МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ПОВОЛЖСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНОЛОГИЧЕСКИЙ
УНИВЕРСИТЕТ»

На правах рукописи

Корелов Сергей Викторович

МЕТОД И АЛГОРИТМ ОБНАРУЖЕНИЯ СПАМА НА ОСНОВЕ ВЫДЕЛЕНИЯ ПРИЗНАКОВ ЭЛЕКТРОННЫХ ПИСЕМ С ИСПОЛЬЗОВАНИЕМ КОНТЕНТНОЙ ФИЛЬТРАЦИИ

Специальность 2.3.6. Методы и системы защиты информации, информационная безопасность

Диссертация
на соискание ученой степени
кандидата технических наук

Научный руководитель:
доктор технических наук,
профессор Сидоркина И. Г.

Йошкар-Ола – 2024

ОГЛАВЛЕНИЕ

Введение

Общая характеристика работы

В настоящее время одним из наиболее распространенных способов повседневной и деловой коммуникации, а также управления являются электронные почтовые сообщения. Однако столь высокая популярность электронной почты сопровождается и рядом проблем. Одним из ставших классическим рисков, связанным с ее использованием, является спам, т. е. анонимные массовые непрошенные рассылки [1]. Указанный вектор рассматривается в мировом сообществе информационной безопасности как один из основных векторов компрометации информационных систем организаций. Общая доля таких сообщений составляет в среднем не менее 50 % от общего количества сообщений электронной почты в трафике [2-6].

Доказано, что спам является угрозой безопасности информации, нейтрализация которой является актуальной задачей. В связи с этим исследование, разработка, создание и внедрение новых и совершенствование существующих решений, моделей, алгоритмов, средств, систем и технологий обеспечения безопасности информационных систем, ориентированных на обнаружение (выявление) спама, является актуальной и практически значимой задачей.

Наиболее распространенные способы, используемые в мировой практике для выявления спама, заключаются [7] в анализе заголовков и содержимого сообщений электронной почты. Существующие методы фильтрации [7, 8], используемые для анализа заголовков, достаточно легко обходятся отправителями спама. При этом наиболее эффективными для выявления спама [7, 9] считаются методы машинного обучения, эффективность которых в решении задач классификации текстов обратила внимание исследователей на обучаемые модели [8], которые и положены в основу второго подхода.

Вопросам составления моделей и анализа различных текстов, их классификации, а также методам машинного обучения посвящены работы российских и зарубежных ученых М. С. Агеева, В. Н. Вапника, К. В. Воронцова,

Б. В. Доброва, Н. Г. Загоруйко, К. Г. Кирьянова, Н. Н. Леонтьевой, Н. В. Лукашевич, Л. Н. Федотовой, В. И. Шалака, Т. Э. Шульги, K. Aas, A. Dasgupta, H. Drucker, C. Manning, F. Sebastiani, A. Uysal и многих других. Ими проведены исследования и предложены теоретические и прикладные подходы к анализу текстов и составлению их моделей, а также классификации текстов с применением различных методов машинного обучения.

Многие исследования последних лет в области выявления спама, направленные на анализ содержимого сообщений электронной почты, посвящены оценке эффективности методов машинного обучения и вопросам формирования признакового пространства электронных писем. Такого рода исследованиям в области обнаружения спама посвящены работы российских и зарубежных исследователей Б. В. Доброва, А. С. Катасёва, М. П. Малыхиной, Е. М. Мезенцевой, А. П. Никитина, А. Н. Розинкина, М. А. Семеновой, П. Б. Хорева, В. А. Частиковой, Е. Н. Чернопрудовой, I. Androutsopoulos, W. Cohen, S. Delany, H. Drucker, K. Junejo, K. Gee, P. Graham, V. Metsis, G. Robinson, M. Sahami, G. Sakkis, H. Shen и многих других. Ими проведены исследования и предложены теоретические и прикладные подходы к решению вопросов:

- обнаружения спама на основе анализа содержимого электронных писем с составлением моделей писем и классификации текстовой информации, содержащейся в электронных письмах, с применением различных методов машинного обучения, таких как наивный Байесовский классификатор, искусственные нейронные сети, деревья решений, искусственные иммунные системы, метод опорных векторов, k-ближайших соседей и некоторые другие;

- оценки в различных условиях эффективности применения методов машинного обучения в задаче обнаружения спама;

- отбора признаков, необходимых для классификации электронных писем.

В качестве базовых признаков, определяющих содержание сообщений электронной почты, используются слова; лексемы; словосочетания слов; термы (как последовательности символов и их устойчивые словосочетания); метрики

читаемости; характеристики жанра и стиля; глобальные статистические закономерности; различные лексические особенности электронных писем; предложения как минимальные семантические единицы. Для оценки значимости этих признаков применяются различные веса.

Для автоматического построения списка слов с их весами могут использоваться методы машинного обучения, входными наборами данных для которых являются спам и легальные[1] письма пользователей. Наибольшие издержки классификации [10] формируются при неправильном отнесении:

- легальных писем к классу спама (ложноположительная классификация);
- спама к классу легальных писем (ложноотрицательная классификация).

При этом целью злоумышленников (отправителей спама) является снижение вероятности выявления спама, для чего содержание спамовых писем наполняется наиболее употребимыми словами легальных сообщений электронной почты [11].

Вместе с тем в основу экспериментов многих из проведенных исследований положены различные и недоступные в открытом доступе наборы электронных писем. Это не позволяет осуществить прямое сравнение эффективности предлагаемых авторами подходов и выбор абсолютно лучшего решения [8].

Результаты проведенного анализа отечественной и зарубежной практики за последние несколько лет в предметной области обнаружения спама показывают, что задача выявления спама решается в основном схожими известными методами классификации [12]. Вместе с тем много внимания уделяется работам по отбору признаков сообщений электронной почты, позволяющих повысить эффективность применения выбранных методов классификации.

Таким образом, можно утверждать, что научным сообществом сформирована достаточно устойчивая система технологий, методов и средств обнаружения спама [13]. При этом очевидная сложность данной задачи, заключающаяся в наличии различных интересов у различных пользователей [12], не позволяет сформировать

[1] Здесь и далее применительно к настоящему диссертационному исследованию под легальным сообщением (сообщением, не относящимся к спаму) понимается электронное сообщение, доставленное абоненту и (или) пользователю с их предварительного согласия и позволяющее определить отправителя этого сообщения, т. е. не подпадающее под определение спама в соответствии с постановлением Правительства Российской Федерации от 10 сентября 2007 года № 575 «Об утверждении Правил оказания телематических услуг связи».

универсальное описание спамовых писем и, как следствие, соответствующее универсальное решение [10, 14]. Кроме того, многие исследования и сформированные в них подходы не оперируют информационными интересами конкретных пользователей и не обеспечивают выявление легальных сообщений электронной почты, что приводит к их ложной классификации.

Сложившаяся ситуация ярко демонстрирует актуальность проблемы выбора признаков сообщений электронной почты, обеспечивающих высокое качество выявления спама и идентификации легальных сообщений, с учетом персональных информационных потребностей пользователя при классификации сообщений электронной почты [11]. Решение показанной проблемы требует разработки новых или совершенствования текущих технологий, методов и средств, учитывающих содержание сообщений электронной почты конкретного пользователя и оценку эффективности применяемых методов [12], что позволит достичь одного из ключевых свойств средств выявления спама – персонализации[2] [11], а также повышению эффективности обнаружения спама.

Поэтому создание модели электронных писем, обеспечивающей выделение признаков электронных почтовых сообщений на основе их содержания, для обнаружения спама является актуальной задачей и представляет научный и практический интерес. Построение модели осуществлялось на базе математических моделей текстов [15, 16] и их последующего анализа с использованием «генетических карт». В основе предлагаемого метода лежит теория структурной идентификации и анализа текстовой информации с помощью базовых параметров, идеологом которой выступил доктор технических наук, профессор К. Г. Кирьянов. Указанная теория применялась в различных научных областях для решения задач идентификации и анализа текстов, однако не применялась ранее в области обнаружения спама.

[2] Здесь и далее применительно к настоящему диссертационному исследованию под персонализацией понимается ориентированность на персональные (пользовательские) особенности (с т. з. информационных потребностей) электронных писем и их содержание применительно к конкретным пользователям (группе пользователей).

Объектом исследования в диссертационной работе являются технологии обнаружения спама.

Предметом исследования являются модели электронных писем и алгоритмы обнаружения спама.

Целью диссертационной работы является повышение эффективности обнаружения спама и достоверности идентификации легальных электронных почтовых сообщений на основе классификации их содержания.

Для достижения поставленной цели в работе решались следующие **задачи исследования**:

1. Анализ современного состояния исследований в области обнаружения спама (соответствует п. 3, 5 паспорта специальности 2.3.6).

2. Разработка модели электронного почтового сообщения, учитывающей содержание электронных писем конкретного пользователя (персонализацию) (соответствует пп. 3, 5, 15 паспорта специальности 2.3.6).

3. Разработка метода классификации электронных писем для обнаружения спама и идентификации легальных электронных писем (соответствует пп. 3, 5, 15 паспорта специальности 2.3.6).

4. Разработка алгоритма классификации электронных писем (соответствует п. 3, 5, 15 паспорта специальности 2.3.6).

5. Разработка архитектуры подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем (соответствует п. 3, 5, 15 паспорта специальности 2.3.6).

Научная новизна

1. Разработана модель электронного почтового сообщения для классификации электронных писем на основе метода «генетических карт», отличающаяся от известных моделей методом выделения значимых последовательностей символов текста (признаков электронных писем на основе их содержания, термов), позволяющим усилить смысловое содержание термов (соответствует пп. 3, 5, 15 паспорта специальности 2.3.6).

2. Разработан метод классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, основанный на положениях задачи классификации текстовых документов, отличающийся использованием разработанной модели электронных писем, применение которого позволяет повысить эффективность обнаружения спама и достоверность идентификации легальных электронных писем, а также снизить количество неклассифицированных писем (соответствует пп. 3, 5, 15 паспорта специальности 2.3.6).

3. Разработан алгоритм классификации электронных писем на основе методов машинного обучения, отличающийся наличием дополнительной процедуры определения «схожести» термов на основе расстояния Левенштейна[3], обеспечивающей вычисление мер принадлежности классифицируемого электронного письма к классам спама и легальных для повышения достоверности идентификации электронных писе, позволяющий осуществить программную реализацию разработанных модели и метода (соответствует пп. 3, 5, 15 паспорта специальности 2.3.6).

4. Разработана архитектура подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем на основе разработанного алгоритма, отличающаяся от известных блоком выделения термов и блоком нечеткой классификации, реализующая предложенные в работе метод и алгоритм, применение которых позволяет повысить достоверность идентификации легальных электронных писем с учетом меняющихся информационных потребностей конкретного пользователя (персонализации) (соответствует пп. 3, 5, 15 паспорта специальности 2.3.6).

[3] Расстояние Левенштейна – минимальное количество операций удаления, вставки и замены символа, необходимое для преобразования одной строки в другую. Используется наиболее часто для вычисления редакционного расстояния (метрики, измеряющей разность между двумя последовательностями символов), а также для исправления ошибок в слове (в поисковых системах, базах данных, при вводе текста, при автоматическом распознавании отсканированного текста или речи), сравнения текстовых файлов утилитой diff и ей подобными, а также в биоинформатике для сравнения генов, хромосом и белков.

Теоретическая значимость работы

Теоретическая значимость полученных результатов заключается в том, что в работе предложены новая модель электронного почтового сообщения, учитывающая содержание электронных писем конкретного пользователя (персонализацию), метод классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, алгоритм классификации электронных писем.

Практическая значимость работы

Практическая значимость полученных результатов заключается в разработке программных модулей исследовательского прототипа подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем. Применение разработанных модели и метода позволяет повысить эффективность обнаружения спама и достоверность идентификации легальных электронных писем с учетом меняющихся информационных потребностей конкретного пользователя (персонализации) с точностью классификации до 0,995 и полнотой классификации до 0,993, а также снизить количество ошибочно классифицированных и неклассифицированных писем.

Методы и методология исследования. Для решения поставленных в работе задач были использованы методы интеллектуального анализа данных и защиты информации, теория систем и системного анализа, теория принятия решений, теория эксперимента, методы контент-анализа, методы машинного обучения, методы теории вероятностей и математической статистики, методы объектно-ориентированного анализа и проектирования.

Положения, выносимые на защиту

1. Модель электронного почтового сообщения, учитывающая содержание электронных писем конкретного пользователя (персонализацию) (соответствует пп. 3, 5, 15 паспорта специальности 2.3.6).

2. Метод классификации электронных писем для обнаружения спама и идентификации легальных электронных писем (соответствует пп. 3, 5, 15 паспорта специальности 2.3.6).

3. Алгоритм классификации электронных писем (соответствует пп. 3, 5, 15 паспорта специальности 2.3.6).

4. Архитектура подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем (соответствует пп. 3, 5, 15 паспорта специальности 2.3.6).

Достоверность и обоснованность научных положений и выводов, полученных в диссертационной работе, подтверждается корректной постановкой задач, применением известных технологий и методов, успешно используемых в других прикладных областях, апробацией разработанных модели, метода, алгоритма и программного модуля. Выводы и положения диссертации научно обоснованы и подтверждены положительными оценками на научных конференциях и результатами экспериментальных исследований автора.

Апробация результатов диссертации. Основные положения и результаты диссертации докладывались и обсуждались на научных конференциях: X, XII, XIV, XV, XXIV, XXV и XXVI научных конференциях по радиофизике в Национальном исследовательском Нижегородском государственном университете им. Н. И. Лобачевского (г. Нижний Новгород, 2006, 2008, 2010, 2011, 2020-2023 годы); Международной научно-технической конференции студентов, аспирантов и молодых ученых «Научная сессия ТУСУР – 2010» (г. Томск, 2010 год); научно-технической конференции «Автоматизированные системы управления и информационные технологии» АСУИТ-2020 (г. Пермь, 2020 год); XII Международной Интернет-конференции молодых ученых, аспирантов и студентов «Инновационные технологии: теория, инструменты, практика» InnoTech-2020 (г. Пермь, 2020 год); VI Всероссийской молодежной научно-практической конференции с международным участием «Информационные технологии обеспечения комплексной безопасности в цифровом обществе» (19-20 мая 2023 г., г. Уфа); Международном конгрессе по интеллектуальным системам и информационным технологиям (2-9 сентября 2023 г., Россия, Черноморское побережье, Геленджик-Дивноморское).

Результаты диссертационной работы внедрены в ООО «Омега Софт» (г. Йошкар-Ола), ООО «ТРЭВЕЛ ЛАЙН СИСТЕМС» (г. Йошкар-Ола) и в учебный процесс кафедры информационной безопасности ФГБОУ ВО «Поволжский государственный технологический университет», г. Йошкар-Ола.

Соответствие паспорту специальности. Результаты диссертационной работы соответствуют следующим пунктам паспорта научной специальности 2.3.6. «Методы и системы защиты информации, информационная безопасность»: п. 3 «Методы, модели и средства выявления, идентификации, классификации и анализа угроз нарушения информационной безопасности объектов различного вида и класса»; п. 5 «Методы, модели и средства (комплексы средств) противодействия угрозам нарушения информационной безопасности в открытых компьютерных сетях, включая Интернет»; п. 15 «Принципы и решения (технические, математические, организационные и др.) по созданию новых и совершенствованию существующих средств защиты информации и обеспечения информационной безопасности».

Публикация результатов работы. Основные результаты диссертации опубликованы в 19 печатных работ, в том числе в 4 статьях в научных изданиях из Перечня рецензируемых научных изданий, рекомендованных ВАК, в 15 статьях в других изданиях.

Структура и объем диссертации. Диссертация включает в себя введение, четыре главы с выводами, заключение, список литературы и приложения. Основной текст работы изложен на 181 странице, содержит 29 рисунков, 37 таблиц, 6 приложений. В список используемой литературы включено 203 наименования, среди которых 89 зарубежных и 114 отечественных публикаций.

Глава 1 Анализ проблемы обнаружения спама

1.1 Спам – угроза безопасности информации

Устойчивая популярность[4] использования электронной почты в бизнес-процессах организаций сформировала классические риски информационной безопасности, заключающиеся в возможности проведения компьютерных атак с использованием спама, т. е. анонимных массовых непрошенных рассылок [1]. Здесь имеет значение каждое включенное в него слово. Анонимная: пользователи электронной почты страдают, в основном, именно от рассылок электронных писем со скрытым или фальсифицированным обратным адресом. Массовая: рассылки такого типа именно массовые, и только они являются настоящим бизнесом для их отправителей и настоящей проблемой для получателей. Непрошенная: очевидно, легальные электронные письма и подписные рассылки не должны попадать под определение спама.

На текущий момент спамовые письма вызывают проблемы для многих операторов связи и организаций, заключающиеся, в том числе, в передаваемых объемах спама. Известная статистика показывает, что его доля в почтовом трафике сохраняется на высоком уровне (2018 – 52,48 % [2], 2019 – 56,51 % [3], 2020 – 50,37 % [4], 2021 – 45,56 % [5], 2022 – 48,63 % [6]).

Спамовые сообщения становятся причиной всевозможных проблем информационной безопасности получателей [например, 7, 9, 10, 14, 18-24]. Также он приводит к серьезному негативному эффекту для экономик стран всего мира. Если же говорить про конкретные финансовые потери от спама, то, например, со слов президента Ассоциации документальной электросвязи [25], ущерб операторов связи и пользователей интернета от рассылки несанкционированных рекламных сообщений в России ежегодно составляет около 55 млн. долларов. По оценке Департамента стратегического анализа ФБК [26] потери российской экономики от спама в 2008 году составили от 31,3 до 47,2 млрд. рублей, а по оценке компании

[4] По оценкам The Radicati Group, Inc. [17], в 2023 г. число пользователей электронной почты превысит 4,3 млрд. с прогнозом более 4,8 млрд. в 2027 году.

НП «РАЭК» – 14,1 млрд. рублей в 2009 году [27]. По информации компании Fastnet SA [28], ежегодный ущерб производительности компаний во всем мире от спама в денежном эквиваленте составляет порядка 1,6 млрд. фунтов стерлингов.

Существенные объемы входящих информационных потоков со значительным преобладанием в них спама создают нагрузку на элементы сетевых инфраструктур операторов связи и организаций. Поиск легальных электронных писем среди всего поступающего объема может быть сопряжен со значительными затратами временных и трудовых ресурсов и зачастую приводит к потерям необходимых и важных электронных писем. Рабочее время, потраченное на удаление спама, будет безвозвратно потеряно, и оно же будет оплачено из кармана работодателя. Так, например, одна компания в 2020 году получила около 300 тыс. писем спама всего за один день, что вынудило ее отключить затронутые учетные записи и сбросить учетные данные [29].

Результаты аналитической деятельности, сформированные в 2010 году компанией «Код Безопасности», по итогам опроса порядка 140 российских организаций, задействованных в государственном и коммерческом секторах экономики, показали, что угроза рассылки анонимных сообщений электронной почты различного содержания лицам, не изъявлявшим желание их получать, является наиболее распространенной (31 % от общего количества угроз) в сфере информационной безопасности [30]. Также компания F-Secure [31] определяет, что с использованием спама реализуется один из основных векторов проникновения в информационные системы организаций. По их мнению, вредоносное программное обеспечение содержится в 23% спамовых сообщений. При этом еще 31 % писем спама содержит ссылки на вредоносные ресурсы [31]. Также велико количество открываемого спама (14,2 % в первой половине 2018 года по сравнению с 13,4 % второй половины 2017 года) [31].

Таким образом, можно констатировать, что выявление спама в отношении пользователей организаций является одной из приоритетных задач любой системы обеспечения информационной безопасности. Этот тезис находит свое подтверждение в документе «Состав технических параметров компьютерного

инцидента, указываемых при представлении информации в ГосСОПКА, и форматы представления информации о компьютерных инцидентах» [32] и стандарте Банка России СТО БР ИББС-1.0-2014 [33], в соответствии с которым меры обеспечения информационной безопасности, позволяющие обеспечить противодействие распространению спама, являются обязательными в организациях банковской системы Российской Федерации. В соответствии с [32, 34-36] спам отнесен к категории компьютерных инцидентов, информацию о которых участник информационного взаимодействия должен передавать в Национальный координационный центр по компьютерным инцидентам и в Центр мониторинга и реагирования на компьютерные атаки в кредитно-финансовой сфере Главного управления Банка России соответственно.

