**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное автономное образовательное учреждение высшего образования**

**«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»**

ИНСТИТУТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И

ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

КАФЕДРА ТЕХНОЛОГИЙ ПРОГРАММИРОВАНИЯ

Направление: 09.03.03 – «Прикладная информатика»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**Разработка алгоритма обработки естественного языка при помощи машинного обучения с использованием нейросетей BERT**

**Работа завершена:**

Студент 4 курса

группы 09-852

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_ 2022 г. \_\_\_\_\_\_\_\_\_\_\_\_ Саидмуродов С.С.

**Работа допущена к защите:**

Научный руководитель

старший преподаватель

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_ 2022 г. \_\_\_\_\_\_\_\_\_\_\_\_ Прокопьев Н.А.

Заведующий кафедрой

канд. экон. наук, доцент

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_ 2022 г. \_\_\_\_\_\_\_\_\_\_\_\_ Вахитов Г.З.

Казань – 2022

**Оглавление**

[Введение 3](#_Toc94970214)

[1. Формирование требований 7](#_Toc94970215)

[1.1. Изучение предметной области машинного обучения и нейронных сетей 7](#_Toc94970216)

[1.2. Анализ йподходов к обработке естественного языка 7](#_Toc94970217)

[1.3. Техническое задание 7](#_Toc94970218)

[2. Проектирование платформы для обучения нейросетевой модели 8](#_Toc94970219)

[2.1. Используемые модели нейронных сетей 8](#_Toc94970220)

[2.2. Формализация алгоритмов основного анализа текста 8](#_Toc94970221)

[2.3. Структура платформы для обучения нейросетевой модели 8](#_Toc94970222)

[3. Экспериментальная реализация алгоритма анализа текста 9](#_Toc94970223)

[3.1. Подготовка данных для обучения 9](#_Toc94970224)

[3.2. Реализация платформы для обучения нейросетевой модели 9](#_Toc94970225)

[3.3. Результаты машинного обучения 9](#_Toc94970226)

[4. Тестирование по результатам обучения модели для основного анализа текста 10](#_Toc94970227)

[4.1. Экспериментальная оценка результатов обучения 10](#_Toc94970228)

[4.2. Целевой оценочный тест реализации алгоритма 10](#_Toc94970229)

[Заключение 11](#_Toc94970230)

[Список использованных источников 12](#_Toc94970231)

Введение

Для понимания текста людьми и программами искусственного интеллекта (ИИ) часто требуется правильное использование заглавных букв и знаков препинания. Простые предложения, состоящие из нескольких слов, и люди, и ИИ могут читать и обрабатывать, даже если использование заглавных букв или знаков препинания отсутствует или неправильно. Например, в случае голосовых команд обработка обычно выполняется только для распознанных слов нижнего регистра. Однако, если объем текста, подлежащего анализу, увеличивается, например, до целых абзацев или страниц, то даже для человека быстрое понимание его смысла становится сложной задачей. Это было изучено Джонсом и др. (2003), которые проанализировали влияние заглавных букв и пунктуации на читаемость расшифровки речи в текст.

Ранние работы рассматривали пунктуацию только как подсказки с точки зрения читателя к возможным просодическим характеристикам и паузам текста (Markwardt, 1942). Нанберг (1990) утверждает, что пунктуация играет гораздо большую роль. Кроме того, знаки препинания классифицируются как разграничивающие, разделяющие и устраняющие неоднозначность. Некоторые метки, такие как запятая, могут принадлежать к нескольким категориям, поскольку они могут выполнять несколько ролей. Джонс (1994) доказывает, что «для более длинных предложений реального языка грамматика, использующая знаки препинания, значительно превосходит аналогичную грамматику, которая их игнорирует». Основываясь на этом и других подобных выводах, современные языковые модели считают пунктуацию частью своего словарного запаса. Сюда входят новейшие модели, такие как BERT, ELMo, OpenAI GPT-2 и GPT-3.

