



МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)»

Институт (Филиал) № 8 «Компьютерные науки и прикладная математика»

Кафедра 806

Группа М8О-406Б-20 Направление подготовки 01.03.02 «Прикладная математика
и информатика»

Профиль Информатика

Квалификация: бакалавр

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

на тему: Интеллектуальная система обнаружения признаков компрометации
корпоративной электронной почты

Автор ВКРБ: Лисин Роман Сергеевич ()

Руководитель: Борисов Август Валерьевич ()

Консультант: Ухов Петр Александрович ()

Консультант: ()

Рецензент: ()

К защите допустить

Заведующий кафедрой № 806 Крылов Сергей Сергеевич ()

____ мая 2024 года

Москва 2024

РЕФЕРАТ

Выпускная квалификационная работа бакалавра состоит из 47 страниц, 43 рисунков, 1 таблицы, 25 использованных источников, 1 приложения.

МАШИННОЕ ОБУЧЕНИЕ, НЕЙРОННЫЕ СЕТИ, КОМПРОМЕТАЦИЯ КОРПОРАТИВНОЙ ЭЛЕКТРОННОЙ ПОЧТЫ, БИНАРНАЯ ТЕКСТОВАЯ КЛАССИФИКАЦИЯ

Объектом исследования в данной работе является электронное письмо из корпоративной корреспонденции.

Цель работы - создание модели, обучающейся по предоставленным письмам автора и способной подтвердить принадлежность тестовых писем данному автору.

Для достижения поставленной цели были рассмотрены и изучены различные алгоритмы векторизации текстов (bag of words, tf-idf) и бинарной классификации текстов (наивный классификатор Байеса, логистическая регрессия, метод опорных векторов, метод k-ближайших соседей, дерево решений, случайный лес, градиентный бустинг, перцептроны и рекуррентные, свёрточные, глубокие нейронные сети).

Результатами работы стала интеллектуальная система обнаружения признаков компрометации корпоративной электронной почты, которая использует глубокую нейронную сеть с архитектурой трансформер BERT.

Также данная работа будет использоваться в продукте по защите корпоративной почты Business Email Protection (BEP) компании F.A.C.C.T., занимающейся кибербезопасностью, для выявления компрометации корпоративной электронной почты.

СОДЕРЖАНИЕ

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ	5
ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ	6
ВВЕДЕНИЕ	7
1 ПОСТАНОВКА И ФОРМАЛИЗАЦИЯ ЗАДАЧИ	11
1.1 Постановка задачи	11
1.2 Требования к набору данных для обучения и тестирования .	11
2 ПОДГОТОВКА И ПРЕДОБРАБОТКА НАБОРА ДАННЫХ ДЛЯ ОБУЧЕНИЯ И ТЕСТИРОВАНИЯ МОДЕЛЕЙ	12
2.1 Обзор данных компании Enron	12
2.2 Подготовка данных для обучения и тестирования	15
2.3 Предобработка данных для обучения и тестирования	17
3 АЛГОРИТМЫ ВЕКТОРИЗАЦИИ ТЕКСТОВ ДЛЯ БИНАРНОЙ КЛАССИФИКАЦИИ ТЕКСТОВ	19
3.1 Задача бинарной классификации текстов	19
3.2 Алгоритм Bag of Words	19
3.3 Алгоритм TF-IDF	20
3.4 Токенизатор BERT	20
4 АЛГОРИТМЫ БИНАРНОЙ КЛАССИФИКАЦИИ ТЕКСТОВ . .	22
4.1 Наивный Байесовский классификатор	22
4.2 Логистическая регрессия	22
4.3 Метод опорных векторов	22
4.4 Метод k-ближайших соседей	23
4.5 Дерево решений	23
4.6 Случайный лес	23
4.7 Градиентный бустинг	23
4.8 Перцептрон	24
4.9 Рекуррентные нейронные сети	24
4.10 Свёрточные нейронные сети	24
4.11 Трансформеры	25
5 РЕЗУЛЬТАТЫ	26
5.1 Выбор метрик для оценки моделей	26
5.2 Сравнение моделей	27
ЗАКЛЮЧЕНИЕ	43

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	44
ПРИЛОЖЕНИЕ А Исходный код	47

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящей выпускной квалификационной работе бакалавра применяются следующие термины с соответствующими определениями:

Спуфинг почтового адреса — имитация адреса отправителя в электронных письмах.

Почтовая мимикрия — метод социальной инженерии, при котором киберпреступники используют поддельные электронные письма, чтобы обмануть потенциальную жертву, выдавая себя за того, кем они не являются.

Пулинг — уменьшение разрешения объекта с сохранением данных, необходимых для классификации.

Векторизация — подход к преобразованию входных данных из их исходного формата в векторы действительных чисел.

Бинарная классификация — определение принадлежности элементов заданного множества к одной из двух групп.

Трансформер — архитектура глубоких нейронных сетей, представленная в 2017 году исследователями из Google.

Стемминг — процесс нахождения основы для заданного исходного слова.

Лемматизация — процесс приведения словоформы к лемме - нормальной (словарной) форме.

Токенизатор — инструмент для разделения текста на токены - слова и другие цепочки символов, которые считаются минимальными линейными единицами текста.

Метрика — способ измерения того, насколько хорошо работает модель.

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

В настоящей выпускной квалификационной работе бакалавра применяются следующие сокращения и обозначения:

ФБР — Федеральное бюро расследований

БЕР — Business Email Protection, защита корпоративной электронной почты

БЕС — Business Email Compromise, компрометация корпоративной электронной почты

SPF — Sender Policy Framework, политика безопасности отправителя

DKIM — DomainKeys Identified Mail, почта идентифицирована с помощью доменного ключа

DMARC — Domain-based Message Authentication, Reporting, and Conformance, аутентификация, отчётность и соответствие сообщений на основе домена

GPT — Generative pre-trained transformer, генеративный предобученный трансформер

CSV — Comma-Separated Values, значения, разделённые запятыми

HTML — HyperText Markup Language, язык гипертекстовой разметки

CSS — Cascading Style Sheets, каскадные таблицы стилей

TF-IDF — Term Frequency and Inverse Document Frequency, частота слова и обратная частота документа

SVM — Support Vector Machine, метод опорных векторов

KNN — K-Nearest Neighbors, метод k-ближайших соседей

LSTM — Long Short-Term Memory, длинная цепь элементов краткосрочной памяти

CNN — Convolutional Neural Network, сверточная нейронная сеть

RNN — Recurrent Neural Network, рекуррентная нейронная сеть

BERT — Bidirectional Encoder Representations from Transformers, двунаправленная нейронная сеть кодировщик

ВВЕДЕНИЕ

В наши дни киберпреступность становится все более изощренной, хитрой и сложной особенно в сфере электронной почты. Электронной почтой чаще всего пользуются коммерческие и некоммерческие организации для ведения деловой переписки. Для неформальных переписок сейчас используются преимущественно мессенджеры и социальные сети. Атаки на почтовые ящики различных компаний представляют из себя весомую угрозу, так как они приносят вред бизнесу. Одной из самых трудно выявляемых атак является Business Email Compromise (BEC) - компрометация корпоративной электронной почты.

В разных источниках понятие BEC может включать в себя спуфинг почтового адреса, мимикрию на домены контрагентов, поставщиков и другое [1]. В данной работе под компрометацией корпоративной электронной почты подразумевается взлом электронной почты сотрудника компании и отправка писем от его имени. Атаки с применением похожих почтовых ящиков, доменов и другие признаки мимикрии можно обнаружить проверкой подписей SPF, DKIM, DMARC и другими эвристическими методами.

Актуальность темы связана с ущербом от компрометации корпоративной электронной почты и частотой таких атак. Данный вид атаки достаточно популярен среди хакеров и других киберпреступников. В 2023 году количество BEC-атак резко возросло. Число ежемесячных атак на 1000 почтовых ящиков увеличилось более чем вдвое и составило 10.77 согласно [2]. Это на 108% больше, чем в 2022 году. Диаграмма продемонстрирована на рисунке 1.

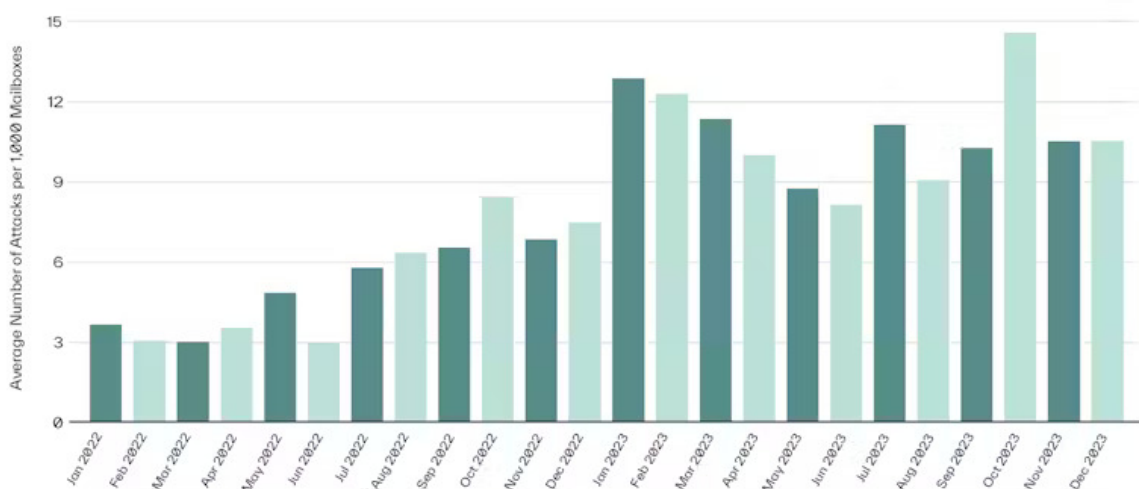


Рисунок 1 – Диаграмма среднемесячного числа ВЕС-атак на 1000 почтовых ящиков

ВЕС представляет угрозу для компаний и организаций, поскольку включает в себя мошеннические схемы, направленные на обман сотрудников и финансовый ущерб для бизнеса. По данным ФБР средний ущерб успешной атаки на корпоративную электронную почту составляет более 125000 долларов. Несмотря на то, что на ВЕС приходится меньший процент всех атак по электронной почте, данный тип атаки может принести огромную прибыль киберпреступникам. Фактически ВЕС является одним из самых разрушительных с финансовой точки зрения киберпреступлений, в результате которого только за предыдущий год убытки составили 2.7 миллиарда долларов. Чаще всего мошенники выбирают в качестве своих целей крупные компании [2].

Одним из ярких примеров компрометации корпоративной электронной почты является ВЕС-атака на американскую международную компанию по производству игрушек Mattel [3]. В 2016 году хакеры взломали почтовый аккаунт недавно назначенного генерального директора компании и написали письмо сотруднику, в котором попросили перевести 3 миллиона долларов новому поставщику. Сотрудник поверил отправителю, так как письмо пришло с настоящей почты генерального директора, и перевёл деньги на счёт мошенников.

С развитием генеративных нейронных сетей (GPT) использование

методов машинного обучения для обнаружения Business Email Compromise становится необходимостью в борьбе с данным типом киберугрозы. Методы машинного обучения позволяют анализировать обширные объемы электронной почты, выявлять аномалии и нештатные ситуации, а также строить модели, способные автоматически распознавать характеристики и признаки атак ВЕС.

В данной работе рассматривается применение алгоритмов автоматической классификации текстов писем для выявления и предотвращения случаев ВЕС с помощью классических моделей и нейронных сетей. Исследуются различные подходы и техники, которые помогают повысить эффективность детектирования и реагирования на подобные угрозы в корпоративном мире.

С научной точки зрения работа актуальна исследованием передовой архитектуры нейронных сетей трансформеров в задачах с небольшими выборками реальных корпоративных почтовых данных.

Цель работы - создание и обучение модели для автоматизации выявления компрометации корпоративной электронной почты.

Для достижения поставленной цели необходимо решить следующие задачи:

- Подготовить тексты писем для обучения и тестирования моделей;
- Выполнить предобработку текстов для обучения и тестирования моделей;
- Реализовать различные модели для интеллектуального анализа деловой корреспонденции отправителя;
- Выявить лучшую модель для обнаружения компрометации корпоративной электронной почты.

Объектом исследования является электронное письмо из корпоративной корреспонденции.

Предмет исследования - подтверждение авторства электронного письма.

Исследование основывается на теории по автоматической классификации текстов. Для работы с моделями используются Jupyter Notebook, язык программирования Python и его библиотеки - pandas, mail-parser, sklearn, numpy, nltk, tensorflow, pytorch, matplotlib, seaborn, tqdm, transformers.

Основным результатом работы является готовая модель для выявления компрометации корпоративной электронной почты.

Оригинальностью работы является подготовка и предобработка текстов для обучения и тестирования.

Практическая значимость работы представляет собой использование результатов исследования в продукте по защите корпоративной почты Business Email Protection (BEP) компании F.A.C.C.T., занимающейся кибербезопасностью, для выявления компрометации корпоративной электронной почты [4].

1 ПОСТАНОВКА И ФОРМАЛИЗАЦИЯ ЗАДАЧИ

1.1 Постановка задачи

Требуется реализовать модели, которые решают задачу бинарной классификации текстов электронных писем. Они обучаются по предоставленным письмам автора и способны подтвердить принадлежность тестовых писем данному автору с процентом верных предсказаний более 50%.

Для обучения и тестирования моделей должна использоваться деловая корреспонденция компании Enron из открытого источника [5]. Модели требуется обучить на датасетах 5 выбранных сотрудников компании Enron. Каждый датасет представляет собой набор электронных писем, написанных самим сотрудником, и письма, написанные другими сотрудниками, в равном соотношении.

Для каждой модели нужно подсчитать следующие стандартные метрики - accuracy (доля правильных ответов), precision (точность) и recall (полнота). После обучения и тестирования всех моделей требуется сравнить модели по средним арифметическим показателям выбранных метрик по 5 датасетам и выбрать лучшую модель.

1.2 Требования к набору данных для обучения и тестирования

Для обучения и тестирования моделей из писем нужно извлечь только текст, написанный самим автором письма. При этом должны быть удалены подписи авторов, обращения и другие слова, явно подчеркивающие авторство письма.

Также требуется сопоставить письма почтовому адресу отправителя, так как у него может быть несколько почтовых ящиков.

2 ПОДГОТОВКА И ПРЕДОБРАБОТКА НАБОРА ДАННЫХ ДЛЯ ОБУЧЕНИЯ И ТЕСТИРОВАНИЯ МОДЕЛЕЙ

2.1 Обзор данных компании Enron

В качестве набора данных для обучения будем использовать деловую корреспонденцию компании Enron, состоящую из 126841 электронных писем на английском языке [5]. Распаковав архив, посмотрим на структуру данных на рисунках 2, 3, 4.