Таким образом, обоснована актуальность задачи нейтрализации угрозы информационной безопасности получения пользователями спама. Следовательно, разработка и совершенствование существующих технологий, методов и средств информационной безопасности, направленных на выявление спамовых сообщений электронной почты, является актуальной и практически значимой задачей.

1.2 Существующие исследования в области обнаружения спама

Большое количество исследований в области выявления спама показывают отсутствие идеального решения [10, 14], обеспечивающего действительно эффективное решение этой задачи. Вместе с тем поиск на 100 % эффективного подхода продолжается [9]. Это связано, в том числе с проблемой отсутствия возможности формализации универсального описания спамового письма в связи с вариативностью информационных интересов конкретного пользователя, что может привести к неправильной классификации писем [12].

Наиболее распространенные способы, используемые в мировой практике для выявления спама, заключаются [7] в анализе заголовков и содержимого сообщений электронной почты. Существующие методы фильтрации [7, 8], используемые для анализа заголовков, достаточно легко обходятся отправителями спама. При этом наиболее эффективными для его выявления [7, 9] считаются методы машинного

обучения, эффективность которых в решении задач классификации текстов обратила внимание исследователей в области обнаружения спама на обучаемые модели [8], которые и положены в основу второго подхода.

Вопросам анализа и составления моделей текстов, их классификации, а также методам машинного обучения посвящены работы российских и зарубежных ученых М. С. Агеева, В. Н. Вапника, К. В. Воронцова, Б. В. Доброва, Н. Г. Загоруйко, К. Г. Кирьянова, Н. Н. Леонтьевой, Н.В. Лукашевич, Л. Н. Федотовой, В. И. Шалака, Т. Э. Шульги, K. Aas, A. Dasgupta, H. Drucker, C. Manning, F. Sebastiani, A. Uysal и многих других. Ими проведены исследования и предложены теоретические и прикладные подходы к анализу текстов и составлению их моделей, а также классификации текстов с применением различных методов машинного обучения.

Исследованиям последних примерно 30 лет в области обнаружения спама посвящены работы многих российских и зарубежных исследователей, например, С. Ю. Блинова, Б. В. Доброва, А. С. Катасёва, М. П. Малыхиной, И. В. Машечкина, Е. М. Мезенцевой, А. Н. Мироненко, А. П. Никитина, А. С. Павлова, А. Н. Розинкина, М. А. Семеновой, П. Б. Хорева, В. А. Частиковой, Е. Н. Чернопрудовой, I. Androutsopoulos, W. Cohen, S. Delany, H. Drucker, K. Junejo, K. Gee, P. Graham, V. Metsis, G. Robinson, M. Sahami, G. Sakkis, H. Shen и многих других. Ими проведены исследования и предложены теоретические и прикладные подходы к решению вопросов:

- обнаружения спама на основе анализа содержимого электронных писем с составлением их моделей и классификации текстовой информации, содержащейся в электронных письмах, с применением различных методов машинного обучения;

- оценки в различных условиях эффективности применения методов машинного обучения в задаче обнаружения спама;

- отбора признаков, необходимых для классификации электронных писем.

Наиболее распространенными методами машинного обучения, используемыми для решения задачи выявления спама, являются: Байесовский классификатор [например, 11, 19, 37-39, 40-43], деревья решений [например, 44, 45,

46], метод опорных векторов [например, 47-50], k-ближайших соседей [например, 51, 52], искусственные иммунные системы [например, 14, 53, 54], нейросетевые классификаторы [например, 13, 48, 55-61].

На рисунке 1.1 в обобщенном виде представлены основные технологии обнаружения спама.

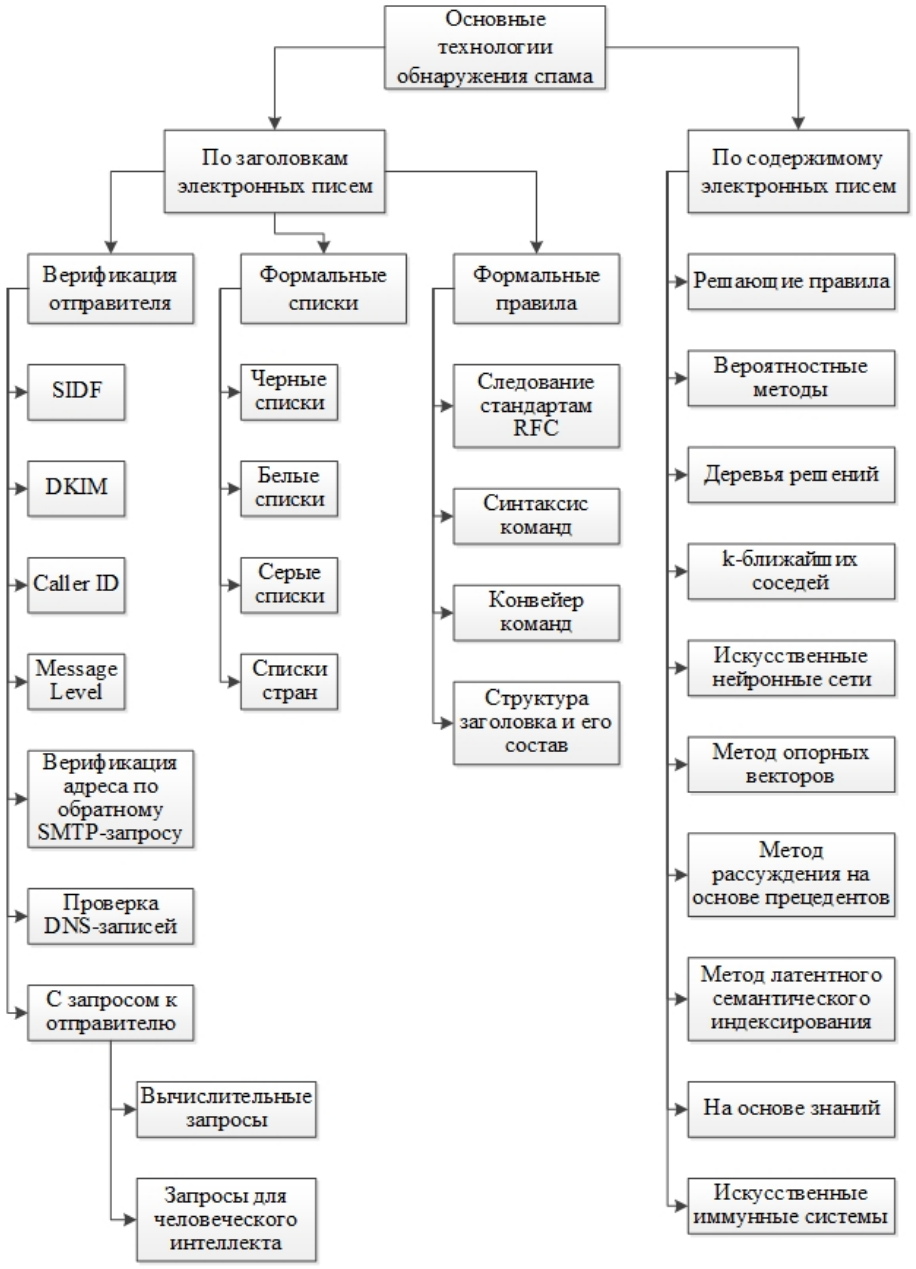

Рисунок 1.1 – Основные технологии обнаружения спама

Основными базовыми признаками электронных писем на основе их содержания являются:

- слова и их словосочетания;
- лексемы;
- термы как последовательности символов и их устойчивые словосочетания;
- метрики читаемости;
- характеристики жанра и стиля;
- глобальные статистические закономерности;
- различные лексические особенности электронных писем;
- предложения как минимальные семантические единицы.

А, например, в [24, 62-64] предложены подходы к выявлению спамовых сообщений посредством представления текстов сообщений электронной почты как последовательности символов. Для выделения значимых признаков электронных писем применяются различные веса.

Кратко остановимся на отдельных из вышеупомянутых исследований и их ключевых позициях, обзор которых в обобщенном виде представлен в таблице 1.1.

P. Graham в 2002-2003 годах опубликовал результаты своих исследований [41, 42], которые послужили определенным толчком в развитии научных исследований в области обнаружения спама на основе анализа содержимого электронных писем. В основе предложенного им Байесовского классификатора лежит статистический подход с использованием значений вероятностей спамности отдельно взятых слов или токенов. С целью выявления спама тексты электронных писем представлялись в виде наборов токенов (последовательности буквенно-цифровых символов, с включением в их состав точек, апострофов и символа «$»). Остальные символы принимались в качестве разделителей между токенами. Токены спама и легальных писем объединялись в соответствующие словари. Для каждого токена осуществлялся подсчет количества его встречаемости в наборах спама и легальных писем. После этого всем токенам сопоставлялась рассчитываемая вероятностью того, что электронное письмо, содержащее его, является спамом. В [42] исследователем предложено усовершенствованное

правило разбиения текста на токены, что, с одной стороны, привело к увеличению словарей спама и легальных писем, а с другой – к большей различимости этих словарей.

Предложенные P. Graham подходы к обнаружению спама и формированию словарей признаков спама и легальных писем на основе специальным образом сформированных токенов позволили на тот момент получить результат в 99,75 % обнаруженного спама для индивидуально полученных писем. При этом значащим явился вывод о том, что по своей сути фильтрация спама является оптимизационной задачей, требующей постоянного анализа пропущенного спама и выяснения, что необходимо предпринять, чтобы их обнаруживать.

G. Robinson в [43] предложил статистический подход, модернизирующий предложенный P. Graham метод Байесовской фильтрации для работы с редко встречающимися словами, что позволило решить проблему недостаточности исторических данных. В дополнение предложен подход, заключающийся в объединении вероятностей слов с применением метода R. Fisher в объединенную вероятность, характеризующую электронное письмо в целом. Кроме этого, предложен способ расчета значения показателя спама на основе комбинирования вероятностей спамности электронного письма и его неспамности. Он рассматривается как обобщенный показатель, значение которого близко к 1 в случае принадлежности электронного письма к спаму, и к 0 – к легальным.

Практическая реализация предложенного G. Robinson подхода нашла свое отражение в исследовании Е. М. Мезенцевой [39], помимо чего автором для оценки качества данного фильтра предложен и обоснован подход на основе анализа подмножества пересечения множеств, распознанных методами Байеса и Фишера по категориям. При этом в качестве признаков, необходимых для классификации электронных писем, Е. М. Мезенцевой предложены отдельные слова с учетом морфологии, а также словосочетания, составленные по разработанному ей алгоритму разбиения, который выделяет редко встречающиеся комбинации словосочетаний в тексте (каждое слово группируется с последним словом текста).

Результаты проведенного автором эксперимента продемонстрировали около 98 % в обнаружении легальных писем и около 87 % – спама.

Этот же подход использован в исследовании М. А. Семеновой [65], предложившей на его основе модель и метод фильтрации спама, позволяющие улучшить качество фильтрации. В них для расчета общей оценки применены разработанные автором способы нахождения количества слов для оценки письма и вычисления коэффициентов спамности на основе частотных словарей слов. Результаты проведенных автором опытно-экспериментальных исследований продемонстрировали возможность обнаружения около 93 % спама и до 92 % – легальных.

Дерево решений – один из методов машинного обучения на основе построения древовидной структуры данных, позволяющих находить потенциальные правила ассоциации между важными атрибутами из существующих данных. В [46] А. С. Павловым для выявления текстового спама разработан и обоснован подход на основе статистической оценки тематического разнообразия текстов и разработанной модели массово порожденных неестественных текстов. Гипотеза о невозможности достоверного воспроизведения всех свойств текстов, написанных человеком, положена им в основу разработанного подхода. К трудно контролируемым авторами текстов и предположительно плохо воспроизводимые искусственными генераторами А. С. Павловым отнесены следующие признаки текстов: глобальные статистические закономерности, характеристики жанра и стиля, метрики читаемости текстов, а также характеристики тематического разнообразия.

В результате исследования [46] его автором:

- разработан алгоритм обнаружения текстовых спамовых сообщений, основанный на анализе большого числа признаков текстов, представляющих их связность, стиль, читаемость, с учетом оценки разнообразия тематик;

- предложен алгоритм машинного обучения, являющийся модификацией одного из алгоритмов построения деревьев решений, заключающейся в построении

ансамбля деревьев решений, каждое их которых построено с помощью базового алгоритма.

В [47] H. Drucker и соавторами предложены подходы классификации на основе метода опорных векторов, где в качестве признаков используются слова и текстовые лексемы совместно с некоторыми нетекстовыми признаками.

А. Н. Розинкиным в [50] разработано решение для почтового сервера на основе метода опорных векторов с обоснованием алгоритмов и методов выбора модели писем. Для повышения устойчивости метода к ошибкам в наборе для обучения им был предложен дополнительный признак в тренировочный набор. Результаты экспериментальной апробации представленного А. Н. Розинкиным комплекса алгоритмов показали лучшие результаты по сравнению с Байесовским классификатором.

Предложенный С. Ю. Блиновым в [49] классификатор также основан на методе опорных векторов. На обучающем множестве документов, размеченных пользователем на классы, им строятся системы линейных неравенств, задающие множества спамовых и легальных точек-документов. Автоматическая классификация документов на спам и легальные в дальнейшем осуществляется с использованием построенного слоя, разделяющего указанные множества. С. Ю. Блиновым выявлено ограничение такого метода, связанное с получением конструктивного оператора проектирования, в результате чего им произведена его замена последовательностью фейеровских отображений. Таким образом, С. Ю. Блиновым предложена [49] модификация построения разделяющей гиперплоскости с использованием фейеровских отображений.

По результатам работы фильтра на основе разработанного С. Ю. Блиновым метода достигнуто около 95 % полноты обнаружения спама.

В [52] G. Sakkis и соавторами предложен подход к фильтрации спама с использованием алгоритма k-ближайших соседей. Его особенностью является использование для классификации ранее полученных экземпляров писем в качестве обучающих примеров без построения уникальных моделей для каждой из категории писем. Анализируемые письма рассматриваются как точки в

многомерном пространстве их признаков и классифицируются путем оценки их сходства с ранее сохраненными обучающими экземплярами. В качестве базовых признаков анализируемых текстов использованы определенные отдельные слова. Результаты эксперимента позволили достичь полноту обнаружения спама более 88 %.

В [51] S. Jiang и соавторами предложен улучшенный алгоритм k-ближайших соседей для категоризации текстов, который строит модель классификации, комбинируя алгоритм кластеризации с ограниченным одним проходом и классификацию текста методом k-ближайших соседей.

Методы на основе искусственных иммунных систем, в основу которых положены свойства и механизмы живых организмов, обеспечивающих их выживание, являются одними из перспективных для решения задачи выявления спама [54]. В рамках создания методики, интегрирующей в себе функции искусственной иммунной системы, М. П. Малыхиной и В. А. Частиковой в [54] предложен подход к решению задачи обнаружения спама.

В [13] Е. Н. Чернопрудовой представлена методика контентной фильтрации на основе нейросетевого классификатора. В основе разработанной ей методики лежит предложенная автором модель электронного почтового сообщения, учитывающая семантику контента электронной почтовой корреспонденции, в векторном пространстве признаков, которое отражает содержание электронного письма с помощью термов и устойчивых словосочетаний из них. Основу модели электронного сообщения составляют термы и их объединения в устойчивые словосочетания. Особенность предложенной модели заключается в применении дополнительных мер, позволивших при сокращении признакового пространства повысить семантическую нагрузку термов при классификации легальных сообщений электронной почты.

Модель электронного почтового сообщения, предложенная Е. Н. Чернопрудовой, основана на матричном представлении содержания писем. Ее отличие от известных векторных моделей текста заключается в сокращенном признаковом пространстве, обеспечивающем семантическую классификацию

сообщений электронной почты в реальном масштабе времени. Оно представлено термами с сопоставленными им весовыми коэффициентами, равными частоте термов в сообщении. При этом Е. Н. Чернопрудовой в качестве меры взвешивания термов предложено использовать величины этих частот в логарифмическом масштабе, что позволило устранить в них эффект больших различий (высокочастотные, среднечастотные и низкочастотные). В связи с большой размерностью получаемой матрицы признаков Е. Н. Чернопрудовой реализован подход по ее сокращению с использованием закона Ципфа.

Эксперименты с применением разработанных алгоритмов, проведенные Е. Н. Чернопрудовой в ходе исследования [13], продемонстрировали следующие лучшие результаты: точность около 96 % при полноте обнаружения около 90 %.

В [56-58] А. С. Катасёвым и соавторами предложены нейросетевая и нейронечеткая модели для решение задачи разработки технологии для обнаружения спама. При этом из множества признаков электронных писем выделены следующие наиболее информативные, влияющие на результат классификации писем на категории спам и легальные:

- частота встречаемости слов верхнего регистра;
- частота встречаемости цифр в письме;
- количество разных цветов в тексте письма;
- размер текста письма;
- количество пустых строк в тексте письма.

Совмещение нейросетевого подхода и метода опорных векторов явилось основой смешанного алгоритма фильтрации, предложенного в [48] А. Н. Мироненко. Целью его разработки явилось уменьшение времени работы фильтра электронных почтовых сообщений при помощи уменьшения объема обрабатываемых данных.

В исследовании А. Н. Мироненко в качестве анализируемых признаков электронных писем использованы термы, за которые приняты последовательности символов, разделенные точками, пробелами и т. п., а в качестве их весов – частоты появления в обоих классах электронных почтовых сообщений. Так как за слово

была принята произвольная последовательность символов, это позволило сделать процесс классификации независимым от какого-либо конкретного языка.

В результате проведенных экспериментов полнота обнаружения спама на специально созданной коллекции составила около 81 %, а на реальном почтовом ящике – около 92 %.

В [66, 67] K. Junejo и соавторами предложен персонализированный двухэтапный алгоритм обнаружения спама на основе разработанных авторами настраиваемой статистической модели электронных писем и дискриминантного классификатора. На первом этапе (обучение) происходит обучение статистической модели на основе слов спама и легальных слов из соответствующих классов писем из обучающего набора. Второй этап (классификация) выполняется в два прохода:

- первоначальная классификация электронных писем по обученной статистической модели и ее обновление с учетом проведенной классификации;
- окончательная классификация электронных писем по обновленной статистической модели.

В основу статистической модели сообщений электронной почты [66, 67] авторами положен расчет обобщенного показателя принадлежности слов, содержащихся в электронном письме, к спамовым или легальным сообщениям, вычисляемый как разница между количеством вхождений слова в словари спама и легальных сообщений. Дополнительно проводится сокращение размерности полученного признакового пространства путем отбора значимых слов, для которых абсолютное значение обобщенного показателя превышает некоторый целочисленный порог. Одновременно каждому слову в модели присваивается вес, основанный на соотношении его встречаемости в спаме и легальных электронных письмах.

В результате проведенного авторами [66, 67] эксперимента с применением статистической модели электронных писем и дискриминантного классификатора среднее значение обнаружения спама составило около 96 %.

Также для обнаружения спама исследована применимость классификации на основе решающих правил [например, 68, 69], метода рассуждения на основе

прецедентов [например, 70], использующих в качестве признаков различные лексические особенности электронных писем. А в [71] описано применение метода латентного семантического индексирования.

