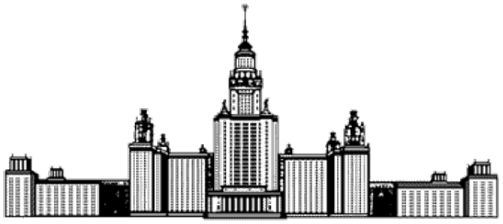

Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Воробьев Сергей Юрьевич

Модели выявления манипуляций и их мишеней в новостных сообщениях
Models for detection of manipulations and their targets in news articles

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:
д. ф.-м. н.
Воронцов Константин Вячеславович

Москва, 2023

Содержание

Аннотация

В моей работе описан комплексный подход к выявлению манипулятивных фрагментов в новостных статьях. Данные представляют собой собранный набор новостных статей, которые были размечены экспертами в области лингвистики, социологии и политических наук. Цель работы — решить проблему субъективности разметки и несогласованности между разметчиками при помощи новой техники сопоставления разметки, основанной на нескольких критериях, а также рассмотреть нейросетевые модели, которые можно использовать для решения задачи. Также в работе представлена новая метрика для измерения относительного качества алгоритма по отношению к людям-экспертам.

1 Введение

В современном мире информация стала основой общественного мнения и влияния. С развитием интернета и социальных сетей информационное пространство стало более открытым и доступным для массового распространения. Однако, такая доступность породила множество проблем, связанных с манипулированием сознанием людей с целью пропаганды, манипуляции общественным мнением и дезинформации. Языковые манипуляции в новостных текстах являются одним из инструментов, которые используются для достижения этих целей[6].

Задача моего исследования заключается в разработке метода для выявления языковых манипуляций в новостных текстах и определении мишеней, на которые направлены эти манипуляции. Мишенью может быть, например, определенный человек, политическая партия или социальная группа. В качестве моделей для разметки текста в задаче выявления языковых манипуляций и определения мишеней могут использоваться языковые нейросетевые архитектуры. Благодаря их способности анализировать и обрабатывать большие объемы текстовых данных, они хорошо подходят для таких задач[7].

Данные для обучения модели на задаче выявления языковых манипуляций и определения мишеней были получены с помощью разметки экспертами в области лингвистики. Эти специалисты обладают глубокими знаниями о языковых структурах, семантике, прагматике и риторических приемах, что позволяет им точно идентифицировать манипулятивные стратегии, используемые в текстах. В процессе аннотирования текстов лингвисты опреде-

ляли наличие и тип языковых манипуляций, а также указывали мишени, на которые эти манипуляции направлены. Такой подход к получению обучающих данных позволяет создать качественный и достоверный набор примеров, на котором модель сможет обучаться и адаптироваться к задаче выявления манипуляций и определения мишеней. Благодаря экспертному анализу текстов, модель будет способна обнаруживать сложные и тонкие манипуляции, которые могут быть незаметны для неподготовленного человека или простых алгоритмов обработки текста.

Одной из основных проблем в разметке данных для обучения модели в задаче выявления языковых манипуляций и определения мишеней является субъективность экспертов при аннотировании текстов. Данный фактор может привести к тому, что один и тот же текст может быть размечен по-разному разными экспертами, а в некоторых случаях их разметки могут даже противоречить друг другу. Такая проблема может снизить надежность обучающих данных и, как следствие, повлиять на качество и эффективность обученной модели. В связи с этим, важным аспектом данной работы является разработка методов и подходов, позволяющих минимизировать влияние субъективности экспертов при разметке данных, а также учесть возможные расхождения в их оценках.

2 Обзор литературы

Обнаружение пропаганды — проблема, аналогичная нашей. Это задача на понимание естественного языка, которая активно решается с 2019 года, по ней проводится множество конкурсов. Задача поиска пропаганды обсуждалась в [5]. В этой статье авторы предложили выявлять пропаганду на разных уровнях – уровне предложений и отдельных токенов. Авторы использовали набор новостных материалов, размеченных по 18 классам пропагандистских приемов. Однако, в работе не использовалось понятие «мишени».

Поставленная задача состоит из двух подзадач — выявление фрагментов манипуляции в тексте и нахождение мишеней этих манипуляций. Нахождение мишеней представляет из себя классическую задачу NER (Named entity recognition) [4]. В датасете, использующемся в моей работе, мишени уже размечены, нужно только сопоставить каждому фрагменту правильную мишень. Подзадача выявления фрагментов является Token Classification задачей. Мы можем использовать дообученную модель BERT[1] для этой задачи. Вторая

подзадача является задачей Relation Extraction, наиболее обещающим подходом является использование билинейных слоев и BERT для этой задачи [2].

В моей задаче есть проблемы с согласованием разметок на одном и том же тексте. Одним из способов решить эту проблему является консолидация экспертов и подготовка финальной разметки. Этот способ описан в [5]. В своей работе, однако, я рассматриваю алгоритмический способ объединения и агрегации разметок для одного и того же текста.

3 Постановка задачи

Целью данной работы является разработка метода для детектирования структур, связанных с языковыми манипуляциями, в новостных текстах. В частности, рассматриваются тройки объектов (**фрагмент**, **класс**, **мишень**):

1. Фрагмент манипулятивного текста: участок текста, содержащий манипулятивные приемы или выражения, направленные на воздействие на мнение читателя.
2. Класс манипуляции: категория манипулятивного приема, использованного в данном фрагменте текста (например, утверждение непроверенных фактов, апелляция к эмоциям, упрощение сложных вопросов и т.д.).
3. Мишень: персона, группа людей или организация, на которую направлена манипуляция, с целью формирования определенного образа или отношения к ним.

