Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

Направление подготовки / специальность: 03.03.01 Прикладные математика и физика
Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и математическое моделирование в экономике

ВЫЯВЛЕНИЕ МАНИПУЛЯЦИЙ В НОВОСТЯХ

(бакалаврская работа)

Студент:
Лукьяненко Иван Андреевич

_____
(подпись студента)

Научный руководитель:
Воронцов Константин Вячеславович,
д-р физ.-мат. наук

_____
(подпись научного руководителя)

Консультант *(при наличии)*:
_____
_____
(подпись консультанта)

Москва 2023

Оглавление

Аннотация

Данная работа посвящена задаче Propaganda Detection. Задача выявления пропаганды в новостях на английском языке упоминалась впервые в 2017[1], в 2019 [2]. В работе [3] задача Propaganda Detection была поставлена как две независимые подзадачи обработки естественного языка: Span Identification и Text Classification, а также были предложены метрики качества и базовые модели. В бакалаврской работе предлагается расширение задачи обнаружения пропаганды - добавление подзадачи Span Targetting, размеченный русскоязычный корпус новостей и модели для решения задачи Propaganda Detection на русском языке.

Задача Span Targetting подразумевает под собой поиск мишени на которую направлен фрагмент манипуляции. Собранный датасет представляют собой набор новостных статей, которые были размечены экспертами в области лингвистики, социологии и политических наук. Разработанные базовые модели для решения задачи Propaganda Detection на русском языке представляют собой большие лингвистические модели на основе архитектуры трансформер дообученные на специализированные задачи обработки естественного языка Span Targetting и Span Identification.

Глава 1

Введение

Манипуляции в тексте – это скрытое психологическое воздействие на восприятие читателя к определенной цели, именуемой мишенью манипуляции. Основной целью манипулятивных фрагментов является формирование оценочного отношения к мишени, который может влиять на поведение человека. Чтобы достичь этой цели, авторы используют манипулятивные приемы на языковом и речевом уровнях, реализующие различные тактические ходы.

Задача выявления манипуляций в новостном потоке впервые была поставлена группой исследователей под руководством Preslav Nakov в 2019 [2]. Данная задача изначально решалась на уровне документов, то есть первое предложенное решение задачи Propaganda Detection можно формализовать как решение задачи классификации текстов. В этом же году была опубликована работа той же группы исследователей [3], в которой авторы детализировали поиск манипуляций, а именно было предложено решение задачи на уровне фрагментов текстов, то есть формально решалась задача сегментации текстов. Несмотря на интерес исследовательского сообщества к Propaganda Detection, ни в одной из работ не ставилась задача поиска мишени манипуляции. Мишень манипуляции позволяет детальнее исследовать пропаганду, как социальное явление.

Целью работы является построение моделей, сбор датасета, методов оценки, сравнение качества решений на английских и русских текстах, а так же расширение задачи Propaganda Detection для русскоязычных новостей.

Задача Propaganda Detection изначально представляет собой две независимые подзадачи: Span Identification и Span(Text) Classification. В первой подзадаче необходимо выделять манипулятивные фрагменты в текстах. С формальной точки зрения каждому токену предложения модель сопоставляет метку принадлежит ли этот токен к манипулятивному фрагменту или нет.

Вторая же подзадача является классической задачей NLP - классификация текста или фрагмента текста. Каждый выделенный фрагмент необходимо отнести к одному из 18 классов - техник манипуляции. Под манипулятивными техниками подразумеваются 18 выделенных техник:

1. прием «после этого не значит поэтому»,
2. вкрапление деприссивов,
3. прием маскировки под ссылку на авторитет,
4. прием моделирования негативного сценария,
5. негативирующая гиперболизация,
6. навешивание ярлыков,
7. эвфемизация,
8. лозунговые слова и словосочетания,
9. ссылки на неопределенный источник,
10. ссылки на свидетельства участников и очевидцев событий, имена и фамилии которых не называются,

11. позитивирующая гиперболизация,
12. прием обесценивания,
13. дисфемизмы,
14. высказывание о состоянии другого,
15. ложное причинно-следственное моделирование,
16. подмена тезиса,
17. антифраз,
18. поставка мишени в один ряд с негативно оцениваемым объектом.

Подзадача Span Targetting, поиск мишени выделенных фрагментов манипуляции, не представлена в основном наборе подзадач Propaganda Detection. Однако, в определении манипуляции заложено изменение восприятия к определенной цели, чем и вдохновлено включение данной задачи в методолгию исследования пропагандистких новостей.

Глава 2

Постановка задачи

В данной главе обсуждаются математические постановки задач Span Identication и Span Targetting.

2.1 Span Identification

В данной работе для решения задачи Span Identification ставится задача классификации токенов. Требуется в новостном тексте выделить все фрагменты, которые содержат манипулятивные техники. Нейросетевые алгоритмы обработки естественного языка используют токенизацию текстов. В данной работе используется токенизатор соответсвующий модели RoBERTa, BPE - токенизатор, который разбивает слова на подслова. Корпус новостей собран и размечен экспертами в области социологии и психологии. Обучающие данные представляют собой набор текстов в виде последовательности токенов и последовательности классов, к которым относится каждый токен. Необходимо выделить фрагменты в тексте относящиеся к одному из видов манипуляций. Поэтому задачу выделения фрагментов манипуляций удобно рассматривать как задачу выделения именнованных сущностей с использования BIO - кодирования. Классификация токенов происходит на 3 класса: B, I, O, начало фрагмента, токен внутри фрагмента и не относя-

щийся к фрагменту. Введем обозначения:

- $p(x^t, \theta)$ - параметрическое семейство моделей для описания распределения вероятности классов каждого токена
- последовательность $\{x_i^t\}_{i=1}^N$ - токенизированный текст $t \in T$, где T - множество всех текстов
- параметры $\theta \in \Theta$, где Θ - пространство параметров модели
- последовательность $\{y_i^t\}_{i=1}^N$ - класс i-го токена формате one-hot в тексте $t \in T$
- $L_1(y, \hat{y}) = \sum_{c \in C} y_c \log(\hat{y}_c)$, где вектор $\hat{y}$ - распределение вероятности классов, $c \in C$ - класс из множества допустимых классов