Таблица 1.1 – Обзор отдельных исследований в области обнаружения спама (обобщение)

Исследование	Используемый классификатор	Признаки писем	Обнаружение спама
P. Graham [41, 42]	Байесовский	Слова, т. н. токены	99,75 %
G. Robinson [43]	Байесовский	Слова	
Е. М. Мезенцева [39]	Байесовский, Фишера	Слова, словосочетания	87 %
М. А. Семенова [65]	Байесовский, Фишера	Слова	93 %
А. С. Павлов [46]	Модифицированный алгоритм построения деревьев решений	Признаки, представляющие связность, стиль, читаемость текстов, с учетом оценки разнообразия тематик документа	98,5 %
H. Drucker и др. [47]	Опорных векторов	Слова, текстовые лексемы	
А. Н. Розинкин [50]	Опорных векторов	Текстовые лексемы, нетекстовые признаки	78,7 %-100 %
С. Ю. Блинов [49]	Опорных векторов	Слова	95,17 %
G. Sakkis и др. [52]	k-ближайших соседей	Слова	88 %
М. П. Малыхина, В. А. Частикова [54]	Искусственные иммунные системы		
Е. Н. Чернопрудова [13]	Нейросетевой	Термы, словосочетания термов	90 %
А. С. Катасёв [56-58]	Нейросетевой	Частота встречаемости слов верхнего регистра, частота встречаемости цифр в письме, количество разных цветов в тексте письма, размер текста письма, количество пустых строк в тексте письма	97,5 %
А. Н. Мироненко [48]	Совмещение опорных векторов и нейросетевого	Слова	91,79 %
K. Junejo и др. [66, 67]	Статистическая модель	Слова	96,66 %

Обзорам, а также сравнительному анализу отдельных методов машинного обучения и алгоритмов классификации применительно к задаче обнаружения спама посвящены, например, исследования [7-9, 18, 20, 22, 72, 73].

Таким образом, результаты анализа проведенных исследований в области обнаружения спама обосновывают наличие устойчивой системы технологий, методов, средств и моделей, предназначенных для выявления спамовых сообщений [13]. В целом можно выделить следующие основные группы методов обнаружения спама:

1. Основанные на черных/белых/серых списках.

2. Основанные на анализе контента (содержания) и его классификации с применением методов машинного обучения.

3. Основанные на контроле массовости рассылок.

4. Основанные на контроле различного рода вложений.

При этом последние исследования, представленные отечественной и зарубежной практикой в области обнаружения спама, показывают, что данная задача решается в основном схожими известными методами классификации или их различными модификациями [73]. Вместе с тем значительное внимание исследователями в работах уделяется отбору признаков сообщений электронной почты, позволяющих повысить эффективность выявления спамовых сообщений с использованием различных выбранных методов.

1.3 Особенности существующих систем обнаружения спама

В настоящее время известно большое количество продуктовых пакетов, предназначенных для обнаружения спама и реализующих вышеописанные подходы и технологии. Возможности большинства из них во многом пересекаются. В данных программных продуктах реализованы развитые пользовательские интерфейсы и запрограммированы богатые функциональные возможности по обработке и анализу данных. Их функционирование обеспечено в клиент-серверной архитектуре с предоставлением доступа к разнообразным источникам данных. Как правило, в этих продуктовых пакетах в полной мере реализовано

обнаружение спама с использованием одновременно нескольких технологий, наиболее распространенными из которых являются верификация отправителя и анализа содержимого электронных писем без учета их смысловой составляющей. Вместе с тем анализ состояния «отрасли» средств борьбы со спамом показывает, что ее уровень недостаточен для современных потребностей общества и государства.

Большинство известных средств выявления спама в сообщениях электронной почты построены на базе списков адресов, ключевых слов и правил, составляемых экспертами и требующих постоянного обновления ввиду очень быстрой потери актуальности. Для таких средств полнота обнаружения составляет порядка 50-70 % [74]. Кроме того, такие средства имеют жесткие связи с конкретной организацией (сервисом) и сильно зависимы от оперативности их обновления. В промежутках времени, в которых спамовые электронные письма уже претерпели изменения, а база знаний еще не обновлена – организация остается незащищенной. Также необходимо отметить, что средства такого рода не имеют персонификации применительно к конкретной организации и ее работникам, т. е. не учитывают сферу деятельности организации и присущей им специфики общения (взаимодействия) с использованием электронной почты. Отсутствие такой специфики не позволяет добиться приемлемой точности выявления.

Вместе с тем существует другой тип средств выявления спамовых писем, построенных на методах машинного обучения. Указанные методы используются для автоматической классификации, и для их применения подготавливаются априорно известные данные о сообщениях электронной почты. Применение методов машинного обучения для обнаружения текстового спама требует представления содержания писем в виде числовых векторов [75].

Вместе с тем методы машинного обучения предполагают наличие подготовительного процесса, заключающегося в их обучении на множестве сообщений электронной почты, которые предварительно были классифицированы экспертом как спамовые или как легальные. Для этого требуется заблаговременная подготовка обучающих наборов и проведение вручную присвоения классов

письмам, включенным в их состав. Классифицированные экспертом электронные почтовые сообщения подлежат математическому преобразованию для построения их модели, которая используется для дальнейшей автоматической классификации новых сообщений электронной почты.

В связи с тем, что большой объем работы по подготовке к использованию таких средств и сопровождению их эксплуатации осуществляет непосредственно сам пользователь (в том числе в части модели электронных писем), решается проблема актуализации базы знаний. Средство обнаружения спама становится независимым от поставщика услуги (сервиса, продукта) и более персонализированным применительно к конкретной организации и ее пользователям. При этом у пользователей появляется возможность самостоятельного принятия решения о принадлежности поступающей к ним электронной почтовой корреспонденции к спаму или легальной рассылке.

В то же время системы обнаружения спама, использующие методы машинного обучения, не лишены недостатков, основными из которых, по мнению автора настоящего исследования, являются следующие:

- проблема обучения с учетом содержания электронных писем;

- чувствительность к качеству и составу обучающих наборов электронных писем;

- ресурсоемкость.

Проведенный автором анализ основных вышеописанных технологий обнаружения спама и признаков сообщений электронной почты, отнесенных к категории спама, показывает несовершенство существующих средств обнаружения спама и невозможность с их помощью адекватно реагировать на постоянно изменяющиеся способы составления спамовых писем [76]. Вместе с тем человеческий интеллект легко распознает спам, используя лишь сведения о собственных интересах пользователя в сопоставлении с содержанием электронных писем. Исходя из этого, видится целесообразным выявлять спамовые письма по аналогии с человеческим интеллектом посредством «распознавания» содержимого сообщений электронной почты и дальнейшей их автоматической классификации.

Приведенные в [77] доводы с их проецированием на исследуемую предметную область, свидетельствуют о том, что автоматическая классификация сообщений электронной почты может осуществляться посредством сопоставления выделенной из них значимой информации с определенным «понятием» какого-либо класса. Следовательно, задача выявления спама может быть сведена к задаче автоматической классификации сообщений электронной почты с использованием значимых признаков на спам и легальные сообщения.

Формально задачу классификации электронных писем можно описать [75] как задачу присвоения булевого значения $\{True, False\}$ каждой паре $\langle el_j,\ c_i\rangle \in EL \times C$,

где EL – множество классифицируемых электронных писем;

$C = \{Spam, Legal\}$ – два класса электронных писем, между которыми их необходимо распределить. $Spam$ – класс спама, $Legal$ – класс легальных электронных писем;

$\{True, False\}$ – булево значение принадлежности электронного письма el_j классу c_i $(True)$ или нет – $False$.

Более формально задачу классификации электронных писем на спам и легальные можно представить как построение функции:

$$\tilde{\phi}_{el}: EL \times C \ \rightarrow \ \{True, False\}, \qquad (1.1)$$

которая описывает процедуру классификации писем.

Эффективное решение задачи выявления спама возможно с применением технологий искусственного интеллекта [54] с применением методов машинного обучения. Человеческий интеллект позволяет легко распознать спам, используя лишь опыт пользователя, сведения о собственных интересах, предпочтениях и знаниях о подписках на рассылки в соответствии с ними. В связи с этим наделение средств борьбы со спамом навыками и качествами, присущими человеку, с одновременным использованием различных оптимизационных механизмов является актуальной задачей при разработке и модернизации таких систем [54].

Для автоматического построения списка слов с их весами используются методы машинного обучения, входными наборами данных для которых являются

спам и легальные письма пользователей. Наибольшие издержки классификации [10] формируются при неправильном отнесении легальных писем к классу спама и наоборот – спама к классу легальных писем. При этом целью злоумышленников (отправителей спама) является снижение вероятности выявления спама, для чего содержание спамовых писем наполняется наиболее употребимыми словами легальных сообщений электронной почты [11].

Необходимо отметить, что в основу экспериментов многих из проведенных исследований по оценке разработанных различными авторами подходов к обнаружению спама положены различные и недоступные в открытом доступе наборы электронных писем. Это не позволяет осуществить прямое сравнение эффективности предлагаемых авторами подходов и выбор абсолютно лучшего решения [8].

Также результаты проведенного анализа показывают, что большинство исследований, направленных на решение задачи выявления спама, построены на идентичных методах классификации с внесением в них незначительных изменений [12, 73]. По-прежнему, внимание исследователей во многом остается прикованным к процессам формирования признакового пространства с задачей выбора качественных признаков, повышающих эффективность применяемых методов классификации сообщений электронной почты.

Автором настоящего диссертационного исследования показано отсутствие универсального описания спама в связи с наличием вариативного изменения информационных потребностей конкретного пользователя, влияющего на классификацию спамовых электронных писем [12]. В связи с этим имеются основания утверждать об отсутствии единственно верного и универсального решения задачи обнаружения спама [10, 14]. Большинство известных методов исключают возможность вариативного поведения пользователя и не имеют функционала идентификации легальных сообщений электронной почты, что обуславливает их классификацию как спамовых.

Вместе с тем известные алгоритмы выявления спамовых сообщений электронной почты в основном используют статистическую обработку данных и не

в полной мере учитывают содержание сообщений электронной почты, информационная направленность которого для конкретного пользователя может изменяться со временем или в соответствии с решаемой задачей. Также необходимо отметить, что существующие модели сообщений электронной почты, как правило, не содержат формализованного описания легальных сообщений электронной почты, что приводит к увеличению ошибок первого и (или) второго рода.

1.4 Постановка цели и задач диссертационного исследования

Для современных средств обнаружения спама, использующих разнообразные технологии и методы обработки информации и ее классификации, а также широкий набор признаков спама, центральным является вопрос создания эффективных механизмов, способных работать с актуальным персонализированным потоком электронных почтовых сообщений и противостоять вновь появляющимся спам-технологиям [78-80].

Для эффективного обнаружения спама при сохранении надежности доставки персональных легальных электронных писем необходимо четко понимать его особенности.

1.4.1 Особенности спама

Необходимо отметить, что не существует единого четкого общепринятого определения термина *спам* и самого понятия *спама*. Проблема же такого определения является важной. От понятной и четкой формулировки данного понятия зависят не только перспективы научного изучения спама, но и развитие юридической базы, ставящей целью законодательный запрет спама.

В [81] выделены следующие особенности спама:

- отправляемые абоненту, который не выразил (в явном или неявном виде) желания их получать;

- содержащие рекламу (в том смысле, как она определена в Федеральном законе от 13 марта 2006 года № 38-ФЗ «О рекламе» [82], то есть, в широком смысле);

- рассылаемые в массовом порядке.

Согласно [83] к особенностям спама отнесены:

- рассылка по воле отправителя предварительно незапрашиваемых получателями электронных сообщений;

- массовая рассылка одного сообщения с использованием значительного количества адресов электронной почты либо многократная одному получателю;

- специфическое содержание сообщения;

- фальсификация адреса отправителя;

- фальсификации заголовков сообщения.

В соответствии с «Нормами пользования сетью» (OFISP-008) [84] рассылка электронных почтовых сообщений (кроме тех, на получение которых дано прямое согласие получателя) и иных, в том числе единичных сообщений, отнесена к спаму в случае, если они:

- являются рекламой, несут коммерческую направленность или являются агитацией к чему-либо;

- содержат выражения и предложения, носящие грубый и оскорбительный характер;

- направлены получателям, которые ранее явно выразили свое нежелание получать такого рода письма.

Согласно «Правилам оказания телематических услуг связи», утвержденных постановлением Правительства Российской Федерации от 10 сентября 2007 года № 575 [85], спам – это телематическое электронное сообщение, предназначенное неопределенному кругу лиц, доставленное абоненту и (или) пользователю без их предварительного согласия и не позволяющее определить отправителя этого сообщения, в том числе ввиду указания в нем несуществующего или фальсифицированного адреса отправителя.

Основываясь на результате проведенного анализа и обобщении приведенных выше определений и признаков, по мнению автора настоящего диссертационного исследования, под спамом можно понимать электронное сообщение, обладающее

наиболее значимыми отличительными особенностями, представленными на рисунке 1.2.

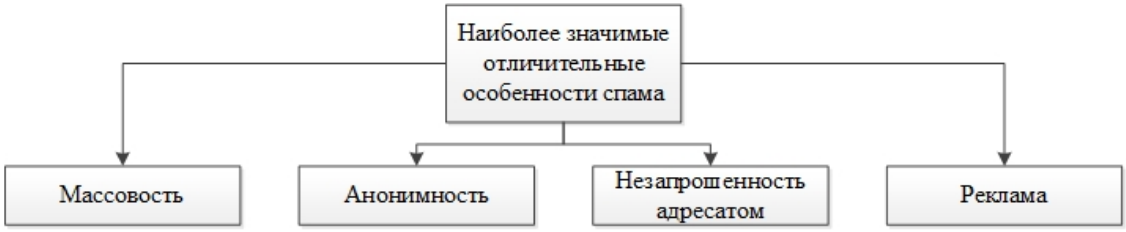

Рисунок 1.2 – Наиболее значимые отличительные особенности спама

Обобщая изложенное при рассмотрении особенностей спама, можно сделать вывод, что главной целью рассылки спама является анонимное доведение незапрошенной информации до неопределенного круга адресатов с целью продажи описываемого продукта или услуги либо с целью открытия получателем вредоносного вложения. При этом текст электронных писем из одной рассылки в независимости от их содержания должны иметь определенный заложенный их автором смысл.

Необходимо дифференцировать понятия «содержание» и «смысл», являющиеся нетождественными, но воспринимаемыми в единстве аспектами для одного и того же текста. Содержание представляет собой мысль о чем-либо, выраженную словарными значениями текста. Содержание, следовательно, явно. Смысл же текста является результатом понимания читателем его содержания.

Следовательно, один конкретный текст может порождать у различных читателей разный смысл в зависимости от ментальной реакции, вызываемой у них содержанием. Для достижения же единой цели при массовости рассылки спамовых писем их отправители должны при формировании текста исходить из необходимости порождения у получателей спама заданного единообразного для всех смысла. Иначе говоря, спамеры могут идти на любые уловки с адресами, включая их фальсификацию, и подгонкой текста электронных писем, но они должны продать продукт, услугу или побудить открыть вредоносное вложение.

Автор диссертации пришел к аналогичному выводу, сделанному P. Graham в его статье «A Plan for Spam» [41]: если содержание электронного письма, полученное в результате вынужденного применения различных уловок отправителями спама, будет непонято получателями или ими неправильно будет распознан его смысл, то и не будет достигнута необходимая цель такой рассылки. В текущей ситуации большого информационного потока вчитываться и искать нужный отправителям смысл никто не будет. Следовательно, содержание спамовых писем должно быть понятно и однозначно, порождая требуемый смысл, призывающий нас к какому-то единственному предопределенному действию.

На основании вышеизложенного обоснован вывод об относительной схожести содержания спамовых сообщений в пределах одной смысловой массовой рассылки.

1.4.2 Основные признаки электронных писем

С целью определения наиболее важных информативных признаков электронных писем автором настоящего диссертационного исследования был проведен анализ имевшихся в распоряжении англоязычного [86] и русскоязычного спама, трех информационных рассылок: xakep.ru, security.nnov.ru и securitylab.ru за период с 28.04.2009 г. по 04.03.2011 г., – а также материалов АО «Лаборатория Касперского» [87].

Полученные результаты подтверждают изложенное в [88-93] в том, что наиболее важные информативные признаки, которые позволяют отнести то или иное электронное сообщение к категории спама, условно можно разделить на формальные и лингвистические. На рисунке 1.3 представлены основные информативные признаки сообщений электронной почты, позволяющие относить их к спамовым или легальным.

Рисунок 1.3 – Основные информативные признаки электронных писем, позволяющие классифицировать их на спам и легальные

Под формальными следует понимать признаки, соотносящиеся с типичными, предопределенными признаками электронных писем.

Например, в соответствии с предназначенным для передачи электронной почты в сетях TCP/IP протоколом SMTP (*аббр. от англ.* Simple Mail Transfer Protocol) в состав электронного письма должны входить его заголовок и тело.

Заголовок содержит в своем составе служебные данные для идентификации электронного письма. В таблице 1.2 представлены названия основных полей заголовка и их назначения. Непосредственно информация, направляемая адресату, содержится в теле сообщений электронной почты.

Таблица 1.2 – Основные поля заголовка электронных сообщений и их назначение

Название	Назначение
Received	Идентификационная информация почтовых серверов, участвовавших в доставке сообщения от отправителя получателю. Каждый сервер добавляет в заголовок электронного письма свое поле Received
Return-Path	Обратный адрес электронной почты для пересылки сообщения в случае его недоставки
Reply-To	Адрес электронной почты для направления ответа
From	Адрес электронной почты отправителя сообщения
Date	Дата и время отправления сообщения
To	Адрес электронной почты получателя сообщения

Любое электронное почтовое сообщение должно обязательно содержать в своем заголовке указанные в таблице 1.2 поля, помимо которых в нем могут присутствовать необязательные, раскрывающие дополнительную детальную информацию об электронном письме.

Таким образом, к основным формальным признакам электронных писем, по результатам анализа которых возможно выявление спама с учетом их особенностей и, как следствие, его обнаружение, могут быть отнесены представленные в таблице 1.3.

Таблица 1.3 – Примеры основных формальных признаков электронных писем и соответствующим им признаков спама

№ п/п	Формальные признаки спама	Формальные признаки электронных писем
1.	Неправильный формат заголовка электронного письма или «непонятный» путь его доставки	Заголовок электронного письма
2.	Определенное значение почтового адреса отправителя или его отсутствие	Электронный почтовый адрес отправителя
3.	Определенное значение IP-адреса или отсутствие в системе интернет-адресов DNS	IP-адрес отправителя
4.	Отсутствие или наоборот наличие большого количества электронных почтовых адресов получателей	Электронный почтовый адрес получателя
5.	Нетипичный для получателя размер и/или формат входящего электронного письма	Размер и/или формат (кодировка) электронного письма

№ п/п	Формальные признаки спама	Формальные признаки электронных писем
6.	Наличие, тип, формат или размер прикрепленных к электронному письму файлов	Прикрепленные к электронному письму файлы

Под лингвистическими следует понимать признаки, отражающие особенности содержания электронного письма. К лингвистическим признакам можно отнести малохарактерные для получателя тематики текстов, наличие в тексте типичных для спама последовательностей символов, а также слов, фраз и предложений, включая их статистические показатели.

Лингвистические признаки электронных сообщений выделяются на основе содержания электронных писем. К ним могут быть отнесены представленные на рисунке 1.3, а также некоторые другие.

Под эвристиками понимают [91] наборы ключевых специфических слов или словосочетаний, которые позволяют отнести данное электронное сообщение к спаму, вместе с их вероятностными и весовыми показателями. Преимущество эвристик заключается в способности обнаруживать спам за счет сравнительно небольшого объема словаря, используемого отправителем для составления текста электронного письма. Их же недостаток заключается в необходимости ручной обработки электронных сообщений, характеризующейся высокой трудоемкостью.

Например, эвристиками могут выступать термины из различных сфер деятельности, предложения о покупках, реклама различных услуг и продуктов и прочее. Сообщение будет относено к категории спамовых в случае превышения заданного порогового значения количества определенных лексических единиц в тексте письма. Применение регулярных выражений расширяет возможности применения эвристик и позволяет создавать для поиска заданных эвристик по большим текстам разнообразные эффективные образцы и правила.

Сигнатуры представляют собой некий образ или признак электронного письма. Они более короткие, чем само письмо, и предназначены для идентификации оригинала. К ним относятся:

- текстовые фрагменты и свертки;

- ключевые и служебные слова, наиболее часто встречающиеся в тексте;

- байтовые контрольные суммы каждых n слов и пр.

