Федеральное государственное автономное образовательное
учреждение высшего образования

Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук

Основная образовательная программа
Прикладная математика и информатика

ПРОЕКТ ПО КУРСУ АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВ

«Clickbait Challenge at SemEval 2023 - Clickbait Spoiling»

Выполнили студенты:
Соколов Ян, группа 194, 4 курс,
Екимов Егор, группа 194, 4 курс
Токкожин Арсен, группа 194, 4 курс
Гвасалия Лукас, группа 194, 4 курс

МОСКВА 2022

СОДЕРЖАНИЕ

1 Краткое описание задачи

1.1 Задача

Посты-кликбейты ссылаются на веб-страницы и рекламируют их содеражание, вызывая у пользователя любопытство, а не предоставляя информативные резюме. Основная задача заключается в классификации типа спойлера, который предупреждает пост-кликбейт.

1.2 Данные

Датасет содержит посты кликбейты, а также предобрабоанные документы, на которыее ссылается данный кликбейт. Также спойлеры делятся на три типа: спойлеры коротких фраз, спойлеры более длинных отрывков и спойлеры нескольких непоследовательных фрагментов текст („phrase“„passage“„multi“).

Всего у нас имеется 3200 постов для обучения и 800 для валидации, а также 1000 постов на которых в конечном итоге будет проверен наш классификатор. Данные представляют собой следующий JSON формат:

- uuid: uuid поста.
- postText: текст поста кликбейта. targetParagraphs: основное содержание связанной с постом страницы, основной текст выделен вручную.
- targetTitle: название связанной с постом страницы.
- targetUrl: url страницы на которую ссылается кликбейт-пост.
- humanSpoiler: сгенерированный человеком спойлер (абстракция) для поста-кликбейта со страницы, на которую дана ссылка.
- spoiler: извлеченный человеком спойлер для кликабельного поста со связанной веб-страницы.
- spoilerPositions: позиция спойлера в тексте, извлеченного человеком для кликабельного поста со связанной веб-страницы.
- tags: тип спойлера (может быть "фраза "отрывок"или "мульти"), который должен быть классифицирован.
- некоторые поля содержат дополнительную метаинформацию о записи, но не используются: postId, postPlatform, targetDescription, targetKeywords, targetMedia.

Ниже приведен пример одного из постов в нашем датасете.


```
{
  "uuid": "0af11f6b-c889-4520-9372-66ba25cb7657",
  "postText": ["Wes Welker Wanted Dinner With Tom Brady, But Patriots QB Had Better Idea"],
  "targetParagraphs": [
    "It’ll be just like old times this weekend for Tom Brady and Wes Welker.",
    "Welker revealed Friday morning on a Miami radio station that he contacted Brady because he’ll be in town for Sunday’s game between the New England Patriots and Miami Dolphins at Gillette Stadium. It seemed like a perfect opportunity for the two to catch up.",
    "But Brady’s definition of \"catching up\" involves far more than just a meal. In fact, it involves some literal \"catching\" as the Patriots quarterback looks to stay sharp during his four-game Deflategate suspension.",
    "\"I hit him up to do dinner Saturday night. He’s like, ‘I’m going to be flying in from Ann Arbor later (after the Michigan-Colorado football game), but how about that morning we go throw?’ \" Welker said on WQAM, per The Boston Globe. \"And I’m just sitting there, I’m like, ‘I was just thinking about dinner, but yeah, sure. I’ll get over there early and we can throw a little bit.’ \"",
    "Welker was one of Brady’s favorite targets for six seasons from 2007 to 2012. It’s understandable him and Brady want to meet with both being in the same area. But Brady typically is all business during football season. Welker probably should have known what he was getting into when reaching out to his buddy.",
    "\"That’s the only thing we really have planned,\" Welker said of his upcoming workout with Brady. \"It’s just funny. I’m sitting there trying to have dinner. ‘Hey, get your ass up here and let’s go throw.’ I’m like, ‘Aw jeez, man.’ He’s going to have me running like 2-minute drills in his backyard or something.\"",
    "Maybe Brady will put a good word in for Welker down in Foxboro if the former Patriots wide receiver impresses him enough."
  ],
  "targetTitle": "Wes Welker Wanted Dinner With Tom Brady, But Patriots QB Had A Better Idea",
  "targetUrl": "http://nesn.com/2016/09/wes-welker-wanted-dinner-with-tom-brady-but-patriots-qb-had-better-idea/",
  "spoiler": ["how about that morning we go throw?"],
  "spoilerPositions": [[[3, 151], [3, 186]]],
  "tags": ["passage"]
}
```


