ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
„ВЫСШАЯ ШКОЛА ЭКОНОМИКИ”»

Шикунов Николай Алексеевич

Применение глубокого обучения к задаче распознавания фейковых новостей

КУРСОВАЯ РАБОТА

Научный руководитель:
преподаватель
Ковалев Евгений Игоревич

Москва, 2020

Применение глубокого обучения к задаче распознавания фейковых новостей

Шикунов Николай Алексеевич

Июнь 2020 г.

Реферат

С каждым годом увеличивается влияние фейковых новостей на общественность. В связи с колоссальным ростом медиа-пространства, растёт скорость распространения ложной информации. В настоящее время процесс факт-чекинга занимает достаточно много времени. Специалисты этой области вручную изучают новости и выявляют факты лжи. В связи с этим появляется потребность в автоматическом, практически моментальном, распознавании фейковых новостей. Чем раньше обнаружена фейковая новость, тем меньше людей попадут под её влияние. Целью данной работы является исследование задачи распознавания фейковых новостей на базе бенчмарк датасета LIAR-PLUS с использованием методов глубокого обучения. В работе представлены результаты экспериментов моделирования для нескольких архитектур нейронных сетей. Использование LSTM+Attention улучшило результат до 0.652 и 0.267 Accuracy для бинарной и многоклассовой классификации соответственно.

Научный руководитель:
преподаватель
Ковалев Евгений Игоревич

Содержание

1 Введение

Фейковые[1] новости - это разновидность жёлтой прессы, цель которой направлена на намеренное введение людей в заблуждение. Мотивами такой деятельности могут быть как мошенничество, так и политические акции, направленные на подрывание общественного порядка. Феномен фейковых новостей появился намного раньше эры глобализации и интернета. 30 мая 1896 года на Ходыском поле состоялось торжество в связи с коронацией Императора Николая II [30]. В этот день бесплатно раздавали еду и царские гостинцы. Но кто-то запустил информацию о том, что якобы подарки уже начали раздавать и их на всех не хватит. В результате появилась ужасная давка, из-за которой погибли свыше тысячи человек. Задача своевременного выявления фейковых новостей усложняется скоростью распространения. Люди, попавшие под влияние такой информации, стараются поделиться этой новостью с другими. Такой процесс можно сравнить с эффектом домино или с вирусной инфекцией. Современный пик активности распространения фейковых новостей приходится на выборы в США 2016. По данным исследования [4], на тему выборов трафик фальшивых новостей в социальных сетях составил 41.8%, что намного больше всего информационного трафика, который проходит через телевидение, газеты и радио. Сегодня общество вновь находится под серьёзным ударом фейковых новостей, по причине COVID-19. Люди активно сжигают вышки 5G, становятся COVID-диссидентами и отказываются от любой вакцинации [6].

Специалисты факт-чекинга[2] при выявлении ложных новостей исследуют большие массивы текстовых данных: новостные заголовки, описание новости, косвенная информация предметной области, информация о докладчике, информация о медиа-ресурсе. Чтобы автоматически выявлять фальшивые новости на таком большом объёме текстовой информации, можно воспользоваться методами глубокого обучения. Например, в работах [29], [23], [27] уже были предложены свои архитектуры нейронных сетей для детекции фейковых новостей.

В этой работе были проанализированы данные фальшивых новостей и предложены несколько архитектур нейронных сетей для их идентификации. В результате работы удалось выяснить, что использование дополнительной информации о предметной области новостного заголовка сильно улучшает результат работы моделей. Для датасета LIAR-PLUS архитектура Bi-LSTM+Attention показало лучший результат: Accuracy на бинарной классификации **0.652**, на многоклассовой классификации **0.267**.

[1]В этой работе используется выражение «Фейковые новости» наряду с «Ложными» и «Фальшивыми» новостями. В 2017 году Collins Dictionary признал термин Fake News словом года. И спустя 3 года этот англицизм всё чаще используется в русской речи

[2]Факт-Чекинг(Проверка фактов) — проверка достоверности сведений, описанных в текстах

2 Данные фейковых новостей

Одной из проблем исследования фейковых новостей является дефицит размеченных данных. Факт-чекинг — это трудоёмкий процесс. Качественную разметку на предмет фейка вручную может сделать только специалист области факт-чекинга. Например, в работах [13] и [14] использовалась разметка краудсорсинговых агенств. Хотя краудсорсинг и является важным подходом к созданию размеченных данных, но в данной области, чтобы избежать сильного шума в разметке, лучше обратиться к работе специалистов. Дефицит размеченных данных является неким ограничителем развития технологий глубокого обучения в этой области. Также, чтобы проводить исследования этой проблемы, надо иметь в общем доступе хотя бы 1 бенчмарк датасет[1]. В работе [11] впервые был представлен размеченный датасет для анализа проблемы фейковых новостей. Но из-за очень маленького объёма данных в 221 новость использовать методы глубокого обучения бесполезно.

2.1 Бенчмарк датасет LIAR

В 2017 году появился новый бенчмарк датасет LIAR [27]. Этот датасет состоит из 12.863 строк данных. В признаках есть текстовый заголовок новости, а также дополнительная информация предметной области. Содержимое датасета LIAR, а также разметка, взято с ресурса POLITIFACT.COM [22]. PolitiFact охватывает широкий спектр политических тем. Специалисты этой организации классифицируют новости на предмет фейка и предоставляют подробные результаты факт-чекинга. Данные PolitiFact содержат новости из разных источников: политические дебаты, телевизионная реклама, посты в FaceBook и Twitter, интервью, новостные релизы и т.д. Основная часть данных LIAR это новости из периода президентской гонки в США 2016.

2.1.1 Признаковое описание датасета LIAR

Разметка

- **pants-fire** - Выдуманная новость;
- **false** - Новость ложная;
- **barely-true** - В целом новость ложная;
- **half-true** - Новость наполовину правдивая;
- **mostly-true** - В целом новость правдивая;
- **true** - Правдивая новость.

[1]Бенчмарк датасет(Benchmark Dataset) — это набор данных предметной области, который служит источником данных для исследований в сфере предметной области

Бинарная разметка

- **is fake** = 1 - новость фейк. (pants-fire или false или barely-true = 1);
- **is fake** = 0 - новость не фейк (half-true или mostly-true или true = 1).

Признаки датасета

- **label** - Таргет новости;
- **statement** - Новостной заголовок;
- **subject** - Тема новостного заголовка (Здравоохранение, выборы, налоги и тд);
- **speaker** - Имя спикера новостного заголовка (Трамп, Клинтон и тд);
- **speaker job** - Профессия спикера новостного заголовка (Президент, губернатор и тд);
- **state** - Штат(город), где появился новостной заголовок (Нью-Йорк, Техас и тд);
- **party** - Политическая партия спикера новостного заголовка (Демократ, республиканант);
- **context** - Контекст новости. Место, где появилась данная новость (Интервью, пресс-релиз, Твиттер и тд);
- Суммарная история новостей спикера (Сумма по всем label). Включая текущий новостной заголовок:
  - **barely true counts** - Сумма “barely true” новостей спикера;
  - **false counts** - Сумма “false” новостей спикера;
  - **half true counts** - Сумма “half true” новостей спикера;
  - **mostly true counts** - Сумма “mostly true” новостей спикера;
  - **pants on fire counts** -Сумма “pants on fire” новостей спикера.

2.1.2 Утечка данных в датасете LIAR

В данных LIAR есть признаки с историей ложных новостей по каждому спикеру. Проблема в том, что в датасете нет привязки ко времени. А значит, что в старых новостях есть информация о будущих новостных заголовках - это утечка данных. По этой причине в работе [23] было предложено не использовать эти признаки в построении модели. В своей работе я аналогично эти признаки не использую.

2.2 LIAR-PLUS

В 2018 году в работе [2] было представлено расширение датасета LIAR под названием LIAR-PLUS. В новом датасете появился новый признак **justification**. Авторы извлекли из факт-чекинг статей POLITIFACT.COM [22] обоснования специалистов при разметке новостей на предмет фейка. Авторы LIAR-PLUS утверждают, что при использовании дополнительной информации и фактов о новости и предметной области, качество предсказания увеличивается. И это логично, ведь во время факт-чекинга специалист акцентирует своё внимание на дополнительных знаниях предметной области. Создатели LIAR-PLUS при выделении обоснований специалистов POLITIFACT.COM почистили данные от возможных утечек. В своей работе я решил выбрать датасет LIAR-PLUS для того, чтобы расширить спектр экспериментов.

Размер датасета LIAR-PLUS	
Train set size	10,269
Validation set size	1,284
Test set size	1,283

2.3 Разведочный анализ данных

label	pants-fire	**label**	true
is fake	1	**is fake**	0
statement	Says Hillary Clinton has even deleted this record of total support (for the Trans-Pacific Partnership trade agreement) from her book.	**statement**	McCain opposed a requirement that the government buy American-made motorcycles. And he said all buy-American provisions were quote 'disgraceful.'
subject	candidates-biography,economy,foreign-policy,history,job-accomplishments,trade	**subject**	federal-budget
speaker	donald-trump	**speaker**	barack obama
speaker job	President-Elect	**speaker job**	President
state	New York	**state**	Illinois
party	republican	**party**	democrat
context	a speech	**context**	a radio ad
justification	Trump said Hillary Clinton "has even deleted this record of total support (for the Trans-Pacific Partnership trade agreement) from her book. "The paperback removed a small reference to the TPP but the two pages that talked about it and why the agreement was important weren't deleted. The paperback edition continues to have text expressing support for the trade deal.	**justification**	I firmly object to all 'buy America' restrictions, as they represent gross examples of protectionist trade policy. "He added, "Furthermore, as a fiscal conservative, I want to ensure our government gets the best deal for taxpayers and with a 'buy American' restriction that cannot be guaranteed. "The McCain campaign did not respond to Obama's charge. But his past words say it all.

