


# Анализ аргументации в отзывах на кинофильмы

Виктория Борисова  
Инна Рабинович  
Анна Сербина



# Предыдущие исследования

# Англоязычные исследования

- **Работа с датасетом IMDb (Internet Movie Database) с помощью разных методов машинного обучения**

Наиболее популярная задача – анализ тональности. Используется tf-idf векторизация и такие методы как логистическая регрессия и метод опорных векторов.

- **Также было исследование, посвящённое связи между упоминанием фильма в соцсетях и его рейтингом на IMDb**

В рамках этой работы в частности использовался подход, принимающий во внимание лексику в записях соцсетей.

# Кинопоиск

- Для Кинопоиска был создан датасет **blinoff/kinopoisk**

Он использовался в рамках Российского семинара по Оценке Методов Информационного Поиска для дополнительного обучения модели и проверки гипотез.

Был применён лексический метод и методы машинного обучения.

Результат: у каждого подхода свои плюсы и методы, поэтому наилучших результатов можно достичь их совокупностью

# Литература

- Blinov, P. D., Klekovkina, M. V., Kotelnikov, E. V., & Pestov, O. A. (2013). Research of lexical approach and machine learning methods for sentiment analysis. Computational Linguistics and Intellectual Technologies, 2(12), 48-58.
- Handayani, Tri. (2023). Unmasking the Hidden Emotions: Sentiment Analysis on IMDb Movie Reviews.  
URL:  
<https://medium.com/@mbaktrihandayani/unmasking-the-hidden-emotions-sentiment-analysis-on-imdb-movie-reviews-9573779c304c>
- Oghina, Andrei & Breuss, Mathias & Tsagkias, Manos & Rijke, Maarten. (2012). Predicting IMDB Movie Ratings Using Social Media. 503-507. 10.1007/978-3-642-28997-2\_51.
- Talibzade, Rustam. (2023). Sentiment Analysis of IMDb Movie Reviews Using Traditional Machine Learning Techniques and Transformers. 10.13140/RG.2.2.29464.16644.



# Анализ аргументации

# Анализ аргументации

- **Анализ аргументации (*argumentation mining*)** – это область компьютерной лингвистики, в которой исследуются методы извлечения из текстов и классификации аргументов и связей между ними, а также построения аргументационной структуры.
- **Аргумент = утверждение (*claim*) + довод (*premise*)**
- Утверждение выражает позицию («за» или «против»)
- Довод – доказательства / мотивация для подтверждения позиции
- Пример: RuArg-2022 – задача анализа аргументации относительно COVID-19



# Датасет для соревнования RuArg-2022



# Структура датасета

<b>train</b>	<b>validation</b>	<b>test</b>	<b><math>\Sigma</math></b>
6717 (70%)	1431 (15%)	1402 (15%)	<b>9550 (100%)</b>

# Структура датасета (train)

- text\_id
- text
- masks\_stance
- masks\_argument
- quarantine\_stance
- quarantine\_argument
- vaccines\_stance
- vaccines\_argument

В 3 текстах есть мнение по трем темам, в 88 – по маскам и карантину, в 55 – по маскам и вакцинам, в 31 – по карантину и вакцинам  
В остальных 6540 текстах есть мнение только по одной теме

## Соотношение классов (маски)

позиция / довод	за (2)	нет аргумента (1)	против (0)	Σ
за (2)	275 (9%)	428 (13%)	1 (0%)	704 (22%)
другое (1)	61 (2%)	1713 (55%)	58 (2%)	1832 (59%)
против (0)	3 (0%)	310 (10%)	281 (9%)	594 (19%)
Σ	339 (11%)	2451 (78%)	340 (11%)	3130 (100%)

## Соотношение классов (карантин)

позиция / довод	за (2)	нет аргумента (1)	против (0)	Σ
за (2)	180 (8%)	405 (20%)	2 (0%)	<b>587 (28%)</b>
другое (1)	37 (2%)	1247 (59%)	57 (3%)	<b>1341 (64%)</b>
против (0)	0 (0%)	104 (5%)	68 (3%)	<b>172 (8%)</b>
Σ	<b>217 (10%)</b>	<b>1756 (84%)</b>	<b>127 (6%)</b>	<b>2100 (100%)</b>