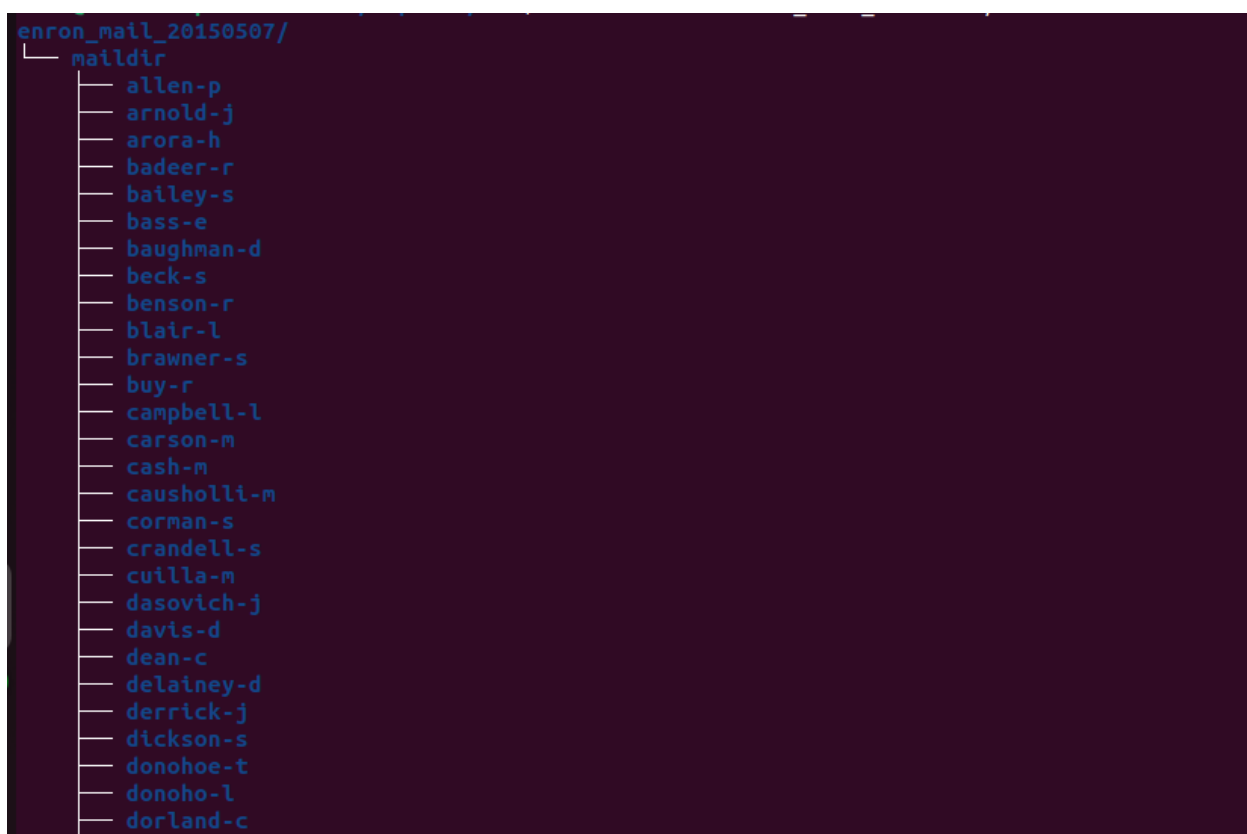


Рисунок 2 – Структура данных компании Enron

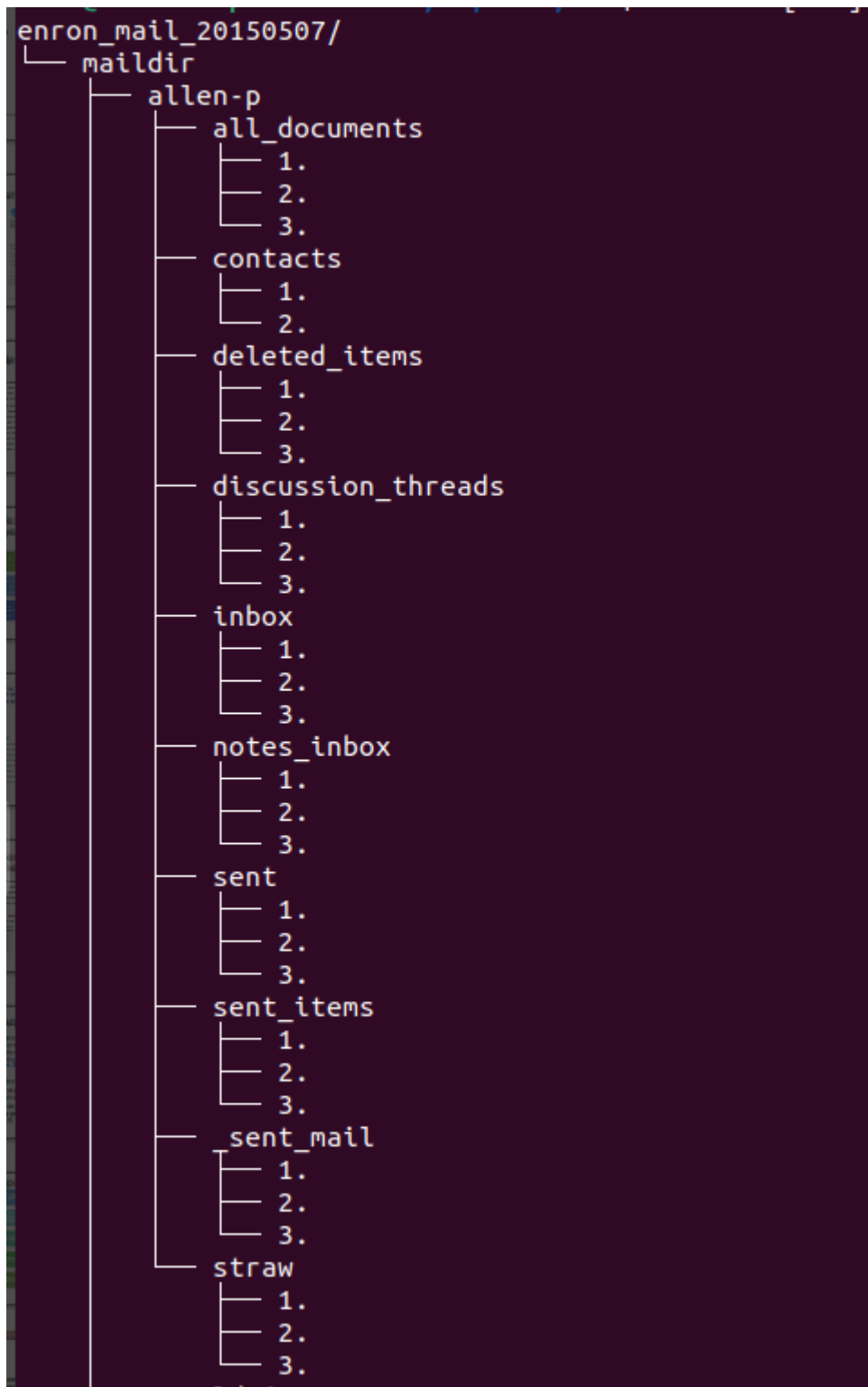


Рисунок 3 – Структура папки с электронными письмами сотрудника компании Enron. Для демонстрации количество писем обрезано до 3

```
Message-ID: <19790540.1075855679828.JavaMail.evans@thyme>
Date: Tue, 12 Dec 2000 04:03:00 -0800 (PST)
From: phillip.allen@enron.com
To: christi.nicolay@enron.com
Subject: Talking points about California Gas market
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: Christi L Nicolay
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Dec2000\Notes Folders\Sent
X-Origin: Allen-P
X-FileName: pallen.nsf

Christy,

I read these points and they definitely need some touch up. I don't
understand why we need to give our commentary on why prices are so high in
California. This subject has already gotten so much press.

Phillip

----- Forwarded by Phillip K Allen/HOU/ECT on 12/12/2000
12:01 PM -----
From: Leslie Lawner@ENRON on 12/12/2000 11:56 AM CST
To: Christi L Nicolay/HOU/ECT@ECT, Joe Hartsoe/Corp/Enron@ENRON, Rebecca W
Cantrell/HOU/ECT@ECT, Ruth Concannon/HOU/ECT@ECT, Stephanie
Miller/Corp/Enron@ENRON, Phillip K Allen/HOU/ECT@ECT, Jane M
Tholt/HOU/ECT@ECT, Richard Shapiro/NA/Enron@Enron
cc:
Subject: Talking points about California Gas market

Here is my stab at the talking points to be sent in to FERC along with the
gas pricing info they requested for the California markets. Let me or
Christi know if you have any disagreements, additions, whatever. I am
supposed to be out of here at 2:15 today, so if you have stuff to add after
that, get it to Christi. Thanks.
```

Рисунок 4 – Пример электронного письма сотрудника компании Enron

На рисунке 2 можно увидеть, что для каждого сотрудника компании есть своя папка с письмами. На рисунке 3 показано, как устроены папки с письмами сотрудников. Для создания датасетов возьмём письма из папок с исходящими письмами, то есть те, в которых встречается слово sent.

На рисунке 4 показан типичный пример электронного письма сотрудника компании Enron.

2.2 Подготовка данных для обучения и тестирования

Из письма будем извлекать почтовый адрес отправителя письма и текст. Для этого используем библиотеку `mail-parser` языка программирования Python, с помощью которой можно анализировать файлы формата `msg` [6].

Из текстов писем отбираем только последнее сообщение, убираем обращения и подписи авторов. Исключаем сообщения с почтовых ящиков, не принадлежащих людям, командные автоматические рассылки и письма менее 11 символов. Пример подготовленного к обучению и тестированию текста электронного письма с рисунка 4 показан на рисунке 5.

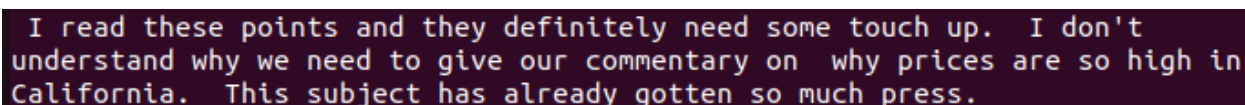
A screenshot of an email body text displayed on a dark background with light-colored text. The text reads: "I read these points and they definitely need some touch up. I don't understand why we need to give our commentary on why prices are so high in California. This subject has already gotten so much press."

Рисунок 5 – Пример подготовленного к обучению и тестированию текста электронного письма с рисунка 4

Для каждого сотрудника собираем отдельный датасет с одинаковым количеством писем, написанных самим сотрудником, и писем, написанных другими сотрудниками. Датасеты представляют собой файлы формата `csv`, состоящие из трех столбцов: первый используется для нумерации, во втором находятся тексты для обучения и тестирования, третий содержит в себе бинарное значение 0 или 1 для определения авторства письма. Тексты со значением 0 не принадлежат автору, а со значением 1 принадлежат.

Для обучения и тестирования моделей в ручном режиме выберем 5 сотрудников. Так как задача чистки писем от отличительных особенностей таких как подписи, обращения нетривиальна, то выберем сотрудников, у которых после подготовки датасетов удалось исключить подписи и обращения, переписка носит преимущественно деловой характер, и у которых больше всего писем. На рисунках 6, 7, 8, 9, 10 представлены выбранные датасеты.

Unnamed: 0	text	label
0	\nI will defer to your expertise on whether we...	1
1	\nso much for extending your business trip to ...	1
2	we could set up an LLC entity to employ those...	1
3	I just called our counsel on this matter this...	1
4	I am working with Mary Joyce on that right now...	1
...
2021	I have lunch plans tomorrow (and I'm not sorry...	0
2022	micha will pick us up at my place saturday @ 1...	0
2023	try having no work to do for two months and th...	0
2024	i think i am going to go to the game.	0
2025	I was not aware that the market closed at 1:00...	0

Рисунок 6 – Датасет Мишель Кэш с 2026 письмами

Unnamed: 0	text	label
0	This is the e-mail where DF objected to my pro...	1
1	Attached is a draft of 2 minor amendments that...	1
2	Sara and I would like to meetwith you sometim...	1
3	suzanne;\nchocolate works!\n\n St. \nEB 3892\n...	1
4	There are a number of factors that determine p...	1
...
2229	I still need the info I requested to comment i...	0
2230	Absolutely.	0
2231	for your help. Due to my lack of a complete u...	0
2232	Brent Price, Bob Hall, Leslie Reeves and I wou...	0
2233	You should try looking at Duke, they seem to ...	0

Рисунок 7 – Датасет Кэрол Клэйр с 2234 письмами

Unnamed: 0	text	label
0	I'm a little confused - 559066 is a Prebon dea...	1
1	Prebon is right on this. Both deals are 25 MW...	1
2	Hey there stranger!\n\nfor the pictures - they...	1
3	Two of these should have fees; two should not...	1
4	Mark's changing deal 581615 to APB - he had it...	1
...
2529	pending the sale of the Wilton Centre unit, I...	0
2530	Same to you! \nAnd I hope you and your family...	0
2531	\n I would appreciate your help in locating fi...	0
2532	\t2- SURVEY/INFORMATION EMAIL - 7/19/01\n\nCur...	0
2533	see if this works! If it does, see "13." w...	0

Рисунок 8 – Датасет Кейт Саймс с 2534 письмами

Unnamed: 0	text	label
0	I won't sent this to you daily, but this is wh...	1
1	\n\nWhen I tried to approve this request (She...	1
2	As discussed - happy reading on the plane...\n...	1
3	I will be in London the week of January 15th. ...	1
4	I have completed internal feedback forms throu...	1
...
2815	glad you were succesful, i am not, getting run...	0
2816	I believe our new Tennessee rep is Sherry Glaz...	0
2817	\n\nI found a copy of the confirmation letter...	0
2818	Deal NN6788.1 with Citibank has been killed. ...	0
2819	It's really easy to assume that companies are ...	0

Рисунок 9 – Датасет Салли Бек с 2820 письмами

Unnamed: 0	text	label
0	Kathryn (EES Logistics) is going to send me s...	1
1	This doesn't have everything on it AND i thoug...	1
2	I faxed a copy of the PSNC invoice for March t...	1
3	not a bad idea.\n\n_61@hotmail.com\t\n2701 Re...	1
4	Should I call you up and sing !?!?!?!?!?! ...	1
...
5929	For your information. please find attached the...	0
5930	Sorry. Don't work. I can do 1:45 California ...	0
5931	\nDo you know anything about the Congress Plaz...	0
5932	Please email me whatever version of the doc yo...	0
5933	I have no such studies.	0

Рисунок 10 – Датасет Крис Германи с 5934 письмами

2.3 Предобработка данных для обучения и тестирования

Предобработка текстовых данных является важным этапом в задачах анализа текста. Она помогает улучшить качество обучения моделей, упростить и ускорить процесс анализа текста, снизить вероятность переобучения и улучшить обобщающую способность моделей. Подготовленные и очищенные данные помогают моделям лучше понимать и извлекать смысл из текста, что приводит к более точным и надежным результатам классификации.

Классические этапы текстовой предобработки включают в себя перевод всех букв в тексте в нижний или верхний регистры, удаление цифр, чисел или замену на текстовый эквивалент, очистку от пунктуации, устранение стоп-слов, стемминг, лемматизацию [7].

Устранение стоп-слов - это исключение общеупотребительных слов, не несущих смысловой нагрузки таких как предлоги, союзы, местоимения, чтобы сосредоточиться на ключевых словах и фразах, которые могут повлиять на классификацию.

Лемматизация и стемминг нужны для приведения слов к их базовой или корневой форме (лемме) с целью уменьшения вариантов написания слова и объединения их одним словом.

Также в предметной области корпоративной электронной почты из писем следует исключать HTML и CSS блоки, эмодзи, почтовые адреса, ссылки, пробельные символы с начала и конца текста, символы переноса строк.