2 случайных объекта из данных.

- Распределения классов по train, val, test сильно не отличаются (Рис. 1). Видно, что правдивых новостей в датасете чуть больше. Самый малочисленный класс фейковых новостей - pants-fire;

- Количество фейковых новостей распределено неравномерно в зависимости от тематики новости subject (Рис. 2). Например, на тему здравоохранения ложных новостей намного больше, а на тематику образования преобладают правдивые новости;

- Вероятность фейковой новости зависит от политической партии, к которой относится спикер (Рис. 3). Республиканцы произносят больше фальшивых новостей, чем демократы. А беспартийные люди, как правило пользователи в социальных сетях, чаще заявляют ложь;

- Среди всех спикеров можно выделить Барака Обаму и Хиллари Клинтон, которые в основном заявляют правду, а также Дональда Трампа, который с огромной разницей в большинстве случаев говорит ложь (Рис. 4). На Рис. 5 видно, что чётко выделяются несколько спикеров, которые являются основными производителями новостей. В связи с этим в своей работе малозначимых спикеров я отношу к одному классу unknown speaker;

- В данных можно выделить 2 типа признаков: мета-данные и текстовые данные. Мета-данные — это вся дополнительная информация, которая идёт вместе с заголовком новости. Текстовых признаков в датасете всего 2: statement и justification. Это те объёмные текстовые признаки, которые нужно по-особому обрабатывать, чтобы не потерять важную информацию.

