

Исследование звукосимволизма с помощью Word Embedding

Заказчик: Max Planck Institute for
Mathematics in the Science

Project X

Булёв Даниил РИ-280022

Духович Александра РИ-280022

Елисеев Ян РИ-280022

Кудрявцев Тимофей РИ-280022

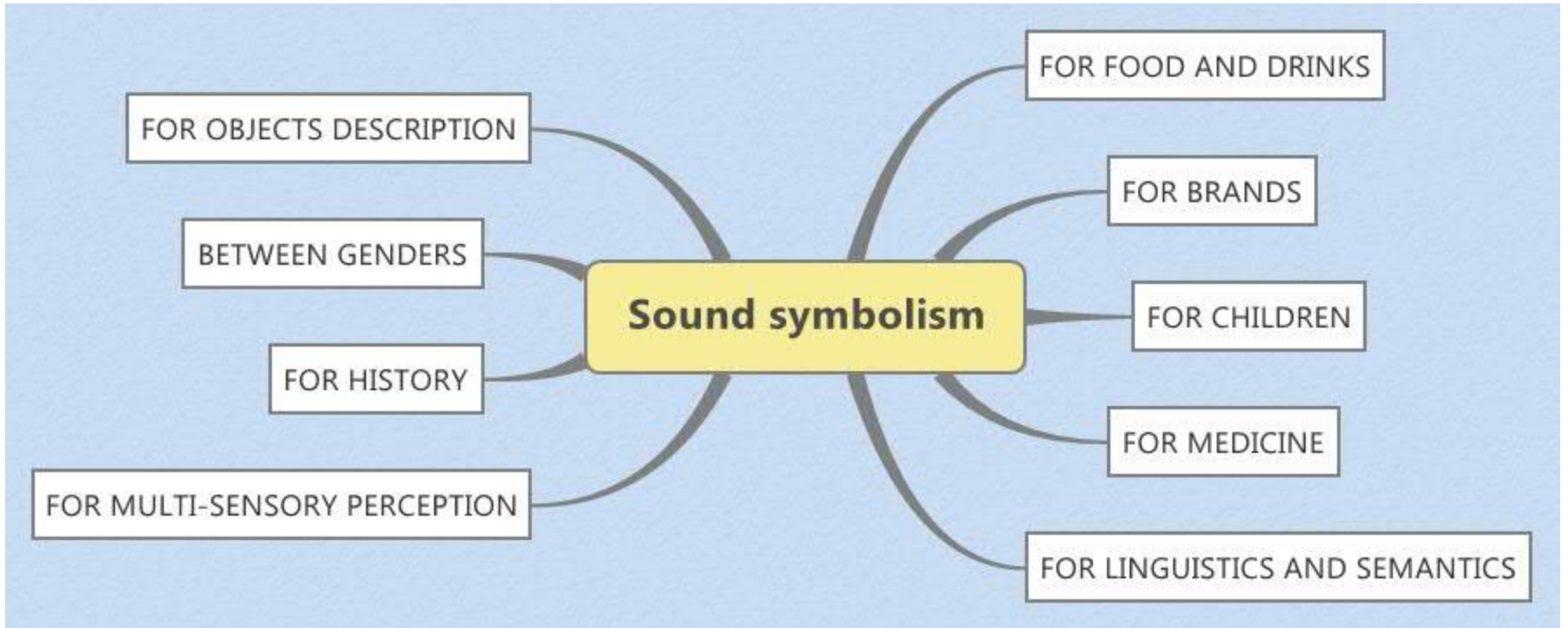
Максимов Филипп РИ-280022

Пелих Владимир РИ-280022

Пивуки

Зубарева Виктория РИ-280018

Пылаева Татьяна РИ-280018



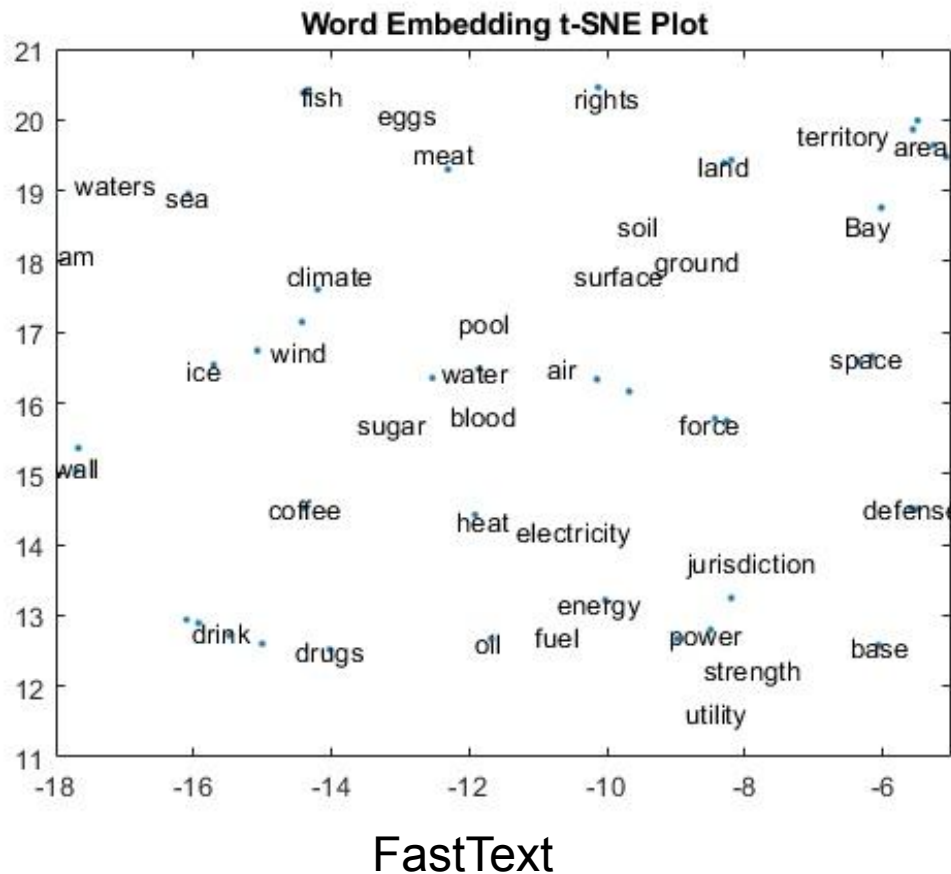
ГИПОТЕЗА

Гипотеза: существует систематическая связь между звуком и смыслом.

ЦЕЛЬ ПРОЕКТА

Цель: подтверждение или опровержение гипотезы о систематической связи между звуком и смыслом.