Инициализация модели происходит предобученной мультиязычной XLM-RoBERTa-base. Данная модель предобученная на 250 террабайт текстов CommonCrawl. Данная модель представитель архитектуры "кодировщик". Такие модели предобучаются на задачи Masked Language Modelling(MLM) и Next Sentence Prediction(NSP). В MLM задаче модель учится "заполнять пропуски в предложениях предсказывая маскированные токены. В NSP задаче модель учится предсказывать, является ли следующее предложение продолжением текущего по смыслу или нет. Такой подход предобучения обусловлен необходимостью в большом количестве данных для обучения лингвистических моделей с миллионами параметров(XLM-RoBERTA-base имеет 270 миллионов параметров). Нет необходимости размечать корпус для предобучения. В процессе предобучения на данные две подзадачи, модель выучивает контекстно-зависимое семантическое векторной представление токенов за счет задачи маскированного языкового моделирования и выучивает векторной представление текстов за счет задачи предсказания следующего предложения. В задачи выделения фрагментов нам важно

векторное представление отдельных токенов, т.к. адаптер-классификатор принимает вектор каждого токена и классифицирует его на 3 класса.

Итоговая оптимизационная задача поиска оптимальных параметров модели выделения манипулятивных фрагментов сформулирована как:

$$\frac{1}{|T|}\sum_{t\in T}\sum_{i\in t} L_1(y_i^t, p(x^t,\theta)_i) \to \min_{\theta\in\Theta}$$

2.2 Span Targetting

В данной работе для решения задачи Span Targeting ставится задача семантического сопоставления текстовых фрагментов. Требуется по данным двум фрагментам текстов решить задачу бинарной классификации - есть ли связь между этими фрагментами или нет.

Обучающие данные представляют из себя набор пар фрагментов манипуляции и их мишеней, а так же бинарный таргет - соотносятся ли данные фрагменты или нет.

- $p(x_i^t, x_j^t, x^t, \theta)$ - параметрическое семейство моделей для описания семантической связи между предложениями, x_i^t, x_j^t, x^t - фрагмент, мишень, полный текст
- параметр $\theta \in \Theta$, где Θ - пространство параметров модели
- $y_{ij} \in [0, 1]$ - связаны ли мишень j и фрагмент манипуляции i
- $L_2(y, \hat{y}) = \sum\limits_{j\in J}\sum\limits_{i\in I} y_{ij}\log(\hat{y}_{ij}) + (1 - y_{ij})\log(1 - \hat{y}_{ij})$, где вектор $\hat{y}$ - вероятность семантической связи фрагмента и мишени.

Как и в задаче выявления манипулятивных фрагментов, модель поиска мишеней манипуляций инициализирована предобученной мультиязычной XLM-RoBERTa-base. Для решения задачи используется смесь архитектур

Cross-Encoder и Bi-Encoder. Cross-Encoder используется для получения совместного векторного представления фрагмента и мишени. Bi-Encoder подход испольуется для добавления контекстной информации о тексте, в котором встретились фрагмент и мишень.

Итоговая оптимизационная задача поиска оптимальных параметров модели поиска мишеней манипулятивных фрагментов сформулирована как:

$$\frac{1}{|I| + |J|} \sum_{i \in I} \sum_{j \in J} L_2(y_{ij}, p(x_i^t, x_j^t, x^t, \theta)) \to \min_{\theta \in \Theta}$$

Глава 3

Обзор литературы

В данной главе обсуждаются языковая модель XLM-RoBERTa, архитектуры моделей, разработанные для решения поставленных задач, метод дообучения языковых моделей с использованием малоранговой апроксимации матриц изменения весов лингвистических моделей.

Также историческая справка развития задачи Propaganda Detection, какие задачи изначально были поставлены и какое качество в их решении было достигнуто.

3.1 XLM-RoBERTa

Предобученные модели языковых моделей являются мощным инструментом для обработки естественного языка[4]. Они обучаются на огромных объемах текстовых данных и способны извлекать высокоуровневые представления языка, которые можно использовать для различных задач обработки текста. Одной из таких моделей является XLM-RoBERTa [5], которая достигает высоких результатов во многих языковых задачах и является одной из наиболее актуальных (State-of-the-Art) моделей.

XLM-RoBERTa-Base является предобученной моделью, основанной на архитектуре Transformer [6] [7] и разработанной для работы с множе-

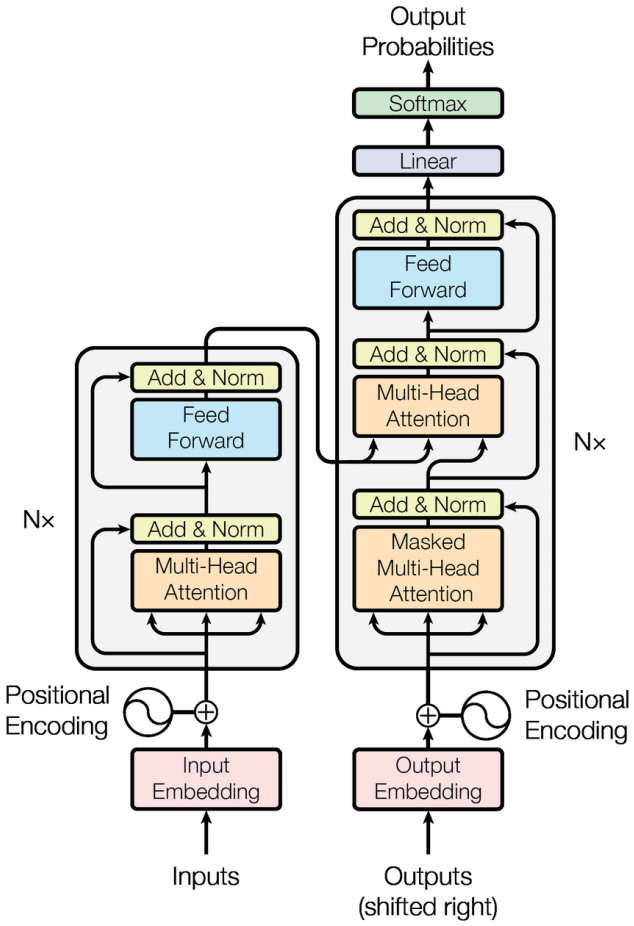

Рис. 3.1: Transformer

ством языков. Она является расширением RoBERTa модели, которая была предварительно обучена на большом корпусе текстов на разных языках. XLM-RoBERTa-Base использует задачи маскирования токенов и предсказания следующего предложения для предобучения модели.