## Соотношение классов (вакцины)

позиция / довод	за (2)	нет аргумента (1)	против (0)	Σ
за (2)	125 (8%)	249 (15%)	0 (0%)	374 (23%)
другое (1)	21 (1%)	791 (48%)	54 (3%)	866 (52%)
против (0)	3 (0%)	198 (12%)	217 (13%)	418 (25%)
Σ	149 (9%)	1238 (75%)	271 (16%)	1658 (100%)



# Датасет отзывов на кинофильмы

# Структура blinoff/kinopoisk (Hugging Face)

- part: top250 или bottom100
- movie\_name – название фильма
- review\_id – id рецензии
- author – автор рецензии
- date – дата написания рецензии
- title – название рецензии
- grade3 – отрицательная / положительная / нейтральная
- grade10 – оценка из 10
- content – текст рецензии

Содержит 36591 рецензию на фильм с сайта «Кинопоиск» (июль 2004 - ноябрь 2012),  
разделения на на train, validation и test нет  
Всего описано 350 фильмов: топ-250 лучших и топ-100 худших по оценкам сайта

<https://huggingface.co/datasets/blinoff/kinopoisk>



Кинопоиск  
24 сентября 2011

## Плакали наши денежки (с)

'Блеф' - одна из моих самых любимых комедий.

Этот фильм я наверно смотрел раз сто, нет я конечно блефую, я видел его куда больше. Не могу не выразить своё восхищение главными действующими лицами этого фильма. Начну с Адриано Челентано для которого как я считаю это лучшая роль в кино. Великолепный актёр, неплохой певец, странно что на его родине в Италии его песни мало кто слушает. Ну я думаю что и итальянцы и французы привыкли к тому, что у нас до сих пор актёры популярней чем даже на своей родине. Да, такой вот парадокс. Челентано конечно профессионал своего дела, комик с серьёзным выражением лица. Он смешон ещё и потому, что одновременно так серьёзен. Адриано браво!...

Dataset Viewer Auto-converted to Parquet API Embed Data Studio

Split (1)  
train · 36.6k rows

Search this dataset

part	movie_name	review_id	author	date
string · classes	string · classes	string · lengths	string · lengths	timestamp
2 values	350 values	1	5	34 e
top250	Блеф (1976)	17144	Come Back	2011-04-24T00:00:00
top250	Блеф (1976)	17139	Stasiki	2008-04-24T00:00:00
top250	Блеф (1976)	17137	Flashman	2007-04-24T00:00:00
top250	Блеф (1976)	17135	Sergio Tishin	2005-17T00:00:00
top250	Блеф (1976)	17151	Фюльгья	2005-20T00:00:00
top250	Блеф (1976)	17142	Marvel	2005-06T00:00:00