WORD EMBEDDING – ГРУППА МЕТОДОВ ПРЕДСТАВЛЕНИЯ СЛОВ В ВИДЕ ВЕКТОРОВ



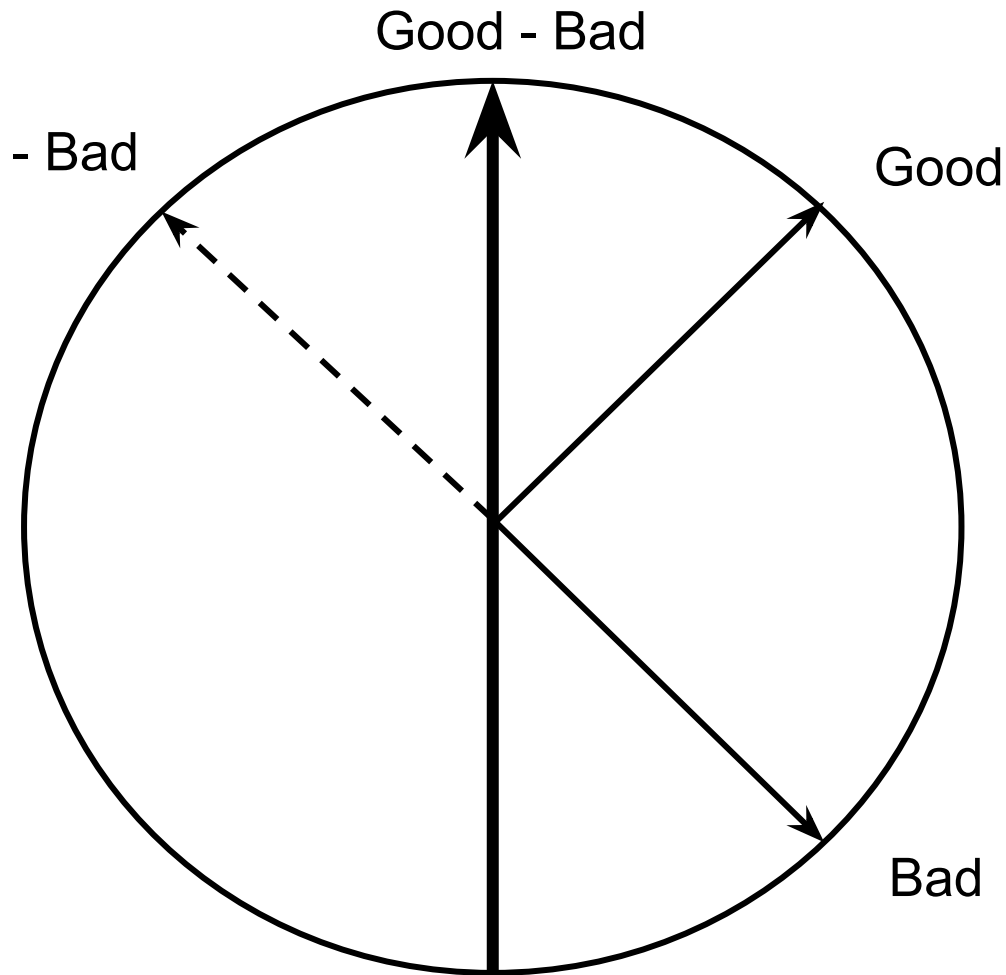
Контекстно-зависимые:
- Bert

Контекстно-независимые:
- Word2vec
- FastText
- Glove

ПЛАН РАБОТЫ