Архитектура модели состоит из нескольких блоков Transformer, которые обеспечивают многоуровневое внимание и моделирование зависимостей между словами в предложении. Каждый блок Transformer состоит из множества линейных полносвязных слоев и слоев само-внимания, которые выполняют операции кодирования и декодирования. XLM-RoBERTa-Base имеет обширный словарь, который включает токены из разных языков.

XLM-RoBERTa-Base обучена на множестве языков, что позволяет ей быть эффективной для обработки текстов на разных языках без необходимости отдельного обучения модели для каждого языка.

Предобучение на большом объеме текстов позволяет модели XLM-RoBERTa-Base извлекать высокоуровневые представления языка, которые могут быть использованы для различных задач обработки текста, включая классификацию, извлечение информации и генерацию текста.

Поскольку модель предварительно обучена, она может быть легко использована для различных задач и сценариев. Это позволяет сэкономить время и ресурсы при разработке и развертывании моделей обработки текста.

XLM-RoBERTa-Base достигает высоких результатов во многих языковых задачах, включая машинный перевод, распознавание именованных сущностей и анализ тональности. Ее использование может значительно улучшить точность и производительность моделей обработки текста.

Модель XLM-RoBERTa-Base представляет собой мощный инструмент для обработки текста на разных языках. Ее многоязычность, представительность, переносимость и высокое качество результатов делают ее привлекательным выбором для решения различных языковых задач.

3.2 LoRA

В последние годы нейронные сети, основанные на языковых моделях, достигают впечатляющих результатов в различных задачах обработки естественного языка. Однако обучение и использование таких моделей требуют значительных вычислительных ресурсов и времени. По мере роста размеров и сложности моделей становится важным разработать методы, позволяющие эффективно адаптировать их к конкретным задачам или доменам без необходимости полного переобучения. [8]

В данном разделе рассмотрен метод файнтюнинга LoRA [9] для адаптации больших языковых моделей с использованием низкоранговой аппроксимации матриц изменения весов. Метод LoRA предлагает ускорение

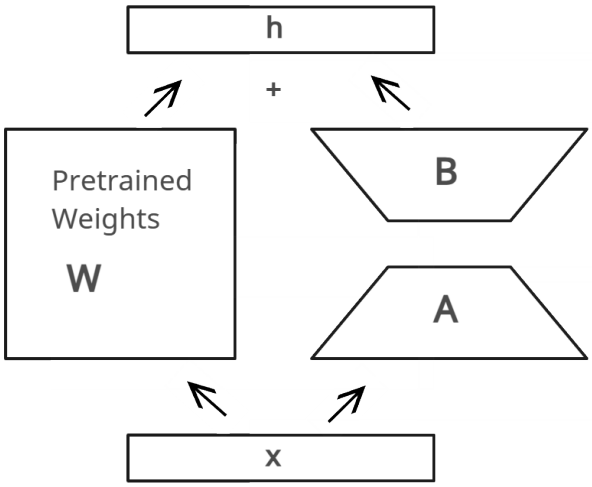

Рис. 3.2: LoRA

процесса адаптации и снижение требований к вычислительным ресурсам без значительных потерь в производительности.

Метод LoRA основан на идее аппроксимации изменения параметров языковой модели с помощью низкоранговых матриц. Метод предполагает, что матрицы измененения весов модели могут быть аппроксимированы более компактными низкоранговыми матрицами без существенной потери обобщающей способности дообученной модели.

В рамках метода LoRA производится обновление весов модели с использованием низкоранговой декомпозиции. Пусть W_0 - предварительно обученная матрица весов размерности $d \times k$. Адаптацию весов осуществляют путем представления обновления ΔW в виде $\Delta W = BA$, где B - матрица размерности $d \times r$, A - матрица размерности $r \times k$, и r - ранг матрицы, такой что из $r << \min(d, k)$.

Во время обучения предварительно обученная матрица W_0 остается замороженной и не получает обновлений градиента, в то время как матрицы A и B содержат обучаемые параметры. Для получения предсказания модели $h = W_0 x$, где x - входной вектор, выполняется модифицированный прямой проход модели: $h = W_0 x + \Delta W x = W_0 x + BAx$.

Метод LoRA может быть применен к любым подмножествам матриц

весов нейронной сети. В архитектуре Transformer данный метод применяется к весам в слоях само-внимания (W_q, W_k, W_v, W_o) и в линейных полносвязных слоях.

Основным преимуществом метода LoRA является снижение потребления памяти и хранилища. При использовании больших языковых моделей Transformer, метод LoRA позволяет сократить использование VRAM на $\frac{2}{3}$ за счет отсутсвия необходимости хранения состояний градиентов в процессе обучения для замороженных параметров.

3.3 XLM-RoBERTa with Adapter

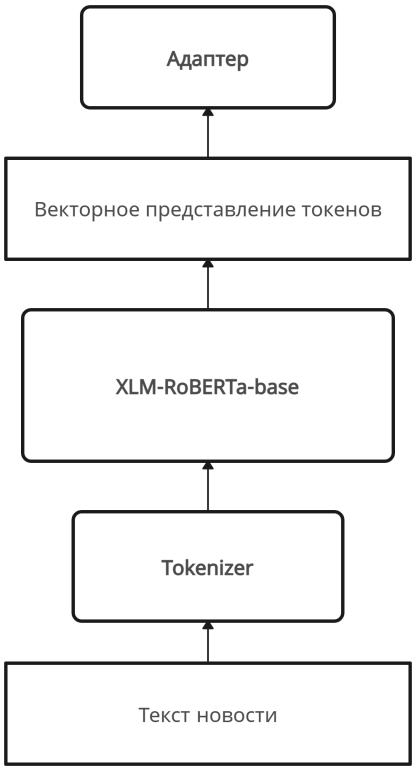

Рис. 3.3: SI-Model Архитектура

Архитектура модели для задачи Span Identification состоит из предобученной модели XLM-RoBERTa-Base и адаптера-классификатора.