Основная задача состоит в разработке и обучении модели, которая сможет автоматически детектировать указанные структуры в новостных текстах. Это включает следующие этапы:

1. Идентификация фрагментов манипулятивного текста в исходном новостном материале.
2. Классификация манипулятивных приемов, используемых в обнаруженных фрагментах текста.
3. Определение мишеней, на которых направлены манипулятивные приемы.

Список возможных мишеней для конкретного текста уже присутствует в данных. Моя задача на 3 этапе состоит в нахождении связей между выделенными фрагментами и мишенями. В качестве модели для выявления таких фрагментов на 1 и 2 этапе можно

применять BERT, обученный на русскоязычном корпусе текстов, например ruBert от SBER[1].

В секции 5 я рассмотрю метрики, которые позволяют оценить качество модели.

4 Разметка

4.1 Сбор данных

Для обучения языковых моделей необходимо разметить данные для fine-tunning. Разметка была выполнена с помощью сервиса Yandex.Toloka ("Яндекс.Толока"). Особенность нашего подхода заключается в использовании профессиональных аннотаторов – экспертов в области лингвистики, социологии и политологии. Для получения более точных результатов один текст размечался тремя разметчиками. В среднем эксперт-разметчик тратит 5 минут на разметку одного текста.

Разметка состоит из выделения фрагмента манипулятивного текста, определения его класса (типа манипуляции) и соединения этого фрагмента с мишенью (целью) в тексте. Цели были размечены заранее с использованием стандартных моделей NER от DeepPavlov[2].

В разметке представлено 18 различных приемов (классов) манипуляции (Рис. 1), к которым могут относиться манипулятивные фрагменты текста. Эти 18 классов сгруппированы в 4 больших класса: негативизация, позитивизация, парологизация и деавторизация. В дальнейшем будет не важно, какой лингвистический смысл скрывается за каждым классом, поэтому я ограничусь описанием больших классов:

- **Негативизация** — группа манипулятивных приемов, приписывающих объекту манипуляции отрицательные качества.
- **Позитивизация** — группа манипулятивных приемов, приписывающих объекту манипуляции позитивные, приукрашенные качества.
- **Деавторизация** — группа манипулятивных приемов, основанных на замалчивании, сокрытии источника информации.

[1] https://huggingface.co/ai-forever/ruBert-base

[2] https://deeppavlov.ai

- **Паралогизация** — группа манипулятивных приемов, основанных на отклонении от формальных законов логики.

Рис. 1: Структура классов.

4.2 Инструкция для разметчиков

Чтобы разметчики выделяли нужные фрагменты и были согласованы между собой, была разработана инструкция, на которую эксперты опирались при разметке текстов. Ее ключевые пункты описаны в приложении A

4.3 Статистика по данным

Рассмотрим количественные показатели получившегося размеченного набора данных для обучения модели. В результате разметки получено 3803 разметки 1421 уникального документа (из них 1165 разметок содержали хотя бы один фрагмент манипуляции). 192 текста размечены тремя экспертами, 458 – двумя, 515 – одним. Всего было получено 5443 манипуляции, состоящие из размеченных троек (фрагмент, класс, мишень).

Если рассматривать объединенные группы приемов, то получаем следующее распределение:

- негативизация - 42,00% (2286 из всех 5443 манипуляций),
- позитивизация - 36,74% (2000 из всех 5443 манипуляций),
- деавторизация - 15,47% (842 из всех 5443 манипуляций),
- паралогизация - 5,79% (315 из 5443 манипуляций).

У разных манипулятивных приемов бывает очень разная длина текстового фрагмента — это является особенностью задачи. Ниже на Рис. 2 представлены распределения длин по классам.