Алгоритмы обработки естественного языка (NLP – Natural language processing), такие как распознавание именованных сущностей (NER – Named entitiy recognition), идентификация части речи, анализ зависимостей, машинный перевод (MT – Machine translation), используют заглавные буквы в качестве признаков обрабатываемого в данный момент слова, в то время как пунктуация используется. как признаки для соседних слов. Например, Stanford Named Entity Recognizer рассматривает признаки на основе формы слова. Это означает построение представления слова на основе типа символов, встречающихся в слове. Было предложено несколько алгоритмов представления формы слова, но общая идея состоит в том, чтобы закодировать прописную букву определенным символом, скажем, «X», строчную букву «x» и цифру «d». В этом случае слово типа «McDonald» станет «ХхХхххх». Работа любых таких алгоритмов возможна только в том случае, если слова правильно представлены в виде прописных и строчных букв.

Особое внимание следует уделить системам автоматического распознавания речи (ASR – Automatiс speech recognition). Первичный вывод таких систем обычно состоит из необработанного текста с использованием одного и того же регистра (нижнего, либо верхнего регистра) и без знаков препинания. В таких ситуациях перед применением дальнейших алгоритмов NLP требуется дополнительная предварительная обработка, чтобы восстановить правильный регистр букв и пунктуацию. Их иногда называют «богатыми транскрипциями». Одна из первых инициатив, касающихся автоматической расширенной транскрипции разговорной речи, началась в 2002 году в контексте программы DARPA «Эффективное, доступное повторное использование речи в текст» (EARS – Effective, Affordable, Reusable Speech-to-text), целью которой было улучшение уровня развития алгоритмов обработки языка. С этой целью NIST (National institute of Standards and Technology) выпустил серию обширных наборов данных для оценки транскрипции, чтобы помочь в оценке таких систем.

Несмотря на то, что большой объем данных, требующих восстановления заглавных букв и пунктуации, поступает из систем ASR, необходимо также учитывать и другие источники. Миллер и др. (2000) идентифицируют другие источники шума в виде текста, полученного с помощью оптического распознавания символов (OCR), или в некоторых газетных статьях. В этих случаях отсутствие надлежащей буквы или пунктуации затрагивает не весь текст, а его части. В случае OCR некоторые знаки препинания могут быть не распознаны, в то время как в случае некоторых статей первое предложение или абзац могут быть написаны только заглавными буквами. Кроме того, в случае коротких текстовых сообщений (SMS), чатов, твитов или других действий в микроблогах люди также могут игнорировать правильный регистр и пунктуацию

Одна из трудностей, при создании человеко-компьютерных интерфейсов с использованием естественного языка, с которыми приходится сталкиваться, связана с непоследовательным использованием пользователем пунктуации и использования заглавных букв. В этом контексте многие подходы пытаются скрыть проблему, удаляя все знаки препинания и заглавные буквы как из данных обучения, так и из входных данных полученных во время работы. Кроме того, Coniam (2014) также проанализировал вывод чат-ботов с точки зрения человека, использующего эти программы для изучения английского как второго языка. Он смог определить проблемы с заглавными буквами и пунктуацией даже в произведенном тексте. Тем не менее, он утверждает, что для коротких предложений, создаваемых чат-ботами, «переход на английский язык за счет все более широкого распространения текстовых сообщений делает спорным вопрос о том, можно ли считать эти проблемы важными в наши дни».

# Формирование требований

Sdas

## Изучение предметной области машинного обучения и нейронных сетей

Sd

## Анализ подходов к обработке естественного языка

Asda

## Техническое задание

Asdada

# Проектирование платформы для обучения нейросетевой модели

Asdada

## Используемые модели нейронных сетей

Asdada

## Формализация алгоритмов основного анализа текста

фвфвф

## Структура платформы для обучения нейросетевой модели

Текст рыба

# Экспериментальная реализация алгоритма анализа текста

Текст рыба

## Подготовка данных для обучения

Текст рыба

## Реализация платформы для обучения нейросетевой модели

Текст рыба

## Результаты машинного обучения

рыыыба

# Тестирование по результатам обучения модели для основного анализа текста

Рыба

## Экспериментальная оценка результатов обучения

Рыба

## Целевой оценочный тест реализации алгоритма

Заключение

Рыба

# Список использованных источников

**Păiş Vasile и Tufis Dan** Capitalization and Punctuation Restoration: a Survey [В Интернете] // ResearchGate. - ResearchGate, 21 ноябрь 2021 г.. - 15 01 2022 г.. - https://www.researchgate.net/publication/356456267\_Capitalization\_and\_Punctuation\_Restoration\_a\_Survey.