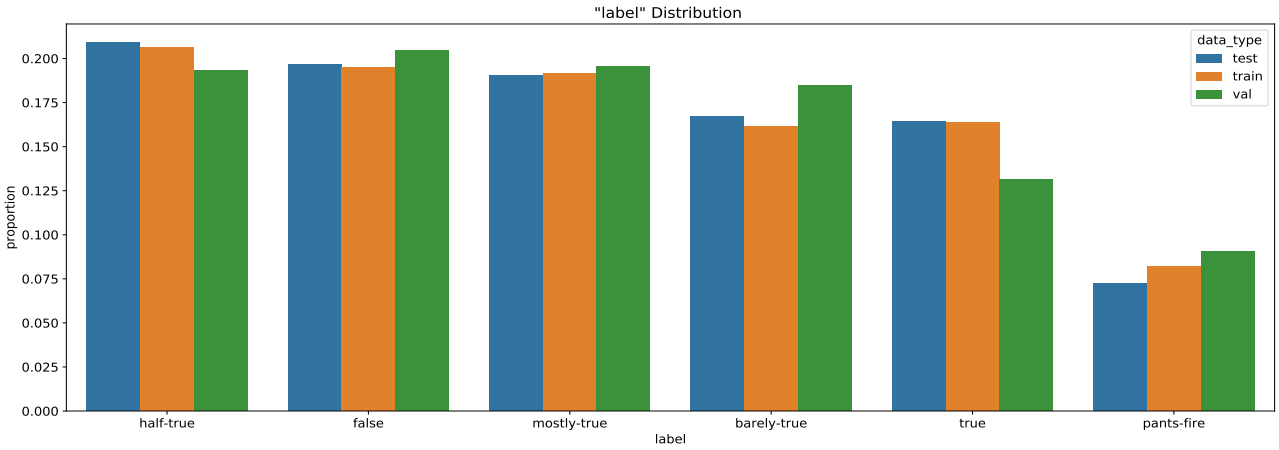

Рис. 1: Гистограмма таргета по train, validation, test датасетам

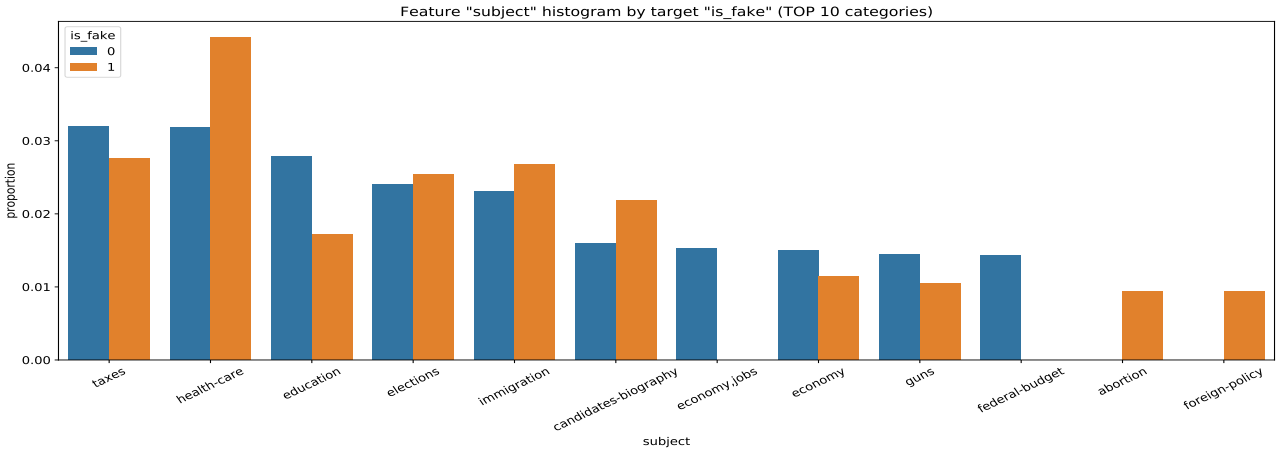

Рис. 2: Гистограмма таргета по группам **subject**

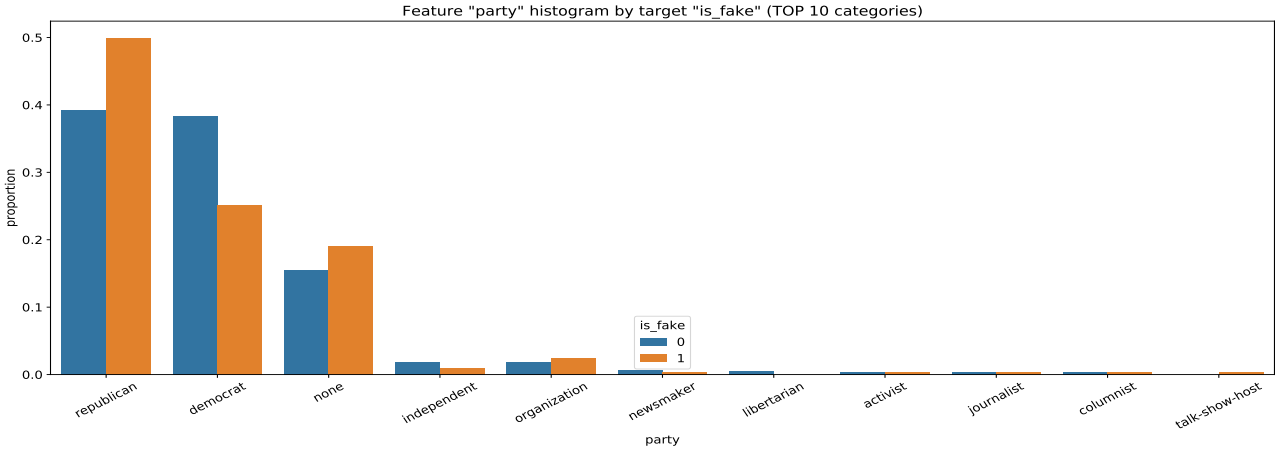

Рис. 3: Гистограмма таргета по группам **party**

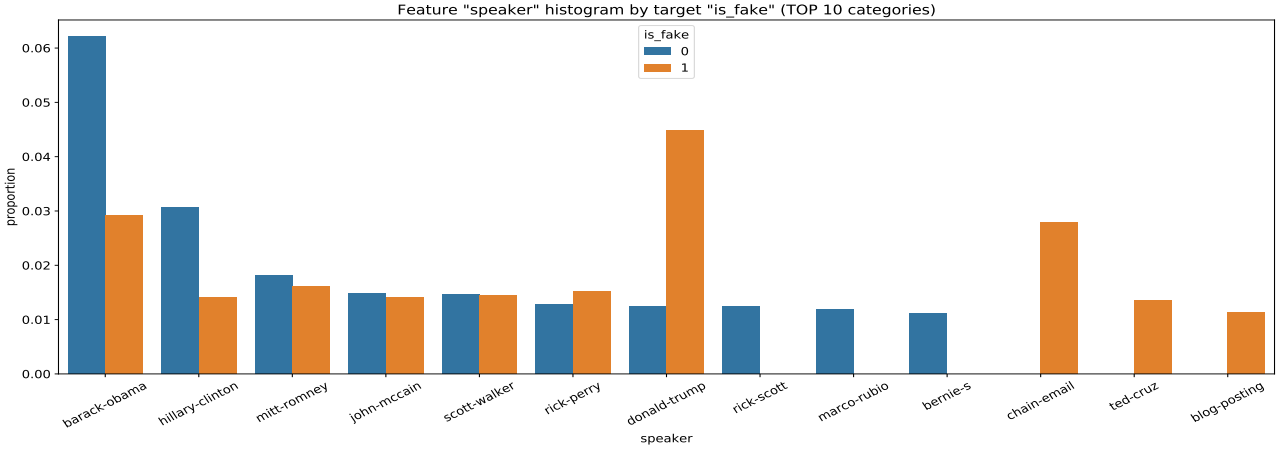

Рис. 4: Гистограмма таргета по группам **speaker**

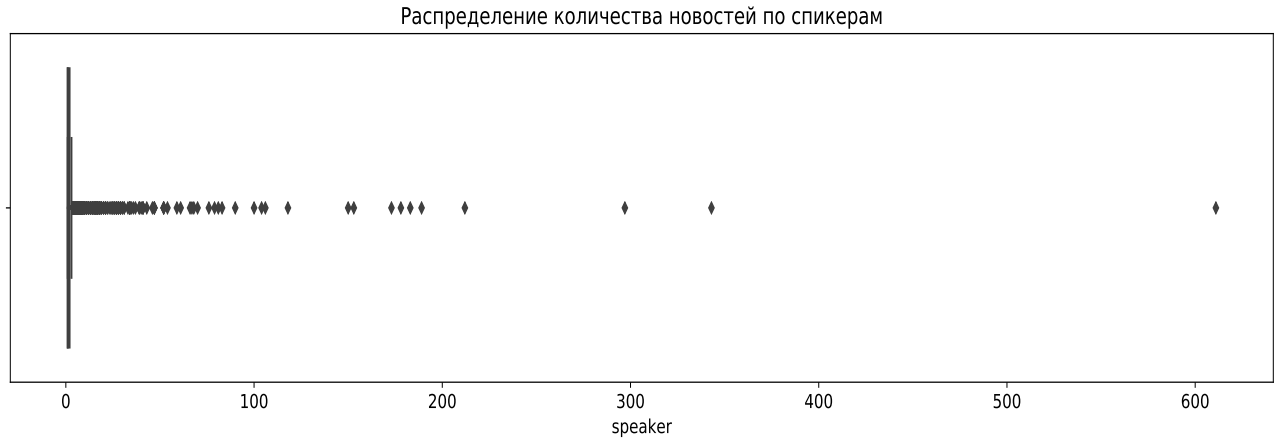

Рис. 5: Boxplot количества новостей в датасете с группировкой по **speaker**

3 Постановка задачи

Цель данной работы — рассмотреть различные методы решения задачи прогнозирования фальшивых новостей с помощью технологий глубокого обучения. А также выяснить, на каком наборе данных результаты прогнозирования выше всего.

Цели:

- Описать все необходимые инструменты для разработки;
- Описать подходы нейронных сетей, которые используются в работе;
- Построить модели для набора данных: **Statement**, **Statement** + **Meta-Data**, **Statement** + **Meta-Data** + **Justification**;
- Построить схемы полученных архитектур;
- Сравнить результаты.

4 Инструменты для разработки

Для реализации данной работы были использованы следующие инструменты:

- **Python** — высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода. Разведочный анализ данных и моделирование было осуществленно на этом языке программирования;
- **PyTorch** — библиотека машинного обучения для языка Python с открытым исходным кодом, созданная на базе Torch. Используется для решения различных задач:компьютерного зрения и обработки естественного языка. С помощью этой библиотеки было реализовано моделирование нейронных сетей;
- **Pandas** — библиотека языка Python для анализа и обработки данных. Вся работа с данными была частично организована на Pandas;
- **Google Colaboratory** — это облачный сервис, направленный на упрощение исследований в области машинного и глубокого обучения. В этом сервисе производилось обучение нейронных сетей с использованием параллельных вычислений на GPU;
- **GitHub** — веб-сервис для хостинга проектов и их совместной разработки, основанный на системе контроля версий Mercurial и Git. В этом сервисе была организована работа с управлением разработки;
- **Overleaf** — веб-редактор LaTeX. С помощью этого редактора была написана работа.

5 Линейные модели

5.1 Основные концепции машинного обучения

Существует 3 основных раздела машинного обучения:

- Обучение с учителем **(Supervised learning)**;
- Обучение без учителя **(Unsupervised learning)**;
- Обучение с подкреплением **(Reinforcement learning)**.

В этой работе использовано обучение с учителем.

Обучение с учителем бывает двух видов:

- Классификация;
- Регрессия.

Задача состоит в том, чтобы по нашим данным (обучающей выборке) $\mathbb{X} = \{(x_1, y_1), ..., (x_l, y_l)\}$, где $x_1, ..., x_l$ — обучающие объекты, l — их количество, $y_1, ..., y_l$ известные ответы (таргеты), построить алгоритм $a : \mathbb{X} \to \mathbb{Y}$.
Выявление фейковых новостей — это задача классификации.

Можно выделить 2 основных типа задачи классификации:

- Бинарная классификация — $\mathbb{Y} = \{0, 1\}$;
- Многоклассовая классификация — $\mathbb{Y} = \{1...K\}$.

В задаче прогнозирования фальшивых новостей с помощью датасета LIAR-PLUS можно выделить 2 задачи:

1. Классификация новости на факт лжи (бинарная классификаци):
   - $y = 1$, если новость фейк (**is fake = 1**);
   - $y = 0$, если новость не фейк (**is fake = 0**).
2. Классификация новости по 6-ти классам POLITIFACT.COM [22] (многоклассовая классификация):
   - $y = 0$, если **label = barely-true**;
   - $y = 1$, если **label = false**;
   - $y = 2$, если **label = half-true**;
   - $y = 3$, если **label = mostly-true**;
   - $y = 4$, если **label = pants-fire**;
   - $y = 5$, если **label = true**.

5.2 Логистическая регрессия

Линейная модель классификации:

$$a(x) = sign\left(\langle w, x\rangle + w_0\right) = sign\left(\sum_{j=1}^{d} w_j x_j + w_0\right) = sign\langle w, x\rangle$$

где x - вектор признаков, а w - вектор весов модели, w_0 — смещение.

Чтобы построить линейную модель классификации, нам надо ввести **критерий качества модели**:

$$Q(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}[\underbrace{y_i\langle w, x_i\rangle}_{M_i} < 0] \leq \frac{1}{\ell}\sum_{i=1}^{\ell}\tilde{L}(y_i\langle w, x_i\rangle) \to \min_w$$

Где M_i - это отступ классификации на объекте i, $\tilde{L}(y_i\langle w, x_i\rangle)$ — пороговая функция

Критерий качества логистической регрессии — $\tilde{L}(M) = \log\left(1 + e^{-M}\right)$ (Logistic Loss) (Рис. 6)

Логистическая регрессия:

- $$Q(a, X) = \sum_{i=1}^{\ell}\log\left(1 + \exp(-y_i\langle w, x_i\rangle)\right)$$
- Логистическая регрессия может оценивать **вероятности принадлежности к классам**: $p(y = 1|x) = \frac{1}{1+\exp(-\langle w,x\rangle)}$

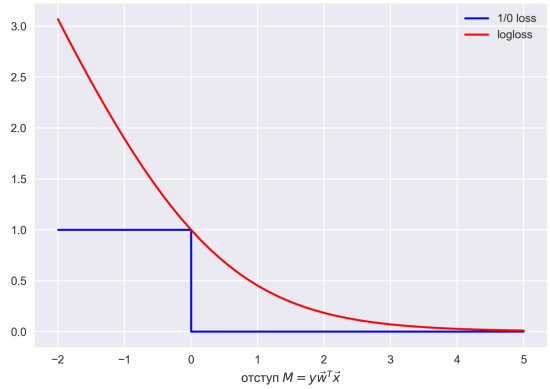

Рис. 6: Logistic Loss

5.3 Обучение моделей машинного обучения

Обучение логистической регрессии происходит с помощью метода оптимизации **градиентный спуск(Gradient Descent)** [7]
Для обучения линейных моделей классического алгоритма градиентного спуска вполне достаточно. Но когда дело доходит до обучения нейронных сетей, то для эффективной оптимизации требуются другие модцификации градиентного спуска.

Во первых, при большом объёме данных, процедура оценивания градиента функции в текущей точке очень долгая. Чтобы ускорить этот процесс, в нейронных сетях применяется способ **Mini-batch Gradient Descent** [21].

Во-вторых, в глубоких нейронных сетях с классической оптимизацией градиентного спуска можно очень легко сойтись либо на плато, либо в локальный минимум. Чтобы избежать этого, были предложены новые адаптивные модификации градиентного спуска. Один из таких алгоритмов — **Adam** [18].
В данной работе обучение нейронных сетей производится с помощью алгоритма **Adam**

5.4 Машинное обучение в задачах обработки естественного языка

Существует несколько типов задач обработки естественного языка **(Natural Language Processing, NLP)**. Основные из них:

- Генерация текста
- Классификация текста

Примером задачи генерации текста могут служить диалоговоые системы или машинные переводчики. Задача детекции фейковых новостей — это задача классификации текста.

Задачи NLP можно решать с помощью классических методов машинного обучения. Основная сложность заключается в преобразовании текста в численный формат. Но когда количество данных слишком большое, то методы обработки текстов по типу Латентного-семантического анализа работают неэффективно. Поэтому в современном NLP активно используют технологии глубокого обучения.

