Министерство образования Республики Беларусь

Учреждение образования

«Белорусский государственный университет
информатики и радиоэлектроники»

УДК 004.738.5:316.472.45

Волосович
Сергей Викторович

ИДЕНТИФИКАЦИЯ ТРОЛЛЕЙ ПО СООБЩЕНИЯМ В СОЦИАЛЬНЫХ СЕТЯХ

АВТОРЕФЕРАТ
диссертации на соискание степени магистра

по специальности 1-40 80 04 – Информатика и технологии программирования

Минск 2024

Работа выполнена на кафедре информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Научный руководитель: **ПИЛЕЦКИЙ Иван Иванович,**
кандидат физико-математических наук, доцент кафедры информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Защита диссертации состоится «28» июня 2024 г. года на заседании Государственной экзаменационной комиссии по защите магистерских диссертаций в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220013, Минск, ул. Платонова, 39, копр. 5, ауд. 308, тел. 293-85-91, e-mail: inform@bsuir.by

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

ВВЕДЕНИЕ

В современном обществе троллинг, как форма агрессивного поведения в Интернете, становится все более распространенной проблемой. Троллинг проникает в различные сферы, влияя на пользователей социальных сетей, политическую атмосферу и экономику.

Развитие технологий и доступность Интернета сделали троллинг массовым явлением, влияющим на общественное мнение и мировую экономику. Дезинформация, ненависть и конфликты, распространяемые через троллинг, могут иметь серьезные последствия.

Эффективным способом решения этой проблемы является применение методов машинного обучения. Эти методы позволяют анализировать данные, выявлять закономерности и прогнозировать поведение. Преимущества машинного обучения включают автоматизацию процесса, обработку больших объемов данных в реальном времени, адаптацию к новым тактикам троллей, снижение ложных срабатываний, масштабируемость и выявление скрытых паттернов.

Диссертационная работа посвящена разработке методов классификации троллинга с использованием нейронных сетей и ПО для систем, позволяющих решать задачи идентификации троллинга в социальных сетях по имени пользователя или тексту сообщений.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Назначение, цель и задачи исследования

Назначением магистерской диссертации является создание инновационной системы классификации текста для приложений, для анализа текстовых сообщений в социальных сетях на предмет троллинга, основанной на большой языковой модели.

Целью магистерской диссертации является разработка алгоритма, который с помощью большой языковой модели позволяет классифицировать текст на предмет троллинга. Также целью является разработка веб-приложения, которое будет предоставлять пользователю возможность классифицировать троллинг по имени пользователя в социальной сети или по тексту сообщений.

Для достижения поставленной цели необходимо решить следующие задачи:

1) изучить предметную область;

2) выполнить анализ больших языковых моделей;

3) определить методы и алгоритмы для обработки и классификации текста;

4) определить функциональные требования к методу классификации;

5) определить архитектуру для реализации метода классификации на основе нейронной сети;

6) произвести сбор и фильтрацию данных из открытых источников для дальнейшего использования;

7) разработать несколько реализаций метода с различными алгоритмами классификации текста;

8) произвести анализ и тестирование реализаций метода, выбрать наиболее точную реализацию для классификации текста;

9) определить и разработать архитектуру приложения для интеграции выбранного метода классификации;

10) реализовать приложение позволяющее идентифицировать троллей по сообщениям в социальных сетях

Объектом исследования являются методы и алгоритмы классификации текста на основе нейронной сети.

Предметом исследования является методы и алгоритмы машинного обучения, большие языковые модели для решения задачи классификации текста сообщений в социальной сети на наличие троллинга.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность использования алгоритмов классификации на основе нейронной сети и больших языковых моделей для решения задачи идентификации троллей по тексту сообщений в социальных сетях.

Личный вклад соискателя

В процессе написания диссертации соискатель проанализировал предметную область, изучил существующие решения, провел ряд исследований, создал и протестировал ПО.

Опубликованность результатов диссертации

По теме диссертации опубликовано 3 работы в сборниках трудов и материалов международной конференций.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трех глав, заключения, списка использованных источников, списка публикаций автора и приложений. В первой главе представлен анализ предметной области, выявлены основные существующие проблемы в рамках тематики исследования, проанализированы методы векторизации и большие языковые модели в NLP. Вторая глава посвящена сбору и обработке данных, выбору и реализации метода классификации текста, а также реализован и проанализирован метод с различными алгоритмами для классификации текста на предмет троллинга. В третей главе разработано приложение с интегрированным методом классификации троллинга и реализованы функции анализа троллинга по имени в социальной сети или по тексту сообщений.

