**ИТМО**

**Факультет Програмной Инженерии и Компьютерных Технологий**

**Лабораторная работа № 1**

“Применение инструментов для исследования и аннотации данных для обучения моделей ИИ”

Выполнили: Распопин Даниил Игоревич, Сыюань Хуан, Р4155

Преподаватель: Старобыховская Анастасия Александровна

Санкт-Петербург

2024

СОДЕРЖАНИЕ

[Задание 3](#__RefHeading___Toc589_3144844761)

[Основные этапы 3](#__RefHeading___Toc591_3144844761)

[Выводы 5](#__RefHeading___Toc593_3144844761)

# Задание

В работе необходимо выбрать набор данных для анализа, написать техническое задание для разметки, выполнить разметку данных, сверить полученную разметку с реальной. По итогам работы описать основные сложности процесса разметки и валидации разметки, проблемы, возникшие при применении инструментов разметки.

# Основные этапы

На начальном этапе работы группа распределила задачи следующим образом:

* Сыюань Хуан – выбор данных, подготовка тестового сабсета, определения требований к аннотации, написание ТЗ, валидация полученной разметки.
* Даниил Распопин – подбор инструмента разметки, разметка тестового сабсета, передача готовой разметки для валидации.

На первом этапе необходимо было загрузить датасет. Выбранный и загруженный датасет доступен по ссылке: https://www.kaggle.com/c/fake-news/data

На втором этапе необходимо было провести анализ данных. Для сбора статистики использовался язык программирования Python, библиотека Pandas.

Набор данных представляет из себя тексты и заголовки новостей. Меткой является признак “label” с двумя значениями – “fake” и “real”. Таким образом, датасет используется в задаче бинарной классификации фейковых новостей.

В полном наборе данных содержится 2095 строк. Из них ~61.7% приходится на фейковые новости. В тестовый сабсет для разметчика вошло 100 первых строк этого датасета. В тестовом сабсете 67% новостных статей были фейковыми.

На третьем этапе работы необходимо было сформировать ТЗ для разметчика. Итоговое ТЗ имело следующую формулировку: “Необходимо выяснить явялется ли новость фейком. Некоторые из экземпляров содержат ссылки, по которым можно найти информацию. Если невозможно проверить является ли новость фейком, в качестве метки оставить пропуск.”

На четвертом этапе необходимо было сформировать требования к инструментам разметки и добавить их в техническое задание. Итоговое ТЗ, переданное разметчику, имело следующую формулировку: “Необходимо выяснить явялется ли новость фейком. Некоторые из экземпляров содержат ссылки, по которым можно найти информацию. Если невозможно проверить является ли новость фейком, в качестве метки оставить пропуск. Размечать можно при помощи любых бесплатных инструментов и сервисов.”

На пятом этапе разметчику необходимо было определиться с инструментами аннотации. В связи с тем, что в аннотационном сабсете содержалось малое число образцов для аннотации, было решено размечать вручную в LibreOffice. В качестве инструментов для разметки использовался браузер с поисковым движком DuckDuckGo.

На шестом этапе разметчику необходимо было выполнить разметку. Выбранный метод разметки характеризуется следующим образом: “в поисковую строку браузера последоватльно вставлялись фрагменты текста новости произвольной длины. Если хотя бы для одного из фрагментов было вхождение в выдаче, новость признавалась реальной.” Проведение разметки сабсета из 100 новостей заняло порядка 4.5 часов реального времени. Наиболее трудозатратными операциями являлись:

* Копирование текста из таблицы в браузер
* Анализ поисковой выдачи на предмет вхождений
* Оценка достоверности поисковой выдачи и проставление метки

Кроме этого, на шестом этапе разметчику требовалось оценить время разметки всего датасета из 2095 строк. Оценка разметчика: “С математической точки зрения, если на 100 строк уходит около 4.5 часов, на 1000 строк уйдет 45 часов, на 2000 – 90 часов. В реальности же необходимо сделать допуск на невозможность концентрировать внимание в течение длительного времени, перерывы на обед, непредсказуемые технические трудности и форс-мажорные обстоятельства. Предполагаю, что на разметку всего датасета уйдет 1.5-2 рабочие недели.”

На седьмом этапе разметчику необходимо было передать размеченный сабсет заказчику. Сабсет был передан.

На восьмом этапе разметчику необходимо было провести валидацию полученной от разметчика разметки. Оценка проводилась с применением языка программирования Python и библиотеки Pandas. Точность разметки составила 53%.

На девятом этапе заказчику и разметчику необходимо было дать оценку проведенной работе, проблемным местам, способам повышения качества разметки.

Анализ проведенной разметки заказчиком.

“Точность разметки на сабсете была равна 72%. Думаю, объяснить это можно несколькими причинами. Во-первых, новости из тестового сабсета были написаны в 2016 году. Возможно, некоторые сайты уже не работают, новости могли быть удалены, поисковые механизмы могли понизить новостные ресурсы в выдаче. Чем новость старее, тем тяжелее ее искать. Во-вторых, нужно было более тщательно определить понятие “фейк”. То, что новость содержится в интернете еще не означает, что она не фейк. Даже новость, содержащаяся на нескольких сайтах сразу, не обязательно будет правдой. Проблема в том, что если для каждой новости требовать глубокую проверку фактов, время на разметку возрастет на порядок. Возможно, в 10 раз.”

Анализ работы разметчиком:

“Помимо того, что механическая работа отнимает много времени и быстро утомляет, часто бывает не понятно, как интерпретировать выдачу браузера по запросу. Например, от всего запроса вхождение было только по половине, трети. Можно ли это считать фейком? Может быть текст является неправильным / измененным цитированием первоисточника? Кроме этого текст новости часто не влезает в поисковую строку целиком и его приходится разбивать на части. Если из всех частей вхождения были только по половине? Или по трети? Неясно, как действовать в такой ситуации. Наконец, понятие “невозможно удостовериться” было для меня размыто. Если я нашел части новости в разных источниках, но не нашел одного относительно полного текста, это можно считать невозможностью удостовериться? Было много техничеких вопросов, вопросов проведения границ между фейком, правдой и невозможностью понять. Думаю, что в ТЗ было бы хорошо видеть четкое описание пограничных случаев, много примеров, и прочих вспомогательных подробностей.”.

# Выводы

При формировании ТЗ необходимо подробно описывать задачу разметки. Для корректного формирования ТЗ важно начать с анализа данных, чтобы получить представление об их виде, характере возможных проблем. При описании задачи разметки важно учесть моменты, в которых между разметчиком или разметчиками возможны разногласия. Порядок разрешения спорных моментов необходимо четко описать. Если это возможно, в процессе разметки необходимо регулярно получать обратную связь от разметчиков на предмет обнаруженных ими спорных ситуаций, трудностей. В случае необходимости процесс разметки нужно остановить и переопределить ТЗ с учетом полученных сведений.