< Previous 1 2 3 ... 366 Next >

Dataset Viewer Auto-converted to Parquet API Embed Data Studio

Split (1)  
train · 36.6k rows

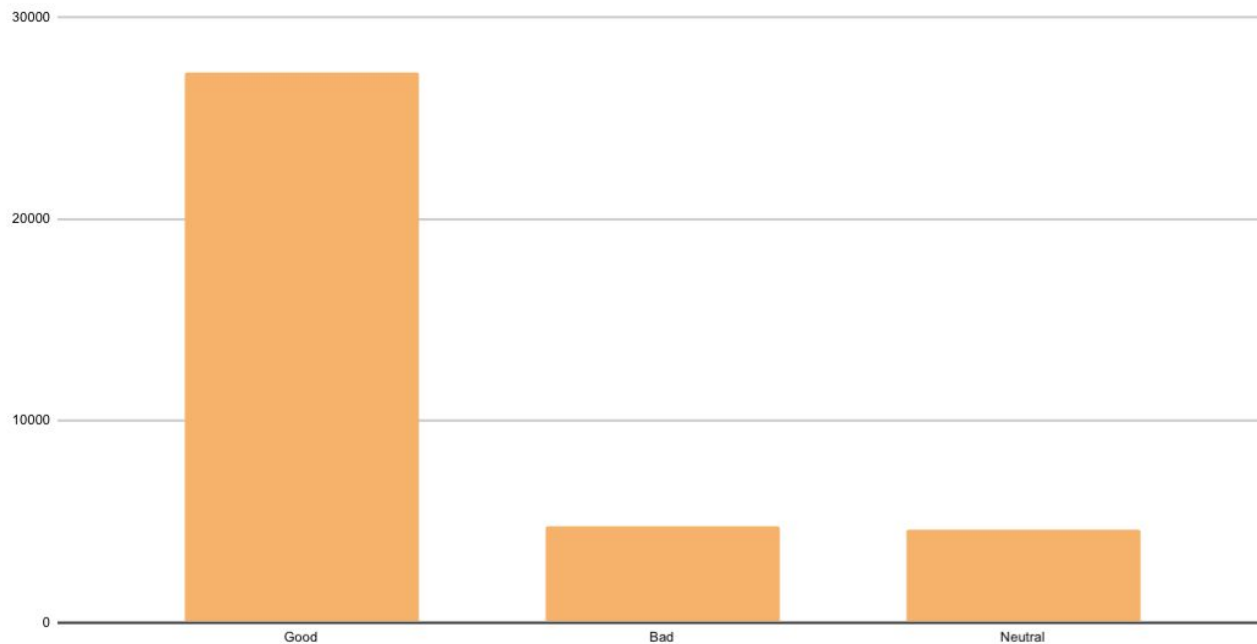
Search this dataset

title	grade3	grade10	content
string · lengths	string · classes	string · classes	string · lengths
1	3 values	69 values	47
Плакали наши денежки	Good	10	"Блеф" — одна из моих самых любимых комедий. Этот фильм
⓪	Good	0	Адриано Челентано продолжает радовать нас своими работ
⓪	Good	10	Несомненно, это один из великих фильмов 80-х...
" Черное, красное, ерунда это все..."	Good	0	Эта фраза на мой взгляд отражает сюжет несомненно
«Он хотел убежать? Да! Блеф, блеф...»	Neutral	7	- как пела Земфира, скорее всего, по совершенно друг
⓪	Good	0	Бесспорный классический шедевр, который и через 2