Общий объем работы составляет 66 страниц, из которых основного текста – 40 страниц, 23 рисунка на 20 страницах, 5 таблиц на 5 страницах, список использованных источников из 35 наименований на 2 страницах и 2 приложения на 12 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** проведен анализ предметной области и методы векторизации текста в задачах NLP. Сформулированы основные определения и описание троллинга, проанализирована статистика по троллингу в интернете (рисунок 1, 2).

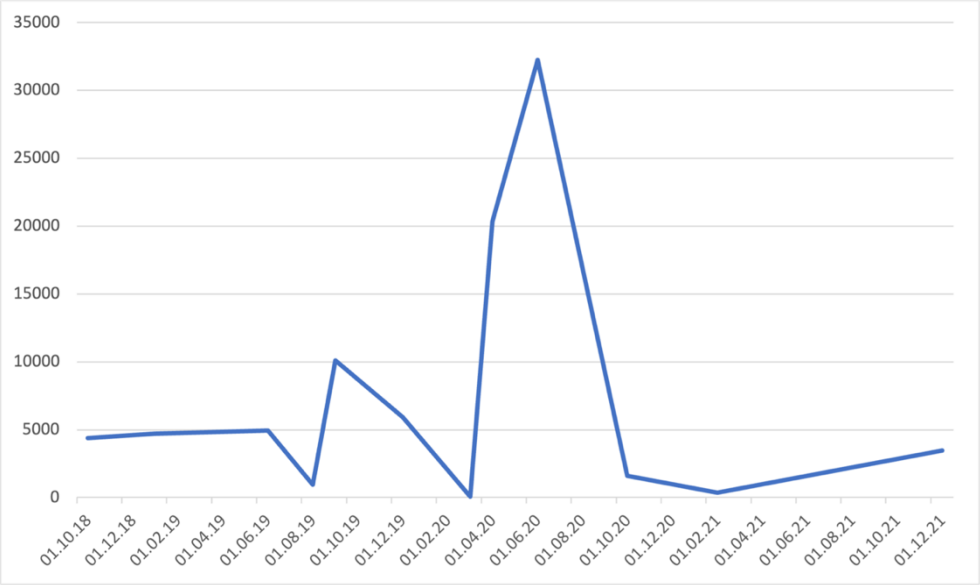

Рисунок 1 – Количество заблокированных пользователей в социальной сети X

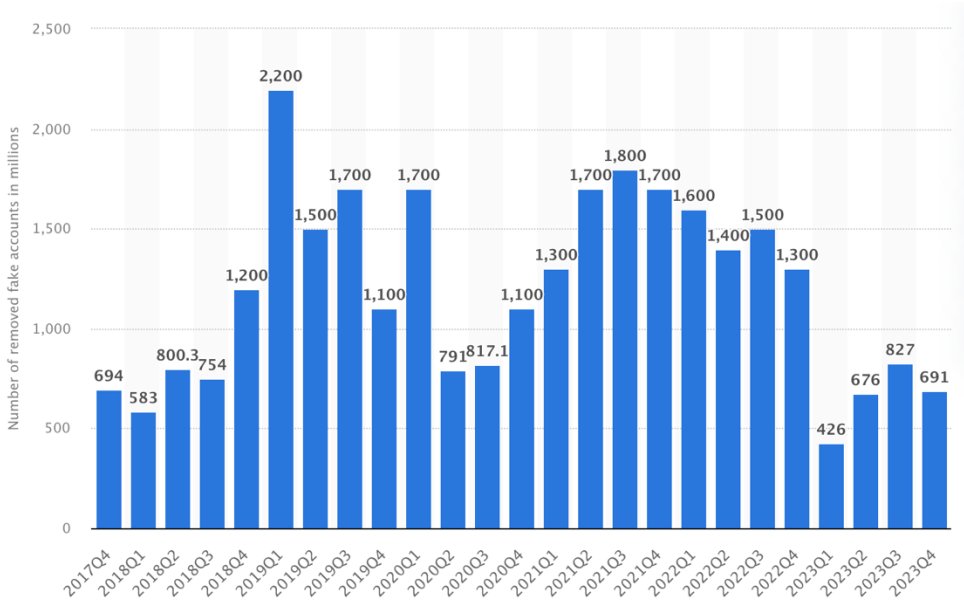

Рисунок 2 – Количество заблокированных фейковых аккаунтов в Facebook

Векторизация является ключевым процессом в обработке естественного языка. Она преобразует текстовые данные в числовые или векторные формы, которые могут быть обработаны алгоритмами машинного обучения. В настоящее время существует несколько популярных методов векторизации:

— мешок слов;
— TF;
— TF-IDF;
— TF-RF;
— TF-IG;
— TF-OR;
— TF-MI;
— TF-ICF;
— Word2Vec;
— Glove.

Однако приведенные методы векторизации уступают большим языковым моделям таким как GPT, BERT, XLNET. Большие языковые модели — это типы нейронных сетей, обученные на огромных объемах текстовых данных для выполнения различных задач, связанных с естественным языком.

Важной особенностью больших языковых моделей является использование трансформеров - сложных архитектур нейронных сетей, которые могут обрабатывать последовательности данных в любом порядке. Это позволяет большим языковым моделям понимать контекст наряду с синтаксисом и грамматикой, что делает их идеальными инструментами для векторизации текста.

Благодаря уникальной архитектуре BERT, основанной на механизме внимания и двунаправленном обучении BERT отлично подходит для решения задач классификации. Процесс начинается с преобразования каждого токена в векторное представление. Затем эти векторы обрабатываются с помощью слоев Transformer, состоящих из слоя внимания (self-attention) и слоя прямого прохода (feed-forward).

Модели, разработанные на базе BERT, сохраняют высокую производительность при значительном уменьшении размера модели и ускорении обучения:

— Distil BERT в 1.67 раза меньше параметров и обучается в 4 раз быстрее, сохраняя 97% производительности BERT;

— RoBERTa использует ту же архитектуру, что и BERT, но обучается на большем объеме данных и показывает лучшие результаты;

— ALBERT в 18 раз меньше параметров, обучается в 1,7 раза быстрее и превосходит по производительности BERT, RoBERTa и DistilBERT.

В качестве наиболее эффективной модели векторизации слов для задач классификации и ограниченных вычислительных ресурсах был выбран DistilBERT.

Вторая глава посвящена реализации метода классификации троллинга на основе нейронных сетей с использованием больших языковых моделей. Разработанная модель была названа “Детектор троллинга”.

Для разрабатываемой модели определены требования: интегрируемость, способность адаптации к различным размерам текстовых сообщений, эффективное использование ресурсов, высокая точность, интерпретируемость результатов, способность к обновлению и адаптации.

Учитывая требования разработана архитектура модели “Детектор троллинга” (рисунок 3):

— входной слой токенизированого текста;

— предварительно обученный слой векторизации данных Distil BERT;

— нейронный средний слой;

— слой исключения;

— выходной слой.

Слой векторизации предоставляет векторное представление каждого слова и их взаимосвязь в n-мерном пространстве. Эти векторные представления затем подаются на вход среднему слою нейронной сети, который передает их через скрытый слой для дальнейшей обработки. После этого применяется слой регуляризации, который помогает предотвратить переобучение модели. Затем векторный вывод пропускается через слой классификации для определения вероятности того, что комментарий является троллингом. В качестве слоя векторизации мы выбрали предобученную модель Distil BERT, поскольку она основана на архитектуре BERT, которая успешно применяется в задачах обработки естественного языка.