Для обнаружения спама требуется вычисление сигнатуры для классифицируемого письма и сравнение ее с сигнатурами обучающей выборки спамовых сообщений. В случае ее совпадения с одной из словаря сигнатур электронное сообщение относится к категории спамовых. Преимущество сигнатур заключается практически в отсутствии ложных срабатываний. Вместе с тем обнаружение спамовых писем становится невозможным в случае отсутствия их сигнатур в базе данных. В то же время слегка модифицированный спам может быть обнаружен с использованием сигнатур с применением методов нечеткого сравнения.

В качестве статистических показателей могут выступать вероятность наличия в спаме определенных слов (словосочетаний) и вероятность присутствия этих же слов в легальных электронных письмах. Баланс этих двух значений и определяет вероятность того, что электронное письмо, в котором встречаются данные слова, является спамом.

Выделение семантических единиц текстов позволяет решать задачу их моделирования и автоматизацию процесса их корректного машинного понимания (результат обработки). Среди семантических единиц текста выделяют [94], например, фреймы и опорные слова.

Указанный перечень признаков, позволяющих отнести электронное письмо к спаму или легальным, не является исчерпывающим. В дополнение к ним можно выделить признаки, при наличии которых невозможно однозначно определить, является ли данное сообщение спамом. Однако в совокупности с иными вышеназванными формальными и лингвистическими признаками они помогают убедиться в том, что письмо действительно относится к категории спама.

1.4.3 Постановка цели и задач диссертационного исследования

Как было показано, разработано значительное количество разнообразных по методам и технологиям подходов к обнаружению спама. Однако несмотря на это с

полной уверенностью можно сказать, что единого универсального и стопроцентного решения этой задачи не существует. Спам крайне разнообразен и постоянно меняется, поэтому в борьбе с ним необходимо использование новых способов и технологий его обнаружения, учитывающих содержание электронных писем.

Также результаты анализа современного состояния исследований в области обнаружения спама показывают, что они не прекращаются [73] в попытке улучшить точность и/или полноту обнаружения спама буквально на каждые 0,01 %. При этом актуальной остается проблема выбора признаков сообщений электронной почты, обеспечивающих высокое качество выявления спама и идентификации легальных сообщений, с учетом персональных информационных потребностей пользователя при классификации сообщений электронной почты [11]. Ее решение требует разработки новых или совершенствования существующих технологий, методов и средств, учитывающих содержание сообщений электронной почты конкретного пользователя и оценку эффективности применяемых методов [12].

Поэтому создание новой модели электронного почтового сообщения, обеспечивающей выделение признаков сообщений электронной почт на основе их содержания с учетом вариативности информационных интересов конкретного пользователя, для обнаружения спама является актуальной задачей и представляет научный и практический интерес, что подтверждается сделанным в [73] выводом. Это позволит придать качество персонализации процессу выделения спамовых сообщений, как одного из ключевых свойств, предъявляемым к системам обнаружения спама [11], а, следовательно, и повышению его эффективности.

Изложенное выше обуславливает актуальность темы и научной задачи диссертационного исследования, заключающейся на содержательном уровне в повышении эффективности обнаружения спама и достоверности идентификации легальных электронных почтовых сообщений на основе классификации их содержания за счет создания модели электронного письма, обеспечивающей выделение признаков электронных писем на основе их содержания с учетом

меняющихся информационных потребностей конкретного пользователя (персонализации). В качестве исходных данных выступают непустые наборы сообщений электронной почты, подлежащих классификации. При этом их класс (спам или легальные) заранее известен.

Учитывая изложенное, построение модели электронного почтового сообщения осуществлялось на базе математических моделей текстов [15, 16] и их последующего анализа с использованием «генетических карт». В основе предлагаемого метода лежит теория структурной идентификации и анализа текстовой информации с помощью базовых параметров, разработчиком теоретических основ которой выступил доктор технических наук, профессор К. Г. Кирьянов. Указанная теория применялась в различных научных областях для решения задач идентификации и анализа текстов, вместе с тем не применялась ранее в области обнаружения спама. Применение данного подхода позволяет выделять участки последовательностей символов в текстах путем нахождения их границ и подразумевает необходимость преобразования исходных текстов в числовую последовательность, что является ключевой особенностью данного подхода.

Применяя для постановки задачи теоретико-множественный подход [95, 96] к описанию, положения системного анализа [96], теории моделировании [96, 97], классификации [97] и принятия решений [97, 98] постановка задачи диссертационного исследования в формальном виде будет выглядеть следующим образом.

Дано:

- задана предметная область;

- задано множество классифицируемых электронных писем EL, представленных в виде их текстовых содержаний на естественном языке;

- заданы два класса $C = \{Spam, Legal\}$, между которыми необходимо распределить электронные письма. $Spam$ – класс спама, $Legal$ – класс легальных электронных писем;

- задана обучающая выборка электронных писем $EL^t \subseteq EL$, для каждого электронного письма $el_j^t \in EL^t$ которой известен его класс $c_i^t \subseteq C$.

Требуется разработать:

1. Модель электронного почтового сообщения Ψ_{el} для классификации электронных писем, отличающуюся от известных моделей методом выделения значимых последовательностей символов текста (признаков электронных писем на основе их содержания, термов), позволяющим усилить смысловое содержание термов за счет применения метода «генетических карт».

$$\Psi_{el} = \langle EL, EL\_PreProc, T\_Proc, T \rangle, \quad (1.2)$$

где $EL\_PreProc$ – процедура предварительной обработки электронного письма (в случае необходимости);

T_Proc – процедура выделения термов – значимых последовательностей исходных символов текста электронного письма;

T – множество термов электронного письма.

2. Метод классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, отличающийся использованием разработанной модели электронных писем.

3. Алгоритм классификации электронных писем, отличающийся наличием дополнительной процедуры определения «схожести» термов на основе расстояния Левенштейна, обеспечивающей вычисление мер принадлежности классифицируемого электронного письма к классам спама и легальных для повышения достоверности идентификации электронных писем.

4. Архитектуру подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, реализующую предложенные в работе метод и алгоритм, применение которых позволяет повысить достоверность идентификации легальных электронных писем с учетом меняющихся информационных потребностей конкретного пользователя (персонализации).

Областью определения модели является текстовое содержание (текст) электронного письма на естественном языке.

Допущения и ограничения:

1. Электронные письма, составляющие множество, должны быть из реальных переписок (отправлений).

2. Исходное множество писем разбивается на обучающие и тестовые выборки, принадлежность писем к спаму и легальным в которых заранее известна.

3. Автор исследования обладает знаниями, обеспечивающими получение максимальных значений полноты и точности неавтоматизированной классификации электронных писем на спам и легальные.

4. Типовой размер одной информационной единицы (документов в массиве, слов в документе) ограничен объемом памяти ЭВМ.

Выводы по 1 главе

Основными результатами рассуждений, представленных в данной главе, являются:

1. Проведен анализ современного состояния исследований в области обнаружения спама.

2. Выявлено, что отдельные средства обнаружения спама не имеют персонификации применительно к конкретной организации и ее работникам, т. е. не учитывают сферу деятельности организации и присущей им специфики общения (взаимодействия) с использованием электронной почты, что отрицательно влияет на их точность.

3. При создании модели электронных писем обоснована целесообразность применения метода выделения термов, позволяющего усилить смысловое содержание термов за счет применения метода «генетических карт». Это позволит учитывать меняющиеся информационные потребности конкретного пользователя и достичь персонализации в обнаружении спама, как одного из ключевых свойств, предъявляемым к системам обнаружения спама, а также повышению его эффективности.

4. Задача обнаружения спама обоснованно сведена к задаче автоматической классификации сообщений электронной почты с использованием их значимых признаков.

5. Постановка задачи диссертационного исследования осуществлена на основе теоретико-множественного подхода. Для ее решения предлагается использовать положения системного анализа, теории моделирования, классификации и принятия решений.

Предложенная последовательность анализируемых научных позиций стала обоснованием перехода к исследовательским материалам второй главы, которая называется «Разработка модели электронного почтового сообщения для классификации электронных писем» и посвящается разработке модели электронного почтового сообщения для классификации электронных писем, отличающегося от известных моделей методом выделения значимых последовательностей символов текста (признаков электронных писем на основе их содержания, термов), позволяющим усилить смысловое содержание термов за счет применения метода «генетических карт».

Глава 2 Разработка модели электронного почтового сообщения для классификации электронных писем

Представляется очевидным, что основной целью научного исследования явления «спам» является создание эффективных методов борьбы с ним, позволяющих с высокой долей точности обнаруживать спам в потоке разнородных электронных писем и каким-либо образом осуществлять реагирование на них.

Как было обосновано в разделе 1.3 главы 1 настоящего диссертационного исследования, обнаружение спама целесообразно осуществлять по аналогии с человеческим интеллектом посредством «распознавания» содержимого сообщений электронной почты и дальнейшей их автоматической классификации. Для этого необходимо разработать модель электронных писем, которая позволит обнаруживать спам на основе контентной фильтрации.

Основными задачами при разработке такой модели при этом становятся выделение признаков, соотносящихся с признаками спама, которые и должны использоваться как ключевые для автоматического обнаружения спама, и их единиц счета.

Однако поскольку речь идет об электронных письмах, создаваемых различными отправителями, которые к тому же используют различные приемы обхода фильтров спама, выделение таких признаков становится нетривиальной задачей. Тем не менее, как показано в первой главе настоящего диссертационного исследования, спамовые сообщения обладают относительной схожестью содержания в пределах одной смысловой массовой рассылки. Этот вывод с учетом меняющихся информационных потребностей конкретного пользователя, несомненно, должен быть использован при разработке модели электронных почтовых сообщений.

Под моделью электронного почтового сообщения применительно к настоящему диссертационному исследованию будем понимать некоторый объект, схожий посредством некоторых отношений с исходными электронными письмами (прототипами) и предназначен для их объяснения и/или описания [96].

Предоставление моделью упрощенного образа, отображающего только существенные для настоящего исследования свойства прототипов, является важнейшим качеством разрабатываемой модели электронных писем.

2.1 Определение базового подхода для разработки модели электронного почтового сообщения для обнаружения спама

Как справедливо отмечено в главе 1, для достижения единой цели при массовости рассылки спамовых писем их отправители должны при формировании текста исходить из необходимости порождения у получателей спама заданного единообразного для всех смысла. Иначе говоря, для достижения своих целей спамеры должны побудить пользователей к определенным действиям. Любые уловки с адресами, включая их фальсификацию, и подгонкой текста электронных писем для противодействия средствам обнаружения спама не должны приводить к искажению заложенного в содержание письма смысла. Это вынуждает отправителей спама учитывать данный фактор при составлении содержания своих писем.

Похожим свойством обладают биологические особи. Так живые организмы способны получать, сохранять и передавать информацию, необходимую для их существования, развития и размножения, из поколения в поколение. Данная информация является основой направления развития биологических особей и на протяжении жизненного цикла она может изменяться, создавая условия для эволюции и развития потомков. При этом формируется и сохраняется новая информация, а ценность самой информации увеличивается, тем самым обеспечивается эволюция организмов с одновременным сохранением у них одинаковых признаков. Передача такой информации, являющейся генетической, осуществляется в закодированном с помощью генетического кода виде, сохраненной в особых функциональных участках молекул ДНК или РНК [99].

Морфологическое строение, рост, развитие особей и др. признаки определяются [99] генетической информацией, обладающей [100] рядом важных свойств:

- дискретность (существование элементарных единиц информации – генов);
- устойчивость (сохранение);
- самовоспроизведение (репликация, копирование);
- реализация (выполнение программы с получением некоторого результата);
- передача из поколения в поколение;
- комбинирование дискретных единиц информации (генов);
- изменение (мутирование) – появление новых дискретных единиц информации (генов).

На основании вышеизложенного и анализа, представленного в главе 1 настоящей работы, можно предположить, что спамовые сообщения также могут содержать в себе специфическую «генетическую» информацию, определяющую смысл сообщения [101].

Таблица 2.1 – Аналогия между биологическими организмами и письмами

Биологические организмы	Электронные почтовые сообщения
Особь	Электронное почтовое сообщение
Генетическая информация	Специфическая «генетическая» информация (смысл)
Молекула	Содержание письма
Ген	Значимая последовательность символов текста, соотносящаяся со спамом
Генетический код	Правила выделения значимых последовательностей символов текста

В соответствии с приведенной аналогией разработка модели сообщения электронной почты осуществлялась автором в парадигме следующих допущений:

1. Спамовые сообщения содержат отличительные признаки в виде отдельных последовательностей символов, отличающие спам от легальных писем.

2. Правила извлечения признаков, соотносящихся со спамом, должны быть универсальными.

3. Основу правил извлечения признаков, соотносящихся со спамом, должны составлять символы, составляющие содержание сообщения электронной почты.

4. Требуется преобразование содержания сообщения электронной почты в последовательность цифровых символов.

Далее в настоящей работе определим, что отдельная значимая последовательность символов содержания сообщения электронной почты, характеризующая класс спама или легальных писем, будет называться «термом». Тогда термы сообщений электронной почты являются элементарными единицами, определяющими отличительные признаки сообщений электронной почты.

Под правилом выделения термов будем понимать совокупность действий с исходной последовательностью символов, составляющих содержание электронных почтовых сообщений и упорядоченных от начала к концу, по выделению конечного числа соответствующих ей последовательностей символов, сформированных определенным образом [15, 16].

Необходимо отметить, что генетический подход нашел свое применение для анализа и диагностирования последовательностей данных [102, 103], исследования, идентификации и измерения параметров потоков в сетях связи и линиях коллективного пользования [104], анализа спектров [105], исследования процессов и объектов различной природы [106-108], решения задач многокритериальной и дискретной оптимизации [109-111], структурной идентификации математических моделей криптосистем [112] и для др. [15].

В [113-115] авторами предложена архитектура системы выявления спамовых сообщений, включающая блоки анализа адреса отправителя, анализа содержимого, сигнатурного анализа и контент-анализа. В основу последнего блока положен метод «генетических карт» [15, 16] применительно к текстовым спамовым сообщениям. Для применения указанного метода на группе писем (рассылках) в [116] предложена модифицированная версия блока контент-анализа.

Результаты применения блока контент-анализа показали [113-116] до 98 % обнаружения спамовых писем, поступивших на адреса электронной почты, объединенных одним доменным именем. В качестве преимущественных сторон применения метода «генетических карт» подчеркиваются отсутствие ограничений:

- по созданию новых и изменению существующих объектов;

- по всевозможным модификациям «генетических карт» спамовых сообщений.

Таким образом, учитывая при принятии решения вышеизложенное, автором настоящего диссертационного исследования для математического моделирования электронных писем в качестве базового выбран метод «генетических карт» [15, 16].

2.2 Разработка базовой модели электронного почтового сообщения

Приведем основные положения описанного в [15, 16] подхода, спроецировав их в модель электронных писем и адаптировав к понятиям исследуемой предметной области.

Для формализации разрабатываемой модели электронного почтового сообщения, учитывающей содержание электронных писем конкретного пользователя (персонализацию), введем следующие условные обозначения:

Ψ_{el} – модель электронного почтового сообщения;

$EL = \{el_i\}$ – множество электронных писем, представленных в виде текстов на естественном языке;

$el = (sym_0, sym_1, \dots, sym_{M-2}, sym_{M-1})$ – электронное письмо, представленное в виде конечной последовательности символов;

$M = |el|$ – длина электронного письма (количество символов в электронном письме);

B – множество десятичных значений байт, соответствующих символам в кодовой таблице, используемой для представления текста в виде числовой последовательности;

b – десятичное значение байта, соответствующего конкретному символу электронного письма в кодовой таблице, используемой для представления текста в виде числовой последовательности;

t_j – терм электронного письма – значимая последовательность исходных символов текста электронного письма;

$T = \{t_j\}$ – множество термов электронного письма;

$l_j = |t_j|$ – длина терма электронного письма (количество символов в терме);

$T^C = \{t_j^C\}$ – множество термов заданного класса электронных писем.

В соответствии с [15, 16, 113, 115, 116] на первом этапе требуется преобразовать исходный текст сообщения электронной почты в числовой вектор с использованием числовых значений кодов символов в соответствии с кодовой таблицей, примененной при составлении текста электронного письма. При этом будет сохранена исходная смысловая информация сообщения за счет возможности обратного преобразования.

Исходя из этого, процедура преобразования содержания сообщения электронной почты в числовую последовательность, принимает следующий вид [15, 16, 113, 115, 116]:

$Conv\_to\_Dig(el)$:

$$el = (sym_0, sym_1, \dots, sym_{M-2}, sym_{M-1}) \xrightarrow{q,\Delta t} el' = (b_0, b_1, \dots, b_{M-2}, b_{M-1}) \quad (2.1)$$

где q – размер кодовой таблицы, используемой для представления текста в виде числовой последовательности;

Δt – шаг выборки символов текста в функции преобразования писем в числовую последовательность.

В результате процедуры (2.1) каждый символ исходного содержания электронного почтового сообщения будет последовательно заменен на соответствующий ему десятичный код.

На следующем этапе из полученной упорядоченной последовательности чисел происходит последовательное выделение ее участков разной длины (термов). Процедура генерации термов в таком виде и алгоритм определения их правых границ являются основой данного шага, ключевым элементом для которых является так называемый «генератор» терма – выборка (последовательность символов заданной длины), на основе которой происходит его выделение [15, 16].

Под выборкой терма s будем понимать такую числовую последовательность

$$s = (b_k, b_{k+1}, \dots, b_{k+n-2}, b_{k+n-1}), \quad (2.2)$$

для которой справедливы условия:

$$\begin{cases} k \in Z, \\ 1 \le k \le (M-n) \end{cases}. \quad (2.3)$$

Выборка является, по сути, частью терма, обладающей следующими ключевыми параметрами:

n – длина выборки (N-граммы – последовательности, порождающей терм);

k – начальная позиция выборки в пределах числовой последовательности, представляющей текст электронного письма.

Исходя из введенного определения, любая числовая последовательность (2.2) длины n, для которой выполняются условия (2.3), может являться кандидатом в выборку.

Следовательно, сообщение электронной почты включает в себя $(M - n)$ последовательностей, потенциально являющихся кандидатами в выборку. В свою очередь, числовая последовательность el' может быть представлена в виде поочередного, упорядоченного (от первого к последнему) наложения выборок. Выборка является параметром алгоритма определения правой границы термов.

Определение правых границ осуществляется с использованием одного из двух алгоритмов разбиения: H_0 и H_1 [15, 16]. Для решения поставленной задачи правой границей для алгоритма H_0 будет являться число, предшествующее последовательности чисел, соответствующей первому повтору выборки терма. Для алгоритма H_1 правой границей будет являться последовательность чисел, соответствующая первому повтору выборки терма. Т. е. для алгоритма H_0 правая граница терма заканчивается перед первым найденным повтором выборки, а для алгоритма H_1 – включает первый найденный повтор «генератора» терма.

Исходя из вышеизложенного, можно сделать вывод об отсутствии цикличности (замыкания на выборку) в последнем терме. Одновременно, окончанием термов в случае применения алгоритма разбиения H_1 всегда (за исключением последнего терма) будет их выборка.

Продемонстрируем на примере работу алгоритма H_1 при $n = 1$ по выделению термов. На первом шаге начало терма начинается с b_0. Для определения правой границы терма в цикле, начиная с b_1, осуществляется сравнение в числовой последовательности el' текущего кандидата в выборку с предшествующими ему символами. Решение о нахождении границы терма принимается при равенстве

текущего кандидата в выборку с одним из предыдущих символов, тем самым происходит выделение терма t_1 электронного почтового сообщения длиной l_1, порожденному расположенным в начале или внутри терма «генератором». Выделение следующего терма t_2 длиной l_2 осуществляется по аналогичной процедуре, начиная с $(l_1 + 1)$-й позиции числовой последовательности el'. Таким образом, осуществляется выделение термов по всей длине последовательности el'.