6 Нейронные сети

6.1 Полносвязные нейронные сети

Нейронные сети [31] — это пример математической модели, которая получилась в результате исследований человеческого мозга. **Полносвязные нейронные сети (или Многослойный перцептрон MLP)**(Рис. 8) — это конструкция взаимодействий нескольких моделей персептронов (Рис. 7).

Персептрон состоит из трёх слоёв: **входной, скрытый, выходной**. Каждый слой состоит их отдельный нейронов. В скрытом слое у каждого нейрона своя взвешенная сумма входного вектора, которая проходит через активационную функцию. Количество выходных нейронов может задаваться вручную. Соответственно, полносвязная нейронная сеть состоит из n последовательных персептронов, где выходы первого персептрона являются входом для второго и т.д.

Алгоритм оптимизации нейронных сетей — **Метод обратного распространения ошибки (Backpropagation)** [10]

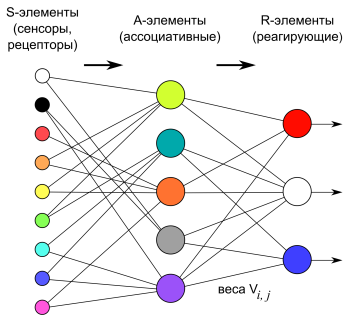

Рис. 7: Математическая модель MLP

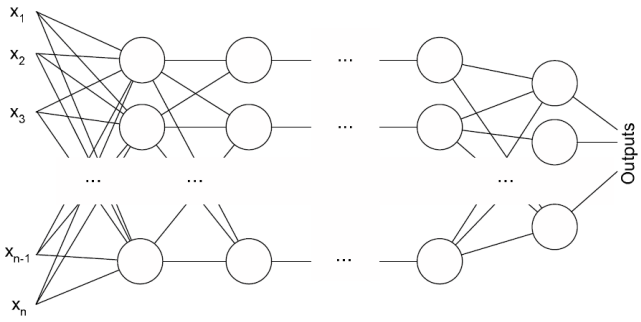

Рис. 8: Модель многослойного персептрона

6.2 Батч-нормализация

Чтобы улучшить и ускорить сходимость оптимизации глубоких нейронных сетей, требуется совершать **Батч-нормализацию** [15] отдельно после каждого слоя. Нормализация ускоряет оптимизацию градиентных методов. Батч-нормализация является слабой формой регуляризацией нейронных сетей. В своей работе я использую Батч-нормализация.

6.3 Регуляризация нейронных сетей

Нейронные сети, как и классические модели машинного обучения, склонны к **переобучению**. Переобучение — это ситуация при которой модель перестаёт искать общие закономерности и зависимости в данных и просто запоминает объекты и таргеты обучающей выборки. Переобучение вызывает падение **обобщающей способности**. То есть, ошибка на train падает, в то время как ошибка на val и test увеличивается. Регуляризация помогает бороться с переобучением.

6.3.1 L_1-L_2 регуляризация

Классическая регуляризация выглядит так: $Q_\alpha(w) = Q(w) + \alpha R(w)$, где R(w) норма весов модели.

- L_1 **регуляризация** — $R(w) = \|w\|_1 = \sum_{i=1}^{d} |w_i|$;
- L_2 **регуляризация** — $R(w) = \|w\|_2 = \sum_{i=1}^{d} w_i^2$.

Разница заключается в том, что L_1 регуляризация склонна к отбору признаков (Рис. 9). То есть, если вес какого-нибудь параметра изначально маленький, то для L_1 выгоднее всего этот вес заменить на 0, в то время как для L_2 выгоднее уменьшать большие веса. В своей работе я использую L_2 регуляризацию в оптимизаторе Adam

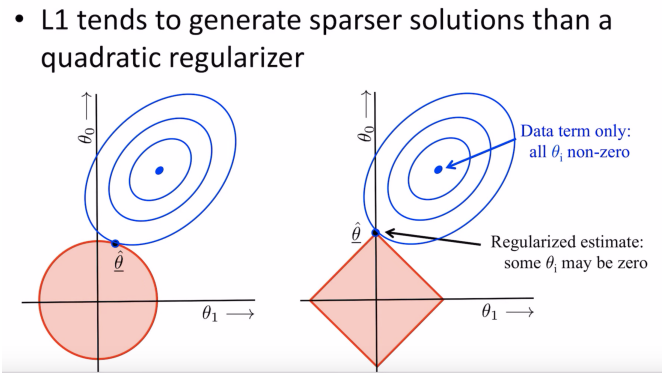

Рис. 9: L_1-L_2 регуляризация

6.3.2 Dropout

Dropout [25] — это мощный способ регуляризации нейронных сетей. Логика заключается в случайном выключении нейронов в процессе обучения (Рис. 10). Dropout эквивалентен усреднению 2^n моделей.

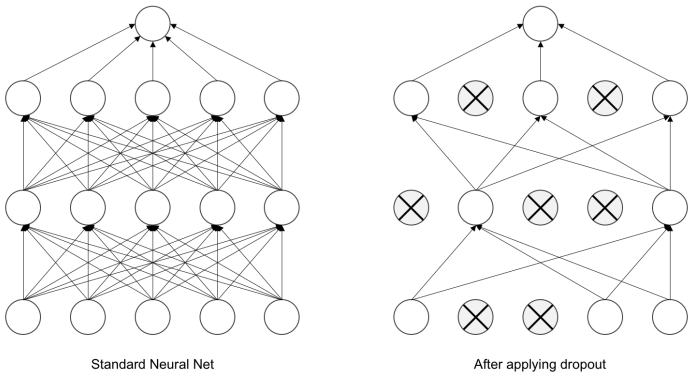

Рис. 10: Dropout

6.4 Эмбеддинговое представление текста

Компьютер не умеет напрямую работать с буквами. По этой причине нужно закодировать слова в цифры. Можно закодировать все тексты с помощью bag-of-words, то есть сделать подсчёт вхождений всех слов в документ. Но цель состоит в том, чтобы не просто перевести слова в цифры, но ещё и сохранить смысловую нагрузку и важность слов.

6.4.1 TF-IDF

TF-IDF (TF — term frequency, IDF — inverse document frequency) — статистическая мера, используемая для оценки важности слова в контексте документа. Существуют различные способы определения точных значений TF и IDF.

- TF (частота слова) — $TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$
- IDF(обратная частота документа) — $IDF_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|+1}$
- TF-IDF — $TFIDF_{i,j} = TF_{i,j} * IDF_i$

TF показывает важность слова в пределах одного документа. IDF — инверсия частоты, с которой некоторое слово встречается в документах коллекции. Таким образом с помощью TF-IDF можно преобразовать текст и применить либо классическую модель машинного обучения, либо нейронную сеть.

6.4.2 Word2Vec

Эмбеддинги (векторные представления) слов нужны для того, чтобы понимать как соотносятся между собой слова в языке. Такое представление слов помогает закодировать смысловую нагрузку слов. Если построить эмбеддинги для всех слов текста, то получится эмбеддинговое пространство.
Один из способов получить эмбеддинговое пространство — Word2Vec[19]. Модель Word2Vec строит такие векторные представления, чтобы векторы похожих слов оказывались близки по косинусному расстоянию. Похожими считаются слова, которые часто встречаются в одном и том же контексте. Существует 2 вида построения эмбеддингов

Word2Vec: CBoW (Continuous Bag of Words) и SkipGram.
Необязательно всегда обучать свои эмбеддинги для текстов из задачи. Можно в качестве начального приближения взять готовые предобученные эмбеддинги. Например, для задачи прогнозирования фейковых новостей можно использовать готовые эмбеддинги Word2Vec, обученные на корпусе текстов GoogleNews

6.5 Разновидности нейронных сетей

6.5.1 Свёрточные нейронные сети

Свёрточные нейронные сети (Convolutional Neural Networks) пользуются большой популярностью среди задач компьютерного зрения (computer vision). Каждый фрагмент изображения перемножается с матрицей свёртки поэлементно, а результат суммируется и записывается в аналогичную позицию выходного изображения. CNN, в отличие от MLP, может находить специфические фрагменты в данных. Если поменять местами входные нейроны MLP, то результат не изменится. А результат CNN после такой перестановки может измениться сильно.

В задачах NLP используются **одномерные свёртки (1d convolutional)** [17](Рис. 11). Ширина такой свёртки равняется длине эмбеддинга слова.

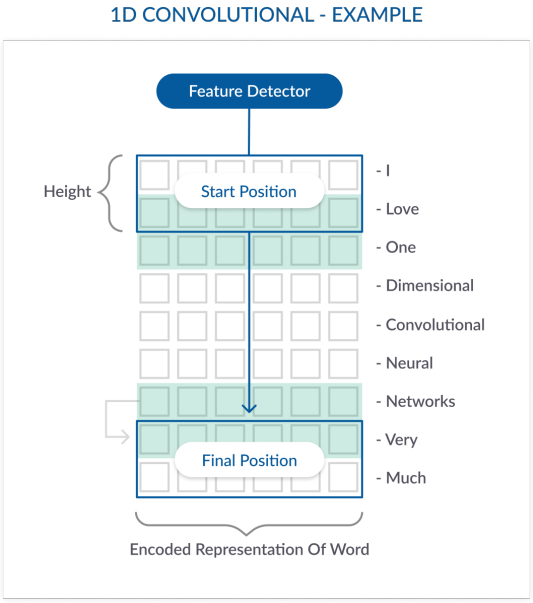

Рис. 11: Одномерная свёртка

6.5.2 Рекуррентные нейронные сети

Рекуррентные нейронные сети (Recurrent Neural Networks, RNN) [24] — это класс нейронных сетей, предназначенный для работы с последовательностями (временные ряды, звук, текст). RNN в NLP работает следующем образом: RNN последовательно (рекуррентно) считывает слова и таким образом запоминает то, что было в тексте (Рис. 11). После каждой итерации RNN передаёт вектор скрытого состояния h_i (hidden state) следующей RNN. Огромный плюс RNN — работа с последовательностями любой длины. Обучение RNN производится с помощью небольшой модификации классического Backpropagation под названием **BackPropagation Through Time**[28]

Из-за специфики алгоритма BackPropagation Through Time у RNN есть 2 больших недостатка [20]:

- **Взрыв градиентов**
- **Затухание градиентов**

С проблемой взрыва градиента мы можем бороться с помощью активационной функцией *tanh* или с помощью gradient clipping. Но избавиться от затухания в классическом RNN, к сожалению, невозможно. Модель просто "забывает"то, что было в самом начале текста.