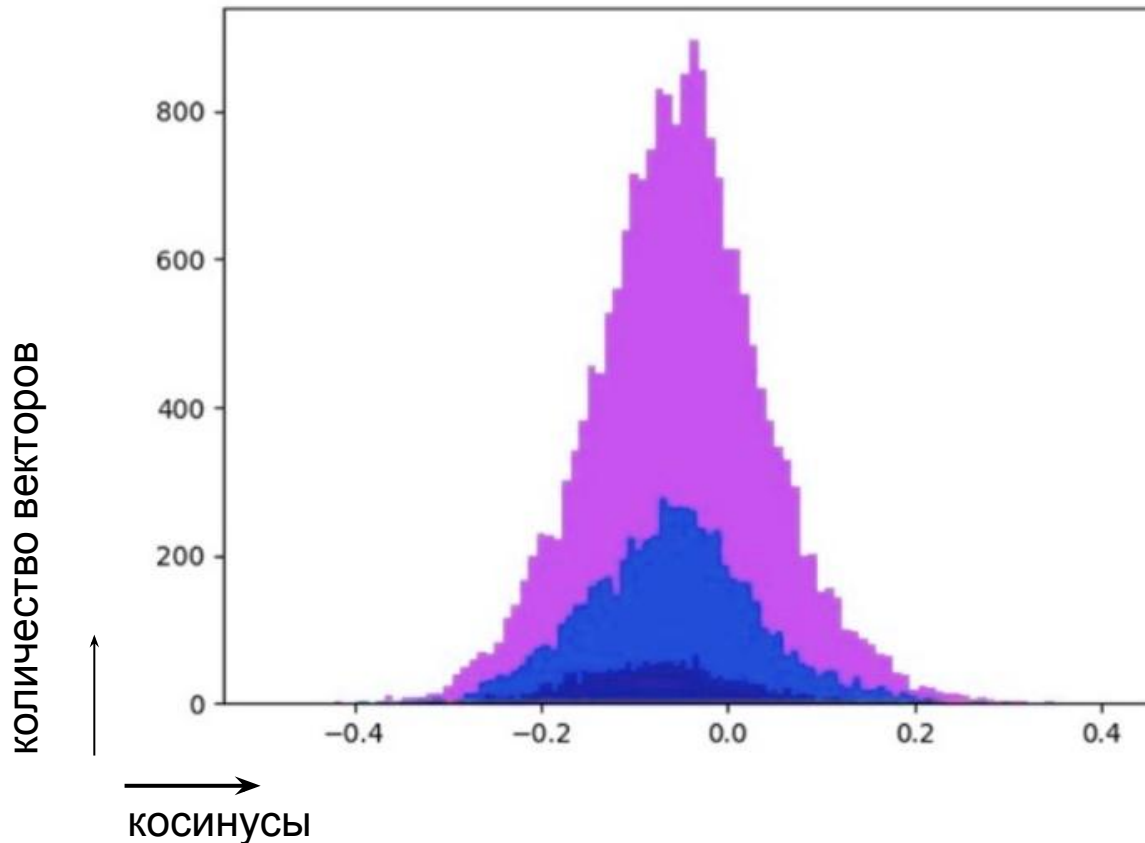
- построить семантические оси в пространстве Word Embedding;
- построить распределение проекции слов;
- найти статистические данные между распределениями;
- проанализировать результаты.