На основе описанных положений метода «генетических карт» и с учетом выбранного подхода к обнаружению спама путем понимания содержания электронных писем в целом модель электронного почтового сообщения может быть представлена в виде кортежа [101]:

$$\Psi_{el} = \langle EL, T\_Proc, T \rangle, \tag{2.4}$$

где T_Proc – процедура выделения термов – значимых последовательностей исходных символов текста электронного письма:

$$T\_Proc = \langle el, Conv\_to\_Dig(el), \mathrm{s}, \{H_0, H_1\} \rangle, \tag{2.5}$$

Поскольку выборка является основополагающей и характерной единицей терма, наиболее рационально использовать алгоритм H_1 для вычисления правых границ термов. Тогда процедура выделения термов сообщений электронной почты (2.5) принимает вид [101]:

$$T\_Proc = \langle el, Conv\_to\_Dig(el), \mathrm{s},\ H_1 \rangle. \tag{2.6}$$

Необходимо отметить, что представленная модель сообщений электронной почты (2.4) позволяет программно реализовать процедуру выделения термов [101, 113-116].

Пример работы процедуры выделения термов на сообщении электронной почты представлен на рисунке 2.1.

$el =$
hi ,
i sent you an email a few days ago , because you now qualify for a new mortgage .
you could get $ 300 , 000 for as little as $ 700 a month !
bad credit is no problem , you can pull cash out or refinance .
please click on this link :
best regards ,
serena
- - - - system information - - - -
particular various publish revision important numbers textual individual
[definition : localized resources be tag believes locale) united
cultures specifying time current pattern goal designers information)

$$\left| \begin{array}{l} Conv\_to\_Dig \\ q = 256, \Delta t = 1 \end{array} \right.$$

$el =$
2 73 32 44 13 10 73 32 83 69 78 84 32 89 79 85 32 65 78 32 69 77 65 73 76 32 65 32 70 69 87 32 68 65 89 83 32 65 71 79 32 44 32 66 69 67 65 85 83 69 32 89 79 85 32 78 79 87 32 81 85 65 76 73 70 89 32 70 79 82 32 65 32 78 69 87 32 77 79 82 84 71 65 71 69 32 46 13 10 89 79 85 32 67 79 85 76 68 32 71 69 84 32 36 32 51 48 48 32 44 32 48 48 48 32 70 79 82 32 65 83 32 76 73 84 84 76 69 32 65 83 32 36 32 55 48 48 32 65 32 77 79 78 84 72 32 33 13 10 66 65 68 32 67 82 69 68 73 84 32 73 83 32 78 79 32 80 82 79 66 76 69 77 32 44 32 89 79 85 32 67 65 78 32 80 85 76 76 32 67 65 83 72 32 79 85 84 32 79 82 32 82 69 70 73 78 65 78 67 69 32 46 13 10 80 76 69 65 83 69 32 67 76 73 67 75 32 79 78 32 84 72 73 83 32 76 73 78 75 32 58 13 10 66 69 83 84 32 82 69 71 65 82 68 83 32 44 13 10 83 69 82 69 78 65 13 10 45 32 45 32 45 32 45 32 83 89 83 84 69 77 32 73 78 70 79 82 77 65 84 73 79 78 32 45 32 45 32 45 32 45 13 10 80 65 82 84 73 67 85 76 65 82 32 86 65 82 73 79 85 83 32 80 85 66 76 73 83 72 32 82 69 86 73 83 73 79 78 32 73 77 80 79 82 84 65 78 84 32 78 85 77 66 69 82 83 32 84 69 88 84 85 65 76 32 73 78 68 73 86 73 68 85 65 76 13 10 91 32 68 69 70 73 78 73 84 73 79 78 32 58 32 76 79 67 65 76 73 90 69 68 32 82 69 83 79 85 82 67 69 83 32 66 69 32 84 65 71 32 66 69 76 73 69 86 69 83 32 76 79 67 65 76 69 32 41 32 85 78 73 84 69 68 13 10 67 85 76 84 85 82 69 83 32 83 80 69 67 73 70 89 73 78 71 32 84 73 77 69 32 67 85 82 82 69 78 84 32 80 65 84 84 69 82 78 32 71 79 65 76 32 68 69 83 73 71 78 69 82 83 32 73 78 70 79 82 77 65 84 73 79 78 32 41 13 10

$$\left| \begin{array}{l} H_1 \\ n = 2 \end{array} \right.$$

$t_1 =$72 **73 32** 44 13 10 **73 32** $s = (73\ 32)$
$t_2 =$83 69 78 84 32 89 79 85 **32 65** 78 32 69 77 65 73 76 **32 65** $s = (32\ 65)$
$t_3 =$32 70 69 **87 32** 68 65 89 83 32 65 71 79 32 44 32 66 69 67 65 85 83 69 32 89 79 85 32 78 79 **87 32** $s = (87\ 32)$
$t_4 =$81 85 65 76 73 70 89 32 70 **79 82** 32 65 32 78 69 87 32 77 **79 82** $s = (79\ 82)$
$t_5 =$84 71 65 71 69 32 46 13 10 89 **79 85** 32 67 **79 85** $s = (79\ 85)$
$t_6 =$76 68 32 71 69 84 32 36 32 51 **48 48** 32 44 32 **48 48** $s = (48\ 48)$
$t_7 =$48 32 70 79 82 **32 65** 83 32 76 73 84 84 76 69 **32 65** $s = (32\ 65)$
$t_8 =$**83 32** 36 32 55 48 48 32 65 32 77 79 78 84 72 32 33 13 10 66 65 68 32 67 82 69 68 73 84 32 73 **83 32** $s = (83\ 32)$
$t_9 =$78 79 **32 80** 82 79 66 76 69 77 32 44 32 89 79 85 32 67 65 78 **32 80** $s = (32\ 80)$
$t_{10} =$85 76 76 32 67 65 83 72 **32 79** 85 84 **32 79** $s = (32\ 79)$
$t_{11} =$82 32 82 69 70 73 78 65 78 67 **69 32** 46 13 10 80 76 69 65 83 **69 32** $s = (69\ 32)$
$t_{12} =$67 **76 73** 67 75 32 79 78 32 84 72 73 83 32 **76 73** $s = (76\ 73)$
$t_{13} =$78 75 32 58 **13 10** 66 69 83 84 32 82 69 71 65 82 68 83 32 44 **13 10** $s = (13\ 10)$
$t_{14} =$83 69 82 69 78 65 13 10 **45 32 45 32** $s = (45\ 32)$
$t_{15} =$**45 32 45 32** $s = (45\ 32)$

t_{16} =83 89 83 84 69 77 32 73 78 70 79 82 77 65 84 73 79 78 32 45 **32 45** s = (32 45)
t_{17} =**32 45 32 45** s = (32 45)
t_{18} =13 10 80 **65 82** 84 73 67 85 76 **65 82** s = (65 82)
t_{19} =32 86 65 82 73 79 85 83 32 80 85 66 76 **73 83** 72 32 82 69 86 **73 83** s = (73 83)
t_{20} =73 79 78 **32 73** 77 80 79 82 84 65 78 84 32 78 85 77 66 69 82 83 32 84 69 88 84 85 65 76 **32 73** s = (32 73)
t_{21} =78 68 73 86 73 68 85 **65 76** 13 10 91 32 68 69 70 73 78 73 84 73 79 78 32 58 32 76 79 67 **65 76** s = (65 76)
t_{22} =73 90 69 68 32 82 **69 83** 79 85 82 67 **69 83** s = (69 83)
t_{23} =**32 66** 69 32 84 65 71 **32 66** s = (32 66)
t_{24} =69 76 73 69 86 **69 83** 32 76 79 67 65 76 69 32 41 32 85 78 73 84 69 68 13 10 67 85 76 84 85 82 **69 83** s = (69 83)
t_{25} =32 83 80 69 67 73 70 89 73 78 71 32 84 73 77 69 32 67 85 82 82 69 78 84 32 80 65 84 84 **69 82** 78 32 71 79 65 76 32 68 69 83 73 71 78 **69 82** s = (73 32)
t_{26} =83 32 73 78 70 79 82 77 65 84 73 79 78 32 41 13 10

Рисунок 2.1 – Пример работы процедуры выделения термов на сообщении электронной почты

Необходимо отметить, что все обучаемые модели предполагают наличие соответствующего модели представления сообщений электронной почты, что лежит в области решения задач обработки естественного языка (NLP, *аббр. от англ.* Natural Language Processing). Наиболее популярное решение по использованию в обработке естественных языков в целом и в решении задач классификации текстов в частности заключается в представлении текстов (содержаний сообщений электронной почты) в виде набора (мешка) отдельных слов (*от англ.* «bag of words») [например, 75], выделенных из спамовых сообщений или легальных электронных почтовых сообщений в совокупности с их количественными характеристиками. Такое решение достаточно просто в реализации, однако при классификации текстов производится без учета контекста появления слов. Исправить данный недостаток позволяют представления текстов в виде n-граммы слов или векторных вложений (*от англ.* «word embedding»), формирующие контекстные связи между словами отдельного предложения или текста в целом. А, например, применение n-грамм символов для представления текстов электронных сообщений положено в основу представленного в [117] подхода к обнаружению спама.

Представленные подходы позволяют провести некую аналогию между выборкой и n-граммой символов. Вместе с тем предложенная модель электронного почтового сообщения (2.4) позволяет определять идентичные среди нескольких писем участки символов с учетом их контекста и позволяет повысить качество обнаружения спама в условиях наличия различных начальных слов и/или символов в сообщении, включения «мусорных» символов и/или слов, наличия орфографических ошибок, и т. п. [118].

Таким образом, предложенная модель является развитием подхода представления «bag of words» и по своей сути представляет собой смешанное представление: частично в виде набора слов и частично в виде векторных вложений, – и может быть использована как соответствующий элемент любого из существующих классификационных пайплайнов в области обнаружения спама. Учет информации в модели осуществляется на различных уровнях: лексическом (слова и их части) и синтаксическом (словосочетания и их части и предложения и их части, в том числе в контексте частей других частей слов, слов, частей предложений и предложений). Такое представление позволяет избежать недостатков представления «bag of words» с одновременным использовании преимуществ «bag of words» и «word embedding». При этом становится практически невозможным осуществить подбор текста под модель, поскольку использование «легальных» слов и их сочетаний в спаме будет нивелировано способом выделения термов.

Основываясь на описанных в [15, 16, 101, 113-116] положениях, в качестве параметров модели (2.4), оказывающих влияние на выделение термов, в [12, 118-120] обоснованы:

q – размер кодовой таблицы, используемой для представления текста в виде числовой последовательности;

Δt – шаг выборки символов текста в функции преобразования писем в числовую последовательность;

n – длина выборки (N-граммы – последовательности, порождающей терм).

Представленные параметры в целом определяют качество выделения термов из текста электронного письма.

Корректность и практическая применимость разработанной базовой модели электронного почтового сообщения (2.1) продемонстрированы в [101, 118-120].

2.3 Уточнение базовой модели электронного почтового сообщения

Как было показано в разделе 2.2 настоящей главы, выделение термов зависит от значений ключевых параметров q, Δt и n предложенной модели электронного почтового сообщения. Предобработка данных также является важным этапом для методов машинного обучения [9, 10, 18, 19].

Указанное обстоятельство свидетельствует об актуальности исследования вопросов предобработки текстов сообщений электронной почты для применения предложенной модели электронного почтового сообщения при решении задачи выявления спама, а также вопросов уточнение модели в части выбора оптимальных числовых значений указанных ключевых параметров. Это целесообразно и возможно осуществить путем проведения экспериментальных исследований [118-123].

2.3.1 Обоснование выбора значений параметров модели, оказывающих влияние на выделение термов

Поскольку смысл текста проявляется в результате понимания читателем его содержания, можно предположить, что смысл определяется сочетаниями букв, цифр и знаков препинания. Следовательно, совместное расположение отдельных символов определенным образом придает тексту определенный смысл и позволяет извлекать соответствующие ему термы.

Тогда для учета всех символов содержания сообщения электронной почты наиболее рационально определить равным единице шаг выборки символов Δt, что приводит формулу (2.1) к следующему виду:

$Conv\_to\_Dig(el)$:

$$el = (sym_0, sym_1, \dots, sym_{M-2}, sym_{M-1}) \xrightarrow{q,\Delta t=1} el' = (b_0, b_1, \dots, b_{M-2}, b_{M-1}) \quad (2.7)$$

Таким образом, в ходе предложенного преобразования осуществляется последовательная замена каждого символа на десятичный код в соответствии с выбранной кодовой таблицей, позволяющая проводить дальнейший анализ сообщения без искажения исходного смысла.

Для обоснования выбора значений других параметров модели электронного почтового сообщения, оказывающих влияние на выделение термов, целесообразно поставить эксперимент. Для его проведения в качестве базового возможно использование уже подготовленного набора электронных писем [86], сформированного Metsis et al. [38] и состоящего из шести поднаборов. Они представляют собой упорядоченные по имени файла англоязычные легальные электронные письма и спам.

Эти сообщения электронной почты являются электронными письмами сотрудников компании Enron, ставшими доступными в открытом виде в начале 2000-х годов в результате правительственного расследования по факту банкротства данной компании [124]. Указанный массив писем включает в себя сообщения электронной почты, действительно сформированные человеком, по различным тематикам. Данный набор писем до сих пор является актуальным и востребованным и используется многими исследователями для валидации предлагаемых ими подходов к выявлению спамовых сообщений.

Экспериментальный набор Enron включает:

- сообщения электронной почты шести сотрудников компании, ведших наиболее активную переписку, представленные в виде упорядоченных по имени файлов;

- спамовые сообщения электронной почты, полученные в рамках проектов SpamAssassin и HoneyPot, из коллекции Bruce Guenter, а также были собраны Georgios Paliouras [38].

Дубликаты из набора не удалялись по причине того, что они составляют естественный поток всех писем (легальных и спама) на почтовый ящик конкретного пользователя. Также указанный набор прошел предварительную обработку со стороны авторов [38], в рамках которой было сделано следующее:

- удалены сообщения с совпадающими адресами отправителя и получателя (письма самому себе);

- из всех писем удалена html-разметка;

- удалены спамовые письма с символами нелатинского алфавита.

Автором настоящего диссертационного исследования также проведены следующие преобразования:

- удалены строки с темами электронных почтовых сообщений;

- удалены письма с пустым содержанием (включали только тему).

Последующий их анализ показал следующее:

- подавляющее большинство символов содержания электронных писем представлено в нижнем регистре (исключение составляют спамовые сообщения, содержащие единожды встречающуюся букву «B» в слове «Binary»);

- отсутствие знаков табуляции.

Таким образом, экспериментальный набор электронных писем представлен текстовыми сообщениями, содержащими строчные и прописные буквы латинского алфавита, цифры, знаки препинания и другие символы, в составе:

- 16 100 легальных сообщений электронной почты, разбитых на 6 групп;

- 16 420 спамовых электронных писем, разбитых на 6 групп.

Количественное разбиение указанных групп писем представлено в таблице 2.2.

Таблица 2.2 – Экспериментальный набор электронных почтовых сообщений[5]

	Поднабор 1		Поднабор 2		Поднабор 3	
	legal1	spam1	legal2	spam2	legal3	spam3
Количество	3 618	1 401	4 189	1 442	3 980	1 452

Таблица 2.2 (продолжение)

	Поднабор 4		Поднабор 5		Поднабор 6	
	legal4	spam4	legal5	spam5	legal6	spam6
Количество	1 500	4 237	1 408	3 551	1 405	4 337

[5] legal – легальные письма, spam – спам.

Вместе с тем дополнительно был сформирован русскоязычный набор сообщений электронной почты: информационные рассылки xakep.ru, security.nnov.ru и securitylab.ru за период с 28.04.2009 г. по 04.03.2011 г. (общее число писем – 1 242) и 2 группы спамовых сообщений, присланных на два электронных почтовых адреса, объединенных одним доменным именем в зоне .ru, (общее число писем – 3 215).

Из них удалена вся html-разметка и сохранено только содержание электронных почтовых сообщений. В тексте присутствуют строчные и прописные буквы, цифры, знаки препинания и иные символы. Последующий анализ подготовленных указанных электронных писем показал, что они содержат:

- повторы пробелов, в особенности, в большом количестве в письмах рассылки security.nnov.ru;

- знаки табуляции, в том числе повторяющиеся;

- кириллические буквы, а также латинские буквы в отдельных письмах.

Проведение экспериментов и оценка их результатов осуществлялись следующим образом.

Для спамовых и легальных сообщений электронной почты каждой группы в наборах осуществлено выделение термов с формированием из них словарей термов, соответствующих группам. Рассчитаны коэффициенты принадлежности каждого письма к легальным сообщениям электронной почты или спаму как сумма термов, содержащихся в письме и встретившихся во всех группах соответствующих классов:

N_T^S – число термов классифицируемого письма, встречающихся в спамовых сообщениях электронной почты;

N_T^L – число термов классифицируемого письма, встречающихся в легальных сообщениях электронной почты.

На основании простейшего решающего правила по принципу большей суммы термов соответствующего класса принималось решение о принадлежности электронного почтового сообщения к спамовым или легальным. Вместе с тем

письмо принималось как неклассифицированное при равном числе термов обоих классов.

Особенностью эксперимента явилась условная имитация очереди поступления писем, заключавшаяся в выделении термов группы, к которой принадлежало классифицируемое письмо, только для сообщений, стоящих перед ним в списке.

Мерами оценки результатов экспериментов приняты полнота, точность и F-мера обнаружения (классификации) [75, 125-128].

Под полнотой R будем понимать соотношение числа всех верно классифицированных сообщений электронной почты к числу писем, которые должны были быть отнесены к соответствующему классу:

$$R = \frac{N_{corr\_a}}{N_{corr\_a}+N_{incorr\_r}}, \quad (2.8)$$

где N_{corr_a} – число электронных почтовых сообщений, правильно отнесенных к заданному классу (истинно положительные результаты или TP, *аббр. от англ.* True Positive);

N_{incorr_r} – число электронных почтовых сообщений, неправильно признанных не относящихся к заданному классу (ложноотрицательные результаты или FN, *аббр. от англ.* False Negative).

Полнота показывает потери при классификации сообщений электронной почты, при этом их уменьшение ведет к увеличению значения полноты. Таким образом, полнота характеризует способность обнаруживать соответствующий класс вообще.

Под точностью P будем понимать соотношение числа всех верно классифицированных сообщений электронной почты к числу писем, которые были классифицированы как принадлежащие соответствующему классу:

$$P = \frac{N_{corr\_a}}{N_{corr\_a}+N_{incorr\_a}}, \quad (2.9)$$

где N_{incorr_a} – число электронных почтовых сообщений, неправильно признанных принадлежащими заданному классу (ложноположительные результаты или FP, *аббр. от англ.* False Positive).

Точность представляет собой часть объектов, признанных принадлежащих соответствующему классу и при этом действительно принадлежащие этому классу. Таким образом, точность характеризует способность правильно обнаруживать заданный класс.

Очевидно, что тем лучше результаты обнаружения спама, чем выше полнота и точность. Вместе с тем на практике, как правило, невозможно одновременное достижение максимальных значений полноты и точности. Поэтому для выбора наиболее эффективного варианта классификации используют сбалансированную F-меру обнаружения [125-128] спама и легальных писем, объединяющую для оценки в единую величину полноту и точность и являющуюся их средним гармоническим:

$$F = \frac{2 \cdot P \cdot R}{P+R}. \quad (2.10)$$

Из данной формулы (2.10) следует, что при полноте и точности, равными единице, F-мера максимальна. В то же время при значении любого из аргументов близкого к нулю, ее значение также приближается к нулю. Таким образом, с помощью F-меры, одновременно учитывающей полноту и точность, становится возможным выбрать наиболее эффективный вариант классификации, т. е. чем больше значение F-меры, тем эффективнее процесс классификации.

После расчетов указанных мер оценки результатов их значения округлялись до тысячных по простому математическому правилу.