XLM-RoBERTa-Base является предобученной многоязычной моделью, основанной на архитектуре Transformer. Она предобучается на большом

объеме текстовых данных на разных языках и способна эффективно кодировать языковые особенности. Результатом работы данной модели получаются контекстно-зависимые векторные представления токенов.

Определение: *Адаптер* - оператор апроксимации векторного представления.

Адаптер - это дополнительный модуль, добавляемый поверх предобученной модели, для дообучения модели под контректную задачу. В данной задаче в качестве адаптера используется линейный слой на вход которому подаются векторные представления токенов, а на выходе распределение вероятностей классов B, I, O. Адаптер является отображением из пространства векторных представлений токенов в пространство классов.

3.4 Cross-/Bi- Encoder

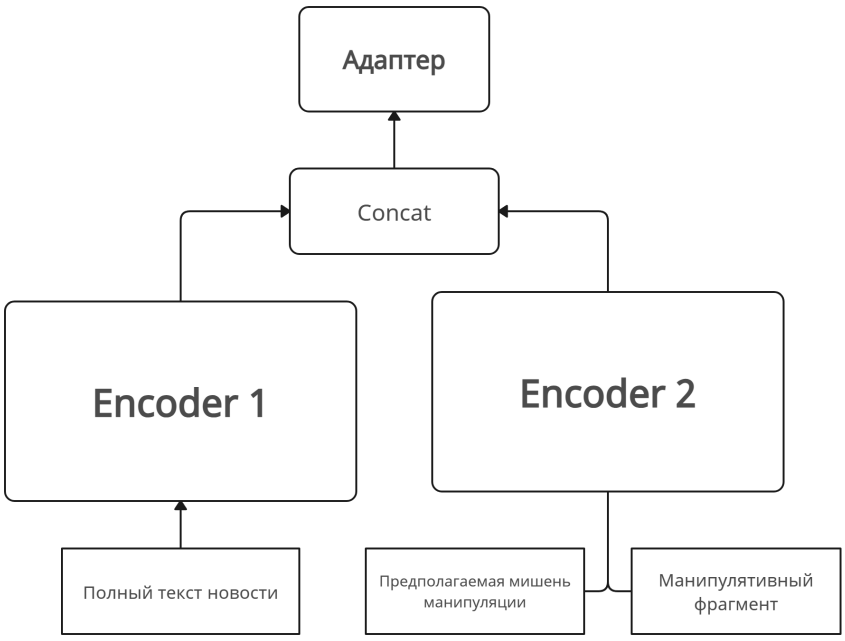

Рис. 3.4: ST-Model Архитектура

Архитектура модели для выявления семантической связи между фрагментами манипуляции и мишенью состоит из двух энкодеров и адаптера-классификатора:

Энкодер фрагментов манипуляции и мишени. Фрагмент манипуляции и мишень конкатенируются вместе и подаются на вход энкодеру. В качестве энкодера используется XLM-RoBERTa-base. Энкодер преобразует полученную текстовую последовательность в совместное векторное представление, содержащее информацию о связи между фрагментом манипуляции и мишенью.

Энкодер полного текста новости. Полный текст новости подается на отдельный вход энкодера. Энкодер преобразует текст новости в векторное представление, содержащее общую информацию о контексте новости.

Конкатенация векторных представлений. Полученный вектор из энкодера фрагментов манипуляции и мишени конкатенируется с вектором из энкодера полного текста новости. Это позволяет объединить информацию о семантической связи между фрагментом манипуляции и мишенью с общей информацией о контексте новости.

Полученное конкатенированное представление подается на вход адаптеру-классификатору. В качетстве адаптера используется полносвязный слой, для определения семантической связи между фрагментом манипуляции и мишенью.

Такая архитектура модели позволяет объединить информацию о фрагменте манипуляции и мишени с контекстуальной информацией о новости для более точного определения семантической связи между ними.

$$
\begin{aligned}
&y_{span,target} = h_1 = \text{Encoder}_1(span, target)),\\
&y_{full\ text} = h_2 = \text{Encoder}_2(full\ text)),\\
&h_3 = \text{concat}(y_{span,target}, y_{full\ text}),\\
&\text{output} = \text{adapter}(h_3)
\end{aligned}
$$

3.5 Propaganda Detection

Исследования в области выявления пропаганды изначально были сосредоточены на уровне документов. По данному тексту новости нужно было предсказать является ли он пропагандистским или нет. [1],[2]. В дальнейшем были сформулированы более общие задачи и требования к данным для построения моделей [3]. Были поставлены задачи Span Identification и Text Classification. Задача Span Identification - задача выявление фрагментов манипуляции. Text Classification - задача классификации манипулятивных фрагментов, необходимо построить модель для предсказания одного из 18 типов манипуляций. Было достигнуто следующее качество в задачи Span Identification:

P	R	F1
44.12	35.01	38.98

Таблица 3.1: Preslav Nakov results

3.6 SemEval

SemEval - это серия международных исследовательских workshop-ов по обработке естественного языка (NLP), основная цель которых заключается в продвижении SOTA-подходов в семантическом анализе и в создании высококачественных аннотированных наборов данных для решения ряда все более сложных задач в области семантики естественного языка. Ежегодный workshop включает в себя набор общих задач, в которых представлены различные прикладные задачи и датасеты для обучения моделей. Сравниваются решения, разработанные разными командами.

В 2020 [10] году одним из заданий было решение задачи Propaganda Detection в классической постановке из двух задач: Span Identification и Text Classification. Обучение и сравнение разработанных решений прохо-

дило на датасете англоязычных новостей. На тот момент были достигнуты следующие State-of-the-Art результаты в задаче Span Identification:

P	R	F1
0.57	0.47	0.52

Таблица 3.2: SemEval20 results

Глава 4

Данные и эксперимент

В данной главе обсуждается процесс сборки данных и информация о собранном датасете, а также описание вычислительного эксперимента. Под вычислительным экспериментов понимается процесс обучения моделей.

4.1 Данные

В данной секции подробно описан процесс сбора данных, внутреннее устрйство данных, а так же поднята проблема субъективности данных и как она решается в рамках данной работы.

4.1.1 Процесс сбора данных

Разметка производилась с помощью сервиса Yandex.Toloka («Яндекс.Толока»). Особенностью подхода является использование профессиональных аннотаторов, экспертов в области лингвистики, социологии и политической науки. Для достижения более точных результатов каждый текст размечался тремя аннотаторами. В среднем, эксперт проводит 5 минут на разметку одного текста.