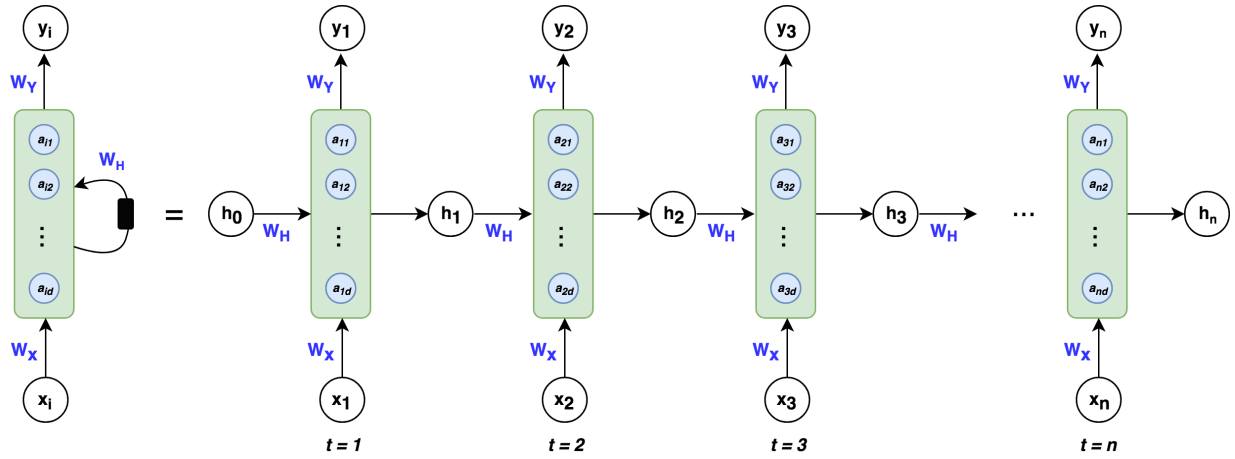

Рис. 12: Recurrent Neural Network

6.5.3 Долгая краткосрочная память

Долгая краткосрочная память (Long short-term memory, LSTM) [24] — это модификация классической рекуррентной нейронной сети. LSTM противодействует затуханию градиента. Вместо одного вектора скрытого состояния у LSTM их 2(c_i и h_i), которые отвечают за долгосрочную и краткосрочную память соответственно. Если записать чему равно c в момент t, то мы получим: $c_t = c_{t-1} + i_t \odot c'_t$. Именно благодаря этому градиенты так сильно не затухают, как в RNN. В каждый момент времени t будет $\frac{\partial c_t}{\partial c_{t-1}} = 1$. Это называется "карусель константной ошибки". У LSTM есть 3 гейта (forget gate, input gate, output gate) (Рис. 13), которые помогают модели отбирать и запоминать важную информацию с предыдущих итераций

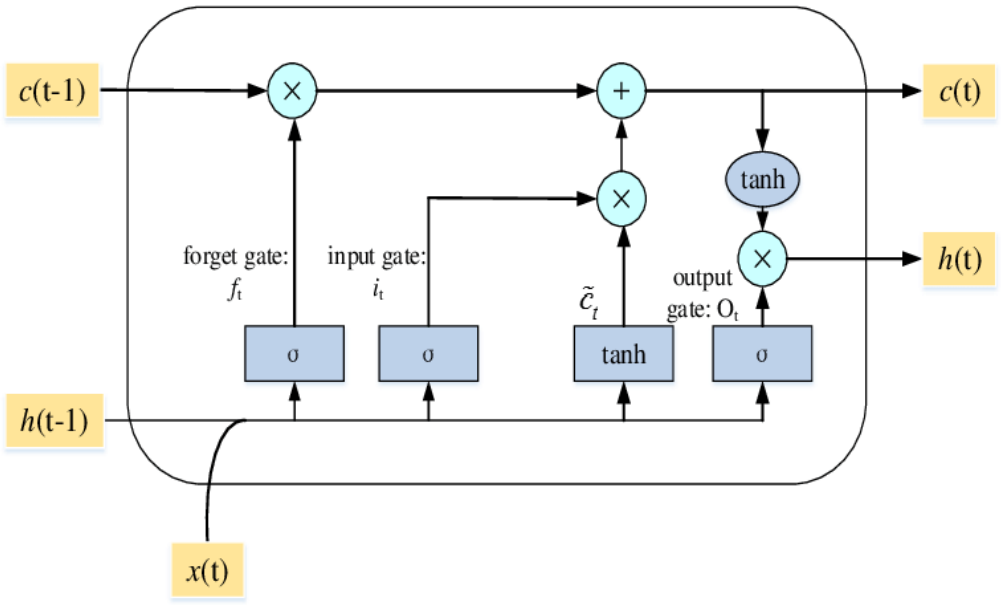

Рис. 13: LSTM

6.5.4 Bidirectional LSTM

Хоть LSTM и борется с затуханием градиентов, но всё равно информация в самом начале последовательности может забываться. Получается, если в самом начале предложения были очень важные словосочетания и длина предложения большая, то модель их может и не учитывать в конечном ответе. **Bidirectional LSTM** (Рис. 14) решает эту проблему. Параллельно с основной LSTM можно параллельно запустить вторую LSTM, которая будет читать последовательность с конца. И конечные вектора скрытых состояний конкатенируются. Таким образом модель будет одинаково учитывать все слова, независимо от их местоположений.

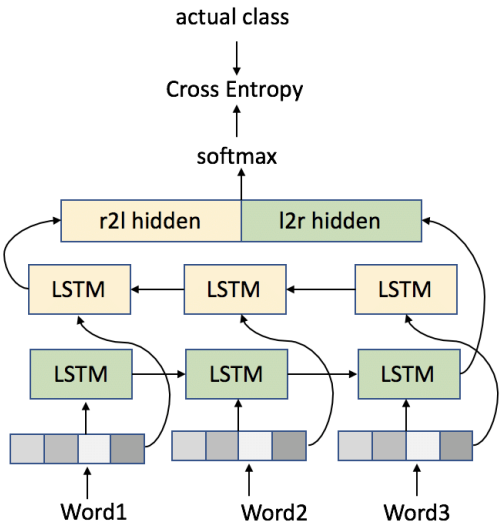

Рис. 14: Bidirectional LSTM

6.5.5 Механизмы внимания (Attention)

Идея механизма Attention состоит в том, чтобы подсказать нейронной сети на какую часть данных нужно обратить своё внимание, чтобы на них произвести детальную обработку. Когда специалист факт-чеккинга исследует новость, он выделяет самые важные элементы всего текста, по которым принимает решение о классификации. Очень часто в тексте есть много шумных слов, которые практически не влияют на то, к какому классу относится этот текст.

Изначально Attention применили к seq2seq моделям машинного перевода[5] (Рис. 15). Аналогично можно применить механизм Attention к задаче классификации с помощью Bi-LSTM[9] (Рис. 16).

$$\alpha(i) = \frac{\exp((h_i, h_f))}{\sum_{j=1}^{n} \exp((h_j, h_f))}$$

$$\mathrm{c} = \sum_{i=1}^{n} \alpha_t(i) h_i$$

h_i - текущее скрытое состояние, h_f - финальное скрытое состояние, $\alpha(i)$ - вероятность "важности"скрытого состояния i;