2.3.1.1 Обоснование выбора значений длины выборки в модели электронного почтового сообщения

Среди статистических показателей, которыми обладают все тексты, встречаются минимальная, максимальная и средняя длина слов, рассчитываемая в символах. Например, средняя длина русскоязычного слова составляет чуть больше 5 символов [129, 130], а англоязычного – чуть больше 4 [130]. Следовательно, выбранное значение n оказывает влияние на то, что будет являться выборкой: от символа к словам, частям предложений и предложениям.

В качестве значений ключевых параметров разработанной модели электронного почтового сообщения при проведении эксперимента приняты:

$q = 256$ – равно числу символов кодовой таблицы Windows-1251;

$n = 1 \ldots 20$.

Результаты эксперимента, представленные в приложении А, показывают [118-120], что полнота и F-мера принимают значения не менее 0,85 при $n = 1$ и $n = 2$. При $n \geq 3$ значения полноты и F-меры уменьшаются более, чем на 0,05, а при $n \geq 4$ и $n \geq 5$ в англоязычном и русскоязычном наборах соответственно (превышение средней длины слова в соответствующих языках) – более, чем на 0,2. В то же время увеличение значения n приводит к увеличению значения точности обнаружения с одновременным уменьшением полноты [118-120].

Также результаты эксперимента демонстрируют более быстрое ухудшение идентификации легальных англоязычных писем по сравнению со спамом. Это подтверждает ранее сделанный вывод о том, что спамовые сообщения обладают относительной схожестью содержания в пределах одной смысловой массовой рассылки. Данный тезис также подтверждается и результатами обнаружения легальных русскоязычных писем, являющихся, по сути, легальными рассылками с отдельными статическими элементами в содержании [118-120].

Таким образом, результаты проведенного эксперимента демонстрируют наилучшие результаты обнаружения с применением разработанной модели электронных почтовых сообщений (2.4) при значениях ключевого параметра $n = 1$ и $n = 2$ [118-120].

2.3.1.2 Обоснование выбора размера кодовой таблицы в модели электронного почтового сообщения

Очевидно, что количество символов, фактически используемых в предложенной модели электронного почтового сообщения для одного и того же содержания электронных писем, будет напрямую зависеть от значения параметра q, используемого при перекодировании по формуле (2.7). Это, в свою очередь, влияет на символьное разнообразие текстов и их похожесть для писем разных

классов в целом по их содержанию [122], что, в конечном счете, будет сказываться на результатах обнаружения спама.

В качестве значений ключевых параметров предложенной модели электронного почтового сообщения при проведении эксперимента приняты:

$q = [256, 224, 192, 160, 128, 96, 64, 32]$;

$n = [1, 2]$ [115-117].

Учитывая нумерацию символов в кодовой таблице ASCII (от 0 до 255), формула перекодировки (расчета десятичного кода символа) при $q \neq 256$ имеет вид:

$$b^q = \left\lfloor \frac{b^{256} \cdot (q-1)}{255} \right\rfloor, \quad (2.11)$$

где b^{256} – номер заменяемого символа по таблице ASCII.

Дополнительно к основным в текущем эксперименте проведены следующие расчеты:

1. Для каждого значения q определено фактическое количество числовых кодов, сопоставляемых символам.

2. Для каждого значения n определено общее количество термов и доля дубликатов среди них.

3. Сформированы случайные последовательности символов с длинами, равными сумме длин в символах англоязычных и русскоязычных писем. Для них произведены расчеты по пп. 1 и 2.

Полученные результаты представлены в приложении Б. Их анализ с учетом содержания сообщений электронной почты показывает следующее [122].

Уменьшение значения q приводит к уменьшению количества реально используемых символов, фактически используемых в предложенной модели электронного почтового оообщения для одного и того же содержания электронных писем. Это снижает символьное разнообразие текстов и повышает их похожесть для сообщений электронной почты разных классов в целом по их содержанию.

Вместе с тем результаты дополнительных расчетов демонстрируют большее символьное разнообразие русскоязычных писем по сравнению с англоязычными

(134 различных символа против 74). Причиной этого является наличие в русскоязычных сообщениях электронной почты наряду с кириллическими буквами латинских. При этом различимость текстов русскоязычных писем (по факту отчасти смешанных) при уменьшении значения q сохраняется дольше по сравнению с англоязычными по причине наличия символов для первых – из обоих половин кодировочной таблицы ASCII, а для вторых – только в первой половине.

Результаты эксперимента демонстрируют при $n = 1$ лучшие результаты классификации при полном символьном разнообразии, т. е. при $q = 256$. Вследствие уменьшения значений q происходит повышение похожести содержания сообщений электронной почты разных классов в целом и снижается эффективность процесса классификации с применением разработанной модели электронного почтового сообщения (2.4). Лучшие результаты классификации при $n = 2$ демонстрируются при иных различных значениях q, отличных от 256, а иногда и превышают лучшие результаты, полученные при $n = 1$.

Это делает возможным проведение расчетов в совокупности для $n = 1 \div 2$ [122]. Вместе с тем данный факт свидетельствует о возможности адаптации (обучения и настройки) процесса обнаружения спама с применением предложенной модели электронного почтового сообщения (2.4) применительно к конкретным текстам сообщений электронной почты конкретных пользователей. Это позволяет учитывать персональные (пользовательские) особенности (с т. з. информационных потребностей) электронных писем и их содержание применительно к конкретным пользователям (группе пользователей), что обеспечивает персонализацию процесса обнаружения спама [122].

Отличие значений общего количества термов и доли дублирующих термов для реальных сообщений электронной почты и случайных последовательностей свидетельствует о неслучайности процесса выделения термов в соответствии с предложенной моделью электронного почтового сообщения (2.4) и результатов классификации электронных писем с ее применением [122].

Также полученные в ходе эксперимента результаты показывают, что разработанная модель электронного почтового сообщения (2.4) не зависит от

конкретного набора символов и их количества (не зависит от кодовой таблицы), т. е. является символонезависимой. Вместе с тем очевидно, что в пределах одной конкретной реализации системы обнаружения спама для обеспечения корректности применения предложенной модели требуются одинаковые исходные данные для всех текстов сообщений электронной почты, анализируемых с ее помощью. Другими словами, все анализируемые электронные сообщения для конкретной реализации системы обнаружения спама должны быть одинаковой кодировки [122].

2.3.1.3 Комбинирование значений параметра n модели электронного почтового сообщения

Результаты проведенного эксперимента, продемонстрированные в разделе 2.3.1.1 настоящей главы, показывают одновременное уменьшение значений полноты обнаружения при увеличении ее точности (число неверно классифицируемых сообщений электронной почты снижается одновременно с увеличением количества неклассифицированных писем). При этом значение точности практически не изменятся при $n \geq 5$.

Учитывая изложенное, автором настоящего диссертационного исследования выдвинута гипотеза о возможном улучшении результатов обнаружения с применением разработанной модели электронного почтового сообщения (2.4) при комбинации значений n (в совокупности по нескольким значениям n). Для ее подтверждения или опровержения поставлен эксперимент.

В качестве значений ключевых параметров предложенной модели при проведении эксперимента приняты:

$q = 256$ – равно числу символов кодовой таблицы Windows-1251;

$n = 1$ и комбинации значений $n = 1 \div 2$; $n = 1 \div 3$; $n = 1 \div 4$; $n = 1 \div 5$.

Эксперимент и оценка его результатов аналогичны описанным в разделе 2.3.1 настоящей главы с учетом следующих изменений [121]:

1. Для всех n для каждого класса спамовых и легальных сообщений электронной почты каждой группы писем рассчитано общее количество термов N_T в письме.

В качестве коэффициента принадлежности сообщения электронной почты к классу спама или легальных принято отношение количества термов соответствующего класса к общему количеству термов N_T в письме:

$$K^L = \frac{N_T^L}{N_T}, \tag{2.12}$$

$$K^S = \frac{N_T^S}{N_T}. \tag{2.13}$$

Значения N_T, N_T^S и N_T^L предварительно были увеличены на единицу с целью избежать возможных случаев равенства нулю соответствующего коэффициента, а также для обеспечения выполнимости расчетов следующего шага.

2. Рассчитано среднее геометрическое полученных значений коэффициентов принадлежности K^L и K^S для принятых для эксперимента комбинаций значений n.

Представленные в приложении В результаты проведенного эксперимента показывают [121] снижение в среднем на 0,015 значения полноты обнаружения спама с применением комбинированного подхода при использовании принятых для эксперимента комбинаций значений n по сравнению со значением $n = 1$.

Вместе с тем применение комбинированного подхода демонстрирует рост полноты идентификации легальных писем и точности обнаружения спама. Лучшая точность выявления спамовых писем достигается при использовании значений $n = 1 \div 2$, демонстрирующая уменьшение неверно классифицированных легальных электронных почтовых сообщений с 735 до 668 штук (более, чем на 9 %).

Таким образом, результаты эксперимента демонстрируют возможность использования комбинаций численных значений ключевого параметра n разработанной модели электронного почтового сообщения (2.4). Это позволяет повысить точность обнаружения спама при одновременном незначительном ухудшении полноты его обнаружения, чем, вероятно, возможно пренебречь в целях снижения количества неверно классифицируемых легальных сообщений электронной почты.

2.3.2 Предварительная обработка текстов электронных почтовых сообщений

Подготовка данных играет значимую роль в машинном обучении [9, 10, 18, 19, 131, 132]. В реальной жизни сообщения электронной почты поступают от разнообразных отправителей, составляющих письма в различных почтовых клиентах и в различных форматах и кодировках. При этом в электронных письмах могут присутствовать различного рода «шумы». К ним относятся [18, 19, 133-135]:

- различные ошибки, опечатки и искажения;
- малозначимые с точки зрения содержания слова;
- малоинформативные элементы, например, символы html-разметки;
- скрипты;
- рекламные вставки и т. п.

Эти «шумы» снижают качество непосредственно самих текстов для анализа и могут привести к снижению эффективности классификации сообщений электронной почты. Поэтому первым значимым этапом в задачах интеллектуального анализа текстов является их предварительная обработка [9, 18, 19, 56, 133-135], реализующая процедуры их очистки и подготовки к классификации [132].

Предварительная обработка электронного почтового сообщения может быть представлена в виде отдельных операций, выполняемые в ходе которых действия будут обрабатывать текст писем различными способами. Основными способами предварительной обработки являются [123, 131, 132, 134-136]:

1. Удаление неинформативных с точки зрения содержания элементов.
2. Удаление стоп-символов (например, знаков препинания).
3. Удаление стоп-слов.
4. Удаление повторяющихся символов пробелов, повторяющихся (всех) символов табуляции, повторяющихся (всех) символов переносов строк.
5. Перевод всех букв в верхний или нижний регистр.
6. Лемматизация.
7. Токенизация.

8. Стемминг (*от англ.* stemming – находить происхождение).

При этом нельзя заранее утверждать, какие из перечисленных способов или их комбинаций однозначно при любых условиях приводят к улучшению результатов классификации применительно к конкретной решаемой задаче [136]. В [123, 132, 136] экспериментально продемонстрировано, что подобранные применительно к конкретным текстам и решаемой задаче способы предварительной обработки могут улучшить качество классификации.

С учетом изложенного, автор настоящего диссертационного исследования приходит к выводу, что использование предварительной обработки сообщений электронной почты в предложенной модели электронного почтового сообщения (2.4) может повысить эффективность решения задачи обнаружения спама. С учетом применяемой в предложенной модели процедуры выделения термов среди вышеуказанных способов предварительной обработки в разработанной модели электронного почтового сообщения (2.14) целесообразно использовать предобработки под номерами 2-5, а также их сочетания [12].

При этом, по большому счету, ни одна из них не будет оказывать существенного влияния на содержание сообщений электронной почты (за исключением, вероятно, стоп-слов) [12]. Переносы строк используются для выделения мыслей и придания тексту логической структуры. Знаки препинания применяются для формирования логических связей. Перевод всех букв в один регистр исключают различия при написании стоящих в разных позициях предложения одних и тех же слов.

При этом целесообразно провести экспериментальные исследования для выбора конкретных способов предварительной обработки в задаче обнаружения спама с использованием разработанной модели электронного почтового сообщения (2.4) [123].

Для экспериментальных исследований приняты следующие способы предварительной обработки и их сочетания.

1. Без предобработки (далее по тексту при упоминании способов предобработки используется нумерация в соответствии с данным списком).

2. Удаление:

а) стоп-символов (под стоп-символами в настоящей диссертационном исследовании понимаются одиночные небуквенные символы, в качестве которых заданы следующие: «-», «—», «`», «^», «~», «<», «=», «>», «|», «_», «,», «;», «:», «!», «?», «/», «.», «'», «"», ««», «»», «(», «)», «[», «]», «{», «}», «@», «$», «*», «\», «&», «#», «%», «+», «№»; в общем случае стоп-символы являются подмножеством более общего понятия стоп-слов);

б) стоп-слов (в качестве англоязычных стоп-слов заданы слова из перечня http://www.antula.ru/noise-word_3.htm, в качестве русскоязычных – частицы, суффиксы, глаголы, причастия, предлоги, союзы, междометия, вводные слова, местоимения и некоторые сочетания букв из перечней http://www.antula.ru/noise-word_2.htm и https://russkiiyazyk.ru/chasti-rechi/spisok-mezhdometiy. html);

в) всех символов табуляции;

г) всех переносов строк с заменой на пробел;

д) всех пробелов.

3.

а) перевод всех буквенных символов в верхний регистр;

б) перевод всех буквенных символов в верхний регистр с удалением стоп-символов;

в) перевод всех буквенных символов в верхний регистр с удалением всех переносов строк с заменой на пробел;

г) перевод всех буквенных символов в верхний регистр с удалением всех пробелов;

д) перевод всех буквенных символов в верхний регистр с удалением стоп-символов и всех переносов строк с заменой на пробел;

е) перевод всех буквенных символов в верхний регистр с удалением стоп-символов и всех пробелов;

ж) перевод всех буквенных символов в верхний регистр с удалением всех переносов строк с заменой на пробел и все пробелы;

з) перевод всех буквенных символов в верхний регистр с удалением стоп-символов и всех переносов строк с заменой на пробел, а также всех пробелов.

В качестве обоснованных значений ключевых параметров предложенной модели при проведении эксперимента приняты:

$q = 256$;

$n = [1, 2]$ [118-120].

Представленные в приложении Г результаты эксперимента заключаются в следующем [12].

1. Предварительная обработка текстов англоязычных сообщений электронной почты в целом не приводит к значимому изменению результатов по сравнению с результатами без предобработок.

2. Предварительная обработка с применением «атомарных» способов 2а–3а также в целом не приводит к значимому изменению результатов по сравнению с результатами без предобработок. Вместе с тем наблюдается существенный рост точности обнаружения (более 0,99) при использовании способов предварительной обработки 3г и 3е. Также значение точности превышает 0,97 практически при применении любых предобработок при условии предварительного удаления повторений пробелов.

3. Результаты зависят от конкретных применяемых способов предварительной обработки, что соответствует выводам, представленным в [47, 123, 132].

Таким образом, учитывая результаты проведенного эксперимента, модель электронного почтового сообщения (2.4) целесообразно дополнить процедурой предварительной обработки [12]:

$$\Psi_{el} = \langle EL, EL\_PreProc, T\_Proc, T \rangle, \quad (2.14)$$

$$EL\_PreProc = \{ws\_reps\_del, tabs\_reps\_del, lb\_reps\_del, up\_case\}, \quad (2.15)$$

где ws_reps_del – процедура удаления повторов пробелов;

$tabs\_reps\_del$ – процедура удаления повторов символов табуляции;

lb_reps_del – процедура удаления повторов переносов строк.

up_case – процедура перевода всех букв в верхний регистр.

В качестве способов предварительной обработки целесообразно рассматривать: удаление повторов символов пробелов, табуляции и переносов строк, а также перевод всех букв в верхний (или нижний) регистр [12].

Иных способы предварительной обработки целесообразно вводить в состав модели электронного почтового сообщения (2.14) только после их экспериментальной оценки и при условии периодической корректировки с течением времени для адаптации процесса классификации применительно к индивидуальным особенностям написания сообщений электронной почты их автором [12], а также меняющегося спама.

Таким образом, обосновано [12], что результаты обнаружения напрямую зависят от выбора способов предварительной обработки, что подтверждается опубликованными в [47, 123, 132] выводами.

2.4 Обоснование неслучайности результатов обнаружения спама с применением разработанной модели

На основе вышеизложенного представляется очевидным, что выделение термов осуществляется всего лишь по одному или последовательности из двух символов, являющихся началом и окончанием терма. В связи с этим автором настоящего диссертационного исследования выдвинута гипотеза о возможном наличии случайности в получаемых результатах обнаружения спама с применением подхода к выделению термов в соответствии с моделью электронного почтового сообщения (2.14). Для ее проверки проведены дополнительные экспериментальные исследования, в качестве значений параметров модели электронного почтового сообщения, для которых приняты следующие значения:

$q = 256$;

$n = [1, 2, 3]$.

Эксперимент и оценка его результатов проведены для англоязычного набора электронных писем аналогично описанным в разделе 2.3.1 настоящей главы с учетом следующих дополнений [137]:

1. Для текстов писем из набора дополнительно было осуществлено выделение термов следующим «псевдослучайным» образом.

На основе результатов основных шагов для каждого письма осуществлялся подсчет длин термов (безотносительно к их символьному составу) с одновременным подсчетом количества термов одинаковой длины. Таким образом, письмо условно можно представить в виде набора групп термов (по длине), например, 5 термов по 10 символов и 3 терма по 5 символов.

Далее случайным образом выбиралась длина терма, присутствующая в наборе, и, начиная с первого символа текста письма, производилось выделение первого «псевдослучайного» терма с уменьшением на единицу значения количества термов с данной длиной (например, выбран терм длиной 5 символов, тогда набор примет вид: 5 термов по 10 символов, 2 терма по 5 символов). Данное действие повторялось до исчерпания всех длин всех термов исходного набора. При этом началом каждого последующего «псевдослучайного» терма являлся символ, следующий в письме за последним символом предыдущего «псевдослучайного» терма.

2. На основе полученных наборов термов (обычного и «псевдослучайного») определялась принадлежность каждого письма к классам спама и легальных писем.

Результаты эксперимента представлены в приложении Д.

В результате анализа полученных результатов эксперимента установлено значимое различие полученных значений при применении подхода к выделению термов в соответствии с моделью электронного почтового сообщения (2.14) с «псевдослучайным» выделением термов. Вместе с тем при $n = 1$ данное различие составляет около 10 %, а при $n = 2$ и $n = 3$ – более 20 %. Это обусловлено применением в экспериментальных исследованиях подхода, учитывающего при классификации все термы, в том числе с суммарным единичным весом.

Одновременно возможное максимальное (предельное) число уникальных термов при бесконечном увеличении числа писем в обучающих наборах при $n = 1$ достигается гораздо быстрее, чем при $n = 2$ и $n = 3$. Дополнительные эксперименты с «псевдослучайным» выделением термов показали, что указанная разница практически не изменяется.

Таким образом, результаты проведенного эксперимента приводят к выводу о неслучайности результатов классификации писем с применением подхода к выделению термов в соответствии с разработанной моделью электронного почтового сообщения и подтверждают его обоснованность [137]. Вместе с тем находит косвенное подтверждение сделанный в разделе 2.3.1.3 настоящей главы и в [121] вывод о возможности использования комбинированного подхода с использованием комбинаций численных значений ключевого параметра n модели электронного почтового сообщения (2.14).

При этом результаты эксперимента показывают, что с целью исключения возможного фактора случайности при классификации электронных писем, обусловленного достижением максимального (предельного) числа уникальных термов при бесконечном увеличении числа обучающих писем целесообразно применять весовые коэффициенты термов, а также производить снижение размерности признакового пространства [137].

Выводы по 2 главе

Основными результатами рассуждений, представленных в данной главе, являются:

1. Определен базовый подход для разработки модели электронного почтового сообщения для классификации электронных писем.

2. Разработана модель электронного почтового сообщения для классификации электронных писем (2.14), отличающаяся от известных моделей методом выделения значимых последовательностей символов текста (признаков электронных писем на основе их содержания, термов), позволяющим усилить

смысловое содержание термов за счет применения метода «генетических карт» [101].