< Previous 1 2 3 ... 366 Next >



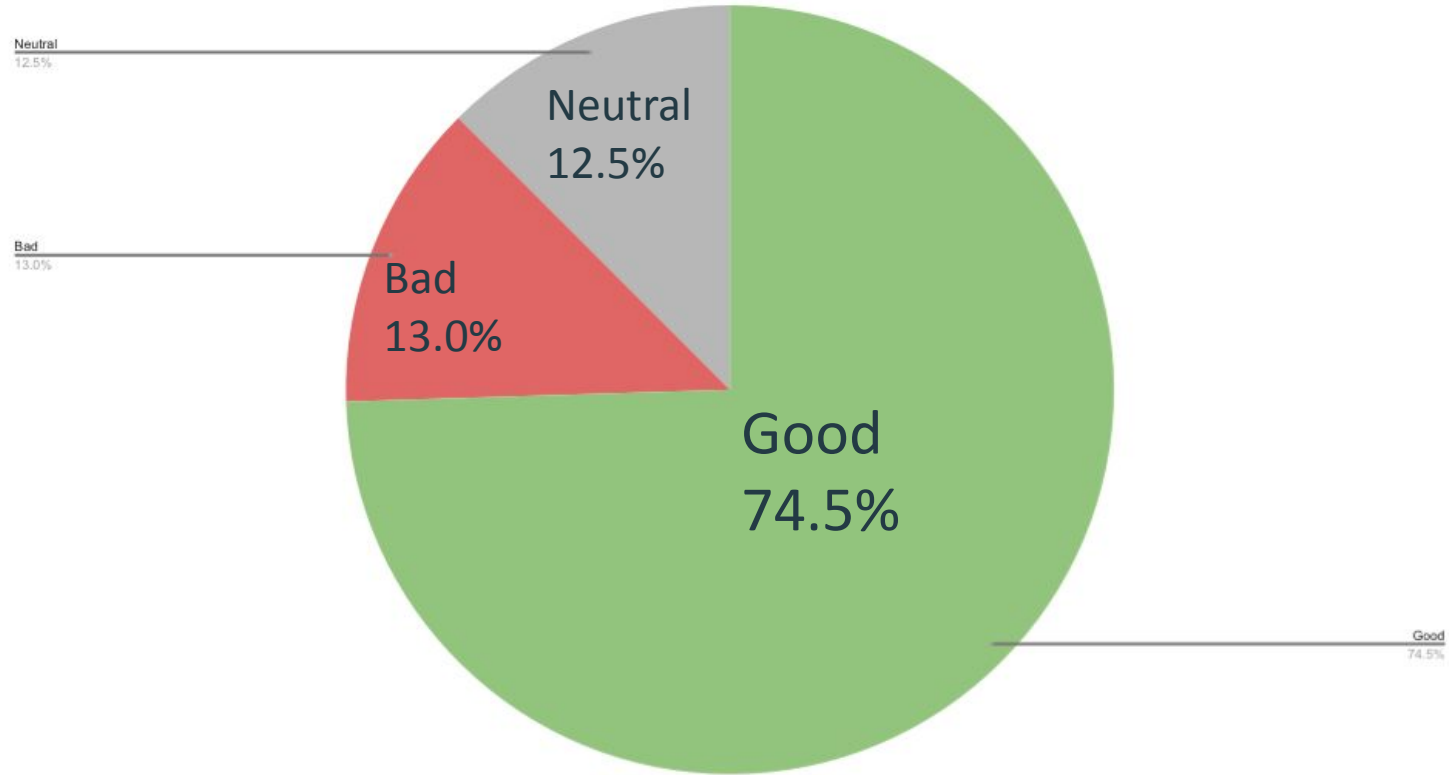
# Соотношение классов



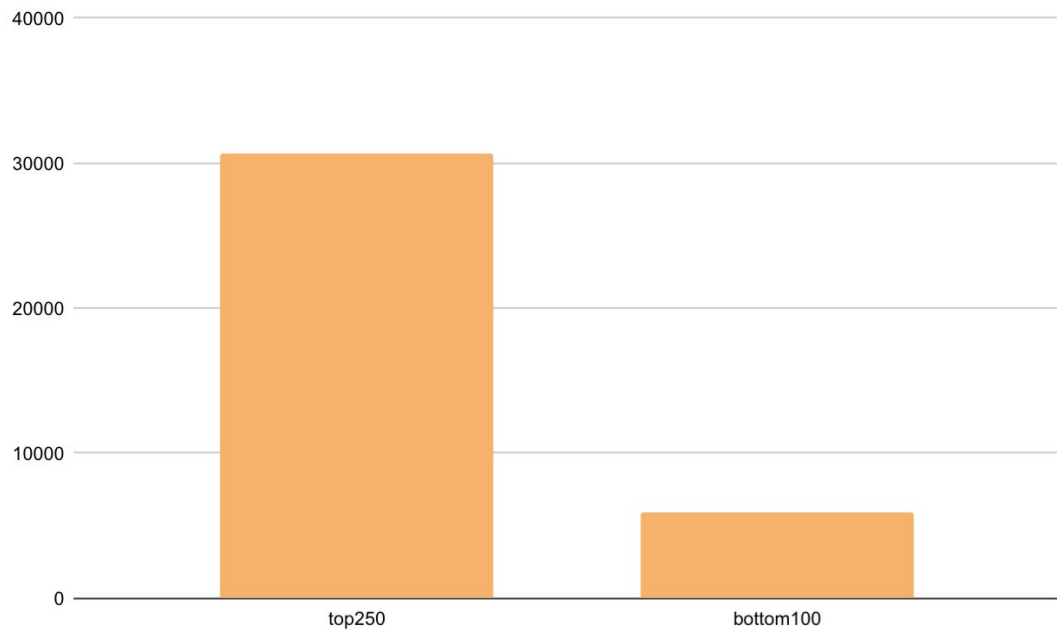
good: 27264

bad: 4751

neutral: 4576



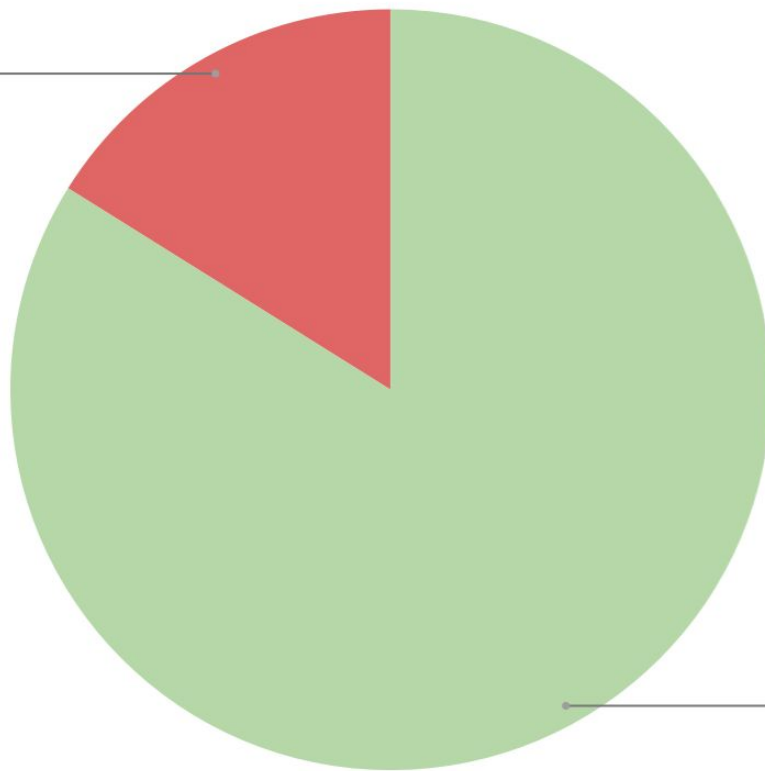
# Соотношение классов



top250: 30695

bottom100: 5896

bottom100  
16.1%



top250  
83.9%



Текущее исследование

# Гипотезы

- Позиция и аргументация будут лучше предсказываться для фильмов из вершины рейтинга
- Конгруэнтные позиции и аргументации могут предсказываться чаще на месте неконгруэнтных, а наоборот – реже
- Позиция и аргументация будут лучше предсказываться для современных фильмов
- *Позиция и аргументация будут лучше предсказываться для русскоязычных фильмов\**

*\* пока мы не оценили сбалансированность таких классов*

# Задачи и метод

## Задачи:

1. Определить подходящий размер датасета и соотношение классов в нем
2. Осуществить ручную разметку доводов
3. Провести дообучение модели на обучающей выборке
4. Оценить результаты работы модели на валидационной и тестовой выборках
5. Провести анализ результатов в соответствии с поставленными гипотезами

## Метод:

Дообучение энкодерной модели

# Какой датасет нам нужен?

- **Примерный объем: train – 2296 (70%), validation – 492 (15%), test – 492 (15%)**

