Разработка и исследование методов обнаружения искаженных данных с использованием мультимодального подхода

А.В. Кузнецов
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
kuznetsoff.andrey@gmail.com

Е.А. Маркина
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
kate.markina2010@yandex.ru

Аннотация – **В работе рассматривается задача обнаружения искаженных данных с использованием мультимодального подхода. Исследуются различные способы классификации. В ходе экспериментальных исследований получено, что наилучшие результаты демонстрирует метод, основанный на предобученной мультимодальной модели RUDOLPH. Наибольшее значение метрики F1 равно 0,841.**

Ключевые слова – мультимодальные модели, искаженные данные, машинное обучение

I. Введение

Задача разработки методов для обнаружения и классификации искаженных данных в современном обществе является актуальной и широко исследуемой. Искаженные данные (фейки) – это потерявшая первоначальный вид информация, которая была модифицирована с целью розыгрыша, введения в заблуждение, побуждения к какому-либо действию и другое [1]. Фейки не являются чем-то новым, но наибольшее внимание получает именно в последние годы. Это связано с развитием информационных технологий и в особенности социальных сетей, которые являются наименее регулируемой частью медиа пространства и предоставляют возможность для постоянного и стабильного распространения информации всеми пользователями. Многие государства на законодательном уровне уделяют внимание влиянию фейковых новостей на различные сферы и общество в целом, защищая население от ложных новостей и сохраняя благополучие регионов.

В статье рассматривается задача обнаружения и классификации искаженных данных, которые представлены в двух модальностях – изображение и текст. Предполагается, что при подаче данных на вход модели они предобработаны и текст сопоставим с изображением. В работе используются предобученные мультимодальные модели, работающие одновременно с текстом и изображением.

II. Описание используемого набора данных

В качестве искаженных данных рассматривается набор синтезированных фейковых новостей [2] – ложных новостей или повествований, позиционирующихся как истинные, также называемых псевдоновостями. Используемый набор данных подходит для мультимодального обучения [3], так как каждая новость представляется в двух модальностях: текст и изображение, хранимое в виде url. Данные собраны с различных интернетплатформ. Поэтому, помимо непосредственно текста и изображения, также есть информация об авторе поста, домен, подтверждение наличия изображения, а также информация о подлинности новости в виде отнесения к какому-либо классу. Общее количество данных в датасете составляет 386881 новостей, среди которых число обучающих равно 330260, а тестовых 56621. Классификация на рассматриваемом наборе данных возможна на два, три и шесть классов. Таким образом, база данных позволяет проводить классификацию как в более широком смысле, разделяя на два или три класса, так и в более узком, выполняя деление на шесть классов. Категории при различных делениях представлены в таблице I.

Таблица I. Возможная классификация новостей в датасете

Кол-во классов	**Возможная классификация**
2 класса	а) истина; б) ложь
3 класса	а) истина; б) истинный только текст; в) ложь
6 классов	а) истина; б) сатира/пародия; в) манипулирующий контент; г) вводящие в заблуждение заголовки; д) ложное соединение; е) «самозваный контент»

III. Предлагаемый подход

Первоначально проводилось исследование качества и разработка методов классификации унимодальных искаженных данных, представленных только текстом. Для этого проводился fine-tuning предобученных моделей BERT и T5. Полученные результаты представлены в таблице II.

Таблица II. Сравнение значений метрики, полученных при использовании разработанных методов, работающих с текстовыми данными

Кол-во классов	**Значение метрики**	
	Предобученная модель T5	***Предобученная модель BERT***
2 класса	0,796	0,62
3 класса	0,775	0,7
6 классов	0,753	0,695

Решение задачи обнаружения и классификации искаженных мультимодальных данных включает в себя предварительную предобработку набора данных и далее работа разработанного метода.

A. *Предобработка набора данных*

При отсутствии семантической связи между рассматриваемыми парами изображение и текст задача обнаружения и классификации искаженных данных может решаться некорректно.

Первоначально из набора были удалены семантически не связанные данные. Это было реализовано с помощью предобученной модели CLIP: представляет собой режим тренировки двух моделей [4]. Архитектура представлена в виде энкодера (encoder), который берёт входную последовательность и приводит её к характеристическому представлению всей последовательности, и декодера (decoder), распаковывающего полученный вектор в целевую последовательность. Мера близости определяется с помощью вычисления косинусного расстояния. Результаты представлены в таблице III.