В данной работе были выполнены все вышеперечисленные этапы предобработки текста и подобран лучший вариант в процессе исключения различных этапов. В приложении в коде представлен лучший вариант предобработки данных по точности предсказания моделей.

3 АЛГОРИТМЫ ВЕКТОРИЗАЦИИ ТЕКСТОВ ДЛЯ БИНАРНОЙ КЛАССИФИКАЦИИ ТЕКСТОВ

3.1 Задача бинарной классификации текстов

Бинарная классификация текстов является одной из распространенных задач в области машинного обучения и анализа текстовых данных. Решение этой задача заключается в том, чтобы отнести каждый текстовый документ к одному из двух классов: положительному (1) или отрицательному (0) в зависимости от целевой переменной или метки класса. В данной работе к положительному классу относятся электронные письма, принадлежащие автору, а к отрицательному - письма, которые не были написаны данным автором.

Эффективная классификация текстов требует использования различных подходов, алгоритмов и методов обработки информации. Алгоритмы машинного обучения хорошо работают с числами и категориальными признаками. Один из ключевых подходов в бинарной классификации текстов – это преобразование текстовых данных в векторное представление, которое позволяет использовать их в алгоритмах машинного обучения. Для этого используются методы векторизации Bag of Words, TF-IDF, Word2Vec и другие. Для моделей с архитектурой нейронной сети процесс векторизации называют токенизацией [8].

В данной работе используются традиционные Bag of Words и TF-IDF в качестве алгоритмов векторизации. Они работают лучше других на небольших текстовых данных и выборках. Для нейросети BERT используется уже обученный токенизатор с сайта HuggingFace.

3.2 Алгоритм Bag of Words

Алгоритм Bag of Words используется для векторизации текстовых данных путем создания словаря из уникальных слов в корпусе текстов и подсчета частоты вхождения каждого слова. При использовании Bag of Words порядок слов игнорируется, а векторное представление каждого документа представляет собой вектор, где каждая компонента соответствует количеству вхождений слова из словаря в соответствующий документ. Данный подход позволяет преобразить тексты в числовые

векторы, которые могут быть использованы в алгоритмах машинного обучения для классификации, анализа тональности, кластеризации и других задач обработки текста. Bag of Words хорошо работает с небольшими наборами данных, но не учитывает семантические и контекстуальные связи между словами, что может быть недостатком при обработке более сложных текстовых данных [9].

3.3 Алгоритм TF-IDF

TF-IDF является классическим методом векторизации текста, который оценивает важность каждого слова в тексте относительно количества его употреблений в данном тексте (TF - Term Frequency) и во всей коллекции текстов (IDF - Inverse Document Frequency). TF измеряет важность слова в контексте отдельного текста. IDF измеряет, насколько уникально слово является по всей коллекции текстов. Слова, которые появляются в большинстве текстов, имеют низкое IDF, так как они не вносят большой информационной ценности. Путём умножения TF на IDF для каждого термина происходит взвешивание значимости терминов, позволяя выделить важные слова в тексте и отфильтровать общие или шумовые термины. Данный алгоритм помогает в выявлении ключевых слов и понимании их контекстуальной значимости для обеспечения точности при анализе и классификации текстовых данных [10].

3.4 Токенизатор BERT

Для нейронной сети с архитектурой трансформер BERT, созданной компанией Google, используется собственный токенизатор, так как модель делает большой упор на контекст. Токенизатор для BERT разбивает входной текст на токены и преобразовывает их в числовые представления, которые понимает модель BERT. Важным шагом в токенизации для BERT является добавление специальных токенов [CLS] в начале предложения и [SEP] между предложениями в задачах с двумя и более предложениями. Токены разбиваются на подтокены с учетом встречающихся в словах специальных символов разбиения, что позволяет BERT обрабатывать незнакомые слова или части слов. Также проводится выравнивание длин последовательностей путем добавления дополнительных токенов до заданной длины, обеспечивая однородность входных данных [11]. Пример

токенизации представлен на рисунке 11.

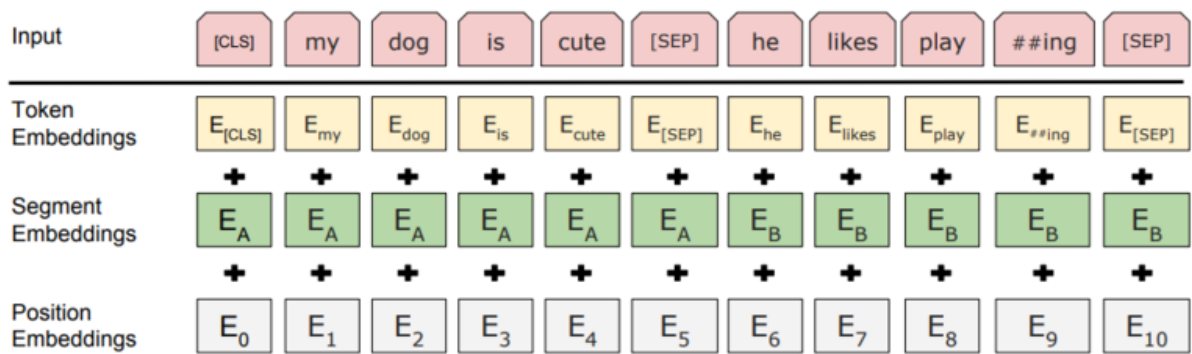


Рисунок 11 – Представление входных данных для модели BERT

4 АЛГОРИТМЫ БИНАРНОЙ КЛАССИФИКАЦИИ ТЕКСТОВ

После векторизации текстов данные готовы для решения задачи бинарной классификации [12]. Эта задача решается при помощи алгоритмов машинного обучения таких как наивный классификатор Байеса, логистическая регрессия, метод опорных векторов (SVM), метод k-ближайших соседей (KNN), дерево решений, случайный лес, градиентный бустинг, перцептроны, рекуррентные (RNN, LSTM), свёрточные (CNN), глубокие (трансформер BERT) нейронные сети [13].

4.1 Наивный Байесовский классификатор

Наивный Байесовский классификатор - это простой и эффективный метод машинного обучения, основанный на теореме Байеса с наивным предположением о независимости признаков. Он оценивает вероятности принадлежности текстов к различным классам, используя подсчет частоты вхождения признаков в обучающем наборе данных и применяя теорему Байеса. Наивный Байесовский классификатор демонстрирует хорошую производительность при правильном подборе признаков и корректном обучающем наборе данных [14].

4.2 Логистическая регрессия

Логистическая регрессия используется для оценки вероятности принадлежности текста к конкретному классу с помощью логистической функции, которая моделирует кривую роста вероятности события по мере изменения управляющих параметров. Данный метод является простым, интерпретируемым и эффективным методом для классификации [15].

4.3 Метод опорных векторов

Метод опорных векторов (Support Vector Machine, SVM) работает путем построения оптимальной гиперплоскости в пространстве признаков, эффективно разделяющей классы данных. SVM стремится максимизировать расстояние между объектами разных классов, что способствует лучшей обобщающей способности модели и устойчивости к шумам в наборе данных. Этот метод хорошо работает с небольшими и средними выборками, демонстрируя высокую точность и устойчивость к

переобучению при правильном подборе параметров модели [16].

4.4 Метод k-ближайших соседей

Метод k-ближайших соседей (k-Nearest Neighbors, KNN) основан на принципе близости в пространстве признаков. Объекту присваивается класс большинства из k его ближайших соседей. Особенность данного метода заключается в том, что он не требует обучения модели, а является «ленивым» алгоритмом, который хранит обучающие данные и осуществляет классификацию на основе расстояния до ближайших соседей [17].

4.5 Дерево решений

Дерево решений представляет собой древовидную структуру с решающими правилами в узлах и метками классов в листьях. При классификации каждый узел дерева оценивает признаки и принимает решение о направлении следования на основе их значений, пока не будет достигнут лист, который определяет принадлежность объекта к классу [18].

4.6 Случайный лес

Случайный лес - это ансамблевый алгоритм машинного обучения, основанный на деревьях решений, где каждое дерево строится независимо на основе случайной подвыборки обучающих данных и случайного подмножества признаков. В процессе классификации объекта каждое дерево прогнозирует результат, а затем решение принимается путем голосования или усреднения прогнозов всех деревьев. Случайные леса эффективны в работе с большими наборами данных, могут автоматически обрабатывать пропущенные значения и выбросы, а также показывают хорошую производительность на практике, что делает их широко используемым методом для решения задач классификации [19].

4.7 Градиентный бустинг

Градиентный бустинг - это мощный ансамблевый метод, который строит композицию более слабых моделей (например, деревьев решений), обучаемых последовательно с целью минимизации функции потерь

градиентным спуском. При классификации каждая новая модель уменьшает ошибки предыдущих моделей, что позволяет улучшить предсказательную способность ансамбля. Градиентный бустинг превосходит многие другие методы своей способностью находить сложные нелинейные зависимости в данных, а также построением моделей высокой точности [20].

4.8 Перцептрон

Перцептрон является простейшей нейронной сетью, состоящей из одного или нескольких слоев нейронов, способных принимать входные данные, вычислять их взвешенные суммы и применять функцию активации для получения выходного значения. Эта модель работает эффективно в задачах с линейно разделимыми данными [21].

4.9 Рекуррентные нейронные сети

Рекуррентные нейронные сети (RNN) и их подтип LSTM (Long Short-Term Memory) широко применяются для классификации текстов благодаря способности учитывать контекстуальные зависимости и последовательность слов в тексте. RNN позволяют передавать информацию от одного временного шага к другому, что особенно важно для анализа текстовых данных. LSTM обладает способностью запоминать и забывать информацию на основе LSTM-модуля, что помогает избежать проблемы затухающего градиента и долгосрочной зависимости между словами [22].

4.10 Свёрточные нейронные сети

Несмотря на своё первоначальное применение в обработке изображений свёрточные нейронные сети (CNN) нашли свое применение и в классификации текстовых данных. Для классификации текстов слова представляются в виде векторов и подаются на вход свёрточным слоям, которые выявляют локальные признаки в тексте. Последующие слои пулинга уменьшают размер пространства признаков, а полносвязные слои позволяют модели сделать окончательное предсказание по выявленным признакам. CNN показывают хорошие результаты благодаря способности распознавать локальные шаблоны в тексте [23].

4.11 Трансформеры

Трансформеры являются самыми современными моделями в области обработки естественного языка, которые показывают хорошие результаты в задаче классификации текстов. Они основаны на механизмах внимания, которые позволяют модели обучаться за счет просмотра всех входных слов одновременно в отличие от рекуррентных моделей. Это позволяет трансформерам эффективно улавливать долгосрочные зависимости в тексте и выражать сложные взаимодействия между словами. В данной работе в качестве трансформера используется модель BERT, разработанная компанией Google [24].

5 РЕЗУЛЬТАТЫ

5.1 Выбор метрик для оценки моделей

Чтобы выбрать лучшую модель из раздела 4 для решения задачи бинарной классификации на датасетах с корпоративной корреспонденцией, нужно выбрать метрики для сравнения моделей. Для каждой модели будем подсчитывать среднее арифметическое метрик accuracy, precision и recall по пяти выбранным датасетам [25].

Также для каждой модели будем строить графики по точности предсказания на обучающей, тестовой выборках и матрицу ошибок (confusion matrix). С помощью этих графиков можно предотвратить переобучение и недообучение моделей, сравнивая графики обучения и тестирования моделей. Пример матрицы ошибок для бинарной классификации представлен на рисунке 12.

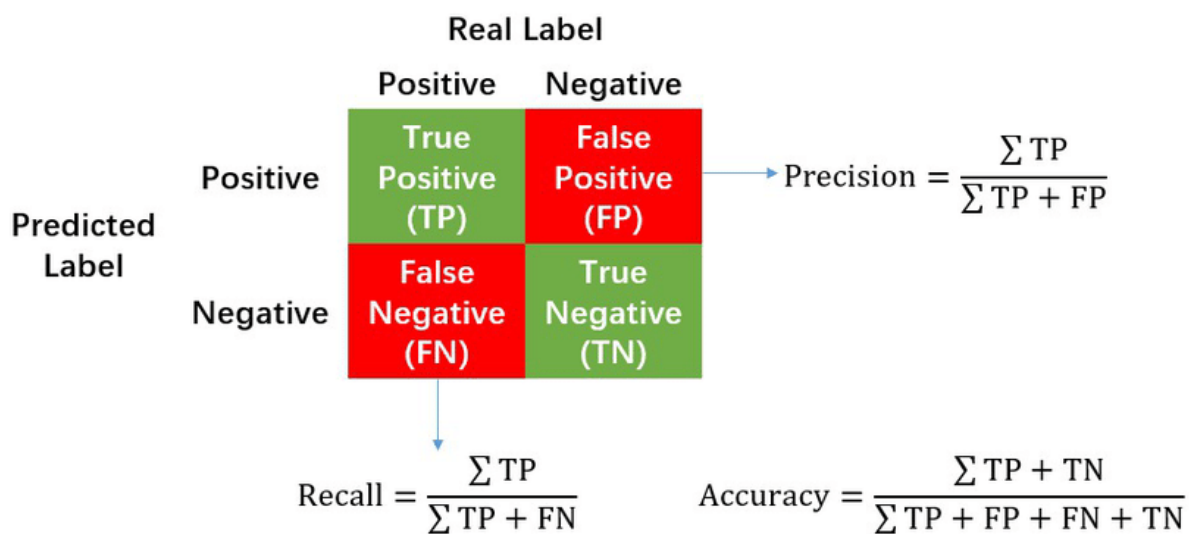


Рисунок 12 – Подсчёт метрик для сравнения моделей, используя матрицу ошибок

Самой важной метрикой для решения задачи определения авторства письма является accuracy. Чем больше количество верно предсказанных случаев, тем лучше модель. Также значимой является метрика recall. С точки зрения продукта по защите почты большое количество ошибочных обвинений о возможном взломе электронной почты хуже противоположных ошибок из-за потенциально меньшего количества писем, написанных злоумышленниками от имени сотрудника компании, по сравнению с

легитимными письмами с этого почтового ящика. Это значит, что при равных долях верных предсказаний предпочтение стоит отдать модели, допускающей меньше ошибок в письмах, которые на самом деле написал автор. Следовательно, нужно уменьшить долю ошибок второго рода, то есть увеличить recall. Таким образом, оценка моделей будет производиться последовательно по метрикам accuracy и recall.

5.2 Сравнение моделей

После обучения и тестирования моделей для бинарной классификации текстов писем, описанных в разделе 4, получились результаты, представленные в таблице 1. С помощью предобработки текста не удалось улучшить точность моделей.

Таблица 1 – Результаты обучения моделей для выявления компрометации корпоративной электронной почты, отсортированные по средним арифметическим показателям метрик accuracy и recall по 5 датасетам

	Accuracy, %	Precision, %	Recall, %
BERT	88	86	90
SVM	87	92	81
Naive Bayes	86	82	92
Logistic Regression	86	90	82
CNN	85	85	86
LSTM	85	86	85
Random Forest	83	89	76
Gradient Boosting	81	86	73
Perceptron	80	90	69
KNN	78	75	86
RNN	77	75	83
Decision Tree	75	76	74

Лучшей моделью с accuracy 88% и второй по доле ошибок второго рода с recall 90% оказалась одна из самых современных нейронных сетей - трансформер BERT. Результаты представлены на рисунках 13, 14, 15, 16, 17, 18, 19, 20, 21, 22.