Суммирование векторов для каждой полуоси



ПРОЕКЦИИ СЛОВ НА ОСИ

$$\overrightarrow{v}_w = \cos(\overrightarrow{v}_w, \vec{A}) = \frac{\overrightarrow{v}_w \cdot \vec{A}}{|\overrightarrow{v}_w| \cdot |\vec{A}|}$$



СТАТИСТИЧЕСКИЕ ДАННЫЕ

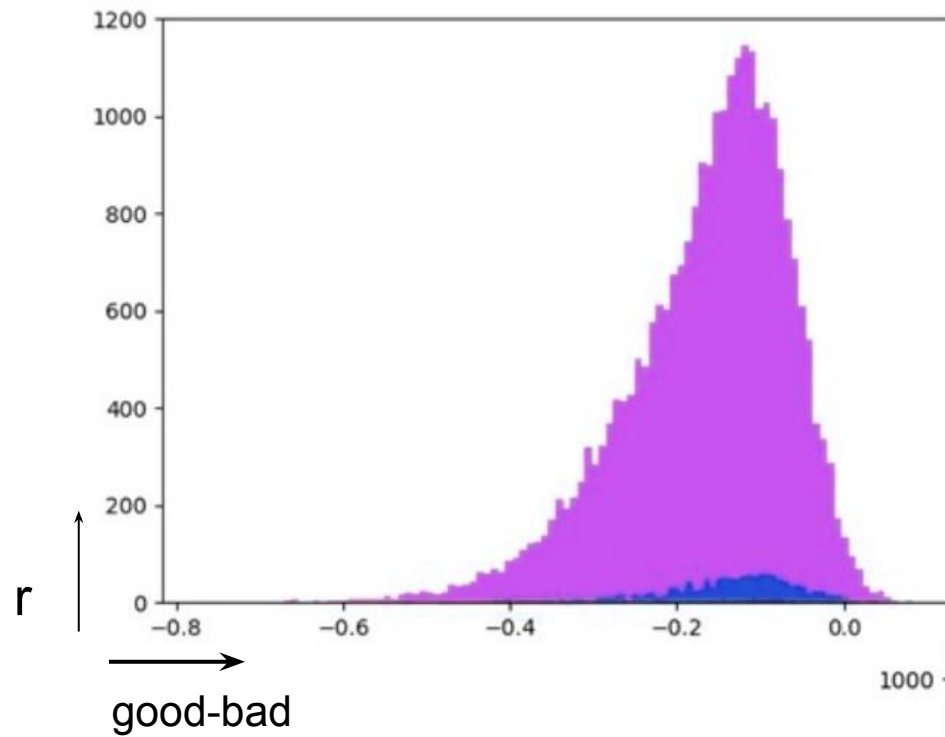
Мы использовали U-тест Манна-Уитни для сравнения двух выборок на идентичность показаний.

p - value = 1 – выборки взяты из одного распределения
 p - value = 0 – выборки взяты из разных распределений