3. Выделены и обоснованы параметры разработанной модели электронного почтового сообщения, оказывающие влияние на выделение термов.

4. Проведен и обоснован выбор значений параметров предложенной модели электронного почтового сообщения, оказывающих влияние на выделение термов.

5. Определены и обоснованы способы предобработки электронных писем, позволяющие повысить эффективность применения разработанной модели в задаче обнаружения спама.

6. Обоснована символонезависимость разработанной модели электронного почтового сообщения, т. е. она не зависит от кодировки текстов писем (от конкретного набора символов и их количества).

7. Обосновано, что результаты обнаружения спама с применением разработанной модели, являются неслучайными.

8. Обоснована целесообразность применения весов термов, а также снижения размерности признакового пространства.

9. Поскольку разработанная модель электронного почтового сообщения по своей сути основана на содержании спама и легальных электронных писем и ориентирована на персональные (пользовательские) особенности (с т. з. информационных потребностей) электронных писем и их содержание применительно к конкретным пользователям (группе пользователей), ее применение предоставляет возможность формировать набор термов с их маркированием по актуальности для конкретного пользователя (персональные особенности входящего потока электронных писем) и на конкретный момент времени (т. е. учитывать текущий «ландшафт» спама, его целевую направленность с т. з. его содержания).

При этом результаты, полученные и изложенные в настоящей главе диссертационных исследований, обосновывают корректность и практическую применимость в различных условиях разработанной модели электронного

почтового сообщения (2.14) для обнаружения спама, а также ее настраиваемость применительно к конкретным текстам электронных писем [12, 101, 118-122, 137].

Предложенная последовательность анализируемых научных позиций стала обоснованием перехода к исследовательским материалам третьей главы, которая называется «Разработка метода и алгоритма классификации электронных писем для обнаружения спама» и посвящается разработке:

1. Метода классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, отличающегося использованием разработанной модели электронных писем.

2. Алгоритма классификации электронных писем, отличающегося наличием дополнительной процедуры определения «схожести» термов на основе расстояния Левенштейна, обеспечивающей вычисление мер принадлежности классифицируемого электронного письма к классам спама и легальных для повышения достоверности идентификации электронных писем.

Глава 3 Разработка метода и алгоритма классификации электронных писем для обнаружения спама

Под классификацией текстов (в зарубежных источниках [75, 125] как синоним употребляется понятие «категоризация текстов») понимается задача распределения исходного множества текстов по нескольким заданным классам (категориям) на основании их содержания [75, 125, 126, 138].

Содержательно постановку задачи классификации текстов можно описать следующим образом [139]. Даны множества предопределенных произвольных классов и предварительно классифицированных текстов. Необходимо на их основе построить классификатор, который для конкретного текста определяет его класс с некоторой степенью точности. Причем классификация должна выполняться только на основе анализа содержимого текста.

3.1 Формирование метода классификации электронных писем для обнаружения спама

Формально задачу классификации текстовых документов можно представить [75, 125, 139] как задачу присвоения булевого значения $\{True, False\}$ каждой паре:

$$\langle d_i, c_k \rangle \in D \times C, \tag{3.1}$$

где D – множество классифицируемых текстовых документов;

d_i – классифицируемый текстовый документ из множества D;

C – множество классов текстовых документов, между которыми их необходимо распределить;

c_k – класс из множества C, сопоставляемый текстовому документу d_i;

$\{True, False\}$ – булево значение принадлежности d_i классу c_k ($True$) или нет – $False$.

Более формально задачу классификации текстовых документов можно определить [75, 125, 139, 140] как задачу аппроксимации неизвестной целевой функции

$$\phi_d : D \times C \rightarrow \{True, False\}, \tag{3.2}$$

описывающей как текстовый документ должен быть классифицирован, через функцию

$$\tilde{\phi}_d: D \times C \rightarrow \{True, False\}, \quad (3.3)$$

описывающую правило (алгоритм) классификации и называемую классификатором (или правилом классификации), такую, что $\tilde{\phi}_d$ и ϕ_d должны совпадать настолько, насколько это возможно. Тогда

$$\tilde{\phi}_d(d_i, c_k) = \begin{cases} True, \text{если } d_i \in c_k, \\ False, \text{если } d_i \notin c_k. \end{cases} \quad (3.4)$$

Поиск требуемого правила классификации $\tilde{\phi}_d$ осуществляется в процессе обучения с помощью некоторого алгоритма и с использованием предварительно классифицированных текстовых документов (обучающего множества).

Проецируя положения задачи классификации текстовых документов на предметную область настоящего диссертационного исследования, необходимо отметить, что в решении задачи обнаружения спама каждому электронному письму может соответствовать только один из двух классов (спам или легальные письма). Это означает, что в настоящем исследовании имеет место бинарная классификация [75, 125], при которой каждое электронное письмо (из множества всех электронных писем) должно однозначно ставиться в соответствие классу «спама» или его дополнению – «легальные письма».

В дальнейшем применительно к задаче классификации электронных писем целесообразно исходить из следующих дополнительных ключевых ограничений [75, 125]:

- классы электронных писем являются по своей сути всего лишь символьными метками и не содержат никакой дополнительной информации, позволяющей перечислить и описать признаки именуемых ими классов (категорий);

- отсутствует какая-либо внешняя (по отношению к электронным письмам) информация, существенная для цели классификации, а значит классификацию электронных писем необходимо осуществлять только по их содержимому (что полностью соотносится с выбранным подходом к обнаружению спама).

Принятые исходные ключевые ограничения позволяют обеспечить независимость предлагаемых автором исследования модели электронного почтового сообщения и метода классификации электронных писем для обнаружения спама и идентификации легальных электронных писем от любой внешней информации об электронных письмах (например, источник и формат электронных писем, дата их получения, адресат и др.) и использовать исключительно их содержание, а также результатов их применения для обнаружения спама [141].

На основании описанных основных положений задачи классификации, выражения (3.4) и с учетом введенных в главе 1 настоящего диссертационного исследования обозначений задачу обнаружения спама можно сформулировать как задачу классификации электронных писем со следующим правилом классификации (классификатором) [141]:

$$\tilde{\phi}_{EL}(el_i, c_k) = \begin{cases} True, \text{если } el_i \in c_k, \\ False, \text{если } el_i \notin c_k. \end{cases} \quad (3.5)$$

где $c_k \in C = \{Spam, Legal\}$ –классы электронных писем, между которыми их необходимо распределить. $Spam$ – класс спама, $Legal$ – класс легальных электронных писем.

Таким образом, задачу классификации электронных писем на спам и легальные с использованием разработанной в главе 2 модели можно графически проиллюстрировать следующим образом [138, 141].

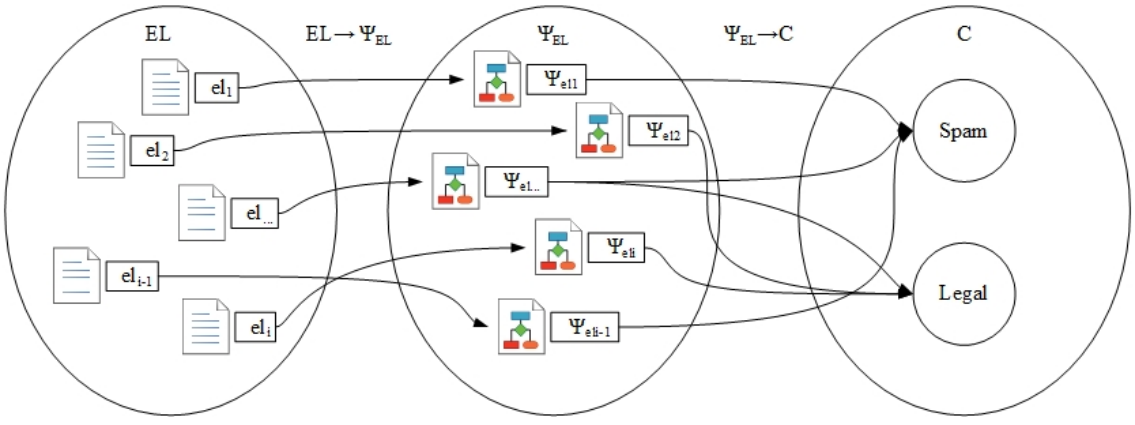

Рисунок 3.1 – Графическая иллюстрация задачи обнаружения спама

На основе изложенного описания задачи становится возможным сформировать метод классификации электронных писем в виде последовательности следующих шести этапов [138, 141, 142]:

1. Предварительная обработка текстов электронных писем EL_PrePoc.

2. Построение признакового пространства (выделение термов) T_Proc.

Изложенное в главе 2 настоящего диссертационного исследования показывает, что данные этапы реализуются через разработанную модель электронных писем Ψ_{el} (2.14).

3. Расчет весов термов – процедура ϕ_T^W.

4. Сокращение размерности признакового пространства – процедура ϕ_T^{DR}.

Данные четыре этапа (без пятого) также используются при формировании базы данных термов спама и легальных писем в режиме обучения на обучающем наборе электронных писем $EL^{tr} \subseteq EL$, для каждого электронного письма $el_i^{tr} \in EL^{tr}$ из которого известен его класс $c_k^{tr} \subseteq C$.

5. Определение классов электронных писем (множество $EL = \{el_i\}$) с использованием задаваемого способа классификации – процедура $\tilde{\phi}_{\Psi_{el}}$.

6. Определение классов неклассифицированных на этапе 5 электронных писем с применением подхода к определению «схожести» термов на основе расстояния Левенштейна (нечеткая классификация), которое можно получить путем расчета элементов матрицы D по следующей формуле:

$$D(i,j) = \begin{cases} 0, i = 0, j = 0 \\ i, j = 0, i > 0 \\ j, i = 0, j > 0 \\ min\{ \\ D(i, j-1) + 1 \\ D(i-1, j) + 1, i > 0, j > 0 \\ D(i-1, j-1) + m(S_1[i], S_2[j]) \\ \} \end{cases}, \quad (3.6)$$

где $d(t_1, t_2) = D(l_1, l_2)$ – редакционное расстояние между термами t_1 и t_2;

i, j – номера строк и столбцов матрицы, где $0 \le i \le l_1, 0 \le j \le l_2$;

$S_1[i], S_2[j]$ – символы термов, соответствующие строкам и столбцам (позициям) i и j;

$m(S_1[i], S_2[j])$ – оператор, $m(S_1[i], S_2[j]) = 1$, если символы $S_1[i]$ и $S_2[j]$ не равны друг другу, и $m(S_1[i], S_2[j]) = 0$ – в противном случае.

По своей сути предложенный метод относится к группе методов, основанных на анализе контента (содержания) и его классификации с применением методов машинного обучения.

3.2 Построение признаковых описаний текстов электронных писем

Под признаковым описанием электронных писем в настоящем диссертационном исследовании будем понимать его термы с их весами, выбор способа расчета которых является важным и оказывает существенное влияние на эффективность решения задачи их классификации.

Необходимо отметить, что разработанная модель электронных писем (2.14) оперирует с преобразованными в числовую последовательность текстами электронных писем и является по своей сути векторной моделью представления текста, что согласуется с необходимой формой представления текстов для применения алгоритмов машинного обучения [75, 142]. В связи с этим целесообразно рассматривать методы расчета весов, применимые для векторных моделей представления текстов и основанные на статистических признаках [143]. Также поскольку выделяемые с использованием модели (2.14) термы электронных писем являются элементарными единицами, являющимися отражением отличительных признаков электронных писем, в которых учитывается связь последовательности символов письма, представляется возможным не задумываться о связях между термами, а значит ограничиться методами расчета весов, применимых к однословным терминам [143].

В основе расчетов весов термов по большей части лежат статистические оценки их вхождения непосредственно в классифицируемый текст и в классифицируемый набор текстов в целом, а также на следующих эмпирических наблюдениях [142]:

- релевантность терма классу конкретного текста тем выше, чем чаще терм встречается в его тексте;

- классифицирующая способность терма уменьшается с увеличением количества текстов классифицируемого набора, в которых встречается терм.

Можно условно выделить следующие группы подходов к вычислению весов термов [144, 145]:

1. По изучаемой коллекции. В основе группы данных способов лежит предположение о том, что информативные термы, как правило, встречаются в наборах гораздо чаще неинформативных. К основным весовым коэффициентам данной группы относятся частотность TF (*аббр. от англ.* Term Frequency), документная частотность DF (*аббр. от англ.* Document Frequency), а также функции класса $TF - IDF$ [например, 146-149], основанные на частотности TF и обратной документной частотности IDF (*аббр. от англ.* Inverse Document Frequency) [например, 150].

2. По классифицируемым и контрастным наборам. В основе подходов данной группы лежит предположение о существенном различии частотности термов в классифицируемом и контрастном наборах. К данной группе весов относятся, например, относительная частотность [например, 144, 151-153] и мера $KF - IDF$ [например, 144, 154, 155], являющаяся разновидностью меры $TF - IDF$.

3. По статистической и контекстной информации из изучаемого набора. Весовые коэффициенты данной группы соединяют в себе частотность термов с данными о контексте их употребления в наборе текстов. К таковым относятся, например, мера $C - Value$ [например, 144, 151-153, 156-158].

Многие работы разных исследователей [например, 144-146, 151-153, 156, 157, 159-168] посвящены экспериментальным сравнениям эффективности различных весов термов на различных наборах текстов. Анализ полученных ими результатов позволили автору настоящего диссертационного исследования прийти к следующим основным выводам [143]:

1. В целом все сравниваемые веса термов дают похожие результаты и различаются в зависимости от экспериментальных наборов данных.

2. Наиболее широко распространены весовые функции $TF - IDF$. Они являются наиболее простыми для вычислений и в сравнении с другими весами показывают относительно лучшую эффективность не только при решении задач классификации, но и поиска информации в массивах различных текстов, в том числе их индексирования и рубрицирования.

3. Комбинации способов могут позволить увеличить эффективность классификации в сравнении с их применением по-отдельности.

Данные выводы делают целесообразным применение весовых функций класса $TF - IDF$[6] при решении задачи классификации электронных писем [143, 169].

Обозначим за tf_{ij} частоту j-го терма в i-м письме – отношение числа его вхождений в текст письма к общему числу всех термов в этом письме. Тогда:

$$tf_{ij} = \frac{n_{ij}}{N_T}, \tag{3.7}$$

где n_{ij} – количество вхождений j-го терма в i-м письме.

Обозначим через idf_j инверсную документарную частоту j-го терма – логарифм отношения числа всех писем к числу писем, в которых встречается j-й терм:

$$idf_j = \log\left(\frac{N_{EL}}{n_j}\right), \tag{3.8}$$

где N_{EL} – количество всех писем:

$$N_{EL} = |EL|, \tag{3.9}$$

n_j – количество писем, в которых встречается j-й терм.

Основываясь на формулах (3.8) и (3.9) в базовом виде формулу расчета веса $TF - IDF$ можно представить следующим образом [144, 146, 151, 159]:

$$w_{ij} = tf_{ij} \cdot idf_j \text{ или} \tag{3.10}$$

$$w_{ij} = \frac{n_{ij}}{N_T} \cdot \log\left(\frac{N_{EL}}{n_j}\right). \tag{3.11}$$

[6] В дальнейшем при описании формул расчета весовых функций описания их параметров будут осуществляться применительно к предметной области настоящего диссертационного исследования.

Особенностью данной меры является то, что вес терма пропорционален частоте его употребления в конкретном письме и обратно пропорционален частоте употребления во всех письмах. Таким образом, можно оценить важность терма в пределах конкретного письма. При этом больший вес получают термы с большей частотой в пределах конкретного письма и с меньшей частотой употребления в других письмах.

У формулы (3.10) существуют следующие наиболее распространенные модификации, которые целесообразно рассмотреть при разработке метода классификации.

Очевидно, что

$$w_{ij} = 0 \text{ при } N_{EL} = n_j. \tag{3.12}$$

Данный случай может наблюдаться при небольшом количестве классифицируемых писем. Во избежание таких случаев целесообразно применение [159] сглаживающего коэффициента для улучшения формулы (3.10):

$$w_{ij} = \log\left(tf_{ij} + 1\right) \cdot \log\left(\frac{N_{EL}+1}{n_j}\right). \tag{3.13}$$

В [159] в дополнение к TF и IDF предложено ввести параметр CF (*аббр. от англ.* Class Frequency), обозначающую частоту терма в пределах заданного класса:

$$\mathrm{c}f_j = \frac{N_{ij}^C}{N_i^C}, \tag{3.14}$$

где N_{ij}^C – количество писем того же класса, что и i-е письмо, в которых встречается j-й терм.

N_i^C – число писем того же класса, что и i-е письмо.

Тогда формула (3.13) примет вид:

$$w_{ij} = \log\left(tf_{ij} + 1\right) \cdot \log\left(\frac{N_{EL}+1}{n_j}\right) \cdot \frac{N_{ij}^C}{N_i^C}. \tag{3.15}$$

В большинстве ситуаций в небольших письмах, как правило, будет присутствовать небольшое количество термов, а в больших наоборот. Это предопределяет необходимость использования коэффициента нормализации, позволяющего устранить эффект больших различий в частотах термов в текстах

писем различной длины. В качестве коэффициента нормализации, как правило, может выступать следующий [146, 159]:

$$norm_i = \frac{1}{\sqrt{\sum_{j=1}^{N_T}(w_{ij})^2}}. \quad (3.16)$$

С учетом коэффициента нормализации унифицированная формула расчета весов $TF - IDF$ примет вид:

$$w_{ij} = w_{ij} \cdot norm_i. \quad (3.17)$$

Тогда формула (3.10) примет вид:

$$w_{ij} = \frac{tf_{ij} \cdot \log\frac{N_{EL}}{n_j}}{\sqrt{\sum_{j=1}^{N_T}\left(tf_{ij} \cdot \log\frac{N_{EL}}{n_j}\right)^2}}. \quad (3.18)$$

а формула (3.13):

$$w_{ij} = \frac{\log(tf_{ij}+1) \cdot \log\left(\frac{N_{EL}+1}{n_j}\right)}{\sqrt{\sum_{j=1}^{N_T}\left(\log(tf_{ij}+1) \cdot \log\left(\frac{N_{EL}+1}{n_j}\right)\right)^2}}. \quad (3.19)$$

Также в [159] предложена следующая модификация формулы (3.19):

$$w_{ij} = \frac{\log(tf_{ij}+1) \cdot \log\left(\frac{N_{EL}}{n_j}\right)}{\sqrt{\sum_{j=1}^{N_T}\left(\log(tf_{ij}+1) \cdot \log\left(\frac{N_{EL}}{n_j}\right)\right)^2}}. \quad (3.20)$$

В ряде исследований [147, 152, 161, 162, 170, 171], посвященных вопросам анализа текстов и информационного поиска, приведен вариант меры $TF - IDF$ в формулировке поисковой системы $INQUERY$ [172]:

$$w_{ij} = \beta + (1 - \beta) \cdot tf_{ij} \cdot idf_j, \quad (3.21)$$

где

$$tf_{ij} = \frac{tf_{ij}}{tf_{ij}+0{,}5+1{,}5 \cdot \frac{N_T}{\overline{N_T}}}, \quad (3.22)$$

$$idf_j = \frac{\log\left(\frac{N_{EL}+0{,}5}{n_j}\right)}{\log(N_{EL}+1)}, \quad (3.23)$$

где $\overline{N_T}$ – среднее число термов в одном письме (в термах),

$\beta = 0{,}4$ [147, 152, 161, 162, 170, 171].

Целесообразно выделить следующие основные значимые свойства весовых коэффициентов $TF - IDF$:

1. При вхождении термов в небольшое количество электронных писем их значения увеличиваются, тем самым усиливая отличие содержащих их писем от других.

2. Их значения снижаются для редких или общеупотребимых термов.

3.3 Сокращение размерности признакового пространства.

Как отмечается в [75, 126, 142], большое количество извлеченных термов, которое может достигать десятков и сотен тысяч, является центральной проблемой в задачах классификации, в том числе при помощи алгоритмов машинного обучения. Некоторые из них не могут принимать такое количество параметров для обучения в качестве исходных, а другие будут требовать значительных вычислительных затрат. Эта проблема в машинном обучении носит название «проклятье размерности». Следовательно, сокращение размерности пространства признаков в задачах классификации представляется актуальной задачей.