Таблица III. РЕЗУЛЬТАТ ВЫЧИСЛЕНИЯ СЕМАНТИЧЕСКОЙ СВЯЗИ МЕЖДУ ТЕКСТОМ И ИЗОБРАЖЕНИЕМ В НАБОРЕ ДАННЫХ

Вид данных	**Количество данных**		
	Всего	***Семантическая связь выявлена***	***Семантическая связь не выявлена***
Train	273 055	264 769	8 286
validation	57 205	57 036	169
test	56 621	56 466	155

B. *Разработанный метод обнаружения и классификации искаженных данных*

При работе с мультимодальными данными использовалась модель, которая может одновременно работать с несколькими модальностями. При разработке метода для решения задачи обнаружения и классификации искаженных данных производился fine-tuning и вносились некоторые изменения в предобученную модель RUDOLPH. Это мультимодальная русскоязычная модель, основанная на transformer, которая архитектурно представляет собой его декодер. Модель работает с сегментированной последовательностью входных токенов: левых – они отвечают за понимание текста и передаваемой текстовой инструкции, изображения – используются для генерации изображения по текстовому запросу или для кодирования переданного на вход изображения и правых – те, в которых генерируются требующие текстовый вывод ответы [5].

IV. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

Для различных задач применяются разные метрики, среди которых наиболее часто используются такие как: accuracy, precision, recall, F1. Решается задача многоклассовой классификации для несбалансированных данных, поэтому для оценки качества работы разработанного метода рассчитывалась метрика *F1*, учитывающая, как распределены данные в используемом наборе данных:

$$F1 = \frac{2 \times presicion \times recall}{precision + recall} \quad (1)$$

Разрабатываемый метод был основан на предобученной модели RUDOLPH. На вход подаются левый тест – сама новость и вспомогательный вопрос, визуальные токены – изображение к рассматриваемой синтезированной новости. Исследования проводились при изменении параметров: размер батча и число эпох. Также учитывалось то, на какое количество классов выполняется классификация.

Наилучшие показатели метрики *F1* были получены при размере батча равном 8. В таблице IV представлены значения метрики *F1* в зависимости от числа эпох, а также время, требуемое на обучение и тестирование.

Таблица IV. ЗНАЧЕНИЯ МЕТРИКИ F1 ПРИ ИСПОЛЬЗОВАНИИ РАЗРАБОТАННОГО МЕТОДА

Кол-во эпох	**Значение**		
	Метрика F1	***Время обучения***	***Время тестирования***
1	0,8247	12ч 10мин	19ч 25мин
2	0,8392	23ч 1мин	19ч 20мин
3	0,841	34ч 9мин	19ч 22мин
4	0,8411	46ч 28мин	19ч 28мин
5	0,836	58ч 8мин	19ч 27мин
6	0,833	70ч 13мин	19ч 23мин

V. ЗАКЛЮЧЕНИЕ

В работе рассматривалась задача обнаружения и классификации искаженных данных с использованием мультимодального подхода. Предложено использовать разработанный метод, основанный на предобученной модели RUDOLPH для решения поставленной задачи. Полученные результаты показывают, что использование разработанного метода позволяет увеличить значение метрики F1 при работе с мультимодальными данными в сравнении с унимодальными или использованием других предобученных мультимодальных моделей.

ЛИТЕРАТУРА

[1] Муратова Н. Fake news: дезинформация в медиа: учебное пособие / Н. Муратова, Н. Тошпулатова, Г. Алимова. – Ташкент: “Innovatsion rivojlanish nashriyot-matbaa uyi”, 2020. –104 с.

[2] Nakamura, K. r/Fakeddit: A New Multimodal Benchmark Dataset for Finegrained Fake News Detection / K. Nakamura, S. Levy, W. Yang Wang // arXiv:1911.03854v2. – 2020. DOI: 10.48550/arXiv.1911.03854.

[3] Akkus, C. Multimodal Deep Learning // C. Akkus, L. Chu, V. Djakovic/ arXiv:2301.04856v1, 2023. – 272 p. DOI: 10.48550/arXiv.2301.04856.

[4] Radford, A. Learning Transferable Visual Models From Natural Language Supervision / A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever // arXiv:2103.00020v1. – 2021. DOI: 10.48550/arXiv.2103.00020

[5] Мультимодальная модель RUDOLPH [Электронный ресурс]. – Режим доступа: https://habr.com/ru/companies/sberbank/articles/733470 (01.05.2023)