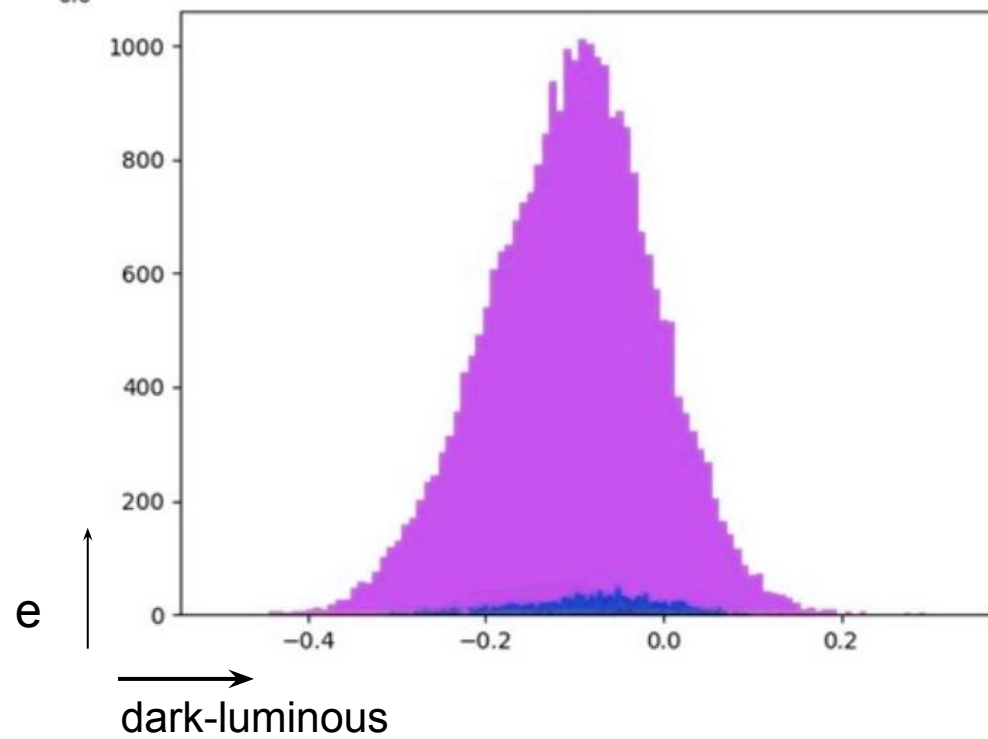
Использовали p - value < 0.0001

FASTTEXT

Звук	Ось	FastText
r	Bad-good	Bad
d	Quiet-loud	Quiet
e	Cold - Hot	Cold
e	Dark - Luminous	Dark
ɪ	Difficult - Easy	Difficult