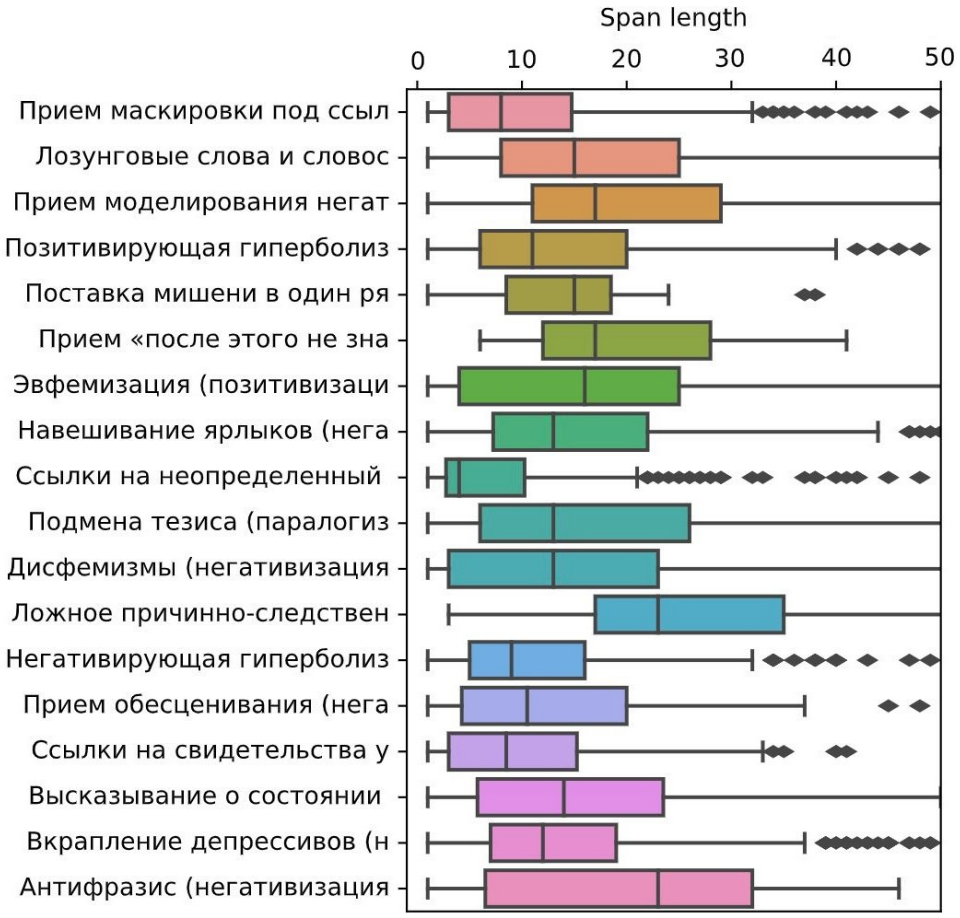

Рис. 2: Длины фрагментов.

4.4 Самые частые мишени в данных

Среди самых частых мишеней присутствовали следующие: россия, украина, рф, москва, сша, казахстан, китай, нато.

4.5 Комбинирование разметок

Одной из основных проблем является согласование нескольких разметок одного текста для более устойчивого процесса обучения и лучших результатов. Такая проблема возникает из-за субъективности манипуляций и особенностей разметки. Когда у 2 или более разметок различные начальные и конечные позиции манипулятивных фрагментов, мишень манипуляции и категория манипуляции, становится неясно, как их согласовать между собой и на чем в итоге обучать модель.

Чтобы преодолеть эту проблему, можно использовать **Markup Matching Loss** (MML)1, разработанный для пар фрагментов разметки, направленный на выбор наиболее оптимального соответствия. Более того, этот лосс применим не только для манипуляции, но и для всех задач Token Classification с метками фрагментов, а также для задач без меток фрагментов – Различия будут просто в установке необходимых констант в нули.

Давайте определим некоторые переменные

- $X||Y$ - одиночные разметки, которые является набором размеченных манипулятивных «троек».
- $x_i \in X||y_i \in Y$ - манипулятивная единица в одной разметке.
- T_i - мишень манипуляции в i «тройке»
- C_i - класс манипуляции в i «тройке»
- $L(x, y)$ - функция потерь между парой найденных манипулятивных «троек»

Основная концепция заключается в определении оптимального соответствия для каждой тройки из одной разметки в другой разметке. Есть опция оставить ее непарной для минимизации совокупных потерь для всех троек.

$$\forall x \in X \quad \exists y \in \{Y \cap \emptyset\} : \sum_{x} L(x, y) = \inf_{x,y} \{\sum_{x} L(x, y)\}$$

Есть 3 параметра, значения которых входят в итоговый показатель согласованности

- Размер пересечения
- Равенство мишеней
- Равенство классов

При разработке функции потерь предполагалось, что пересечение является наиболее критическим параметром. Таким образом, мы устанавливаем правило, что когда между единицами нет пересечения, предпочтительнее не сопоставлять их друг с другом.

В конечном итоге, определим функцию потерь между парами манипулятивных троек. Минимизируя эту функцию потерь для каждой пары разметок, мы получим наиболее релевантное соответствие для манипулятивных троек.

$$L(x,y) = \mathbb{1}\{J(x,y)=0\} + \mathbb{1}\{J(x,y)>0\}\{-J(x,y) - const_T \mathbb{1}\{T_x = T_y\} - const_C \mathbb{1}\{C_x = C_y\}\} \to \min$$

$$\text{MML} = \inf_{x\in X, y\in Y}\{\sum_{x,y} L(x,y)\} \tag{1}$$

Где $J(x,y) = \frac{|x \cap y|}{|x \cup y|} \in [0,1]$ это метрика Жаккара, $const_C$ и $const_T$ - константы, подбираемые по сетке. В моих эксперимeтах я установил их по `0.1`.

Этот подход использовался для объединения разметки для устойчивого обучения моделей. Детали построения набора данных приведены в Приложении Б.

5 Метрики

5.1 Классические метрики для текстов

Для оценки качества модели на задаче детектирования языковых манипуляций в текстах и определения их мишеней, можно применять классические метрики, используемые в анализе текста. Однако, эти метрики могут не всегда точно отражать результаты работы модели из-за особенностей разметки и разнообразия текстов. Рассмотрим те, которыми будем пользоваться для оценки качества:

F1-мера: гармоническое среднее между точностью (precision) и полнотой (recall). F1-мера является популярной метрикой для оценки качества классификации и детектирования объектов, но она может быть чувствительна к различиям в разметке между разными разметчиками и не учитывать специфику манипулятивных текстов.

Расстояние Жаккара: метрика сходства между множествами, определяемая как отношение мощности пересечения множеств к мощности их объединения. Расстояние Жаккара может быть использовано для сравнения разметки текста разными разметчиками, однако оно не учитывает порядок слов и структуру текста, что важно при анализе манипуляций.