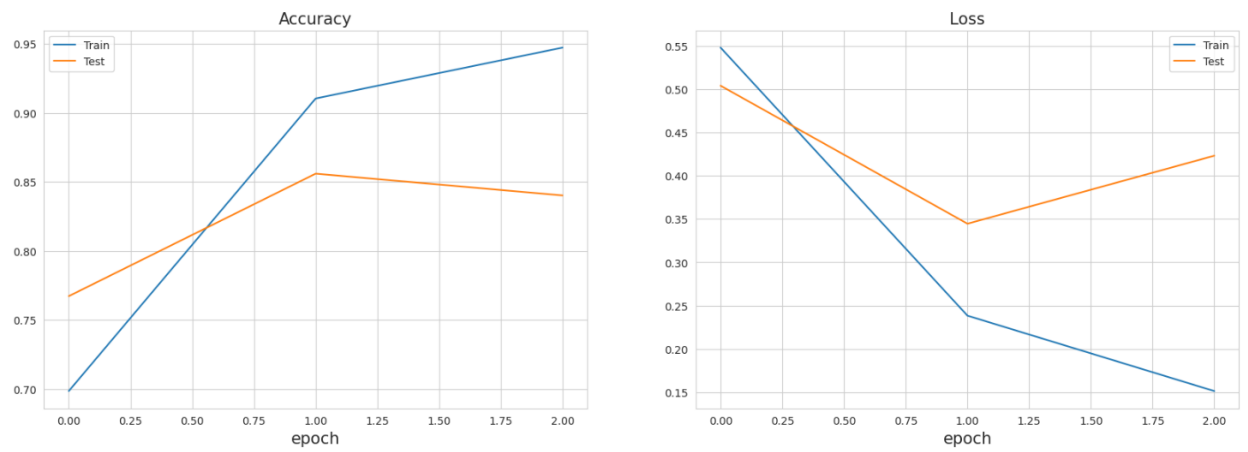


Рисунок 13 – Графики ассурасу и loss по эпохам во время обучения и тестирования модели BERT на датасете Мишель Кэш

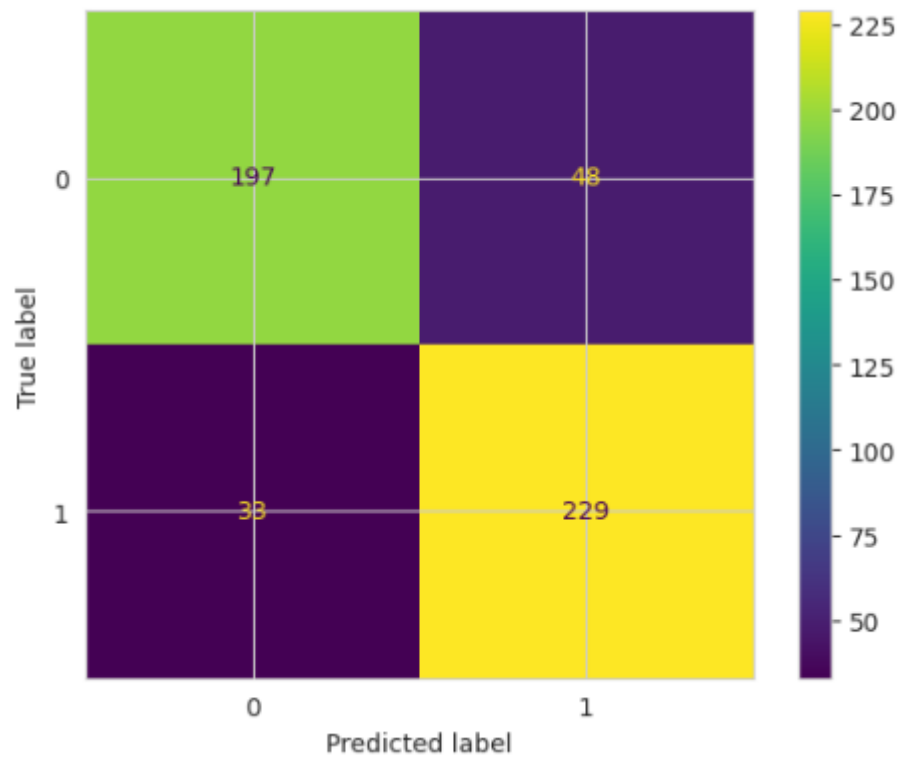


Рисунок 14 – Матрица ошибок для модели BERT на датасете Мишель Кэш

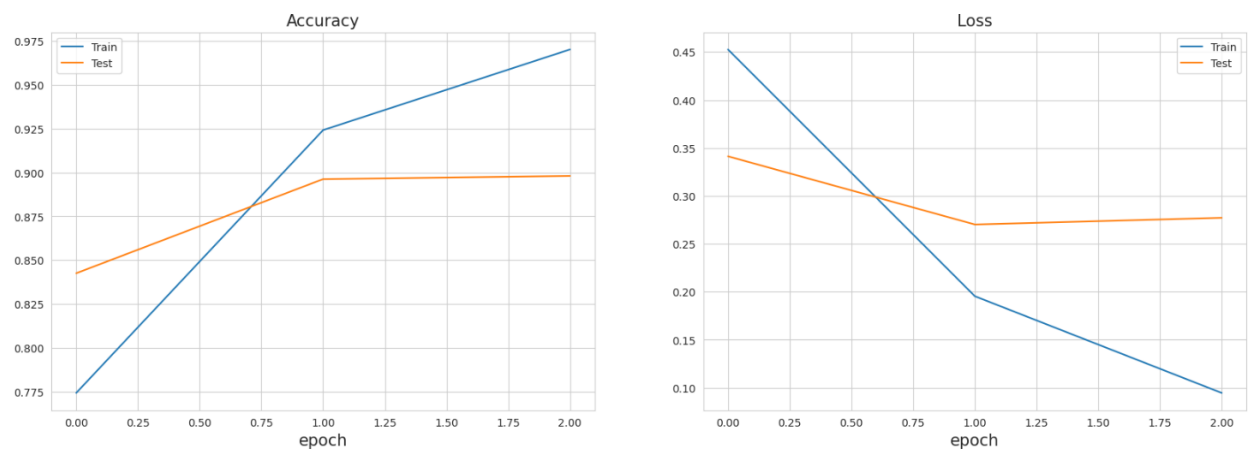


Рисунок 15 – Графики ассурасу и loss по эпохам во время обучения и тестирования модели BERT на датасете Кэрол Клэйр

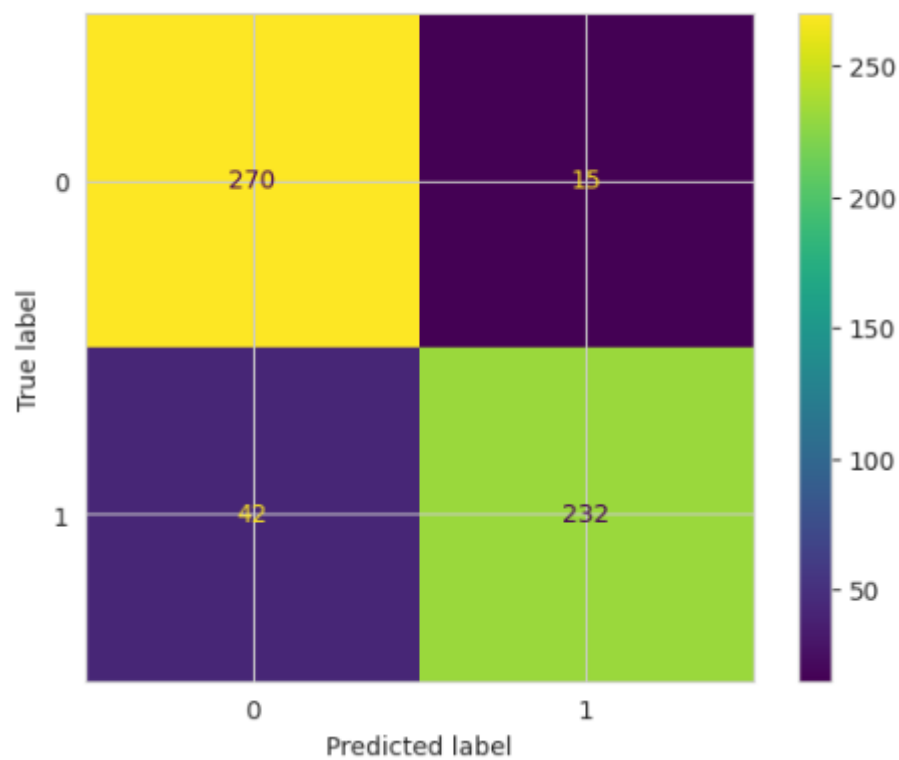


Рисунок 16 – Матрица ошибок для модели BERT на датасете Кэрол Клэйр

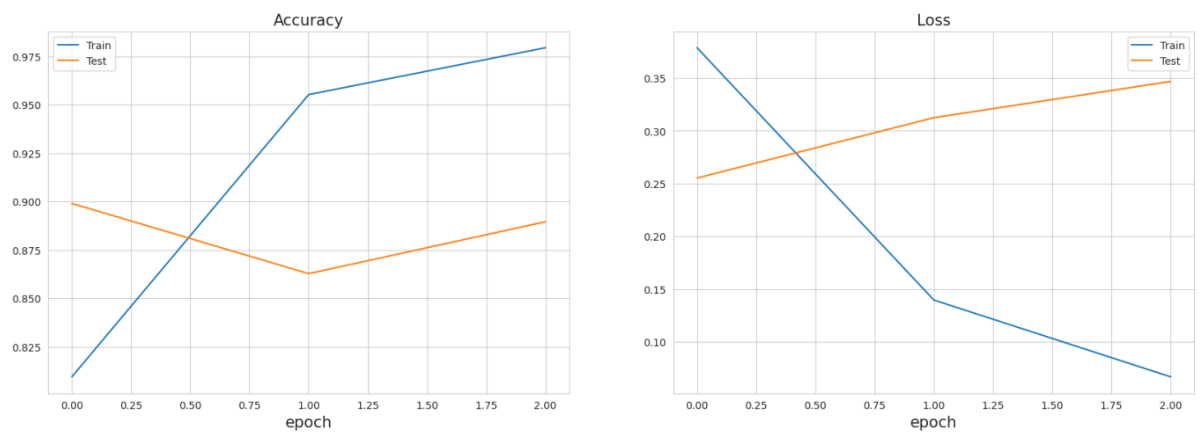


Рисунок 17 – Графики ассурасу и loss по эпохам во время обучения и тестирования модели BERT на датасете Кейт Саймс

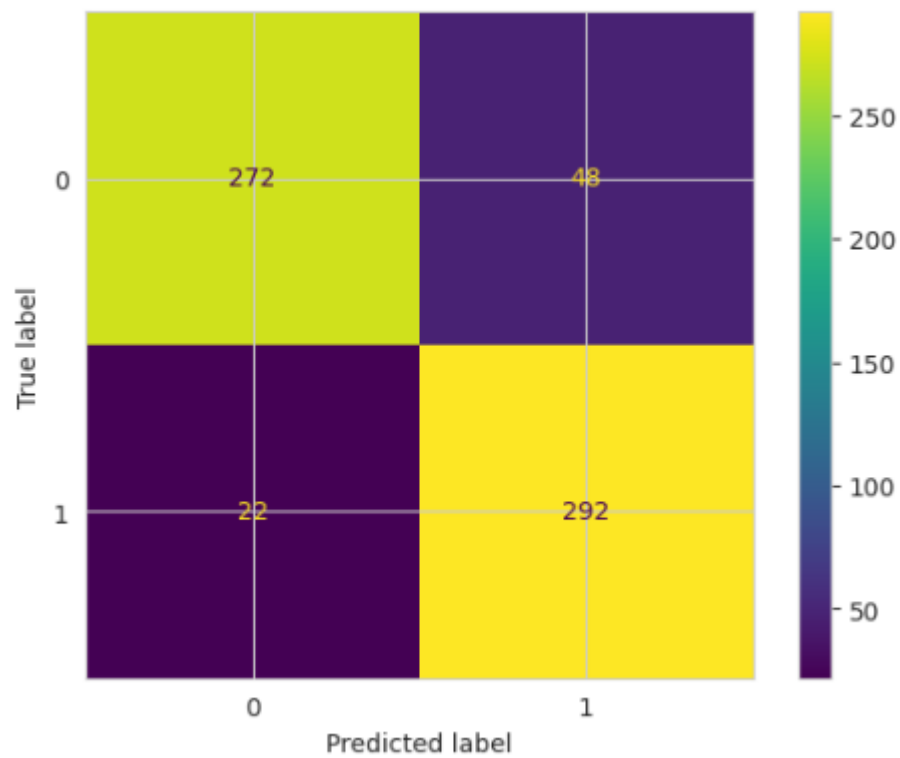


Рисунок 18 – Матрица ошибок для модели BERT на датасете Кейт Саймс

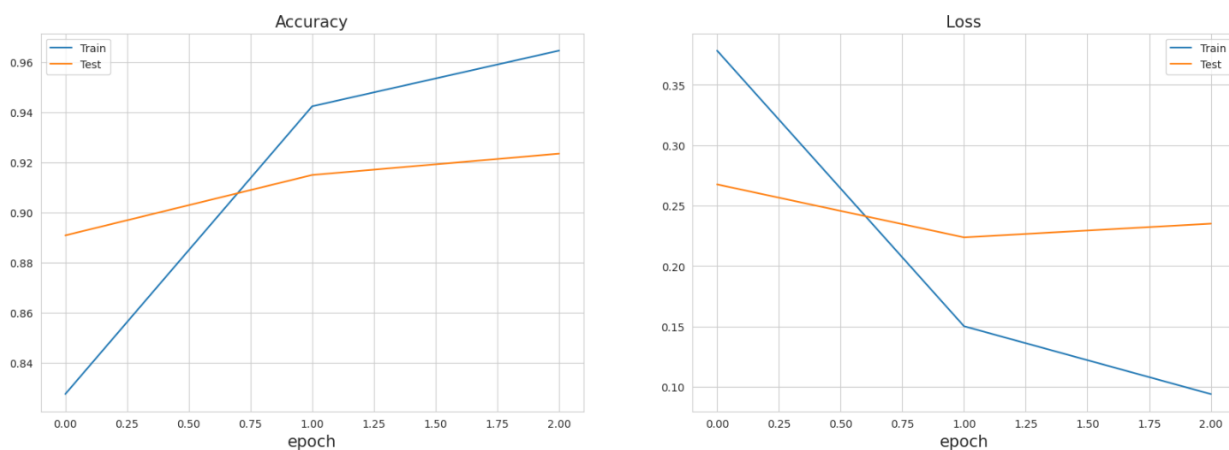


Рисунок 19 – Графики ассурасу и loss по эпохам во время обучения и тестирования модели BERT на датасете Салли Бек