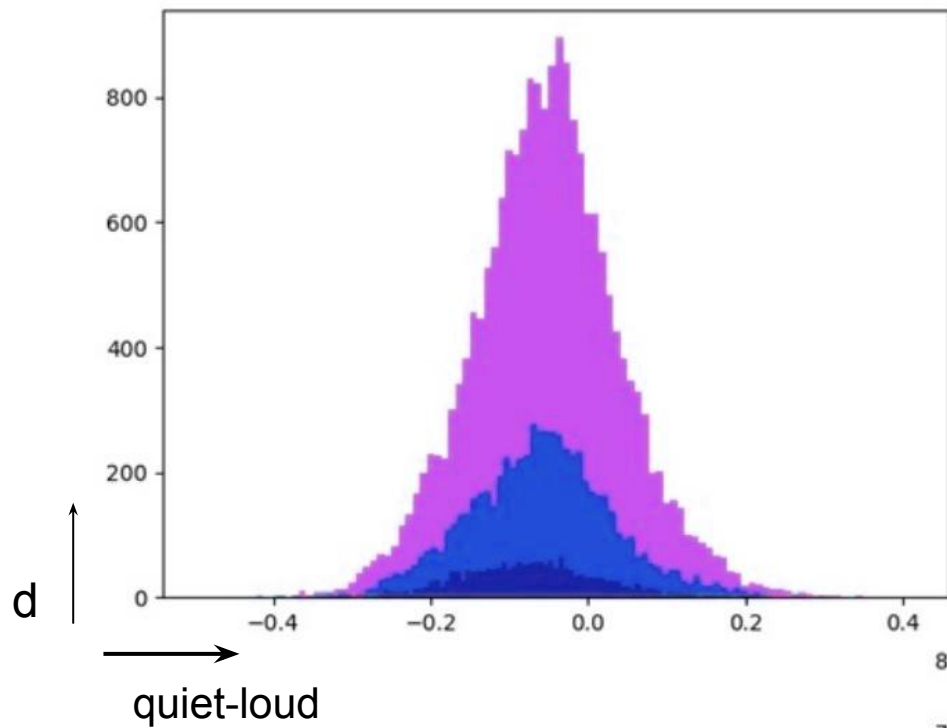


● – with
● – without

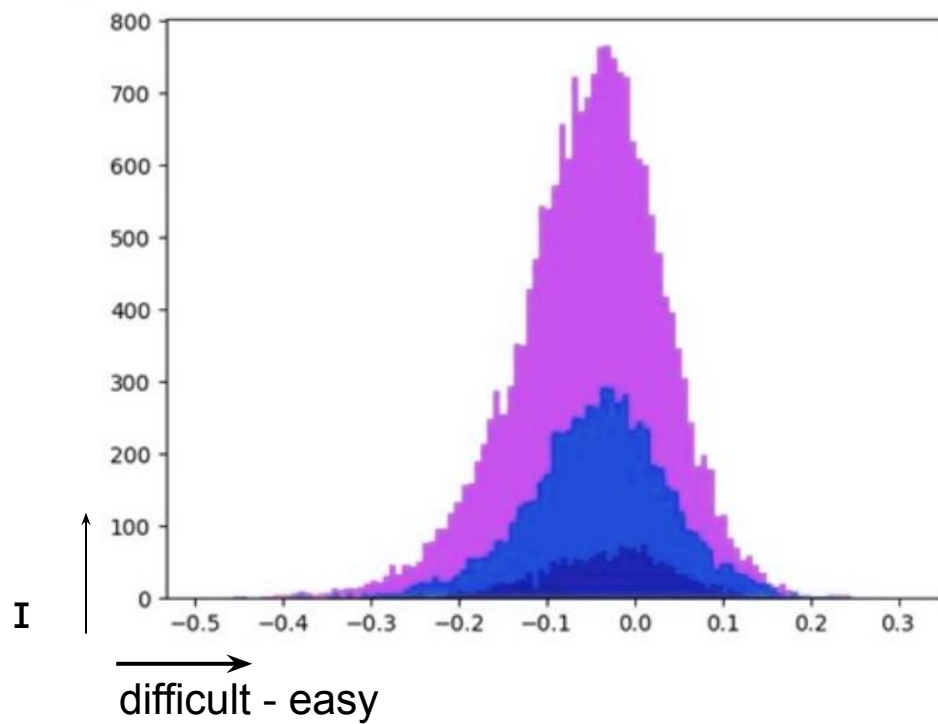


WORD2VEC

Звук	Ось	Word2vec
d	Quiet-loud	Quiet
e	Cold - Hot	Hot
e	Dark - Luminous	Luminous
I	Difficult - Easy	Difficult
I	Evil - Benign	Evil