4.1.2 Основные инструкции для аннотатора

Разметка состоит в выборе фрагмента манипулятивного текста, определении его класса (тип манипуляции) и цели. Цели были определены заранее с использованием стандартных моделей именованных сущностей (NER).

Разметка представляет 18 различных техник манипуляции (Рис. 4), к которым относится каждый из отмеченных фрагментов. Эти 18 классов объединены в 4 большие классы: негативизация, позитивизация, парадоксализация и деавторизация.

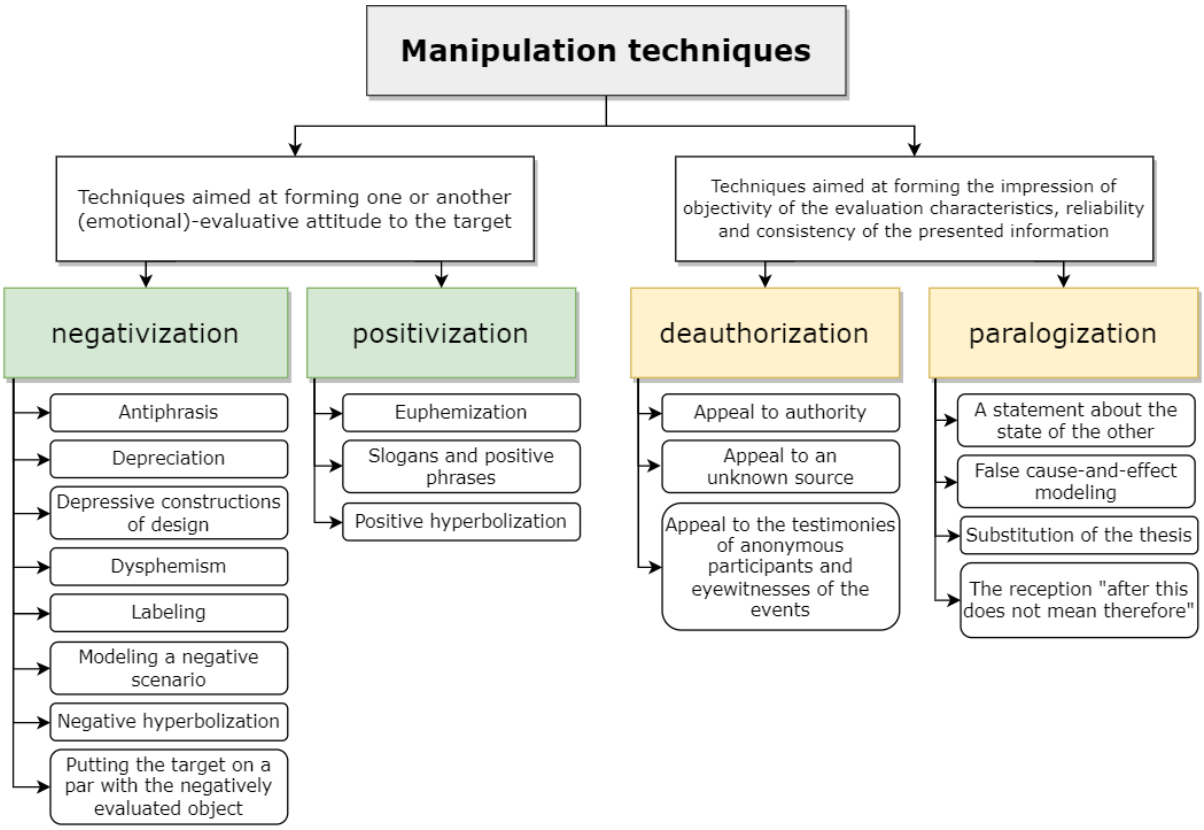

Рис. 4.1: Структура классификации манипулятивных техник.

Негативизация и позитивизация являются классами, направленными аксиологически противоположно: негативизация основана на внесении негативного отношения к цели в сознание получателя. Позитивизация направлена на создание положительного отношения к цели, вплоть до ее превознесения, прославления, героизации. Негативизация и позитивизация способствуют эмоциональному и оценочному тону текста.

Третья и четвертая группы включают техники, усиливающие оценочное воздействие путем создания впечатления объективности, достоверности информации и логических выводов. Техники деавторизации информации помогают скрыть источники информации и вызвать впечатление объективности. Техники парадоксализации основаны на отклонении от законов формальной логики, но направлены на создание вида логического рассуждения и полученных выводов.

Манипуляцию следует не путать с тональностью текста (наличием слов с положительными/отрицательными коннотациями или эмоционально окрашенными словами) по отношению к цели. Разметка манипуляции дополняет способность оценивать тональность текста, но не заменяет ее.

4.1.3 Статистики данных

Рассмотрим количественные показатели полученного набора размеченных данных для обучения модели. В результате разметки было получено 3803 маркировки 1421 уникальных документов (из которых 1165 маркировок содержали хотя бы один фрагмент манипуляции). 192 текста были размечены тремя аннотаторами, 458 - двумя, 515 - одним. Всего было получено 5443 манипуляции, состоящие из отмеченных троек «фрагмент - класс - цель».

Топ мишеней (число демонстрирует количество манипуляций направленных на мишень):

- Россия – 507,
- Украина – 263,
- США – 139,
- Москва – 93,

- ОДКБ – 77,
- НАТО – 70,
- Запад – 65,
- Китай – 51,
- Северная Корея – 40.

4.1.4 Субъективность оценки

Ввиду сложности задачи выявления манипуляций и наличия нескольких разметок, одной из основных проблем, с которой мы столкнулись, является выравнивание нескольких разметок для оптимизации результатов. Причина этой проблемы заключается в специфики задачи. Когда у двух и более разметок есть различные начальные, конечные позиции фрагментов манипуляции, разные мишени и тип манипуляции, возникает неопределенность в том, как совместить разметки для получения .

Для того, чтобы справиться с этой проблемы предлагается использовать **Markup Matching Loss** (MML) — метод, использующий функционал на парах фрагментов разметки, который направлен на выбор наиболее оптимального соответствия. Более того, он применим не только для задач манипуляции, но и для всех задач NER.