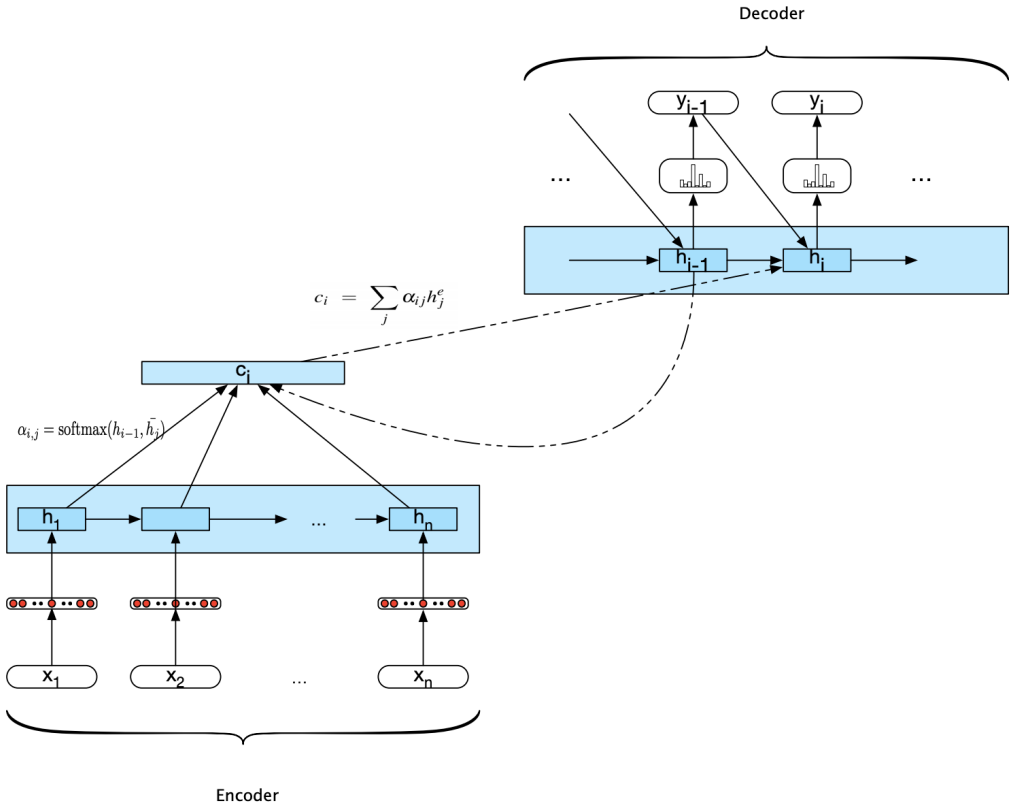

Рис. 15: Seq2Seq модель с механизмом Attention

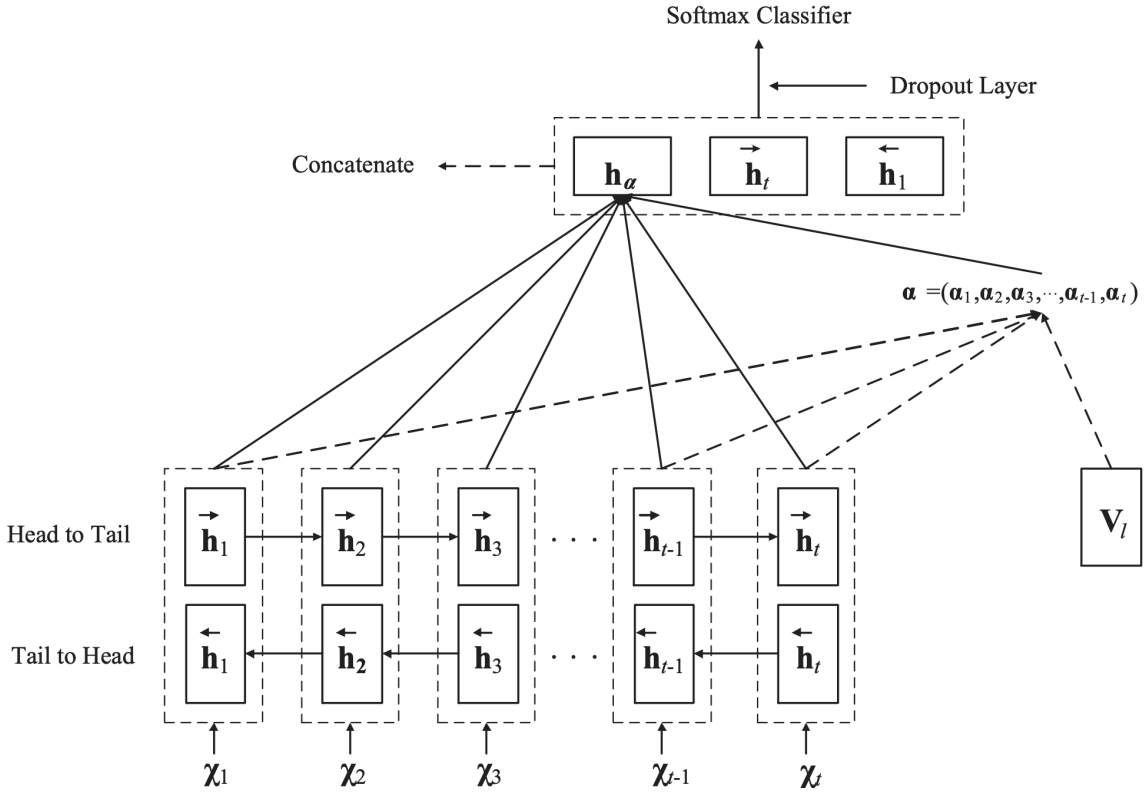

Рис. 16: Bi-LSTM Attention для классификации

7 Эксперименты

Датасет LIAR-PLUS[2] состоит из трёх частей: train, val, test. Основное сравнение моделей происходит по метрике Accuracy на test выборке. Accuracy считается, как доля правильных ответов модели. На данных LIAR-PLUS принято считать именно метрику Accuracy[29][2][27][23]. Эксперименты были произведены на трёх наборах данных: Statement, Statement+Meta-Data, Statement+Meta-Data+Justification. Параллельно решаются 2 задачи: бинарная классификация и многоклассовая классификация. В качестве бейзлайна были сделаны 2 модели: выбор большинства и логистическая регрессия на TF-IDF[16]. Все числа были заменены на токен /NUM/. Нейронные сети обучались с оптимизатором Adam[18] с параметрами 0.9 и 0.999 на протяжении 10 эпох. Размер батча обучения — 128. Loss — кросс-энтропия. Начальный learning rate составил 0.001. Каждую эпоху learning rate рекурсивно умножается на 0.4. Во всех моделях нейронных сетей используется L_2 регуляризация с параметром 0.00001. В качестве активационной функции используется ReLU[1]. Между слоями используется слой BatchNorm[15] и Dropout[25] с параметром 0.5. В качестве начального приближения эмбеддингов были использованы предобученные эмбеддинги Word2Vec GoogleNews[19][12]. Размер эмбеддинга 300. В процессе обучения эмббединги обучаются. Весь код экспериментов можно найти по ссылке — `https://github.com/nashikunov/Fake_News_Detection`

В таблице 1 показаны метрики моделирования. Из результатов видно, что модели можно одинаково сравнивать на двух задачах классификации. Добавление дополнительных данных к новостным заголовкам сильно улучшает результат. Самые эффективные модели построены на всех данных с использованием признака justification. В связи с малым объёмом данных, нейронные сети достаточно быстро переобучаются. Из-за этого пришлось применять сильную регуляризацию, а также уменьшение шага обучения в процессе. Самая сильная архитектура на двух задачах классификации — **Bi-LSTM+Attention** (Рис. 21)

Таблица 1: Результаты экспериментов. Метрика Accuracy

Многоклассовая классификация

Models	Val	Test
Statement models		
Majority vote	0.193	0.209
LR TF-IDF	0.260	0.252
Bi-LSTM (Рис. 17)	0.242	0.246
1-d CNN (Рис. 18)	0.227	0.23
Statement+Meta-Data models		
Bi-LSTM (Рис. 19)	0.273	0.257
Statement+Meta-Data+Justification models		
Bi-LSTM (Рис. 20)	0.279	0.266
Bi-LSTM+Attention (Рис. 21)	0.272	**0.267**

Бинарная классификация

Models	Val	Test
Statement models		
Majority vote	0.520	0.563
LR TF-IDF	0.630	0.614
Bi-LSTM (Рис. 17)	0.6	0.631
1-d CNN (Рис. 18)	0.608	0.625
Statement+Meta-Data models		
Bi-LSTM (Рис. 19)	0.639	0.638
Statement+Meta-Data+Justification models		
Bi-LSTM (Рис. 20)	0.637	0.645
Bi-LSTM+Attention (Рис. 21)	0.643	**0.652**