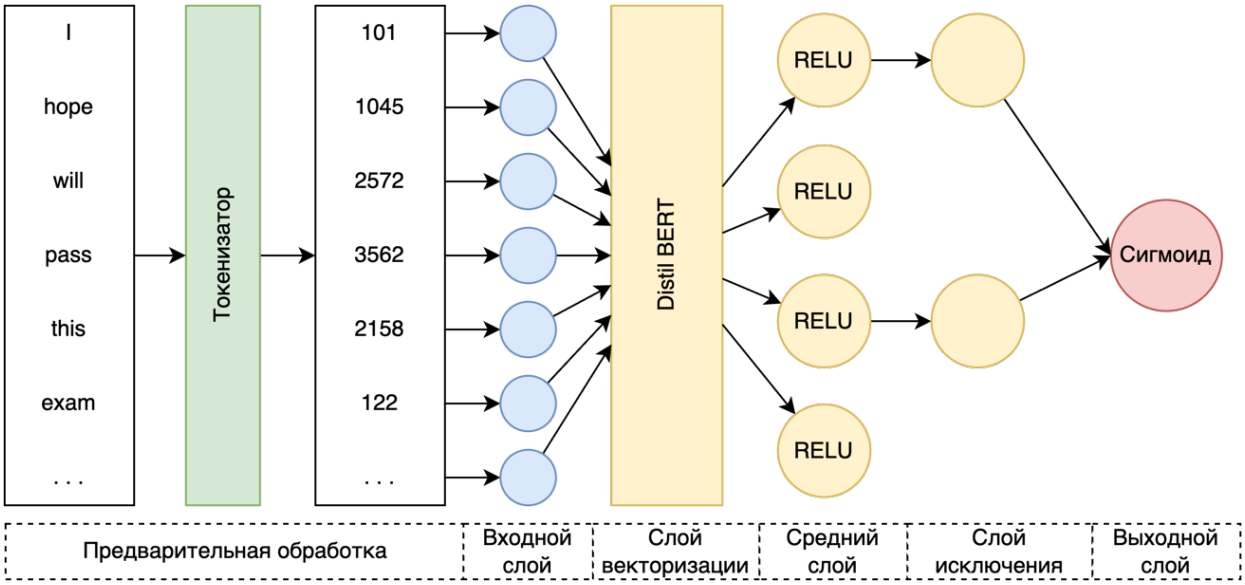

Рисунок 3 – Архитектура модели “Детектор троллинга”

В качестве функции активации для среднего слоя выбран ReLU:

$$R(z) = \max(0, z), \tag{1}$$

где z – входное значение.

Для слоя классификации выбрали сигмоидную функцию:

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad (2)$$

где z - входное значение;

e – значение Эйлера.

В качестве функции метрики была выбрана бинарная accuracy (точность), так как данные, с которыми мы работаем, являются сбалансированными и задача бинарной классификации:

$$Binary\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (3)$$

где *TP* – количество предсказанных истинно положительных примеров;

TN – количество предсказанных истинно отрицательных примеров;

FP – количество предсказанных ложно положительных примеров;

FN – количество предсказанных ложно отрицательных примеров.

С целью нахождения наиболее точной модели “Детектор троллинга”, были обучены со следующими средним слоем и слоем исключения модели:

— полносвязный слой размером 64 нейрона

— сверточный слой размером 64

— рекуррентный слой LSTM размером 40

Для оценки эффективности обученных моделей проведен анализ на данных для тестирования, где оценивалась не только точность (accuracy), но и площадь под ROC-кривой (таблица 1). После проведения анализа моделей было выявлено, что нейронная сеть со средним сверточным слоем обладает наиболее точными предсказаниями при одновременном минимальном потреблении ресурсов.

Таблица 1 – Результаты оценки моделей на данных для тестирования

Нейронная сеть	Точность	AUC
1	2	3
Полносвязная	0.953	0.9903
Сверточная	0.9564	0.9925
Рекурентная	0.9380	0.9855

В **третьей главе** рассмотрена практическая реализация ПО “Антитролль” с интегрированной моделью “Детектор троллинга”. Учитывая существующие типы архитектуры и характер разрабатываемого приложения, было принято решение использовать сервис-ориентированную архитектуру, при которой различные сервисы предоставляют определённые бизнес-функции и взаимодействуют по стандартизированным протоколам (рисунок 4).