Введем некоторые обозначения:

- $X||Y$ - единая разметка, представляющая собой набор единиц манипуляции.
- $x_i \in X||y_i \in Y$ - единица манипуляции в отдельной разметке.
- T_i - мишень манипуляции в i единице манипуляции.

- C_i - класс манипуляции в i единице манипуляции.
- $L(x,y)$ - функционал потерь между найденными парами единиц манипуляции.

Основная идея предполагает определение оптимального соответствия для каждой единицы разметки в другой разметке или остуствия какого либо сопоставления для минимизации функционала соответсвия.

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

3 параметра, которые влияют на значения функционала:

- Размер пересечения фрагментов
- Соответсвие мишеней манипуляций
- Соответсвие классов манипуляций

Пересечение манипулятивных фрагментов является наиболее важным параметром. Таким образом, когда между единицами разметки нет пересечений, предпочтительнее не сопоставлять единицы измерения друг с другом.

Минимизируя эту функцию потерь для каждой пары разметки, мы получим наиболее релевантное совпадение для единиц манипуляций.

$$L(x,y) = \mathbb{I}\{J(x,y) = 0\}+$$
$$+ \mathbb{I}\{J(x,y) > 0\}\{-J(x,y) - \mathbb{I}\{T_x = T_y\} - \mathbb{I}\{C_x = C_y\}\}$$

$$L(x,y) \to \min$$

Где $J(x,y) = \frac{|x \cap y|}{|x \cup y|} \in [0,1]$ - индекс Джакарта.

Несмотря на обьеденную разметку, собранные данные сильно зашумлены, ввиду субьективности размекти манипуляций. Для того, чтобы оценить качество выделения манипуляций человеком, были посчитаны межассессорские метрики качества решения задачи Span Identification.

P	R	F1
0.224	0.223	0.223

Таблица 4.1: Качество выделения фрагментов манипуляции экспертами

4.2 Эксперимент

В ходе исследования различных архитектур для решения поставленных задач проводился ряд экспериментов с различными вариантами архитектурами. Использовались различные предобученые языковые модели: rubert-tiny, ruBert-base. Проводились эксперименты с разными постановками задач.

Стоит отметить ряд экспериментов для решения задачи Span Targeting как задачу relation extraction (RE), подробное описание которого приведено в секции про эксперименты.

Все эксперименты по обучению запускались на видеокарте Nvidia 1660 with Max-Q, 6GB видеопамяти. Использование малорангового приближения матриц изменения весов мотивировано невозможностью дообучать языковую модель в условиях ограниченного железа. Вычисления и хранение полных градиентов языковой модели невозможно ввиду нехватки вычислительных мощностей, в то время как использование LoRA позволило дообучить языковую модель и улучшить качество решаемых задач.

4.2.1 Метрики SI

Для оценки качества и сравнения моделей в задаче Span Identification необходимо ввести метрики качества решения задачи.

Пусть M - множество токенов, выделенных моделью, E - множество токенов, выделенных экспертом. Введем точность и полноту.

$$\begin{aligned} C(m,e,h) &= \frac{|m \cap e|}{h}, \\ P(M,E) &= \frac{1}{|M|} \sum_{m \in M,\ e \in E} C(m,e,|m|) \\ R(M,E) &= \frac{1}{|E|} \sum_{m \in M,\ e \in E} C(m,e,|e|) \end{aligned} \tag{4.1}$$

4.2.2 Span Identification

Модель для задачи выделения фрагментов обучалась в двух режимах.

Первый - заморозка слоев. XLM-RoBERTa-base и обучения только адаптера.

Второй - XLM-RoBERTa-base обучалась с использованием малорангового приближения весов и адаптер в стандартном режиме.

Для обоих режимов обучения модели использовался следующий набор гиперпараметров:

- Оптимизатор AdamW [11]
- Темп обучения $1 \cdot 10^{-5}$
- Функция ошибки: $L_1(y, \hat{y}) = \sum_{c \in C} w_c y_c \log(\hat{y}_c)$, $c \in C = [\mathrm{O,I,B}]$, $w_c = [1, 2, 2]$
- Шедулер: ExponentialLR с параметром затухания 0.99
- Параметр Dropout-a: 0.2
- 12 Encoder - слоев в XLM-RoBERTa-base

Анализ процессов обучения и моделей полученных после сходимости процесса обучения:

- Модель с замороженными слоями после сходимости процесса обучения обладает низкой обобщающей способностью.
- Модель, у которой происходило дообучение энкодера XLM-RoBERTa-base, после сходимости процесса обучения, обладает заметно лучшей способностью выделять фрагменты манипуляций. Достигнутая полнота выделения манипулятивных фрагментов превышает человеческие показатели, посчитанные между разметчиками обучающего корпуса.