5.2 Несогласованность экспертов

Основная проблема с применением классических метрик в данной задаче заключается в том, что разметчики могут размечать тексты по-разному. Так, одни и те же манипулятивные приемы и мишени могут быть идентифицированы в разных частях текста или с разной степенью точности. Это может привести к низким показателям качества модели при использовании классических метрик, даже если модель корректно выявляет языковые манипуляции и их мишени. Для примера давайте посмотрим на Рис. 3 F1-меры, измеренной между экспертами на одних и тех же текстах в задаче классификации токенов. Видно, что большая часть разметок очень плохо согласуются между собой по F1-мере.

Говоря более конкретно про субъективность разметки, было обнаружено, что только 58% манипулятивных единиц в разных версиях разметки имели ненулевое пересечение. Более того, среди всех сопоставленных пар с ненулевым пересечением, только 44% мишеней и 46% классов совпадали. Эти результаты свидетельствуют о том, что менее половины сопоставленных разметок имели совпадающие классы и мишени, что подчеркивает значительную степень субъективности.

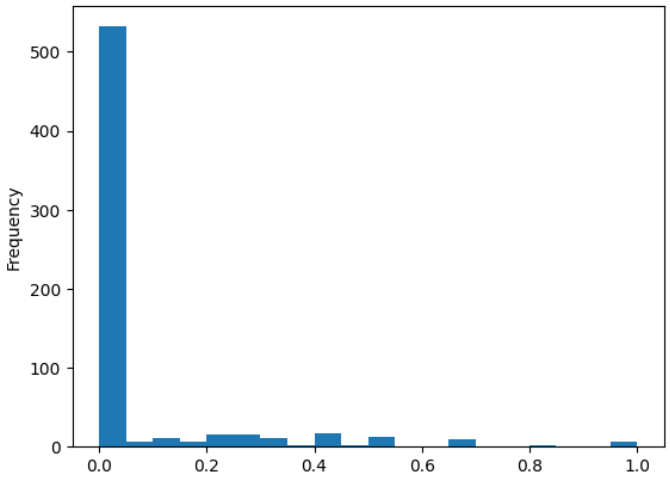

Рис. 3: Согласованность разметок по F1.

5.3 Относительные метрики

Перед нами встает следующая проблема: если эксперты противоречат между собой, то как мы можем ожидать, что наш алгоритм будет не противоречить всем экспертным разметкам одновременно? Чтобы решить проблемы, связанные с противоречивостью

экспертных разметок, рассмотрим новую метрику, учитывающую субъективность и сложность задачи. Основная идея такой метрики заключается в измерении относительного качества алгоритмов по отношению к экспертам-разметчикам. В частности, определяется критерий $M(X, Y) \in \mathbb{R}$, который измеряет согласованность между двумя разметками для данного текста. В качестве $M(X, Y)$ может выступать как MML1, определенный выше, так и обычный F1-score. Сравнивая согласованность алгоритмов с объединенной разметкой и согласованность экспертов-разметчиков между собой, мы можем получить более точную оценку эффективности моделей для выявления манипулятивных фрагментов.

Далее определим:

- Среднюю точность алгоритма $MAA = \frac{1}{N}\sum_{n=1}^{N} M(\hat{Y}, GT^*)$, где $\hat{Y}$ - выход алгоритма, а GT^* - эталонная разметка, полученная из сопоставления разметок (секция 4.5).
- Среднюю точность человека $MHA = \frac{1}{N}\sum_{n=1}^{N} M(X^*, Y^*)$, где X^* и Y^* - две разметки, оптимизирующие MML1.
- Относительную точность $RA = \frac{MAA}{MHA} * 100\%$, которая показывает качество алгоритма с учетом согласованности экспертов для данного текста.

Следует отметить, что при оценке относительной точности, когда эксперты не согласны друг с другом и метрика для их разметок равна нулю, это указывает на то, что уровень согласованности не может быть определен. Такие тексты не берутся для валидационного набора данных.

Я считаю, что этот подход измерения относительного качества между алгоритмами и экспертами-разметчиками является наиболее подходящим для оценки моделей по выявлению манипулятивных фрагментов, учитывая высокую сложность и субъективность данной задачи. Более того, предложенный критерий M позволяет проводить множество возможных экспериментов и оценок, которые могут быть адаптированы к другим исследовательским задачам и данным.

6 Модели

6.1 Общий подход

Наша задача разметки состоит из двух подзадач: во-первых, мы должны найти в тексте фрагменты, содержащие манипуляции (Tagging); во-вторых, мы должны связать

эти фрагменты с соответствующими мишенями в тексте (Relation Extraction). Информация о возможных мишенях уже есть в нашей разметке. Для обеих задач существует множество готовых решений[4][1][2], моя цель — адаптировать эти решения к имеющимся данным и справиться с субъективностью разметок.

Схема моего решения представлена на Рис. 4.