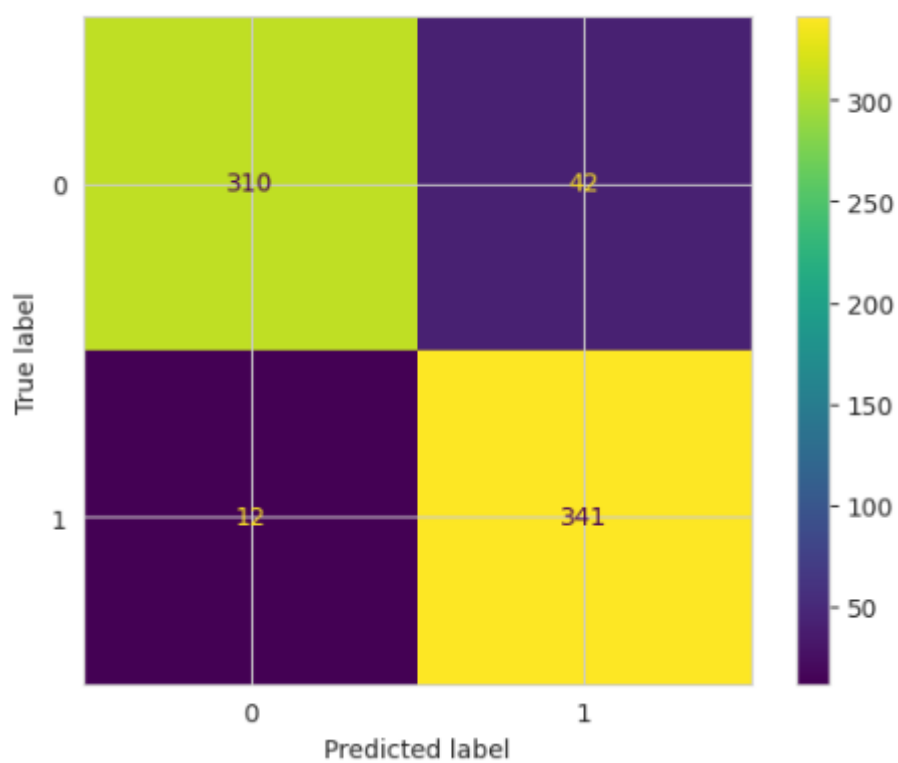


Рисунок 20 – Матрица ошибок для модели BERT на датасете Салли Бек

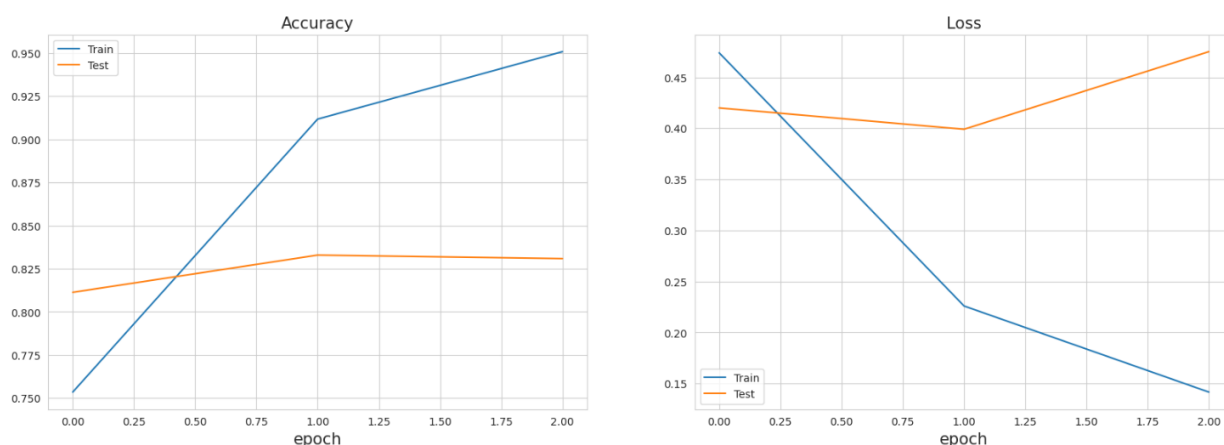


Рисунок 21 – Графики ассурасу и loss по эпохам во время обучения и тестирования модели BERT на датасете Крис Германи

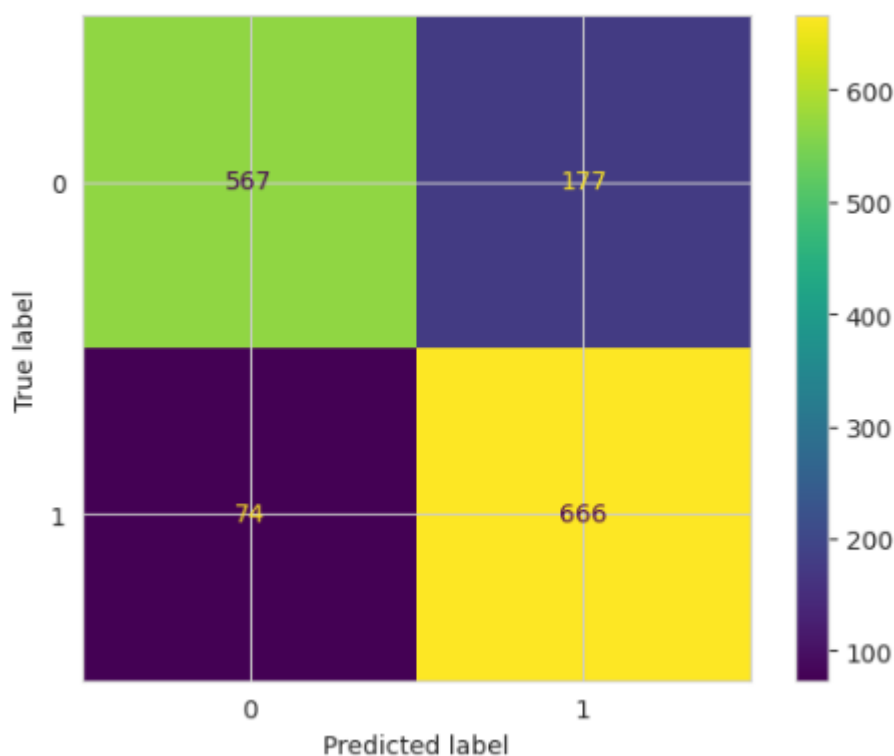


Рисунок 22 – Матрица ошибок для модели BERT на датасете Крис Германи

На втором месте по ассурасу метод опорных векторов (SVM) с 87%, но у данного метода далеко не лучший recall - 81%. Также SVM долго обучается, что делает достаточно проблематичным использование данной модели с автоматизированным обучением для выявления компрометации корпоративной электронной почты. Результаты представлены на рисунках 23, 24, 25, 26, 27, 28, 29, 30, 31, 32.

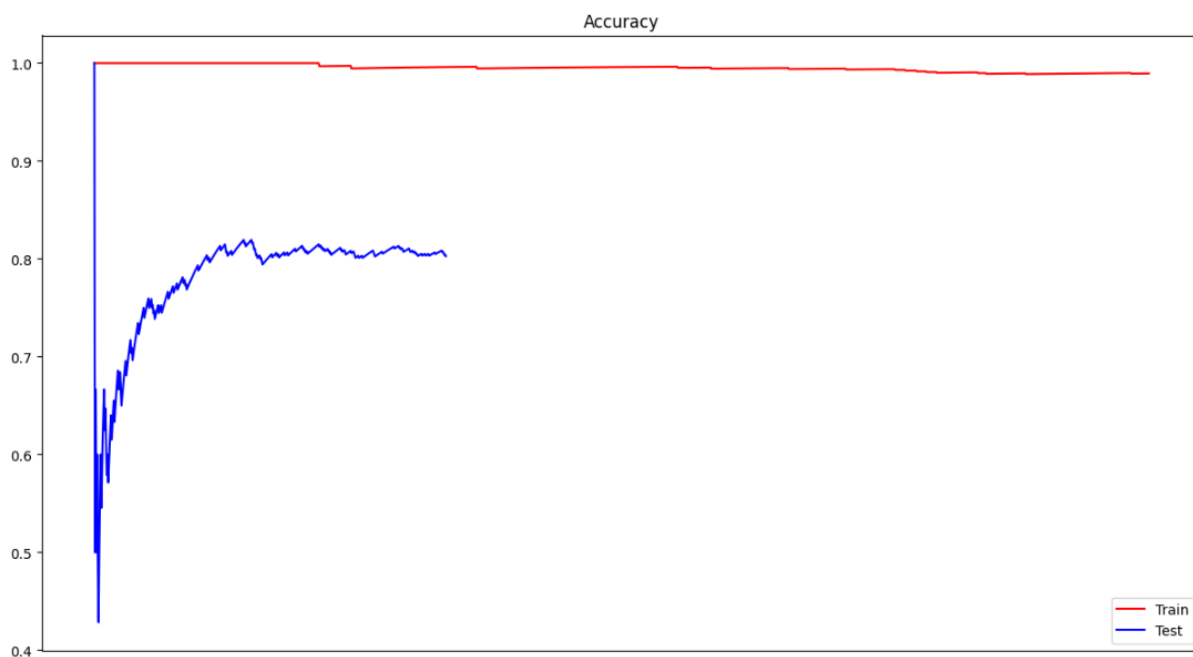


Рисунок 23 – График ассигасу по количеству предсказаний во время обучения и тестирования модели SVM на датасете Мишель Кэш

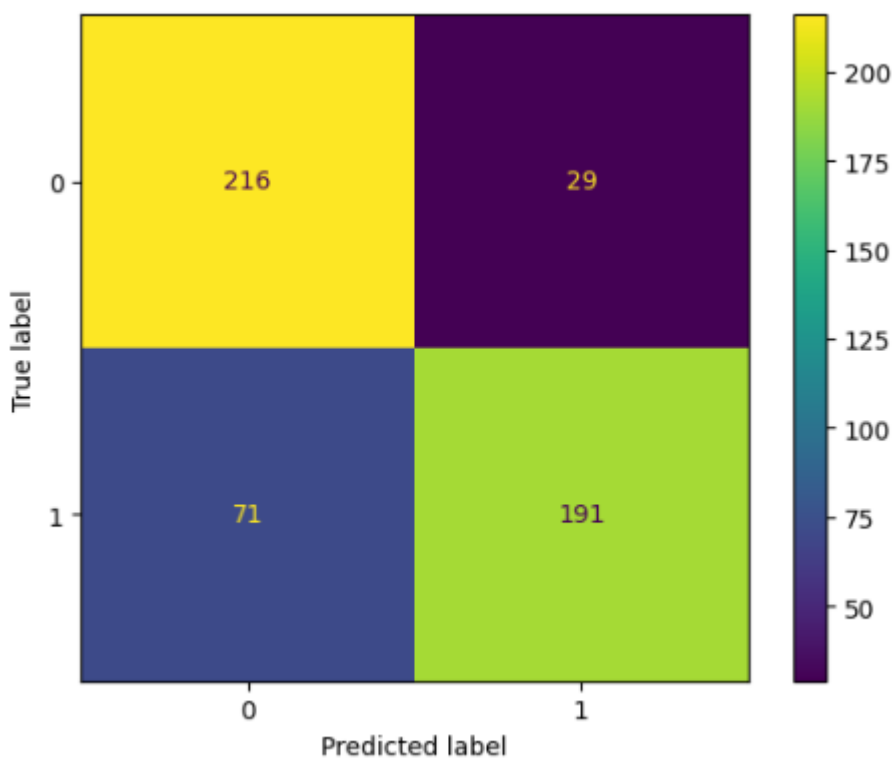


Рисунок 24 – Матрица ошибок для модели SVM на датасете Мишель Кэш

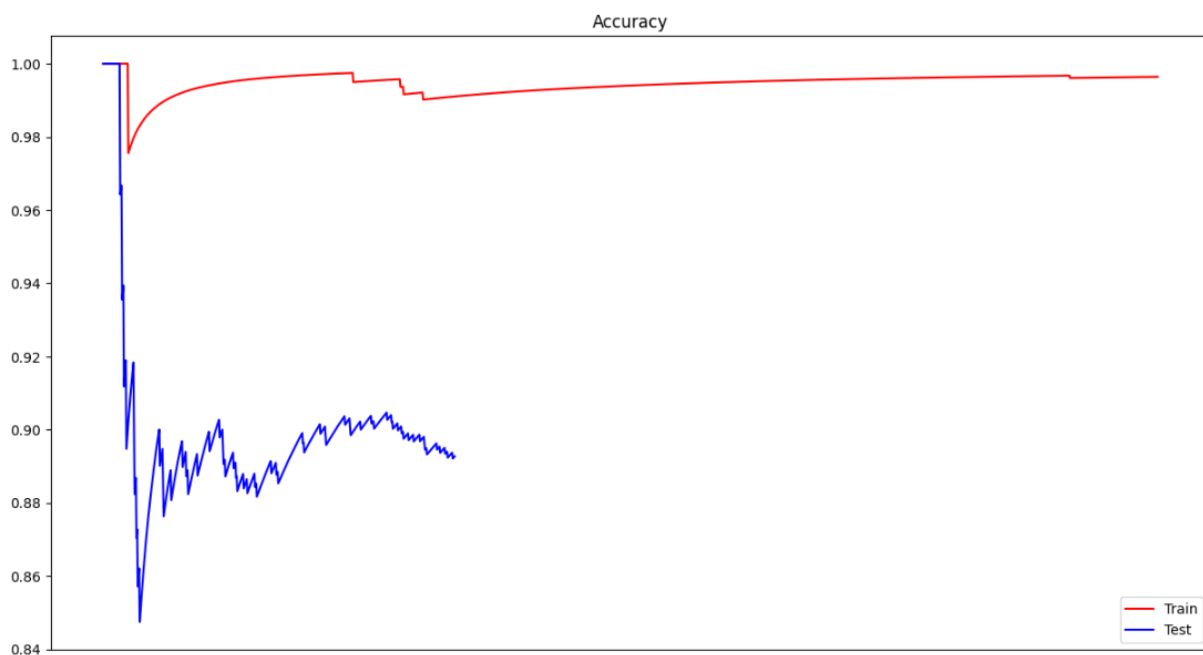


Рисунок 25 – График ассигасу по количеству предсказаний во время обучения и тестирования модели SVM на датасете Кэрол Клэйр

