корпуса НКРЯ
- Английский – Upper Intermediate

Email: vzykova2001@gmail.com
Телефон: +79622058180

УЛЬЯНА КАЗАКОВА

" Совместный бакалавриат НИУ ВШЭ и ЦПМ (Математика)"

- Член жюри устной Московской олимпиады школьников по математике 2021
- Организатор Математического праздника 2020
- Английский – Upper Intermediate

Email: kazakovau64@gmail.com
Телефон: +79164341790