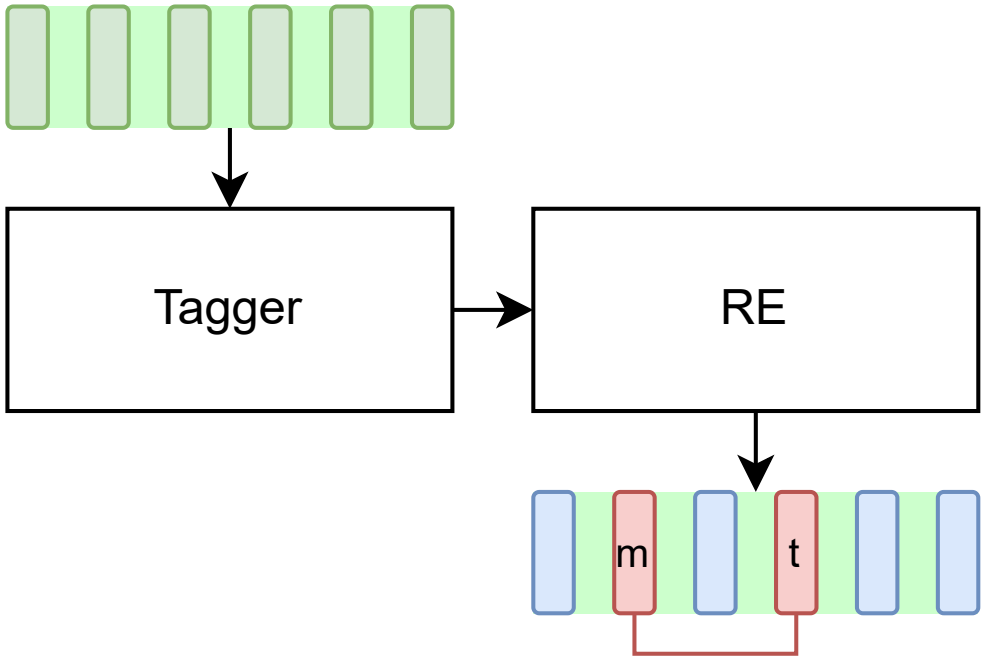

Рис. 4: Архитектура двухэтапной модели. Слева на схеме находится теггер, который отвечает за выделение фрагментов манипуляций, а справа — модуль, отвечающий за выделение связей этих фрагментов с мишенями в тексте.

6.2 Модуль для выявления манипуляций

Задача нашего теггера — каждому слову (токену) сопоставить класс, который соответствует классу манипуляции, или указать что манипуляции нет. Более формально: пусть у нас есть текст, который мы представляем как последовательность токенов $\text{text}_i = [t_1, \dots t_{n_i}]$. Наша модель должна сопоставить каждому токену свой класс: $t_i \to c_j$. Это задача классификации токенов, для которой разработано куча готовых решений. Ниже я опишу суть базового решения. Для начала нам нужно получить из токенов контекстные вектора, для этого лучше использовать архитектуры на базе механизма внимания типа BERT [1]. Есть версии этой архитектуры для русского языка. После прохождения через

BERT, вектора подаются в линейный классифицирующий слой, который выдает вероятности того или иного класса, выбор делается в пользу класса с наибольшей вероятностью.

6.3 Модуль для нахождения мишеней

Задача извлечения связей является бинарной классификацией: либо связь между фрагментом манипуляции и его мишенью существует, либо нет. Для адаптации BERT к этой задаче, к предобученной модели добавляется классификационный слой, отвечающий за определение наличия или отсутствия связи. Входные данные для модели представляются в виде пар сущностей (манипулятивный фрагмент и мишень) вместе с их контекстом. В качестве пар выступает все возможные комбинации «фрагмент»-«мишень» для данного текста. Подробнее эта модель описана в [2].

7 Результаты

7.1 Постановка экспериментов

В предыдущем разделе я описал базовые модели для решения поставленной задачи. Также было предложено разделить задачу на две подзадачи, чтобы сделать ее проще для решения. Пока разметки алгоритма получаются хуже, чем разметки людей по предложенной метрике, поэтому я решил обучать и тестировать модели для двух подзадач отдельно, так как это приводит к лучшему пониманию процесса обучения и проблем на этих подзадачах.

7.2 Метрики

В Секции 5.3 я описал как можно сравнивать качество алгоритма с качеством людей с учетом субъективности задачи. В качестве метрики $M(\cdot, \cdot)$ можно использовать любую метрику качества для текста. В моих экспериментах я буду использовать следующие:

- RA (Related Accuracy) — средний MLL между разметкой алгоритма и объединенной разметкой по всем текстам, деленный на средний MLL между экспертами до объединения по всем текстам.

- RJac (Related Jaccard) — Среднее расстояние Жаккара для фрагментов между алгоритмической разметкой и объединенной разметкой по всем текстам, деленное на среднее расстояние Жаккара между экспертными разметками до объединения по всем текстам.
- RClass (Related Classification) — средняя точность совпадения классов между алгоритмической разметкой и объединенной разметкой по всем текстам, деленная на аналогичную метрику для экспертов.
- RelF1 (Related F1) — Аналогичная величина, только с F1 в качестве метрики.

7.3 Этап 1: Выделение фрагментов

В своей работе я более подробно остановился именно на первом этапе, для второго этапа я привел только финальные метрики и выводы по результатам. На первом этапе мы хотим, чтобы наша модель правильно выделяла фрагменты с манипуляцией и верно указывала их класс. Эту задачу можно решать как Token Classification task, а можно разбить на две подзадачи, это: 1) Выделение самого фрагмента (бинарная классификация) и 2) Классификация этого фрагмента (Sequence classification). Качество для подзадач я рассмотрю в секции 7.6 анализа ошибок, это позволит лучше разобраться почему модель ошибается;

Для начала давайте посмотрим на относительные метрики для всего этапа выделения манипулятивных фрагментов. В Таб. 1 мы видим, что результаты пока не очень хорошие, однако, многообещающие. Так, Related Accuracy имеет значение 0.440, что означает, что наша модель работает с качеством 44% от экспертного. Эта цифра кажется не очень большой, однако, если брать в расчет сложность задачи, получается многообещающе. Проверим также гипотезу о том, что большее число параметров модели способствует лучшему качеству. В Таб. 1 также есть сравнение качества, полученного при базовых моделях tiny-BERT[3] и BERT-base — в первом случае у модели существенно меньше параметров. Качество позволяет нам сделать вывод о том, что большие языковые модели учатся лучше.