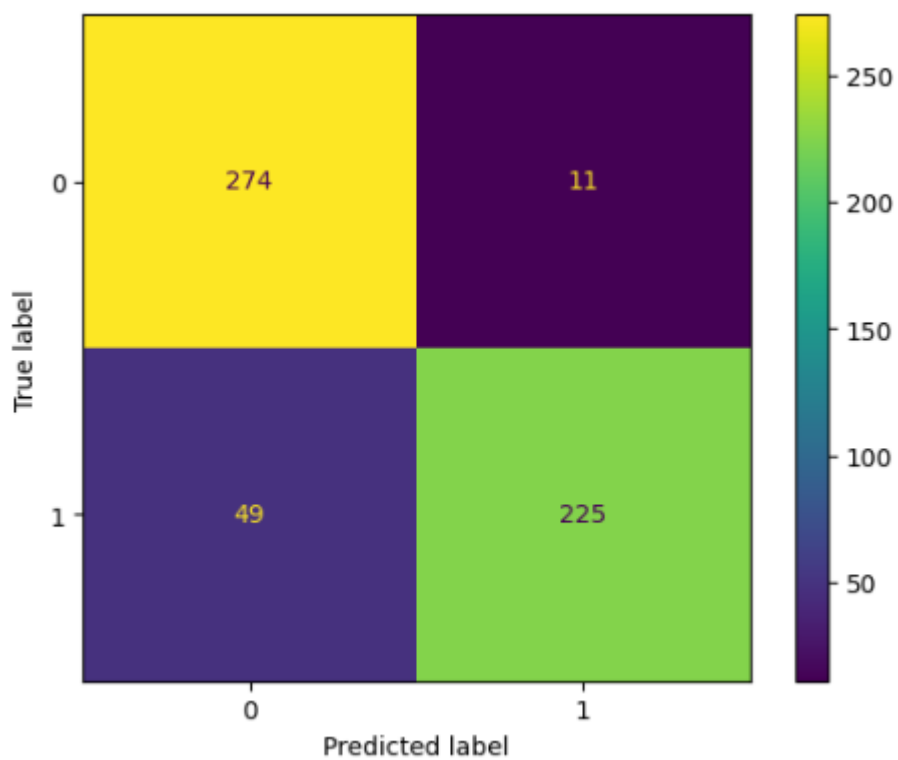


Рисунок 26 – Матрица ошибок для модели SVM на датасете Кэрол Клэйр

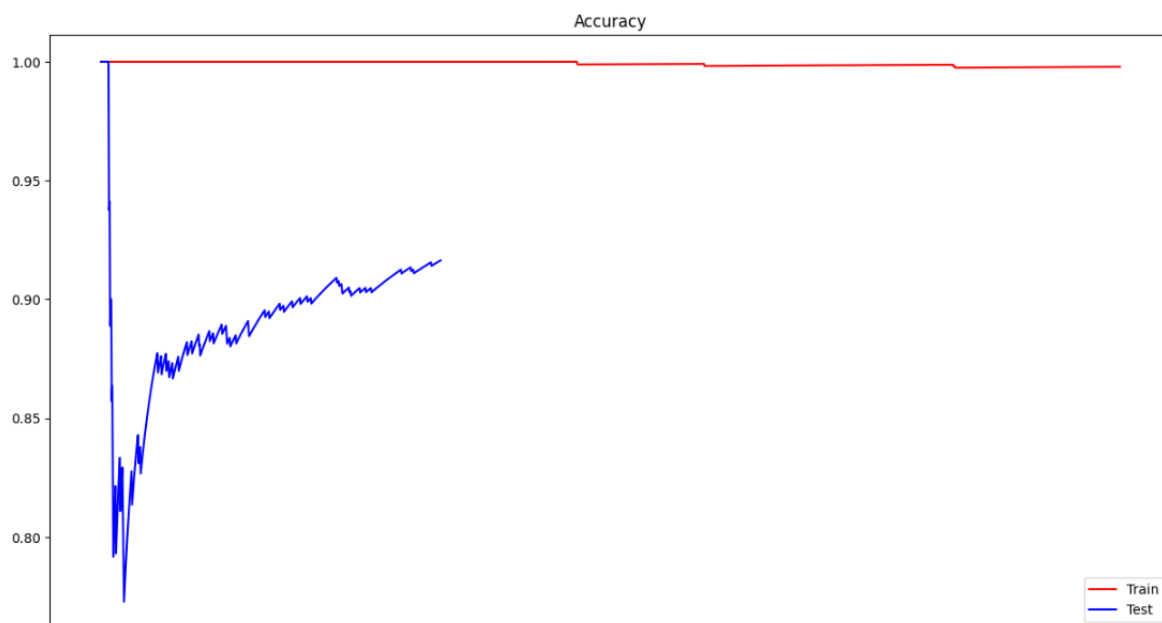


Рисунок 27 – График ассигасу по количеству предсказаний во время обучения и тестирования модели SVM на датасете Кейт Саймс

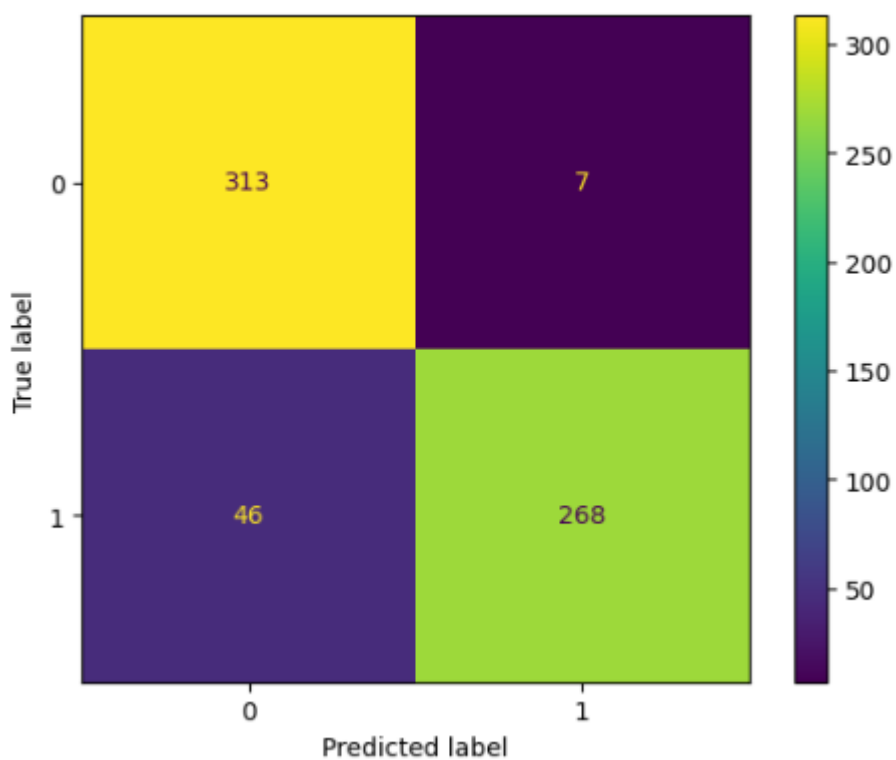


Рисунок 28 – Матрица ошибок для модели SVM на датасете Кейт Саймс

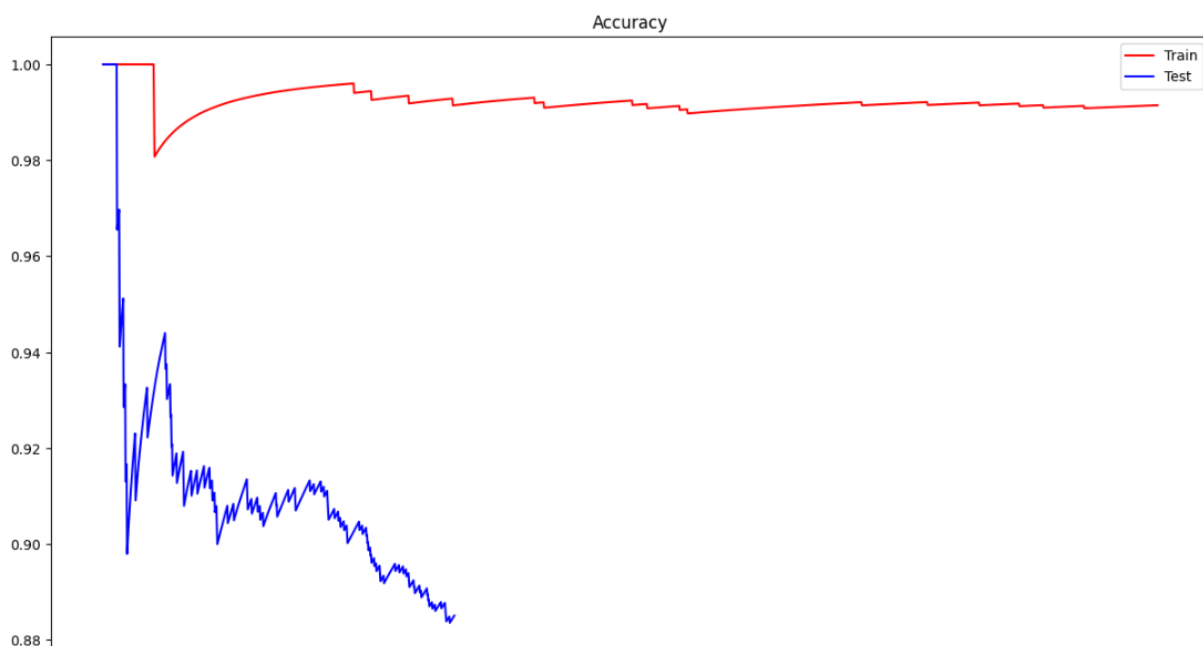


Рисунок 29 – График ассигасу по количеству предсказаний во время обучения и тестирования модели SVM на датасете Салли Бек

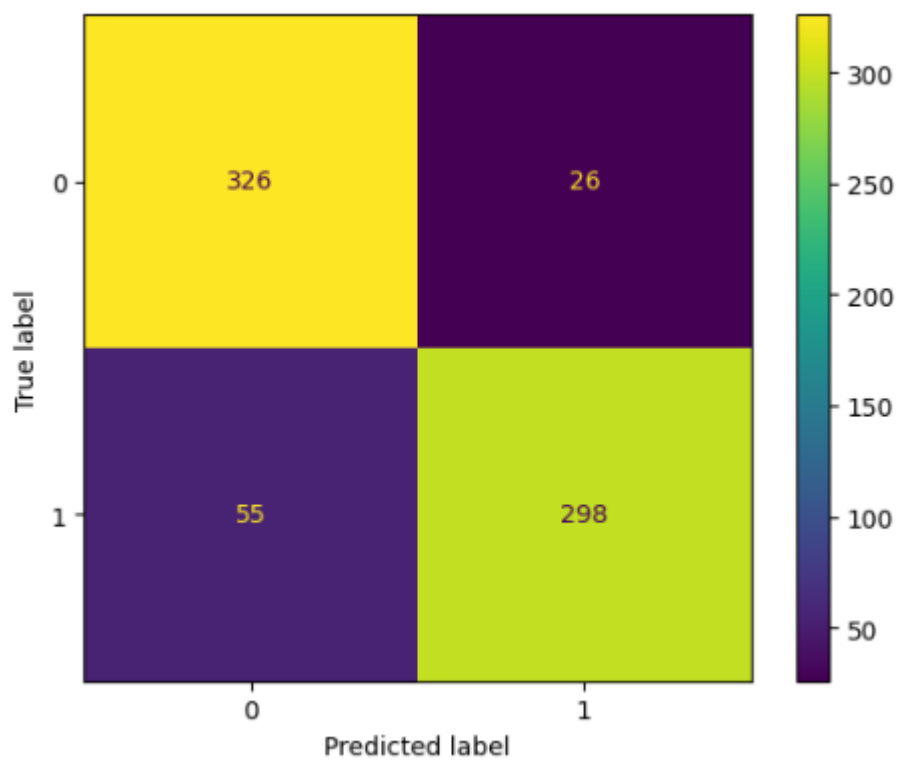


Рисунок 30 – Матрица ошибок для модели SVM на датасете Салли Бек

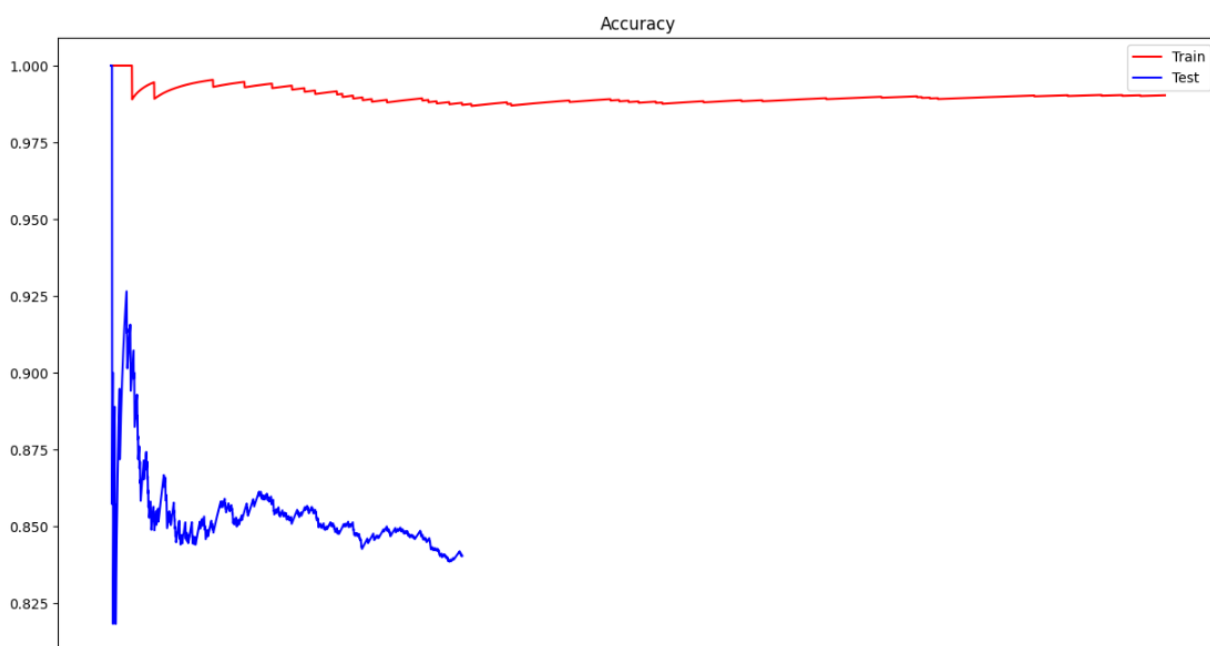


Рисунок 31 – График ассигасу по количеству предсказаний во время обучения и тестирования модели SVM на датасете Крис Германи

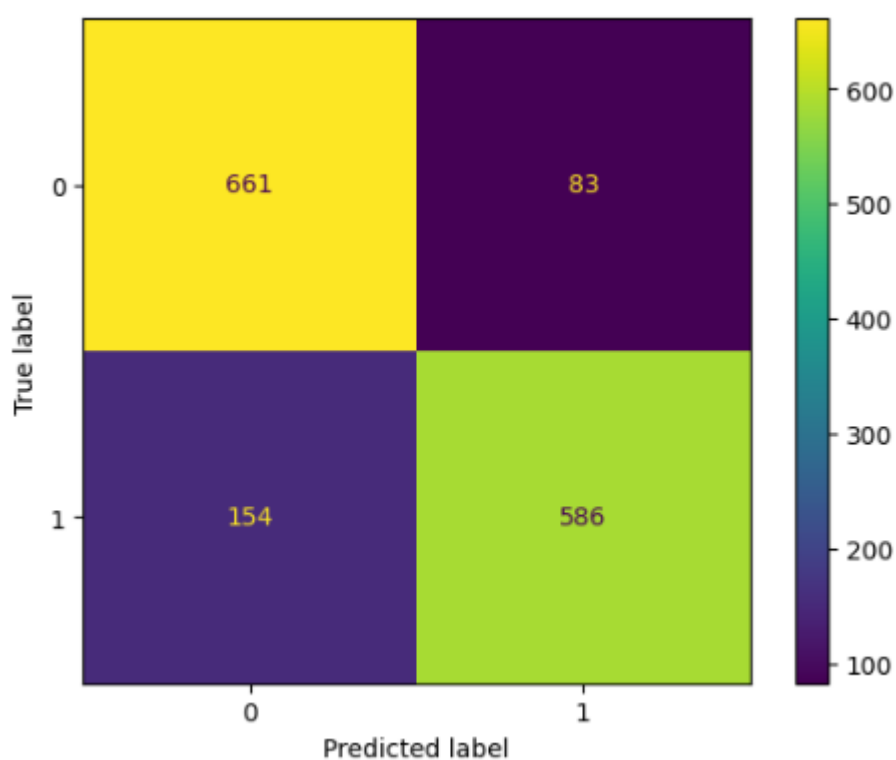


Рисунок 32 – Матрица ошибок для модели SVM на датасете Крис Германи

Также хочется отметить расположенный на третьем месте наивный Байесовский классификатор с ассигасу 86% и лучшим среди всех моделей

recall - 92%. По точности предсказаний данная модель проигрывает трансформеру BERT, но скорости обучения и предсказания этой модели являются одними из самых лучших. Результаты представлены на рисунках 33, 34, 35, 36, 37, 38, 39, 40, 41, 42.