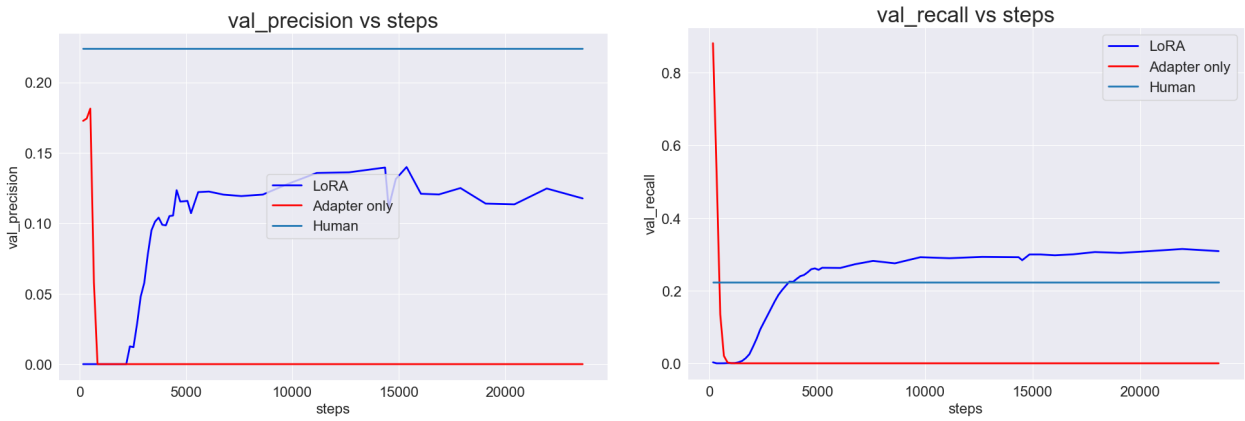

Рис. 4.2: Precision and Recall

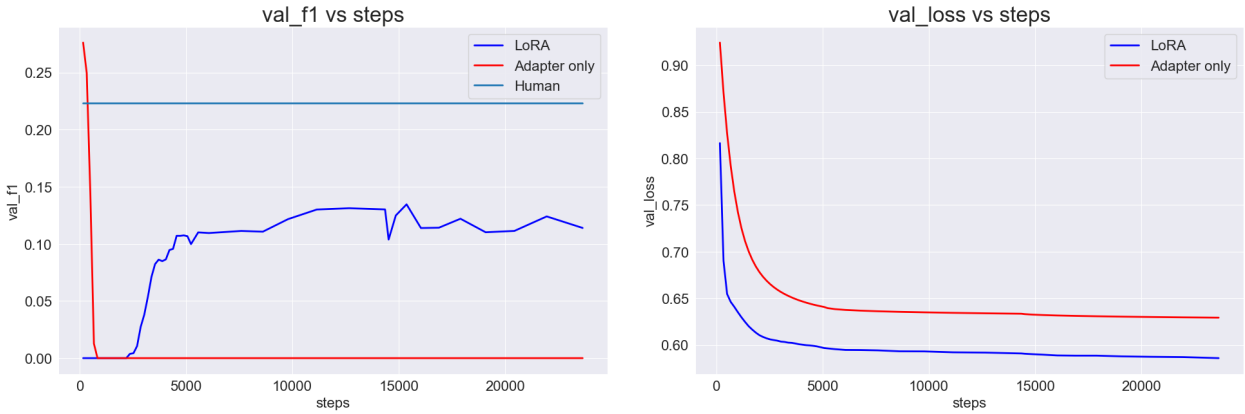

Рис. 4.3: F1 Score and Loss

.

4.2.3 Span Targeting

Также как и в задаче Span Identification модель для задачи связи поиска мишени манипуляции обучалась в двух режимах.

Первый - заморозка слоев энкодера XLM-RoBERTa-base и обучения только линейных слоев классификатора.

Второй - XLM-RoBERTa-base обучалась с использованием малорангового приближения весов и адаптер-классификатор в стандартном режиме.

Для задачи использовался следующий набор гиперпараметров:

- Оптимизатор AdamW [11]
- Темп обучения $1 \cdot 10^{-5}$
- Шедулер: ExponentialLR с параметром затухания 0.99
- Параметр Dropout-a: 0.2
- 12 Encoder - слоев в XLM-RoBERTa-base

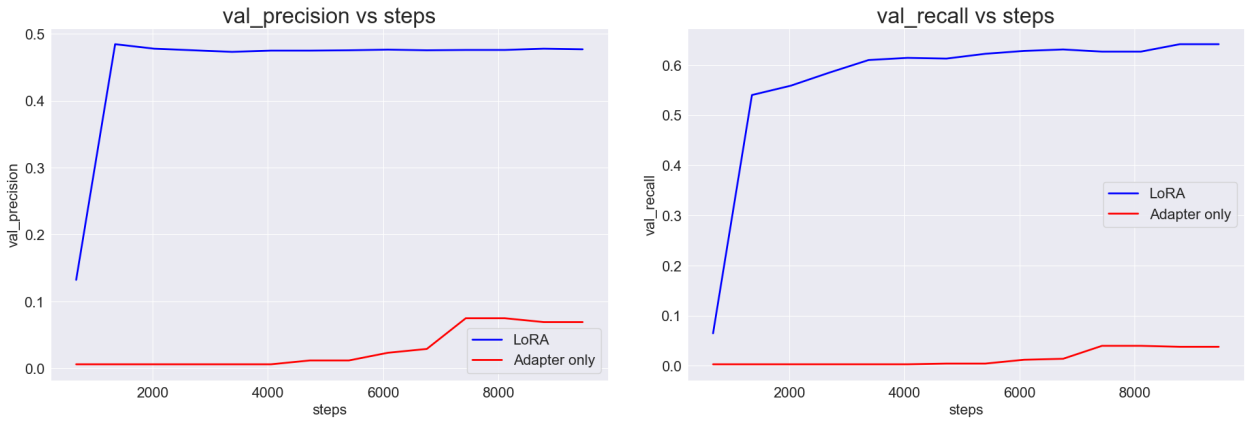

Рис. 4.4: Precision and Recall

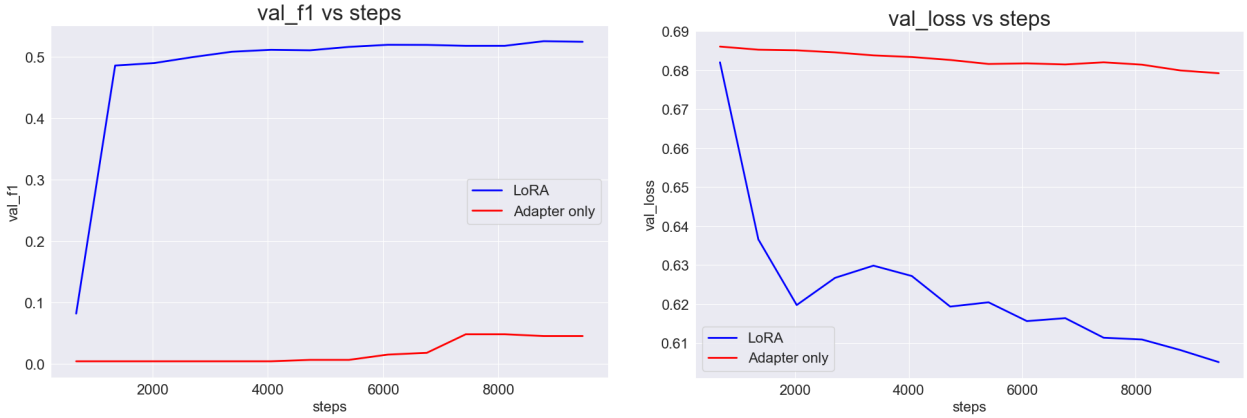

Рис. 4.5: F1 Score and Loss

4.2.4 Анализ эксперимента

Выделение манипулятивных фрагментов и поиск мишеней манипуляции относятся к классу задач NLP, в которых исследуемые тексты обладают сложным семантическим устройством языка. Дообучения языковых моделей является необходимым для качественного решения таких задач.