Метрика	BERT-base	tiny-BERT
RA	0.440	0.283
RJac	0.578	0.363
Rclass	0.669	0.608
RelF1	0.678	0.491

Таблица 1: Значения относительных метрик качества.

7.3.1 Объединение разметок

Я попробовал два подхода к обучению — с объединением разметок (Приложение 8 и без объединения. Результаты на отложенной выборке представлены в Таб. 2. Очевидным выводом здесь является то, что чистка данных и устранение сильной противоречивости между экспертами помогает существенно увеличить качество.

	Объединенная	Полный датасет
RA	0.440	0.0374
RJac	0.578	0.0388
RClass	0.669	0.0731
RF1	0.678	0.1267

Таблица 2: Сравнение метрик моделей, обученных на объединенных с помощью MML тройках и на всем датасете. Очевидно, что стратегия объединения разметок позволяет улучшить качество.

7.4 Этап 2: Соединение фрагментов с мишенями

На втором этапе мы хотим, чтобы наша модель верно соединяла манипулятивные фрагменты с мишенями. Для упрощения отладки и исследования я подавал на вход верно размеченные фрагменты, чтобы избежать суммирования ошибки с первым этапом. В задаче соединения фрагментов с мишенями качество получилось следующим: `0.179 precision, 0.243 recall, 0.206 f1`. Это довольно низкое качество. Прямой анализ разметки позволяет предположить, что такое плохое качество могло получиться из-за плохо размеченных мишеней (секция 7.6.2). Также причиной может стать объективная

сложность задачи, так как определение мишеней является нетривиальной задачей даже для экспертов.

7.5 Переобучение моделей

Одной и проблем, возникающих в задачах с малым количеством данных является переобучение. На Рис. 5 представлен график функции потерь для модели выявления манипуляций.

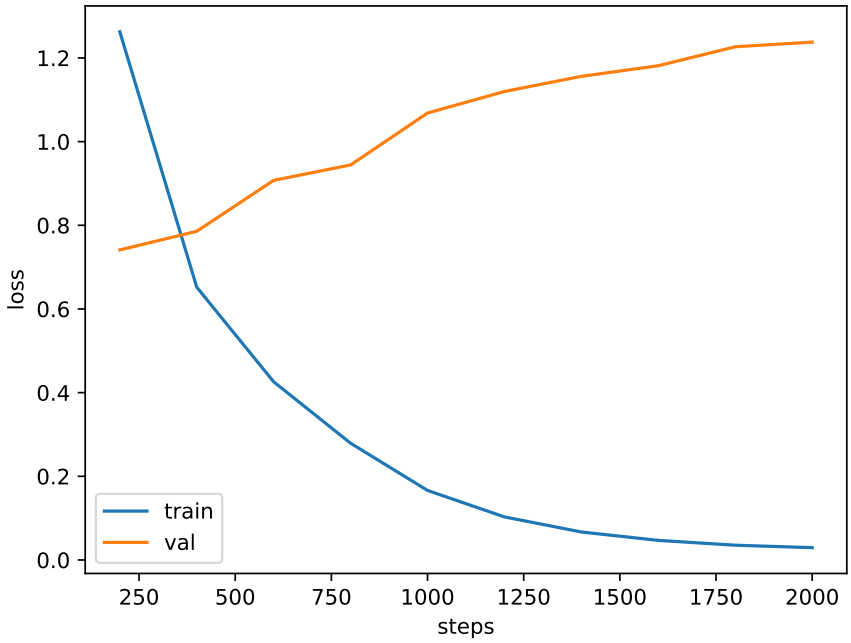

Рис. 5: График функции потерь. Видно, что наша модель склонна к переобучению как и все языковые модели на малом наборе данных.

Можем заметить, что лосс на обучении падает, однако, на валидации он незначительно растет. Давайте посмотрим на график качества на Рис. 6. Можем заметить, что, несмотря на увеличение лосса на валидации, качество растет. Можем сделать следующий вывод: обучение модели заключается в заучивании паттернов, которые она наблюдает в тренировочной выборке. Таким образом, важную роль в итоговом качестве играет размер датасета – увеличивая количество и разнообразие примеров с манипуляцией мы обучаем модель лучше распознавать эти паттерны на новых данных.