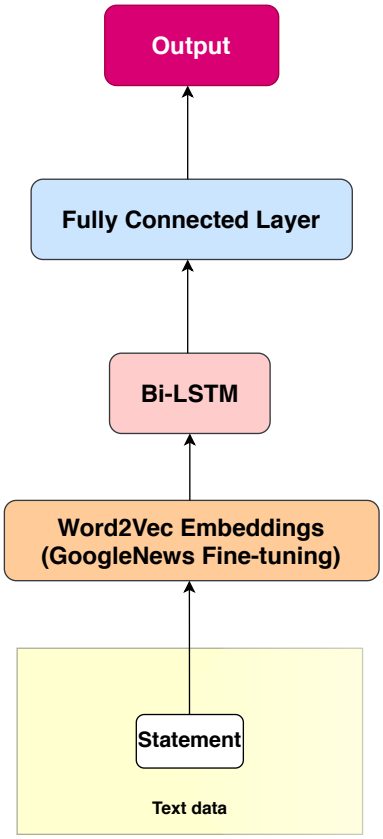

Рис. 17: Модель Bi-LSTM на признаке Statement

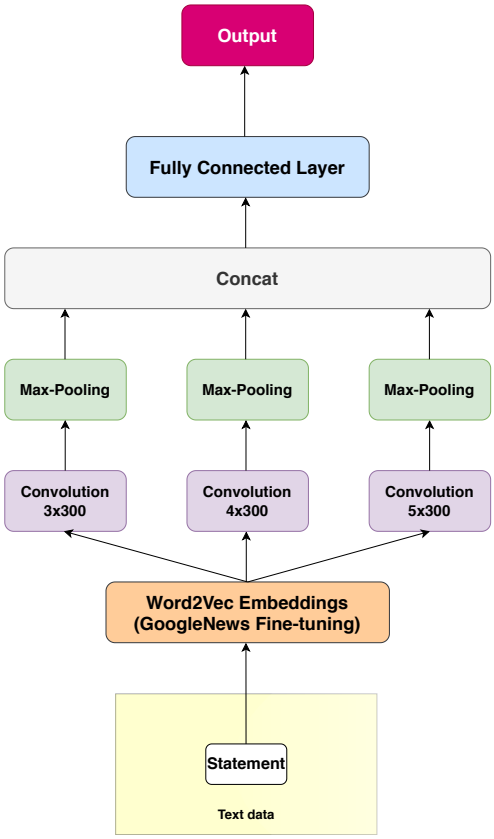

Рис. 18: Модель CNN на признаке Statement

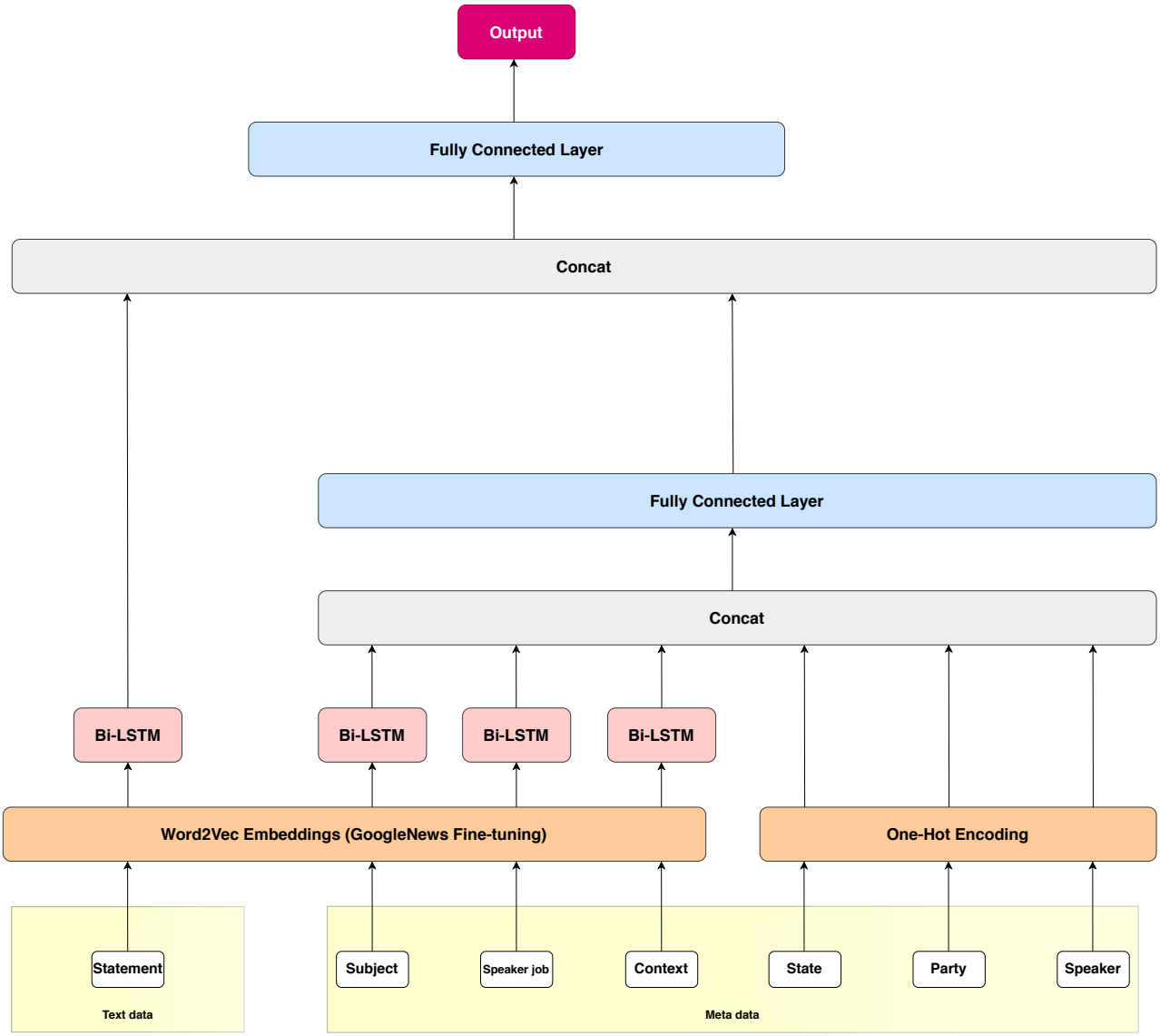

Рис. 19: Модель Bi-LSTM на признаках Statement и Meta-Data

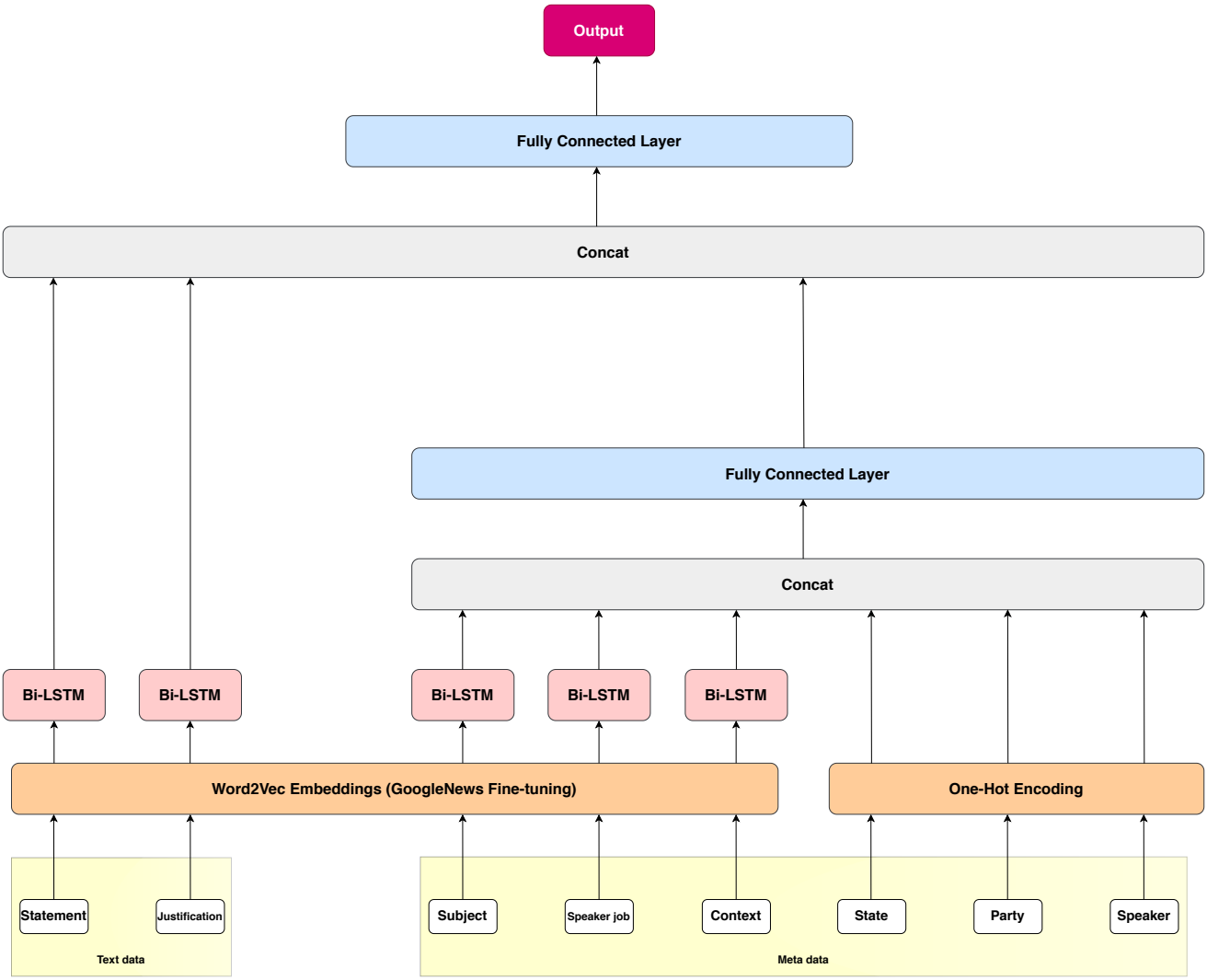

Рис. 20: Модель Bi-LSTM на всех признаках

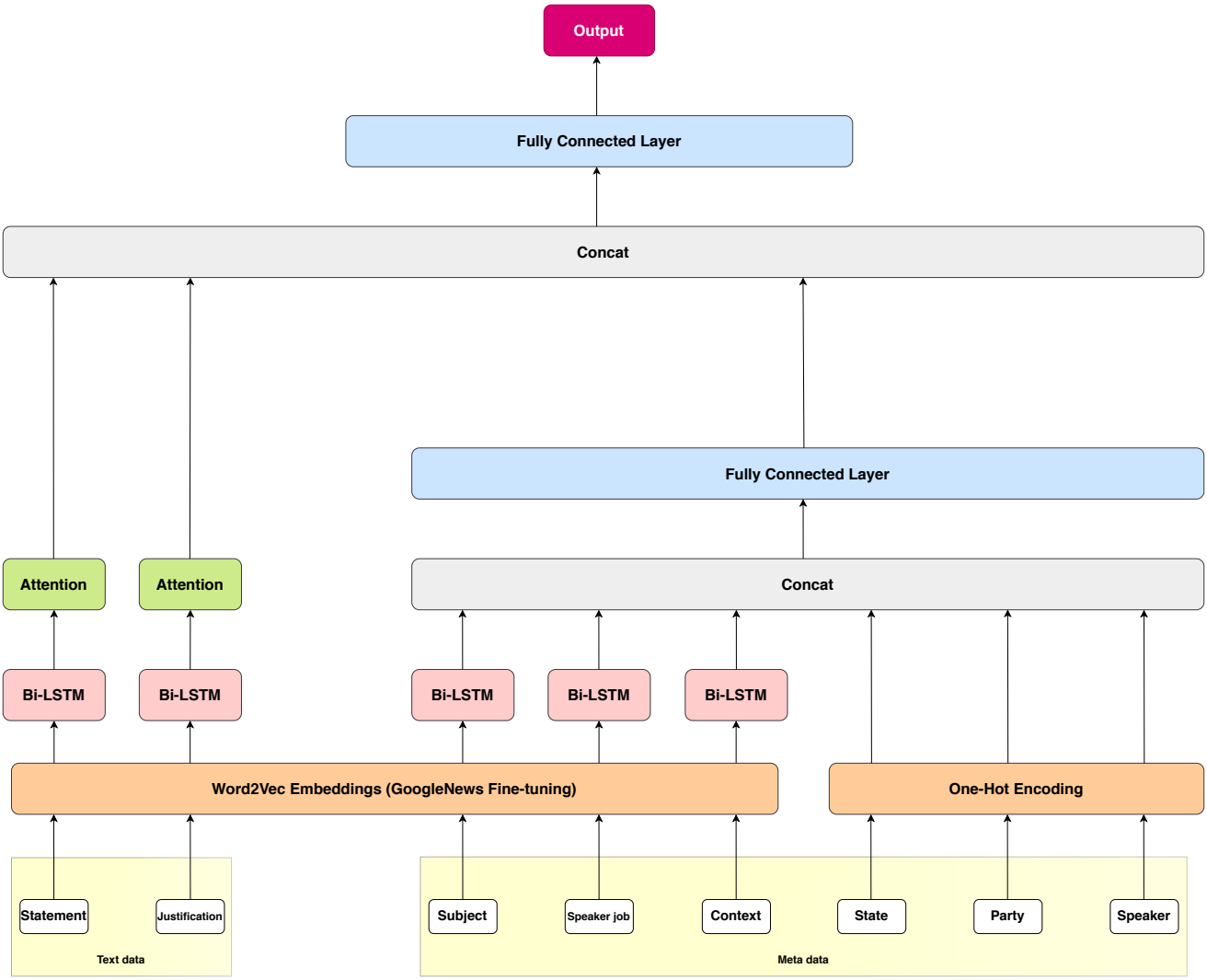

Рис. 21: Модель Bi-LSTM-Attention на всех признаках

8 Выводы

В этой работе были использованы методы глубокого обучения для задачи выявления фейковых новостей. Архитектура нейронной сети **Bi-LSTM+Attention** (Рис. 21) оказалась самой эффективной: Accuracy на бинарной классификации **0.652**, на многоклассовой классификации **0.267** (Таб. 1). Использование дополнительной информации в качестве **justification** сильно увеличивает результат: Accuracy увеличивается на 0.1 в многоклассовой классификации и на 0.14 на бинарной классификации. Этот эффект был также описан в работе[3]. Результаты классификации на 6 классов POLITIFACT.COM[22] хуже, нежели на бинарной классификации на общее наличие лжи в новости. Это связано с тем, что в детальной классификации POLITIFACT.COM присутствует небольшой шум в виде субъективности экспертов факт-чекинга. Для улучшения полученных моделей нужно увеличить обучающую выборку, а также следует расширить круг экспериментов с использованием современных архитектур: Transformer[26] и BERT[8]. Как уже было сказано, подходы глубокого обучения особенно эффективно работают на больших массивах данных.

Результаты показали, что подходы глубокого обучения действительно эффективно работают для задачи прогнозирования ложных новостей. Модель нейронной сети можно использовать для автоматического выявления фейковых новостей в социальных сетях и медиа. Таким образом, можно сильно сократить распространение фальшивых новостей между людьми. Это позволит уменьшить нагрузку на организации, которые занимаются процедурами факт-чекинга. Специалисты будут сосредоточены на действительно нетривиальных кейсах.

Список литературы

[1] Abien Fred Agarap. «Deep learning using rectified linear units (relu)». в: *arXiv preprint arXiv:1803.08375* (2018).

[2] Tariq Alhindi, Savvas Petridis и Smaranda Muresan. «Where is your Evidence: Improving Fact-checking by Justification Modeling». в: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. 2018, с. 85—90.

[3] Tariq Alhindi, Savvas Petridis и Smaranda Muresan. «Where is your Evidence: Improving Fact-checking by Justification Modeling». в: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. 2018, с. 85—90.

[4] Hunt Allcott и Matthew Gentzkow. «Social media and fake news in the 2016 election». в: *Journal of economic perspectives* 31.2 (2017), с. 211—36.

[5] Dzmitry Bahdanau, Kyunghyun Cho и Yoshua Bengio. «Neural machine translation by jointly learning to align and translate». в: *arXiv preprint arXiv:1409.0473* (2014).

[6] BBC. «Social media firms fail to act on Covid-19 fake news». в: ().

[7] Kartik Chandra и др. «Gradient Descent: The Ultimate Optimizer». в: *arXiv preprint arXiv:1909.13371* (2019).

[8] Jacob Devlin и др. «Bert: Pre-training of deep bidirectional transformers for language understanding». в: *arXiv preprint arXiv:1810.04805* (2018).