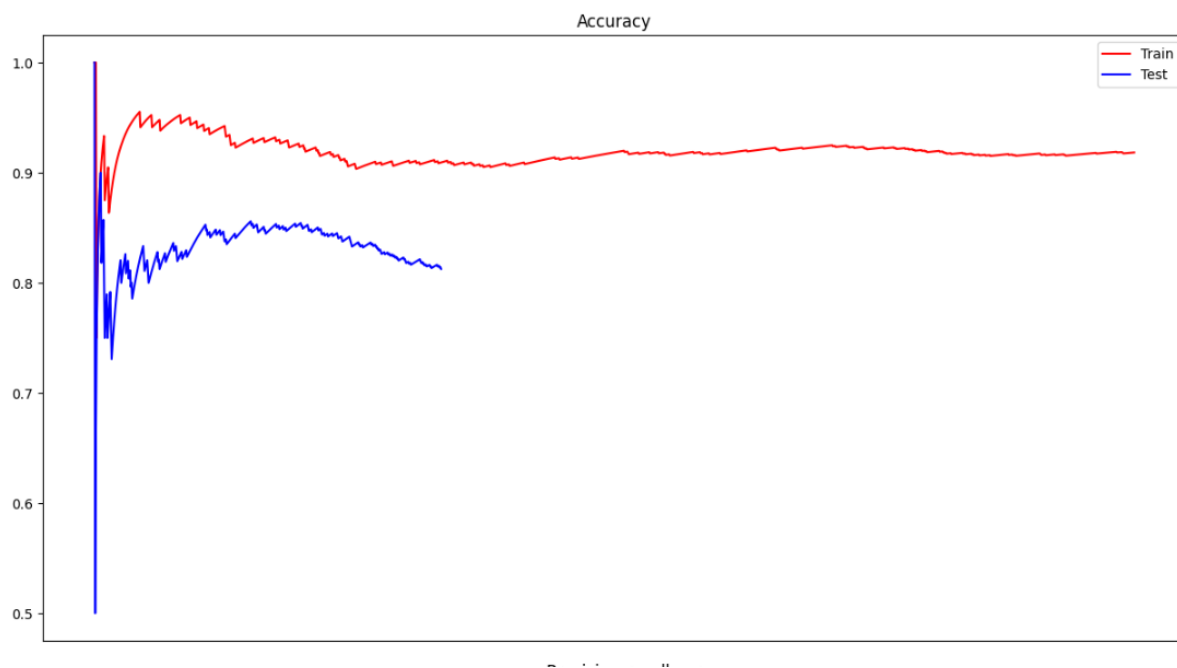


Рисунок 33 – График ассигасы по количеству предсказаний во время обучения и тестирования наивного Байесовского классификатора на датасете Мишель Кэш

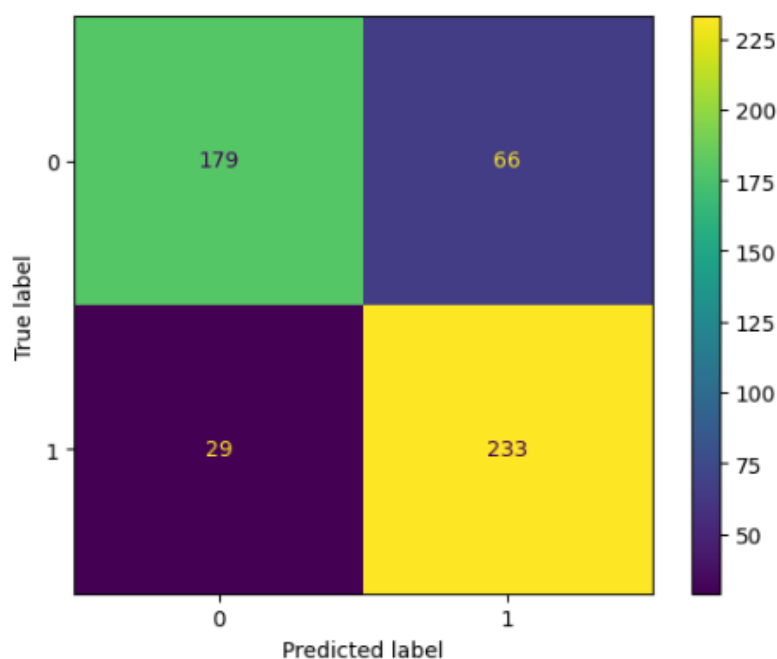


Рисунок 34 – Матрица ошибок для наивного Байесовского классификатора на датасете Мишель Кэш

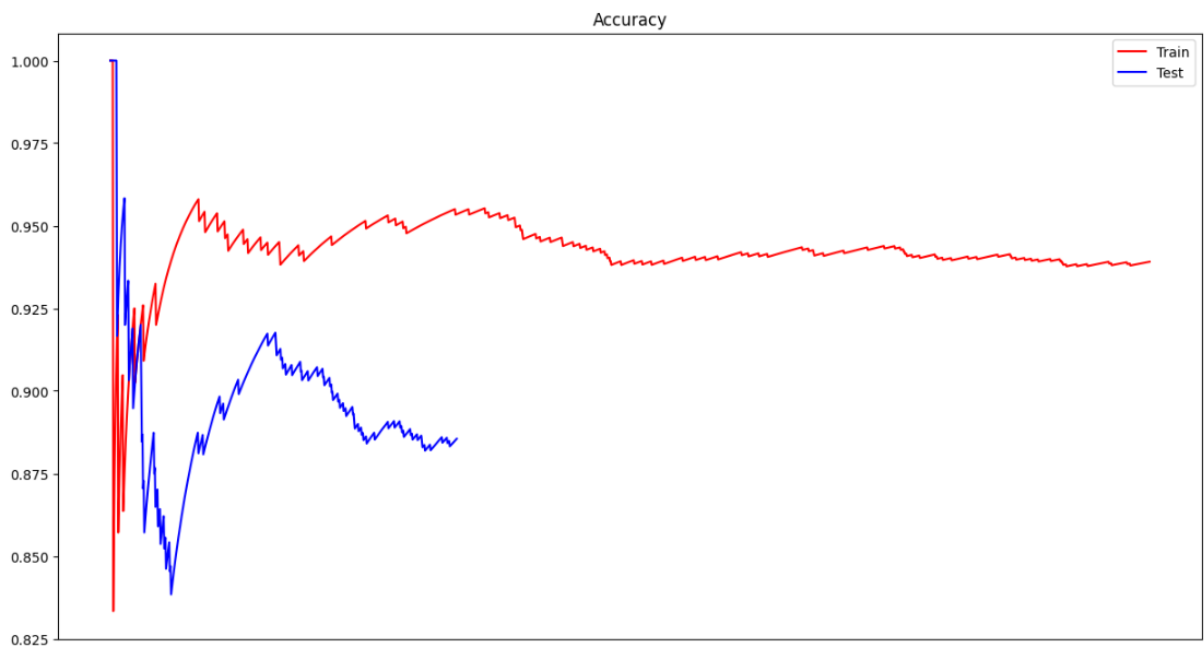


Рисунок 35 – График ассигасы по количеству предсказаний во время обучения и тестирования наивного Байесовского классификатора на датасете Кэрл Клэйр

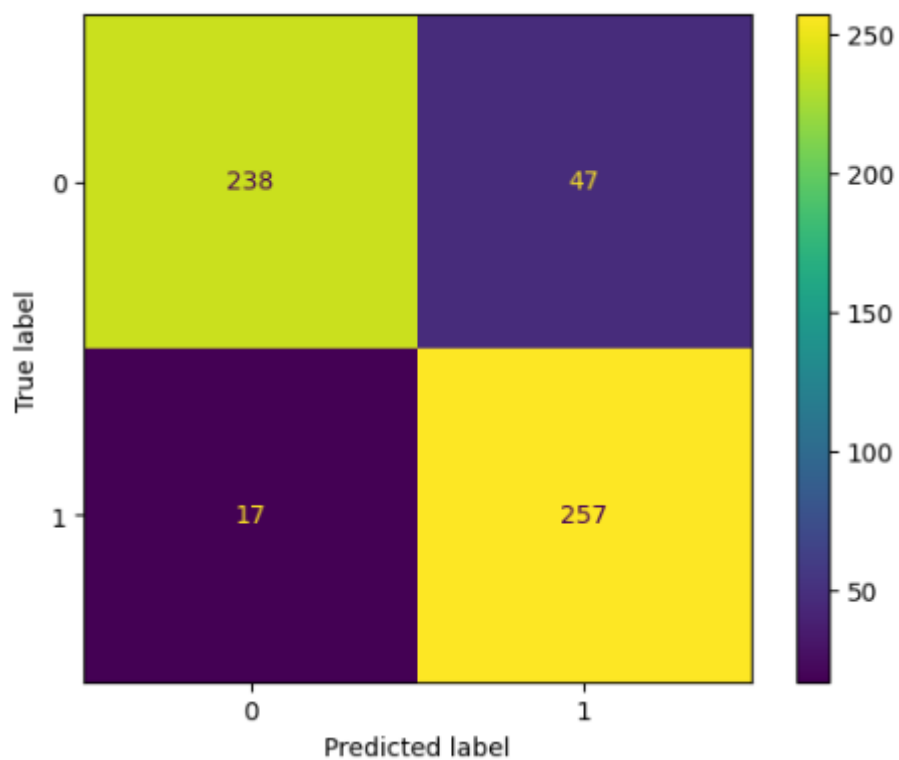


Рисунок 36 – Матрица ошибок для наивного Байесовского классификатора на датасете Кэрл Клэйр

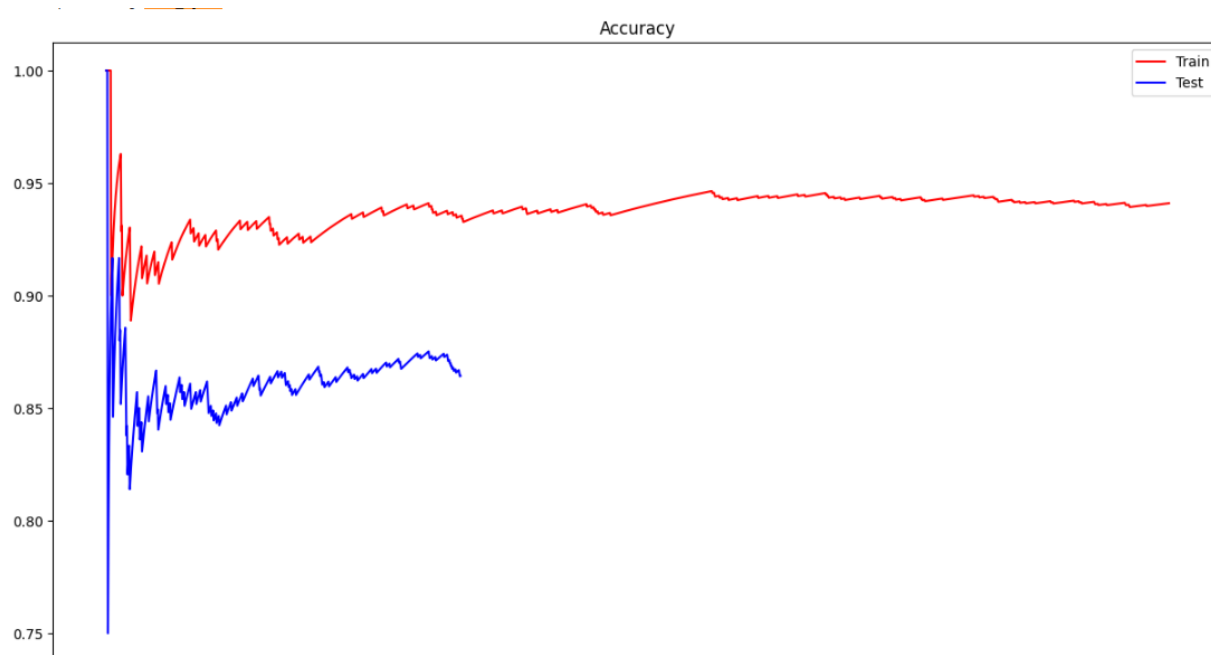


Рисунок 37 – График ассигасы по количеству предсказаний во время обучения и тестирования наивного Байесовского классификатора на датасете Кейт Саймс

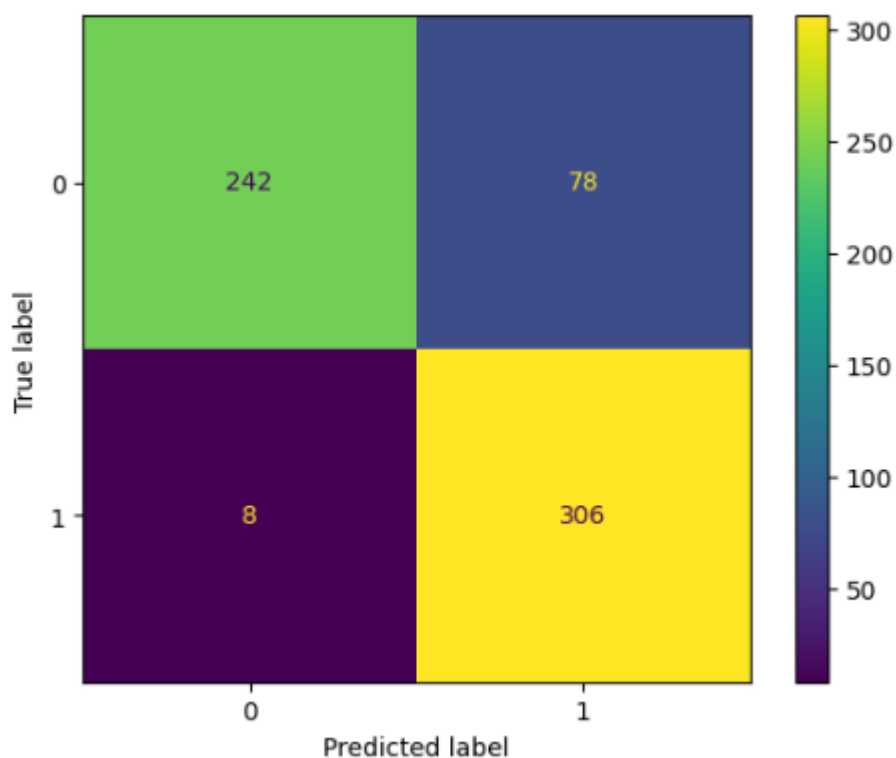


Рисунок 38 – Матрица ошибок для наивного Байесовского классификатора на датасете Кейт Саймс