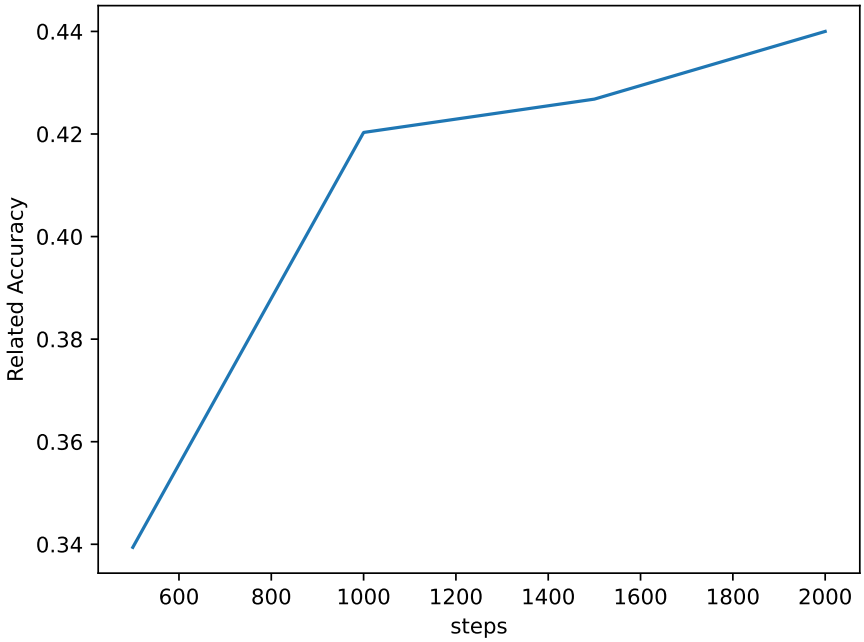

Рис. 6: График качества. Видно, что несмотря на переобучение, относительная метрика качества на валидации растет.

7.6 Анализ ошибок

7.6.1 1 этап

Попробуем разобраться, почему качество алгоритма хуже качества экспертной разметки. Разберем подробнее первый этап (Token Classification). Здесь уже будем пользоваться классическими метриками качества. Вспомним, что f1-score между экспертами очень низкий (Рис. 3). Для анализа разобьем наш первый этап на две подзадачи: выделение фрагметов с манипуляцией и классификация фрагмента. Измерим качество на этих подзадачах, в таблице 3 я посчитал классические метрики для текстов. Здесь в качестве правильной разметки была объединенная разметка. Можно заметить, что с выделением манипулятивных фрагментов (бинарная классификация) наша модель справляется сильно хуже чем с классификацией этих фрагментов на 18 классов.

Давайте взглянем на матрицу ошибок второй подзадачи и попробуем разобраться, почему иногда происходит неверная классификация. На Рис. 7 показана такая матрица ошибок.

	F1 score
Детектирование фрагментов	0.084
Классификация фрагментов	0.687

Таблица 3: F1 на двух подэтапах 1 этапа. Можно сделать выводы, что больше всего страдает задача выделения фрагментов. Она является более сложной из-за субъективности разметок и большого количества экспертов (3 на текст).

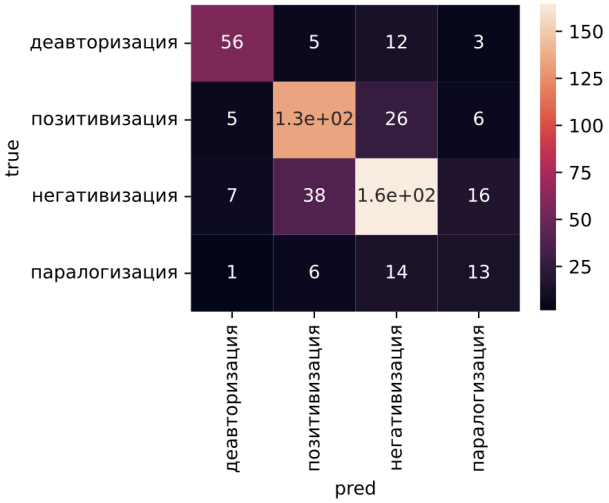

Рис. 7: Матрица ошибок классификации фрагментов на агрегированных классах.

Обратим внимание на то, что некоторые классы представлены очень незначительным количеством примеров в тестовой выборке. Давайте посмотрим на пару типичных ошибок. В Таб 4 представлено несколько примеров ошибочной классификации. Если мы внимательно изучим примеры, то заметим общую черту. Чаще всего модель обращает внимание на определенные слова и фразы, которые она запомнила на тренировочных данных и делает выводы основываясь на этом. Например, во втором примере модель могла посмотреть на слова «мародерство», «преступления» и поставить класс «негативизация». Аналогичная ситуация и с другими примерами, где модель определила «негативизацию». В последнем примере скорее всего модель предсказала «паралогизацию», поскольку в тексте нет указания источника информации и модель неоднократно встречалась с такой конструкцией в тренировочных данных.

Текст фрагмента	**Верный класс**	**Предсказанный класс**
Москва видит замыслы Запада спустить переговоры по гарантиям безопасности в некие абстрактные дискуссии, ждет взрослой реакции со стороны США	паралогизация	негативизация
Правоохранители задержали в Алма-Ате более 2 тыс. человек за участие в незаконных акциях, мародерство и другие преступления.	позитивизация	негативизация
был инфицирован омикрон-штаммом коронавирусной инфекции, но не заболел	паралогизация	негативизация
Набиуллина назвала кардинальным изменением трендов ситуацию с ростом инфляции в мире. Рост инфляции в мире является кардинальным изменением трендов	негативизация	позитивизация
Так, 73 млн клиентов «Сбера» пользуются мобильным приложением «Сбербанк Онлайн», в то время как общее число розничных клиентов банка составляет 104 млн человек.	позитивизация	паралогизация

Таблица 4: Примеры ошибок классификации.

7.6.2 2 этап

На втором этапе мы получаем довольно низкое качество из-за объетивной сложности задачи и из-за проблемы в данных. Так, были выделены следущие мишени: "ЧТЕНИЕ ПРОЕКТА О "СЕРТИФИКАТАХ ИЗ-ЗА "сдумы р "союз

7.7 Выводы

На основе имеющихся графиков и таблиц можем сделать выводы об обучении моделей на нашу задачу.