В тренировочном датасете RuArg-2022 в среднем 2296 текстов на тему

- **Примерное соотношение позиций: good – 787 (24%), neutral – 1902 (58%), bad – 591 (18%)**

Совпадает со средним соотношением по трем темам в RuArg-2022

- **Примерное соотношение доводов: как получится!**



# Какой датасет нам нужен?

- **Как выбрать несколько тысяч предложений из большого датасета?**

Можно выбрать  $n$  самых коротких рецензий, что сблизит нас с датасетом RuArg-2022 (например, в blinoff/kinopoisk 3242 рецензии, чья длина  $< 94$  слов), учесть оценку от 1 до 10, учесть разнообразие фильмов (старые / новые, отечественные / зарубежные), выбрать рандомно

- **Как разметить?**

Воспользоваться инструкцией для разметчиков доводов с RuArg-2022 и набраться терпения :)

# Какая модель нам нужна?

- Так как мы планируем использовать метод дообучения энкодерной модели и работать с русскоязычными текстами, кажется неплохим решением воспользоваться моделью **DeepPavlov/rubert-base-cased**, с помощью которой было создано базовое решение организаторов RuArg-2022, а затем моделью **ai-forever/ru-Roberta-large**, учитывая опыт исследований, связанных с IMDb

Спасибо за внимание!