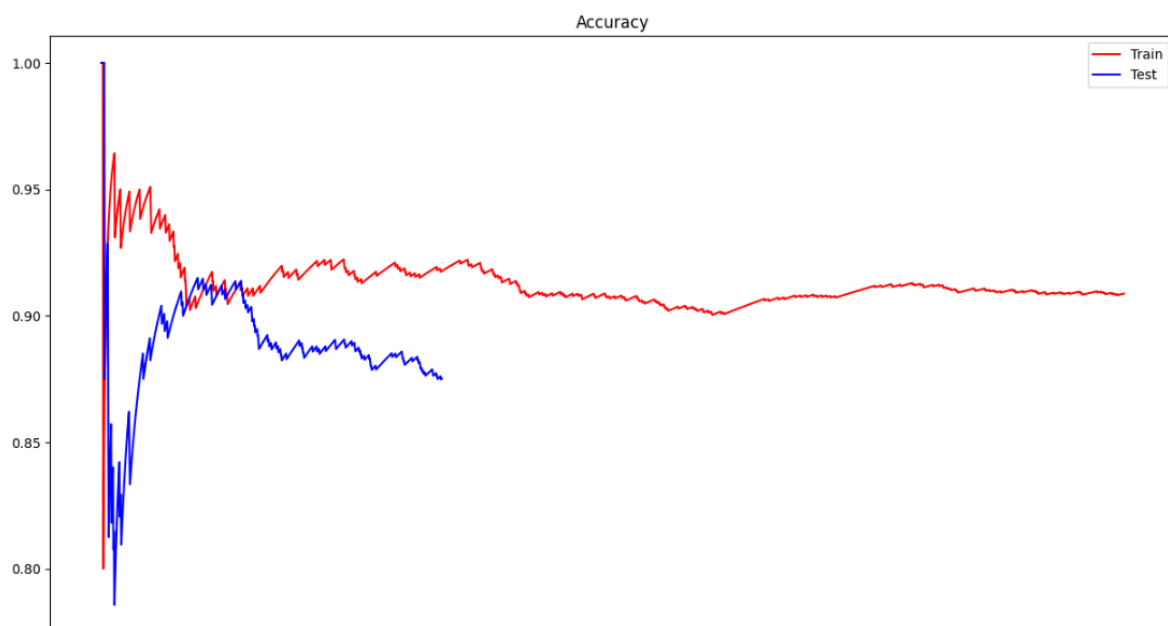


Рисунок 39 – График ассигасы по количеству предсказаний во время обучения и тестирования наивного Байесовского классификатора на датасете Салли Бек

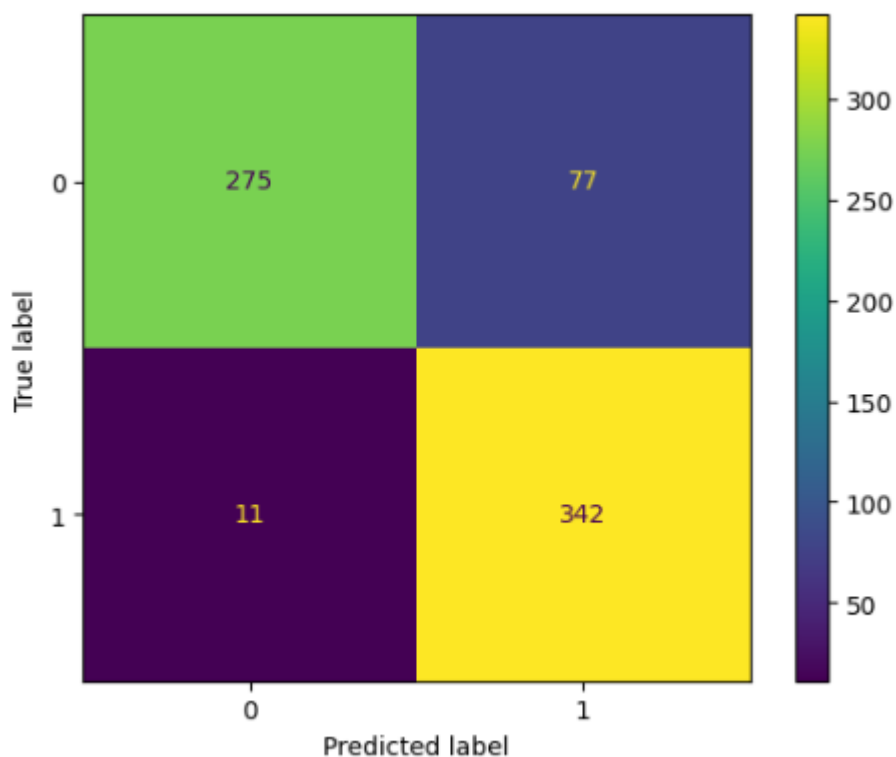


Рисунок 40 – Матрица ошибок для наивного Байесовского классификатора на датасете Салли Бек

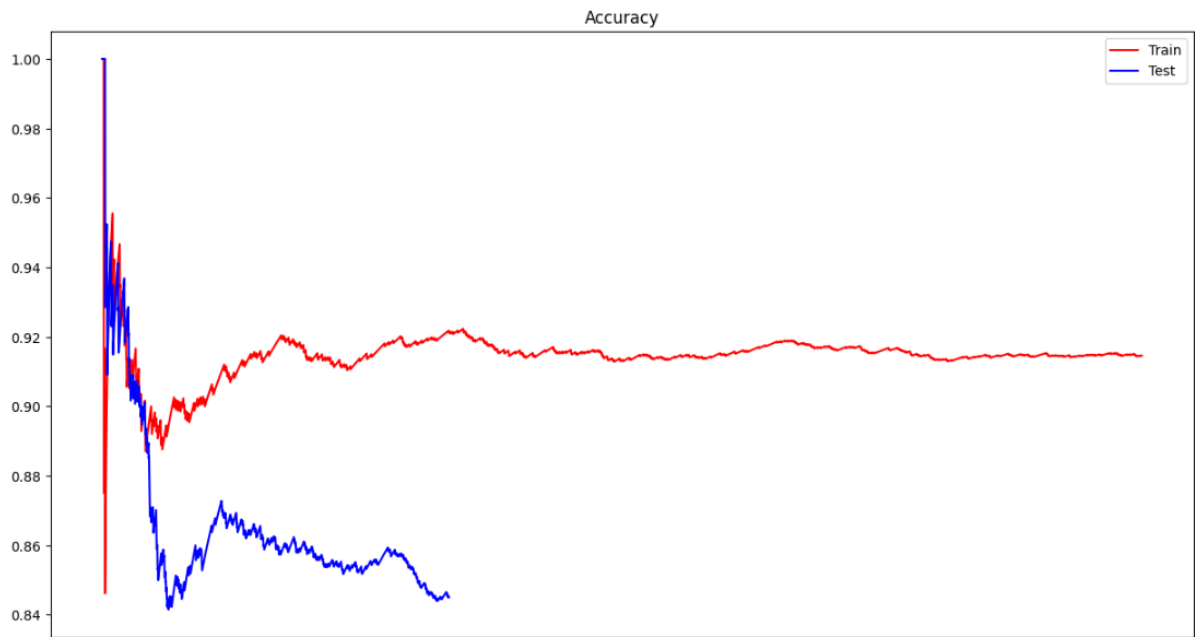


Рисунок 41 – График ассигасы по количеству предсказаний во время обучения и тестирования наивного Байесовского классификатора на датасете Крис Германи

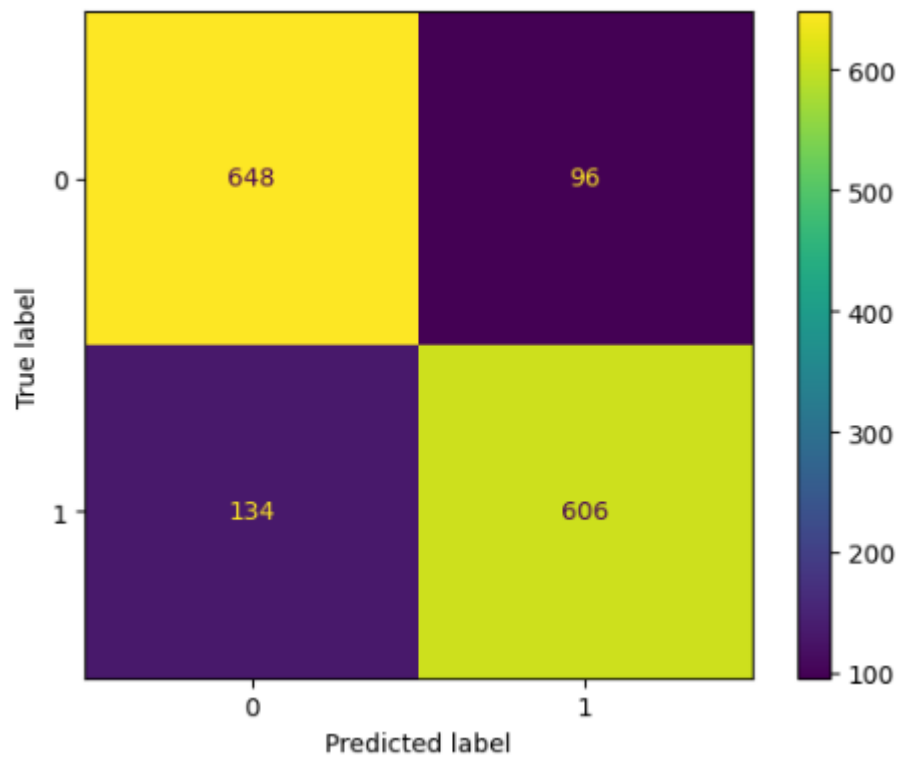


Рисунок 42 – Матрица ошибок для наивного Байесовского классификатора на датасете Крис Германи

ЗАКЛЮЧЕНИЕ

В выпускной квалификационной работе бакалавра была рассмотрена задача выявления признаков компрометации корпоративной электронной почты. Для этого было проведено сравнение различных методов машинного обучения для решения задачи бинарной классификации текстов реальных корпоративных писем.

Были рассмотрены следующие методы машинного обучения для классификации текстов писем по принадлежности определённому автору: наивный классификатор Байеса, логистическая регрессия, метод опорных векторов (SVM), метод k-ближайших соседей (KNN), дерево решений, случайный лес, градиентный бустинг, перцептроны, рекуррентные (RNN, LSTM), свёрточные (CNN), глубокие (трансформер BERT) нейронные сети.

В результате лучшей моделью, способной подтвердить принадлежность писем автору для решения задачи обнаружения признаков компрометации корпоративной электронной почты, является трансформер BERT. Нейронная сеть правильно определяет принадлежность письма автору в 87% случаев, что является очень хорошим результатом в связи со сложностью данной задачи из-за небольших размеров текстов деловой корреспонденции и схожими письменными стилями сотрудников.

Данная модель позволит создать надёжную интеллектуальную систему обнаружения признаков компрометации корпоративной электронной почты, которая сможет обучаться перед защитой определённого сотрудника компании или в автоматическом режиме. После этого будет анализировать исходящую корреспонденцию данного сотрудника. Эта система будет использоваться в продукте компании F.A.C.C.T. по защите корпоративной электронной переписки.

В качестве дальнейшего улучшения данной интеллектуальной системы можно использовать модели машинного обучения для выделения последнего сообщения, удаления подписей и обращений из текстов писем. Это предположительно может улучшить количество выявления случаев компрометации корпоративной электронной почты, так как злоумышленники скорее всего будут подписываться и отвечать на предыдущие письма как автор письма, а обычные эвристические методы не всегда способны выделить нужное содержание из текстов писем.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Сикорский О. Что такое BEC-атака и как ей противостоять. — URL: <https://www.kaspersky.ru/blog/what-is-bec-attack/27623/> (дата обращения 29.03.2024).
2. Callie H. B. Report on BEC and VEC attacks. — URL: <https://abnormalsecurity.com/blog/bec-vec-attacks> (дата обращения 29.03.2024).
3. Business email compromise attack: CEO's hacked email account Results in \$3M in Stolen Funds. — URL: <https://www.certifid.com/article/business-email-compromise-how-big-is-that-phish> (дата обращения 02.04.2024).
4. Безопасность корпоративной электронной почты. — URL: <https://www.facet.ru/products/business-email-protection/> (дата обращения 29.03.2024).
5. Enron email dataset. — URL: <https://www.cs.cmu.edu/~enron/> (дата обращения 03.04.2024).
6. Mail-parser documentation. — URL: <https://pypi.org/project/mail-parser/> (дата обращения 03.04.2024).
7. Обработка естественного языка. — URL: https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0_%D0%B5%D1%81%D1%82%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0 (дата обращения 06.04.2024).
8. Булыга Ф. С. и Курейчик В. М. Сравнительный анализ методов векторизации текстовых данных большой размерности // Известия ЮФУ. Технические науки. — 2023. — № 232. — С. 212—226.
9. Bag of words. — URL: <https://www.ibm.com/topics/bag-of-words> (дата обращения 06.04.2024).
10. Извлечение признаков из текстовых данных с использованием TF-IDF. — URL: <https://habr.com/ru/companies/otus/articles/755772/> (дата обращения 06.04.2024).

11. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin [и др.] // North American Association for Computational Linguistics. — 2019. — С. 4171—4186.
12. Батура Т. В. Методы автоматической классификации текстов // Программные продукты и системы. — 2017. — Т. 23. — С. 85—99.
13. Виды нейронных сетей. — URL: <https://iis.guu.ru/blog/vidy-neironnih-setey/> (дата обращения 06.04.2024).
14. Wickramasinghe I., Kalutarage H. Naive Bayes: applications, variations and vulnerabilities // Soft computing. — 2021. — Т. 25. — С. 2277—2293.
15. Raj A. How to build a logistic regression model for classification. — URL: <https://builtin.com/articles/logistic-classifier> (дата обращения 07.04.2024).
16. Метод опорных векторов SVM. — URL: <https://scikit-learn.ru/1-4-support-vector-machines/> (дата обращения 07.04.2024).
17. K-Nearest Neighbor (KNN) algorithm for machine learning. — URL: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning> (дата обращения 07.04.2024).
18. Decision trees. — URL: <https://scikit-learn.org/stable/modules/tree.html> (дата обращения 07.04.2024).
19. Random forest algorithm. — URL: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm> (дата обращения 07.04.2024).
20. Nelson D. Gradient Boosting Classifiers in Python with Scikit-Learn. — URL: <https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/?ref=blog.paperspace.com> (дата обращения 09.04.2024).
21. Митина О. А и Ломовцев П. П. Перцептрон в задачах бинарной классификации // Национальная ассоциация учёных. — 2021. — № 66. — С. 39—44.
22. Sherstinsky A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network // Physica D: Nonlinear Phenomena. — 2020. — Т. 404. — С. 1—43.

23. Воробьев Е. В. и Пучков Е. В. Классификация текстов с помощью сверточных нейронных сетей // Молодой исследователь Дона. — 2017. — № 6. — С. 2—7.

24. Салып Б. Ю. и Смирнов А. А. Анализ модели BERT как инструмента определения меры смысловой близости предложений естественного языка // StudNet. — 2022. — № 5. — С. 3509—3518.

25. Основные метрики задач классификации в машинном обучении. — URL: <https://webiomed.ru/blog/osnovnye-metriki-zadach-klassifikatsii-v-mashinnom-obuchenii/> (дата обращения 11.04.2024).

ПРИЛОЖЕНИЕ А

Исходный код

Репозиторий с исходным кодом расположен по адресу https://github.com/RomaMaster228/bec_detection. Ссылка в формате QR-кода представлена на рисунке А.1.



Рисунок А.1 – QR-код с ссылкой на репозиторий