1.3 Анализ данных

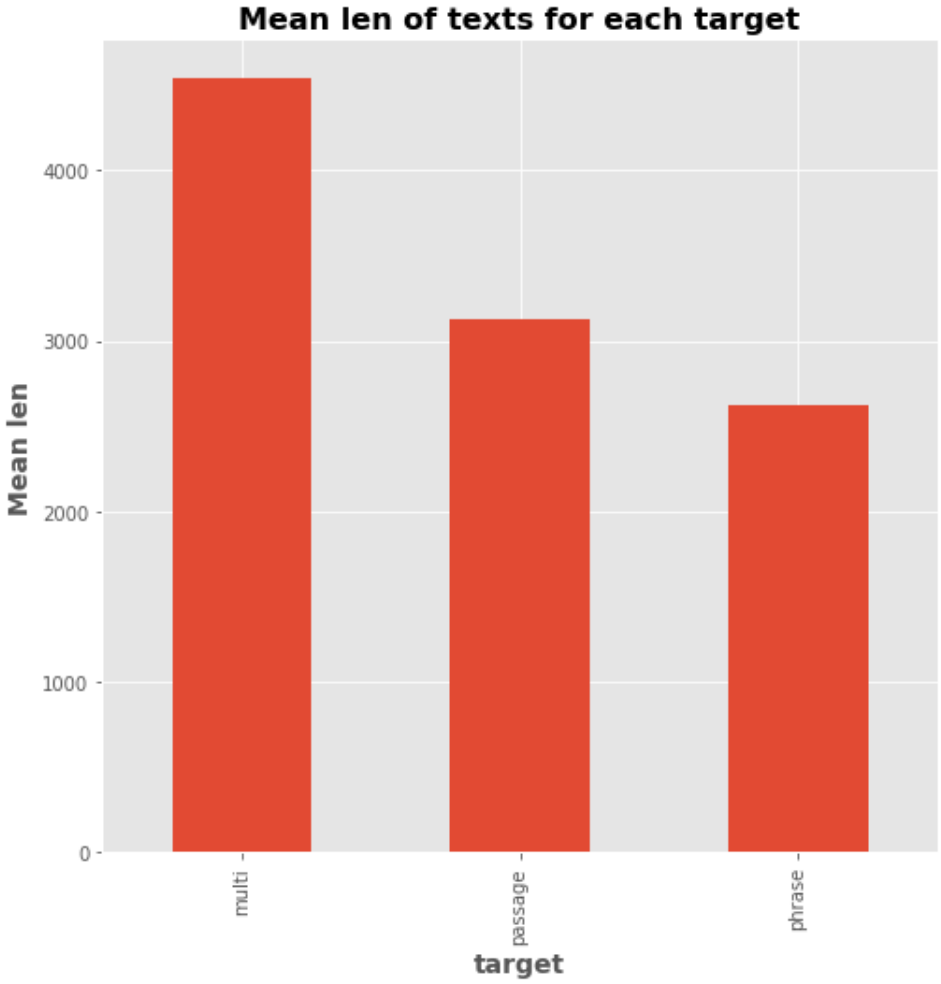
В приложенной к этому отчету тетрадке мы провели анализ предоставленных данных, заметили, что количество классов не сбалансировано:

- phrase -> 1367
- passage -> 1274
- multi -> 559

Рассмотрели топ 5 часто встречающихся слов в каждом из классов:

target		text_tokenized
multi	's	2542
	one	1403
	n't	1205
	people	1177
	like	1071
passage	's	4477
	one	1986
	said	1871
	inc.	1854
	like	1656
phrase	's	4587
	one	1920
	said	1915
	n't	1533
	like	1412

Построили график средней длинны текста в зависимости от класса его спойлера:



Средняя длина текста ментше в фразах и больше в multi, что вполне закономерно. Опечаток в текстах постов, а также примеры некорректной разметки мы не нашли.

2 Ход работы

2.1 Baselines

В соревновании организаторами был предоставлен код для получения двух baseline-ов:

- Baseline1 - предсказние одного класса "passage"
- Baseline2 - предсказания, полученные с помощью fine-tuned трансформера

В качестве метрики для оценивания качества мы решили использовать F-меру.

f-мера для baseline1 = 0.2

f-мера для baseline2 = 0.74

2.2 Fasttext

В качестве начальной модели мы решил попробовать fasttext при этом предварительно предобработали данные: применили токенизацию, учли при токенизации стоп слова, определенные знаки препинания, а также пунктуацию. Также мы осуществили подбор гиперпараметра который отвечат за количество эпох, которое будет обучаться наш классификатор. Итоговый скор на f-мере получился 0,47. Таким образом нам удалось побить первый baseline.

2.3 Краткий обзор литературы

Мы рассмотрели следующую статью. В ней авторы предлагают рассмотреть несколько подходов к решению этой задачи: с помощью трансформера или с помощью классических feature-based моделей. В качестве трансформеров можно взять BERT, DeBERTa, RoBERTa. А если рассматривать классические модели, то тут скорее речь идет о логистической регрессии и методе опорных векторов.