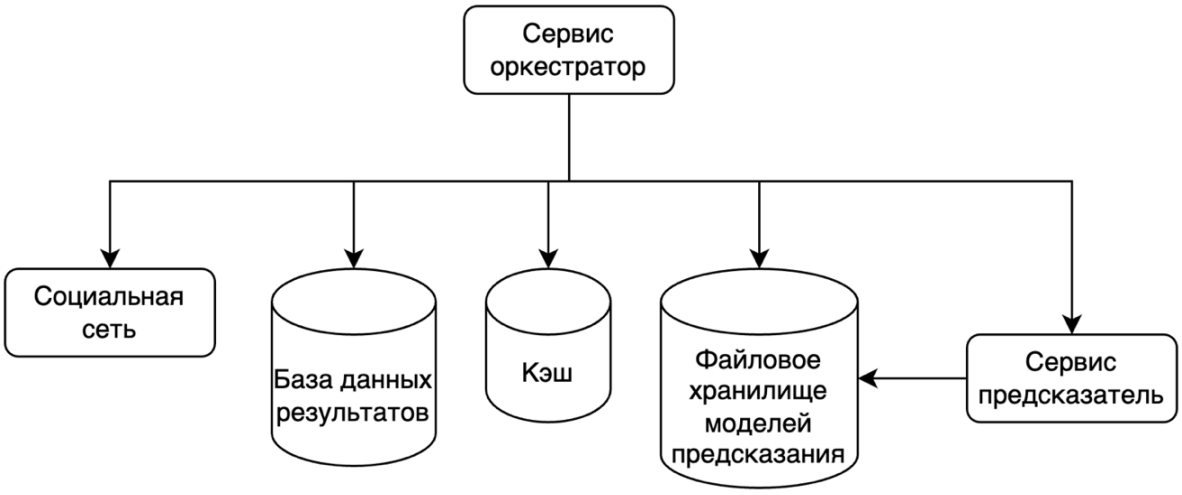

Рисунок 4 – Архитектура компонентов приложения “Антитролль”

Построенная архитектура содержит следующие компоненты, которые выполняют определенную роль:

1 Сервис оркестратор. Служит для получения запросов от пользователей и оркестрации других сервисов для выполнения этих запросов.

2 Сервис предсказатель. Можно назвать ядром данного приложения и выполняет функции для определения, является ли текст троллингом

3 База данных результатов. Служит для хранения результатов предсказаний текста и выполняет следующие функции:

4 Кэш. Служит для увеличения производительности приложения.

5 Файловое хранилище моделей предсказания. Служит для хранения различных версий обученных моделей и их дальнейшего использования в сервисе предсказатель.

6 Социальная сеть. Служит для загрузки последних сообщений проверяемого пользователя для анализа троллинга.

Для разрабатываемого приложения наиболее важным функционалом является классификация сообщения на наличие троллинга, sequence диаграмма функционала представлена на рисунке 5.