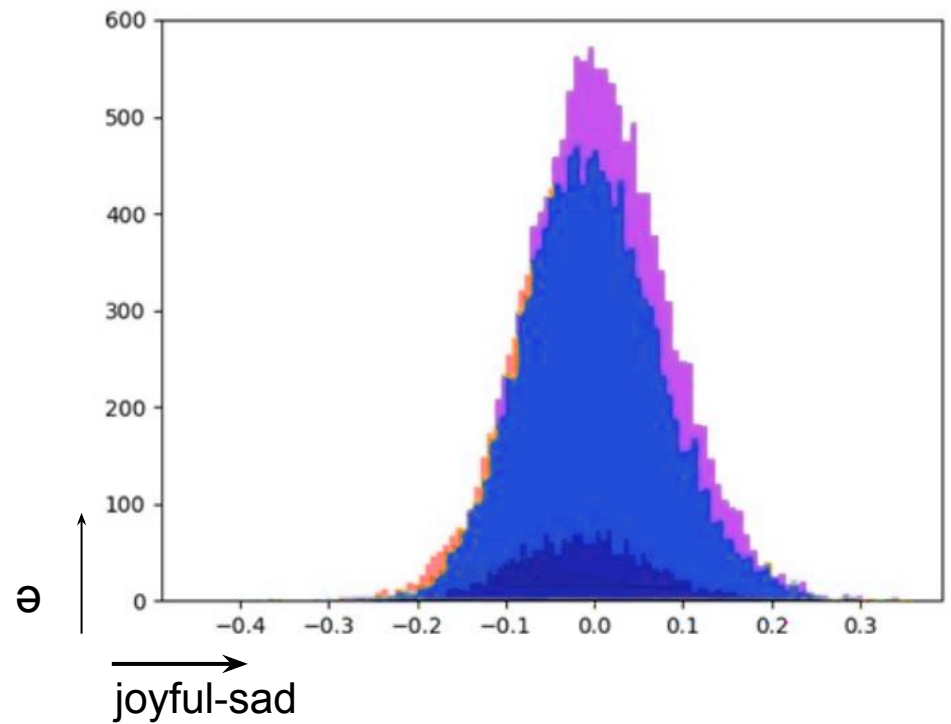
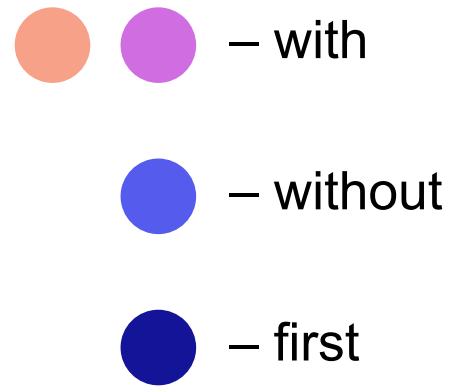
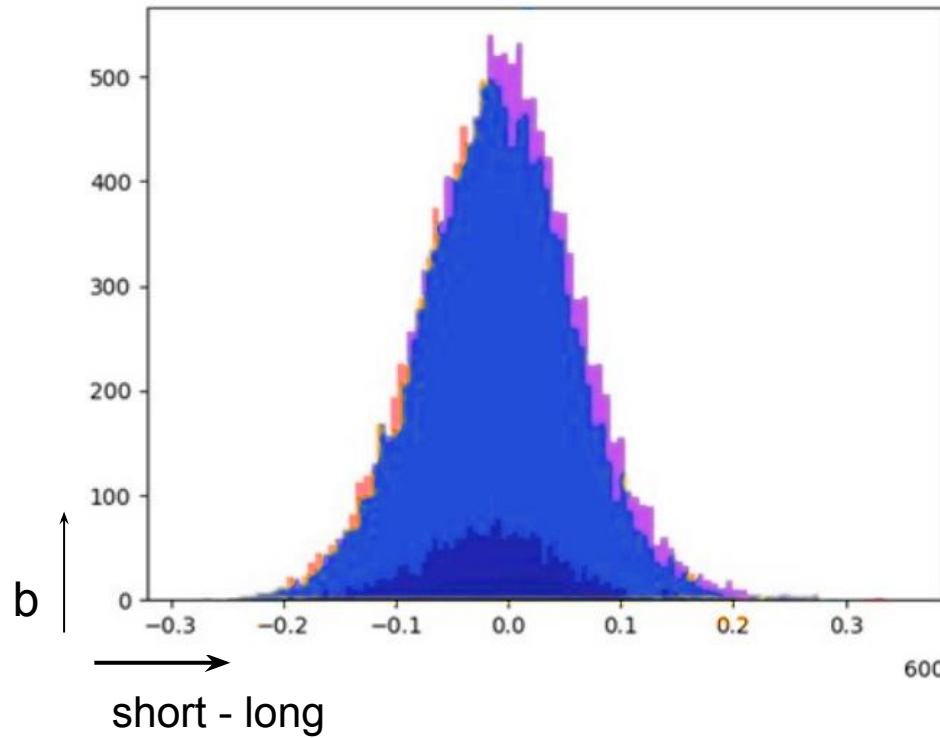