[9] Changshun Du и Lei Huang. «Text classification research with attention-based recurrent neural networks». в: *International Journal of Computers Communications & Control* 13.1 (2018), с. 50—61.

[10] Fenglei Fan, Wenxiang Cong и Ge Wang. «Generalized backpropagation algorithm for training second-order neural networks». в: *International journal for numerical methods in biomedical engineering* 34.5 (2018), e2956.

[11] William Ferreira и Andreas Vlachos. «Emergent: a novel data-set for stance classification». в: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. 2016, с. 1163—1168.

[12] Google. *GoogleNews Word2Vec*. 2013. URL: `https://code.google.com/archive/p/word2vec/`.

[13] Zhen Hai и др. «Deceptive review spam detection via exploiting task relatedness and unlabeled data». в: *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016, с. 1817—1826.

[14] Zhen Hai и др. «Deceptive review spam detection via exploiting task relatedness and unlabeled data». в: *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016, с. 1817—1826.

[15] Sergey Ioffe и Christian Szegedy. «Batch normalization: Accelerating deep network training by reducing internal covariate shift». в: *arXiv preprint arXiv:1502.03167* (2015).

[16] Thorsten Joachims. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* тех. отч. Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[17] Yoon Kim. «Convolutional neural networks for sentence classification». в: *arXiv preprint arXiv:1408.5882* (2014).

[18] Diederik P Kingma и Jimmy Ba. «Adam: A method for stochastic optimization». в: *arXiv preprint arXiv:1412.6980* (2014).

[19] Tomas Mikolov и др. «Efficient estimation of word representations in vector space». в: *arXiv preprint arXiv:1301.3781* (2013). eprint: 1301.3781.

[20] Razvan Pascanu, Tomas Mikolov и Yoshua Bengio. «On the difficulty of training recurrent neural networks». в: *International conference on machine learning*. 2013, с. 1310—1318.

[21] Michael P Perrone и др. «Optimal Mini-Batch Size Selection for Fast Gradient Descent». в: *arXiv preprint arXiv:1911.06459* (2019).

[22] *PolitiFac*. URL: https://www.politifact.com/.

[23] Ekagra Ranjan. «"Fake News Detection by Learning Convolution Filters through Contextualized Attention». в: (2019).

[24] Alex Sherstinsky. «Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network». в: *arXiv preprint arXiv:1808.03314* (2018).

[25] Nitish Srivastava и др. «Dropout: a simple way to prevent neural networks from overfitting». в: *The journal of machine learning research* 15.1 (2014), с. 1929—1958.

[26] Ashish Vaswani и др. «Attention is all you need». в: *Advances in neural information processing systems*. 2017, с. 5998—6008.

[27] William Yang Wang. «"liar, liar pants on fire": A new benchmark dataset for fake news detection». в: *arXiv preprint arXiv:1705.00648* (2017).

[28] Paul J Werbos. «Backpropagation through time: what it does and how to do it». в: *Proceedings of the IEEE* 78.10 (1990), с. 1550—1560.

[29] Jiawei Zhang, Bowen Dong и S Yu Philip. «Fakedetector: Effective fake news detection with deep diffusive neural network». в: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE. 2020, с. 1826—1829.

[30] С. В. Куликов. «Ходынская катастрофа». в: *Большая российская энциклопедия* (1995).

[31] Сергей Николенко, Артур Кадурин и Екатерина Архангельская. *Глубокое обучение*. "Издательский дом Питер", 2017.