1. Чистка и агрегация данных существенно помогает улучшить качество в нашей задаче.
2. В задачах с маленьким набором данных неизбежно возникает переобучение. Важно смотреть на реальное качество модели на отложенной выборке а не на функционал потерь во время обучения.
3. Модель прежде всего обучается запоминанием манипулятивных конструкций тренировочного датасета. Это означает, что существенного улучшения качества можно достичь в том числе увеличением датасета.

8 Заключение

В данной работе мы рассмотрели сложную и актуальную задачу выявления манипулятивных фрагментов и их мишеней в текстах. Мы обсудили теоретические основы этой проблемы и рассмотрели существующие подходы и методы для ее решения.

В ходе работы были предложены новые методы оценки качества моделей на основе относительной точности (RA) и функции потерь для сопоставления разметок (MML). Эти методы позволяют преодолеть сложности, связанные с субъективностью задачи и способствуют более точной оценке эффективности разработанных моделей.

Мы провели исследование и анализ различных моделей, в основе которых лежат архитектуры, такие как BERT, для извлечения связей и классификации манипулятивных фрагментов. Эти модели были обучены и протестированы на специально подготовленном наборе данных с манипулятивными текстами.

В результате проведенных экспериментов были получены следующие результаты:

1. Модели на базе BERT показали неплохую эффективность в решении задачи выявления манипулятивных фрагментов. В задаче соединения с мишенями все еще есть проблемы в данных.
2. Разделение исходной задачи на две подзадачи позволило упростить обучение моделей и получить лучшее понимание проблем, возникающих на каждом этапе.
3. Предложенные методы оценки качества моделей, такие как RA и MML, позволили получить более точные результаты и обеспечить адекватное понимание способностей моделей.

В заключение, предложенные в данной работе методы и модели способствуют решению актуальной и сложной задачи выявления манипуляции в текстах. Однако, существует потенциал для дальнейших улучшений моделей и методов, включая улучшение разметки, разработку новых архитектур и улучшение алгоритмов оценки качества.

Список литературы

[1] Jacob Devlin и др. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». В: *arXiv preprint arXiv:1810.04805* (2018).

[2] John Giorgi и др. *End-to-end Named Entity Recognition and Relation Extraction using Pre-trained Language Models*. 2019. arXiv: 1912.13415 [cs.CL].

[3] Xiaoqi Jiao и др. *TinyBERT: Distilling BERT for Natural Language Understanding*. 2020. arXiv: 1909.10351 [cs.CL].

[4] Guillaume Lample и др. «Neural Architectures for Named Entity Recognition». В: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, июнь 2016, с. 260—270. DOI: 10.18653/v1/N16-1030. URL: https://aclanthology.org/N16-1030.

[5] Giovanni Da San Martino и др. «Fine-Grained Analysis of Propaganda in News Articles». В: *In Proceedings of the 2019 Con- ference on Empirical Methods in Natural Language Processing and the 9th International Joint Confer- ence on Natural Language Processing, EMNLP- IJCNLP 2019, Hong Kong, China.* (2019).

[6] Estela Saquete и др. «Fighting post-truth using natural language processing: A review and open challenges». В: *Expert Systems with Applications* 141 (2020), с. 112943. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2019.112943. URL: https://www.sciencedirect.com/science/article/pii/S095741741930661X.

[7] Thomas Wolf и др. «Transformers: State-of-the-Art Natural Language Processing». В: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, окт. 2020, с. 38—45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

Приложение А. Инструкция для разметчиков

Примерная инструкция для разметчиков содержит следующие пункты:

1. Внимательно изучите инструкцию и приведенные примеры.
2. При разметке текстов строго следуйте инструкции. Не добавляйте ничего, чего нет в инструкции.
3. Не отмечайте одной техникой манипуляции близкие по смыслу фрагменты, которые тем не менее я оговорены в инструкции, так как такие интуитивные знания невозможно встроить в модель.
4. Сначала ознакомьтесь с полным текстом, а затем определите техники манипуляции. Например, в антироссийском тексте вряд ли появятся положительные техники манипуляции в отношении России. Скорее всего, это будет «Навешивание ярлыков».
5. Размечающему необходимо выбирать полный фрагмент с манипуляцией. Если техника используется многократно, но семантически эти предложения не связаны по смыслу, необходимо сделать несколько разметок. (Пример: В Москве замечен неизвестный человек с оружием, и за год в Москве было зафиксировано X случаев вооруженных нападений).
6. В случае текста, на котором не работает инструкция, размечающий должен пропустить текст или не размечать фрагмент.

Приложение Б. Построение набора данных

Полученное сопоставление манипулятивных троек позволяет создавать наборы данных разными способами. Может использоваться несколько политик выбора троек для окончательной версии набора данных, которые перечислены ниже:

- Включение несопоставленных: эта политика определяет, следует ли включать в итоговый набор данных манипулятивные тройки, которые были сопоставлены с пустым множеством.
- Взятие длинных: эта политика определяет, следует ли включать в полученный набор данных длинные или короткие последовательности из сопоставленной пары.
- Политика класса: эта политика определяет, какой класс присвоить результирующей тройке, когда они различны.
- Политика мишени: эта политика определяет, какую мишень присвоить результирующей тройке, когда они различны.

В данной работе я применил стратегию включения несопоставленных пар и выбора более длинных последовательностей для решения проблемы ограниченного размера набора данных и увеличения доступного объема данных. Кроме того, для политик класса и цели я решил случайным образом присваивать их, чтобы обеспечить устойчивость и учесть предположение об одинаковой «экспертности» среди разметчиков.