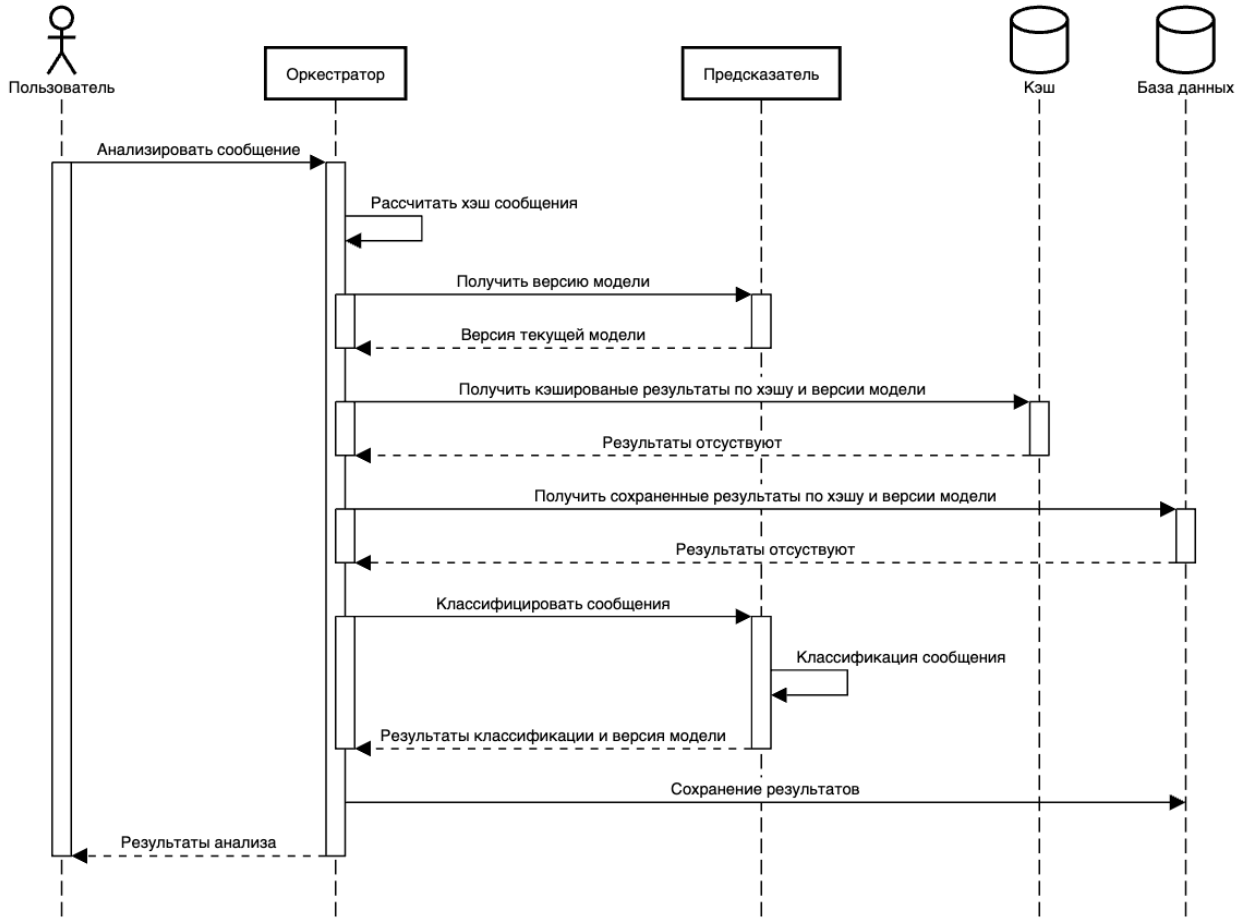

Рисунок 5 – Sequence диаграмма классификации троллинга по сообщениям

Для быстрой реализации компонентов были использованы следующие инструменты доступные в открытом доступе:

— сервис оркестратор реализован как веб сервис с помощью языка программирования Java и фреймворка Spring;

— сервис предсказатель разработан как серверное приложение на Python используя фреймворк FastAPI для реализации быстрого веб приложения с использованием JSON в качестве формата передачи данных;

— Redis использовался в качестве кэша как наиболее популярный продукт для кэширования данных;

— в качестве файлового хранилища моделей предсказания был выбран MinIO;

— база данных результатов выбран Elasticsearch так как является одним из наиболее популярных инструментов для полнотекстового поиска.

ЗАКЛЮЧЕНИЕ

Главным результатом данной диссертации стало обучение модели нейронной сети “Детектор троллинга” для классификации сообщений на предмет троллинга и разработка приложения “Антитролль” для использования модели с поддержкой различных её версий. Модель “Детектор троллинга” была разработана с использованием предобученной модели DistilBERT, что позволило достичь более эффективного предсказания троллинга.

В процессе написания диссертации соискатель провёл всесторонний анализ предметной области, изучил схожие проекты и статьи, а также результаты исследований в сфере обработки естественного языка и классификации. Были проведены исследования различных архитектур нейронных сетей, и их эффективность была сравнена для выбора окончательной архитектуры.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1 Волосович С.В. АНАЛИЗ МОДЕЛИ BERT ДЛЯ ЗАДАЧ КЛАССИФИКАЦИИ ТЕКСТА // Научное сообщество студентов XXI столетия. ТЕХНИЧЕСКИЕ НАУКИ: сб. ст. по мат. CXXXVIII междунар. студ. науч.-практ. конф. № 6(136).

2 Волосович С.В. МЕТОДЫ ВЕКТОРИЗАЦИИ ТЕКСТА В ОБРАБОТКЕ ЕСТЕСТВЕННОГО ЯЗЫКА / С.В. Волосович // Инновационные подходы в современной науке: сб. ст. по материалам CLXVII Международной научно-практической конференции «Инновационные подходы в современной науке». – № 11(167). – М., Изд. «Интернаука», 2024.

3 Волосович С.В. ЗАДАЧИ ГЛУБОКОГО ОБУЧЕНИЯ В СФЕРЕ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА / С.В. Волосович // Инновационные подходы в современной науке: сб. ст. по материалам CLXVII Международной научно-практической конференции «Инновационные подходы в современной науке». – № 11(167). – М., Изд. «Интернаука», 2024.