- – with
- – without
- – first



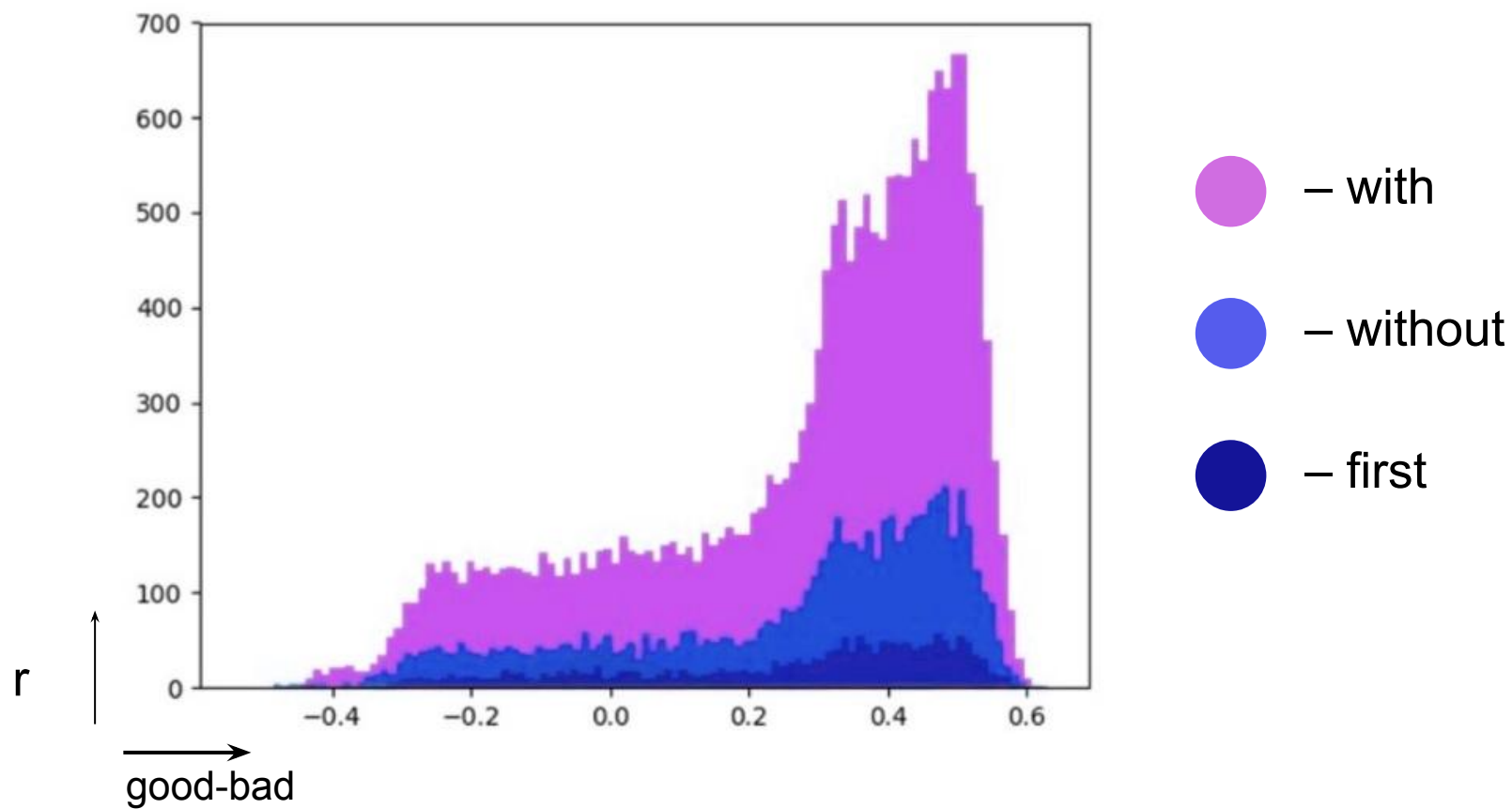
GLOVE

Звук	Ось	Glove
ɪ	Evil - Benign	Evil
æ	Complex - Simple	Complex
ə	Sad - Joyful	Joyful
ʌ	Dangerous - Secure	Dangerous
b	Short - Long	Long



BERT

Звук	Ось	Bert
г	Bad-good	Bad



Звук	Значение звука
e	Tender
k	Complex
k	Joyful
d	Complex
d	Feminine
d	Rough_or_not_smooth
d	Ugly
n	Small
ʌ	Bad
ʌ	Dangerous
ʌ	Dark
ʌ	Evil
ʌ	Feeble

Звук	Значение звука
ʌ	Masculine
ʌ	Sad
ɪ	Complex
ɪ	Dangerous
ɪ	Difficult
ɪ	Evil
ɪ	Feeble
ɪ	Slow
ə	Tender
ə	Joyful
ɑ:	Brave
ɑ:	Secure
r	Difficult

ВЫВОД

Мы подтвердили гипотезу о существовании систематической связи между звуком и смыслом.

Различные методы Word Embedding могут отлавливать одинаковые закономерности.

Для данной задачи подходят такие методы, как Word2vec, FastText и Glove.

Метод Bert не подходит для этой задачи.