Несмотря на зашумленность данных, получилось добиться качества по некоторым показателям превосходящих результаты экспертов.

4.2.5 Relation Extraction

Relation Extraction включает в себя идентификацию и категоризацию семантической связи между объектами или фрагментами в данном тексте.

В задаче Span Targetting я работал со связями между манипулятивными фрагментами и мишенями манипуляции, поэтому наши отношения не имеют категорий. Задача состоит в том, чтобы предсказать, связаны ли фрагменты текста с точки зрения манипулирования или нет. Одним из наиболее многообещающих подходов для извлечения связей является использование билинейных слоев [12].

Векторные представления токенов из XLM-RoBERTa-base подавались на вход билинейного слоя. В данном случае билинейный слой выступал в качестве оператора отображения из произведения пространств векторного представления токенов в отрезок [0, 1]

$$f : H \times H \to [0, 1]$$

Математически билинейный слой без функции активации:

$$h = x_1^\top A x_2 + b \tag{4.2}$$

В билинейном слое два входных сигнала отображаются в одно векторное пространство.

Этими входными данными могут быть текстовые описания, изображения или аудиозаписи. В своем случае я использовал векторные представления токенов для обоих входов. Сеть учится сравнивать векторные представления и определять связи между ними. Это достигается с помощью “тензорного произведения”, которое вычисляет поэлементное умножение двух входных векторов. Каким-то образом он изучает скалярное произведение в новом преобразованном векторном пространстве. Это помогает модели идентифицировать тонкие взаимодействия между двумя входными данными и научиться обнаруживать и идентифицировать значимые взаимосвязи.

Математически модуль Relation Extraction представляет собой:

$$r = \text{BERT}(\text{ввод\_text})$$
$$h = \text{ReLU}(\text{билинейный}(r, r))$$
$$o = \sigma(\text{линейный}(h))$$

Достигаемое качество задачи Span Targetting в постановке Relation Extraction:

P	R	F1
0.017	0.179	0.029

Таблица 4.2: Качество попарной классификации токенов

Глава 5

Заключение

Данная работа предлагает расширение задачи Propaganda Detection за счет добавления подзадачи Span Targetting, математическую постановку данной задачи, базовые модели для решения задач Span Identification и Span Targetting, датасет для дальнейших исследований. Предложенная задача расширяет подход исследования пропагандистких новостей, опираясь на определение манипуляции - как социологического явление, направленного на изменение мнения читателя относительно некоторой мишени. В процессе работы над сборкой датасета были разработаны и предложены методы сопоставления зашуленных разметок. В ходе данной работы также была продемонстрирована эффективность использования малорангового приближения матриц изменения параметров языковых моделей в условиях отсутствия инфраструктуры с большими вычислительными мощностями.

Помимо прочего в рамках исследования было проведено большое количество вычислительных экспериментов для поиска оптимальных гиперпараметров, процесса обучения и архитектуры моделей. Весь код экспериментов выложен в открытый доступ и может быть воспроизведен исследователями в этой области.

В последующих работах планируется сконцентрировать свое внимание на улучшении качества данных и работы с экспертами. На данный

момент очевидно, что текущие наработки достигают хорошего результата, который сопоставим с качеством экспертов, но есть необходимость в улучшении качества данных, а именно согласованности разметчиков.

Литература

[1] Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking / Hannah Rashkin, Eunsol Choi, Jin Yea Jang et al. // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. — Copenhagen, Denmark: Association for Computational Linguistics, 2017. — . — Pp. 2931–2937. https://aclanthology.org/D17-1317.

[2] Proppy: Organizing the news based on their propagandistic content / Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Martino, Preslav Nakov // *Information Processing & Management*. — 2019. — 05. — Vol. 56.

[3] Fine-Grained Analysis of Propaganda in News Articles / Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño et al. // *CoRR*. — 2019. — Vol. abs/1910.02517. http://arxiv.org/abs/1910.02517.

[4] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // *CoRR*. — 2018. — Vol. abs/1810.04805. http://arxiv.org/abs/1810.04805.

[5] Unsupervised Cross-lingual Representation Learning at Scale / Alexis Conneau, Kartikay Khandelwal, Naman Goyal et al. // *CoRR*. — 2019. — Vol. abs/1911.02116. http://arxiv.org/abs/1911.02116.

[6] Attention Is All You Need / Ashish Vaswani, Noam Shazeer, Niki Parmar et al. // *CoRR*. — 2017. — Vol. abs/1706.03762. http://arxiv.org/abs/1706.03762.

[7] Transformers: State-of-the-Art Natural Language Processing / Thomas Wolf, Lysandre Debut, Victor Sanh et al. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. — Online: Association for Computational Linguistics, 2020. — . — Pp. 38–45. https://aclanthology.org/2020.emnlp-demos.6.

[8] Low-rank matrix factorization for Deep Neural Network training with high-dimensional output targets / Tara N. Sainath, Brian Kingsbury, Vikas Sindhwani et al. // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. — 2013. — Pp. 6655–6659.

[9] LoRA: Low-Rank Adaptation of Large Language Models / Edward J. Hu, Yelong Shen, Phillip Wallis et al. // *CoRR*. — 2021. — Vol. abs/2106.09685. https://arxiv.org/abs/2106.09685.

[10] SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles / Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth et al. // *CoRR*. — 2020. — Vol. abs/2009.02696. https://arxiv.org/abs/2009.02696.

[11] *Loshchilov, Ilya.* Fixing Weight Decay Regularization in Adam / Ilya Loshchilov, Frank Hutter // *CoRR*. — 2017. — Vol. abs/1711.05101. http://arxiv.org/abs/1711.05101.

[12] End-to-end Named Entity Recognition and Relation Extraction using Pre-trained Language Models / John Giorgi, Xindi Wang, Nicola Sahar et al. — 2019. https://arxiv.org/abs/1912.13415.