Действительно, применительно к решаемой в настоящем диссертационном исследовании задаче есть причина, обусловливающая необходимость уменьшения размера признакового пространства (сокращения количества термов). Очевидно, что учет всех термов приводит к значительному увеличению их количества, хотя некоторые из них будут оказывать незначительное влияние на обнаружение спама. Высокая размерность в свою очередь может приводить к вычислительной погрешности и низкой скорости работы, а также потребует недопустимо больших вычислительных ресурсов и времени.

Целью сокращения признакового пространства является не только уменьшение его размерности и повышение скорости вычислений. Среди термов могут встретиться так называемые «шумовые», которые уменьшают точность классификации. Такая ситуация возникает, потому что при построении модели в обучающем наборе обнаруживаются некоторые случайные закономерности, которые будут отсутствовать в генеральной совокупности. Например, в

обучающем наборе какой-либо терм окажется только в спаме, но при этом не будет нести никакой информации о данном классе. В таком случае мы рискуем получить правило классификации, ошибочно классифицирующее все электронные письма, содержащие этот терм, как спам.

Такое некорректное обобщение случайных закономерностей из обучающих данных называется переобучением, когда классификатор будет приводить к хорошим результатам на письмах из обучающего набора и к плохим – на письмах, не участвовавших в обучении. Сокращение размерности признакового пространства (количества термов) может помочь решить эту проблему. При этом эффективность классификации на новых данных все равно может быть ниже, чем на данных, использовавшихся для обучения [173].

В общем виде применительно к настоящему диссертационному исследованию задачу сокращения размерности пространства признаков электронных писем можно сформулировать как отбор из начального множества термов T подмножество наиболее информативных термов T^{DR} ($T^{DR} \ll T$), обладающих наилучшими разделяющими свойствами между классами спама и легальных.

Таким образом, решение задачи сокращения размерности пространства признаков необходимо в следующих основных целях [97, 142, 174]:

- исключение дублирования информации, возникающего из-за наличия сильно взаимосвязанных термов;

- исключение неинформативных термов, мало меняющихся при переходе от спама к легальным письмам;

- решение проблемы переобучения классификатора;

- повышение точности классификатора;

- упрощение и снижение погрешности вычислений;

- упрощение интерпретации результатов;

- получение большей наглядности результатов;

- снижение вычислительных затрат.

При решении задачи сокращения размерности пространства термов необходимо выполнение ряда требований [97] (вместе взятых или по отдельности), среди которых применительно к задаче настоящего диссертационного исследования целесообразно отметить следующие:

- должна быть сохранена основная (значимая) часть исходной информации пространства термов;

- термы после сокращения размерности пространства не должны коррелироваться между собой или быть слабо коррелированными;

- новое пространство термов должно давать наибольшую информативность с точки зрения эталонной классификации.

Формально задача сокращения размерности пространства термов может быть представлена как отображение:

$$T^C \xrightarrow{\phi_T^{DR}} T^{C(DR)}. \quad (3.24)$$

где T^C – множество термов электронных писем определенного класса (описание класса);

$T^{C(DR)}$ – новое множество термов электронных писем определенного класса ($T^{C(DR)} \ll T^C$);

Для решения задачи сокращения размерности признакового пространства можно выделить [75, 97, 126, 140, 142] следующие основные классические подходы:

- факторный и компонентный анализ;

- выбор информативных признаков из существующих.

Факторный и компонентный анализ осуществляются с целью перехода от описания объектов с большим количеством исходных признаков к их описанию с помощью меньшего числа новых специально сформированных обобщенных признаков, заменяющих с достаточной точностью группы исходных признаков [97, 142].

Отбор информативных признаков из существующих осуществляется с целью исключения неинформативных признаков из исходного множества по каким-либо

задаваемым критериям. В значительном количестве научных работ [например, 75, 97, 126, 127, 140, 142, 174-192] их авторами приводятся и дается оценка применения методов выбора информативных признаков с применением различных мер, наиболее распространенными из которых являются следующие: сравнение веса термина с порогом, прирост информации (*IG, аббр. от англ.* Information Gain), критерий χ^2, взаимная информативность признаков (*MI, аббр. от англ.* Mutual Information), индекс Джини (*GI, аббр. от англ.* Gini Index) и др.

3.3.1 Прирост информации

Данная мера зачастую применяется в области машинного обучения как критерий «качества» значимых признаков [174]. Значение данной меры для терма показывает разницу бит информации, необходимых для классификации текста с использованием этого признака и без его использования. Чем больше величина прироста информации, тем сильнее разделяющая способность терма.

Удобным способом расчета прироста информации является следующее представление данной меры [183], которое с учетом предметной области исследований будет иметь вид:

$$IG(t_j) = \left(-\sum_{c_k}\left(\frac{\left|EL_{c_k}^{tr}\right|}{|EL^{tr}|}\right)\cdot\log_2\left(\frac{\left|EL_{c_k}^{tr}\right|}{|EL^{tr}|}\right)\right) +$$

$$+\sum_{c_k}\left(\frac{\left|EL_{t_j,c_k}^{tr}\right|}{|EL^{tr}|}\right)\cdot\sum_{c_k}\left(\frac{\left|EL_{t_j,c_k}^{tr}\right|}{\left|EL_{t_j,c_k}^{tr}\right|+\left|EL_{t_j,\overline{c_k}}^{tr}\right|}\cdot\log_2\left(\frac{\left|EL_{t_j,c_k}^{tr}\right|}{\left|EL_{t_j,c_k}^{tr}\right|+\left|EL_{t_j,\overline{c_k}}^{tr}\right|}\right)\right)+$$

$$+\sum_{c_k}\left(\frac{\left|EL_{\overline{t_j},c_k}^{tr}\right|}{|EL^{tr}|}\right)\cdot\sum_{c_k}\left(\frac{\left|EL_{\overline{t_j},c_k}^{tr}\right|}{\left|EL_{\overline{t_j},c_k}^{tr}\right|+\left|EL_{\overline{t_j},\overline{c_k}}^{tr}\right|}\cdot\log_2\left(\frac{\left|EL_{\overline{t_j},c_k}^{tr}\right|}{\left|EL_{\overline{t_j},c_k}^{tr}\right|+\left|EL_{\overline{t_j},\overline{c_k}}^{tr}\right|}\right)\right), \qquad (3.25)$$

где $\left|EL_{c_k}^{tr}\right|$ – количество электронных писем обучающего набора, принадлежащих классу c_k;

$\left|EL_{t_j,c_k}^{tr}\right|$– количество электронных писем обучающего набора, содержащих терм t_j и принадлежащих классу c_k;

$\left|EL_{t_j,\overline{c_k}}^{tr}\right|$ – количество электронных писем обучающего набора, содержащих терм t_j и не принадлежащих классу c_k;

$\left|EL^{tr}_{\overline{t_j},c_k}\right|$ – количество электронных писем обучающего набора, не содержащих терм t_j, но принадлежащих классу c_k, т. е.:

$$\left|EL^{tr}_{\overline{t_j},c_k}\right| = \left|EL^{tr}_{c_k}\right| - \left|EL^{tr}_{t_j,c_k}\right|; \tag{3.26}$$

$\left|EL^{tr}_{\overline{t_j},\overline{c_k}}\right|$ – количество электронных писем обучающего набора, ни содержащих терм t_j, ни принадлежащих классу c_k, т. е.:

$$\left|EL^{tr}_{\overline{t_j},\overline{c_k}}\right| = |EL^{tr}| - \left|EL^{tr}_{c_k}\right| - \left|EL^{tr}_{t_j,\overline{c_k}}\right|. \tag{3.27}$$

Учитывая, что по результатам классификации электронному письму может быть не присвоен ни один из классов c_k, формула (3.25) примет следующий вид:

$$IG(t_j) = \left(-\sum_{c_k}\left(\frac{\left|EL^{tr}_{c_k}\right|}{|EL^{tr}|}\right)\cdot\log_2\left(\frac{\left|EL^{tr}_{c_k}\right|}{|EL^{tr}|}\right)\right) +$$
$$+\frac{\left|EL^{tr}_{t_j}\right|}{|EL^{tr}|}\cdot\sum_{c_k}\left(\frac{\left|EL^{tr}_{t_j,c_k}\right|}{\left|EL^{tr}_{t_j}\right|}\cdot\log_2\left(\frac{\left|EL^{tr}_{t_j,c_k}\right|}{\left|EL^{tr}_{t_j}\right|}\right)\right) +$$
$$+\frac{\left|EL^{tr}_{\overline{t_j}}\right|}{|EL^{tr}|}\cdot\sum_{c_k}\left(\frac{\left|EL^{tr}_{\overline{t_j},c_k}\right|}{\left|EL^{tr}_{\overline{t_j}}\right|}\cdot\log_2\left(\frac{\left|EL^{tr}_{\overline{t_j},c_k}\right|}{\left|EL^{tr}_{\overline{t_j}}\right|}\right)\right), \tag{3.28}$$

где $\left|EL^{tr}_{t_j}\right|$ – количество электронных писем обучающего набора, содержащих терм t_j, т. е.:

$$\left|EL^{tr}_{t_j}\right| = \left|EL^{tr}_{t_j,c_k}\right| + \left|EL^{tr}_{t_j,\overline{c_k}}\right|; \tag{3.29}$$

$\left|EL^{tr}_{\overline{t_j}}\right|$ – количество электронных писем обучающего набора, не содержащих терм t_j, т. е.:

$$\left|EL^{tr}_{\overline{t_j}}\right| = |EL^{tr}| - \left|EL^{tr}_{t_j}\right|. \tag{3.30}$$

Таким образом, на основе обучающего набора имеется возможность рассчитать прирост информации отдельно для каждого терма. Сокращение размерности признакового пространства осуществляется путем удаления из него всех тех термов, величина прироста информации которых ниже заранее заданного порогового значения.

На основе вышеизложенного правило принятия решения в процедуре сокращения размерности признакового пространства (термов) электронных писем ϕ_T^{DR} будет иметь следующий вид:

$$\phi_T^{DR}: IG(t_j) \geq P, \tag{3.31}$$

где P – заданное пороговое значение.

Описанная мера является одной из самых популярных для оценки значимости термов.

3.3.2 Взаимная информативность признаков

Данная мера для отбора значимых термов также широко распространена при решении задач классификации наряду с вышеописанными мерами. Ее понятие произошло из теории информации и предоставляет возможность оценить связь между термами и классами.

Взаимная информативность между термом и классом может быть определена по следующей формуле [174, 177, 183]:

$$MI(t_j, c_k) = \log_2\left(\frac{\left|EL_{t,c_k}^{tr}\right| \cdot |EL^{tr}|}{\left|EL_{c_k}^{tr}\right| \cdot \left|EL_{t_j}^{tr}\right|}\right). \tag{3.32}$$

Мера взаимной информативности $MI(t_j, c_k)$ показывает, насколько много информации (в смысле теории информации) о классе c_k содержит терм t_j. Значение меры взаимной информативности $MI(t_j, c_k)$ терма t_j и класса c_k достигает максимума, если терм t_j является идеальным индикатором для заданного класса c_k, т. е. если терм присутствует в электронном письме в том случае, когда электронное письмо принадлежит классу c_k. При этом если распределение терма t_j в классе c_k совпадает с распределением терма t_j во всем наборе электронных писем, то значение меры взаимной информативности $MI(t_j, c_k)$ становится равным 0 ($MI(t_j, c_k) = 0$).

Следовательно, за значение взаимной информативности для терма t_j целесообразно принимать его наибольшее значение из рассчитанных применительно к различным классам c_k [174, 177, 183]:

$$MI(t_j) = \max_k \{MI(t_j, c_k)\}. \quad (3.33)$$

Таким образом, сокращение размерности пространства термов электронных писем на основе меры взаимной информативности MI можно производить по следующему алгоритму:

1. Для каждого терма t_j рассчитывается значение его взаимной информативности $MI(t_j, c_k)$ с каждым из классов C.

2. Для каждого терма определяется наибольшее значение его взаимной информативности $MI(t_j)$ по правилу (3.33).

3. После шага 2 термы сортируются по убыванию наибольших значений их взаимной информативности $MI(t_j)$.

4. За информативные для электронных писем принимаются термы со значениями их взаимной информативности $MI(t_j)$ выше заданного порога.

На основе вышеизложенного правило принятия решения в процедуре сокращения размерности признакового пространства (термов) электронных писем ϕ_T^{DR} будет иметь следующий вид:

$$\phi_T^{DR}: MI(t_j) \geq P. \quad (3.34)$$

3.3.3 Критерий χ^2

Еще одной популярной величиной среди мер для отбора значимых признаков [184, 191] является критерий χ^2. Данная мера призвана измерять степень зависимости (а вернее, отсутствие независимости) между термом t_j и классом c_k [178, 188, 191]. Измерения с использованием этого критерия также известны как тест независимости [191].

Математическое представление критерия $\chi^2(t_j, c_k)$ между термом t_j и классом c_k в виде формулы выглядит следующим образом [126, 183, 191]:

$$\chi^2(t_j, c_k) = |EL^{tr}| \frac{\left(\left|EL^{tr}_{t_j,c_k}\right| \cdot \left|EL^{tr}_{\overline{t_j},\overline{c_k}}\right| - \left|EL^{tr}_{t_j,\overline{c_k}}\right| \cdot \left|EL^{tr}_{\overline{t_j},c_k}\right|\right)^2}{\left(\left|EL^{tr}_{t_j,c_k}\right| + \left|EL^{tr}_{\overline{t_j},c_k}\right|\right) \cdot \left(\left|EL^{tr}_{t_j,\overline{c_k}}\right| + \left|EL^{tr}_{\overline{t_j},\overline{c_k}}\right|\right) \cdot \left(\left|EL^{tr}_{t_j,c_k}\right| + \left|EL^{tr}_{t_j,\overline{c_k}}\right|\right) \cdot \left(\left|EL^{tr}_{\overline{t_j},c_k}\right| + \left|EL^{tr}_{\overline{t_j},\overline{c_k}}\right|\right)}. \quad (3.35)$$

Значение критерия $\chi^2(t_j, c_k)$ достигает максимума, если терм t_j является наиболее информативным для заданного класса c_k. При этом если терм t_j не зависит от класса c_k, то значение критерия $\chi^2(t_j, c_k)$ становится равным 0 ($\chi^2(t_j, c_k) = 0$) [184, 191]. Следовательно, за значение критерия $\chi^2(t_j, c_k)$ для терма t_j (по аналогии с мерой взаимной информативности) целесообразно принимать его наибольшее значение из рассчитанных применительно к различным классам c_k [126, 174, 177, 191]:

$$\chi^2(t_j) = \max_k\{\chi^2(t_j, c_k)\}. \quad (3.36)$$

Таким образом, сокращение размерности пространства термов электронных писем на основе критерия χ^2 можно производить по следующему алгоритму:

1. Для каждого терма t_j рассчитывается значение его критерия $\chi^2(t_j, c_k)$ с каждым из классов C.

2. Для каждого терма определяется наибольшее значение его критерия $\chi^2(t_j)$ по правилу (3.36).

3. После шага 2 термы сортируются по убыванию наибольших значений их критериев $\chi^2(t_j)$.

4. За информативные для электронных писем принимаются термы со значениями их критериев $\chi^2(t_j)$ выше заданного порога.

На основе вышеизложенного правило принятия решения в процедуре сокращения размерности признакового пространства (термов) электронных писем ϕ_T^{DR} будет иметь следующий вид:

$$\phi_T^{DR}: \chi^2(t_j) \geq P. \quad (3.37)$$

Необходимо отметить, что критерий χ^2 и мера взаимной информативности MI являются различными способами измерения корреляции между термами и классами. Основным отличием между критерием χ^2 и мерой взаимной информативности MI состоит в том, что критерий χ^2 является нормализованной величиной [174]. Следовательно, эти значения могут подвергаться сравнению между различными термами одного класса.

3.3.4 Индекс Джини

Еще одной из наиболее используемых мер для сокращения размерности признакового пространства является мера, известная как индекс Джини, позволяющий оценить разделяющую способность терма и призванный измерить вероятность неправильной классификации конкретного случайно выбранного объекта в эксперименте.

Индекс Джини в классическом представлении [192] для решения задач машинного обучения применительно к области настоящего диссертационного исследования можно представить как:

$$GI(EL^{tr}) = 1 - \sum_{c_k} \left(\frac{\left| EL_{c_k}^{tr} \right|}{|EL^{tr}|} \right)^2. \quad (3.38)$$

Учитывая, что по результатам классификации электронному письму может быть не присвоен ни один из классов c_k, а также адаптируя формулу (3.38) применительно к наличию или отсутствию термина t_j в электронных письмах, получим [192]:

$$GI(t_j) = \frac{\left| EL_{t_j}^{tr} \right|}{|EL^{tr}|} \cdot \left(1 - \sum_{c_k} \left(\frac{\left| EL_{t_j, c_k}^{tr} \right|}{\left| EL_{t_j}^{tr} \right|} \right)^2 \right) +$$

$$+ \frac{\left| EL_{\overline{t_j}}^{tr} \right|}{|EL^{tr}|} \cdot \left(1 - \sum_{c_k} \left(\frac{\left| EL_{\overline{t_j}, c_k}^{tr} \right|}{\left| EL_{\overline{t_j}}^{tr} \right|} \right)^2 \right) \quad (3.39)$$

или

$$GI(t_j) = 1 - \frac{\left| EL_{t_j}^{tr} \right|}{|EL^{tr}|} \cdot \sum_{c_k} \left(\frac{\left| EL_{t_j, c_k}^{tr} \right|}{\left| EL_{t_j}^{tr} \right|} \right)^2 - \frac{\left| EL_{\overline{t_j}}^{tr} \right|}{|EL^{tr}|} \cdot \sum_{c_k} \left(\frac{\left| EL_{\overline{t_j}, c_k}^{tr} \right|}{\left| EL_{\overline{t_j}}^{tr} \right|} \right)^2. \quad (3.40)$$

Как следует из формулы (3.39) (3.40) значение индекса Джини может принимать значения в интервале от 0 до 1, где:

- 0 означает, что все элементы множества принадлежат только одному классу и мы имеем максимум полезной информации об объектах;

- 1 означает, что элементы произвольным образом распределены между различными классами и мы не имеем никакой полезной информации об объектах;

- 0,5 означает, что элементы множества равнозначно распределены между всеми классами.

Таким образом, чем меньше значение коэффициента Джини, тем сильнее разделяющая способность термина.

В связи с тем, что распределение по классам в целом может быть несбалансированным, рассчитываемый по формуле (3.39) (3.40) индекс Джини не всегда точно будет отражать разделительную способность признаков. Для устранения этого недостатка в [192] предложено улучшение формулы (3.39) (3.40), которая с учетом введенных обозначений будет иметь следующий вид:

$$GI(t_j) = \sum_{c_k} \left(\left(\frac{\left| EL^{tr}_{t_j,c_k} \right|}{\left| EL^{tr}_{c_k} \right|} \right)^2 \cdot \left(\frac{\left| EL^{tr}_{t_j,c_k} \right|}{\left| EL^{tr}_{t_j} \right|} \right)^2 \right). \quad (3.41)$$

Учитывая, что по результатам классификации электронному письму может быть не присвоен ни один из классов c_k, формула (3.41) примет следующий вид:

$$GI(t_j) = \sum_{c_k} \left(\left(\frac{\left| EL^{tr}_{t_j,c_k} \right|}{\left| EL^{tr}_{c_k} \right|} \right)^2 \cdot \left(\frac{\left| EL^{tr}_{t_j,c_k} \right|}{\left| EL^{tr}_{t_j} \right|} \right)^2 \right) + \left(\frac{\left| EL^{tr}_{t_j,\bar{C}} \right|}{\left| EL^{tr}_{\bar{C}} \right|} \right)^2 \cdot \left(\frac{\left| EL^{tr}_{t_j,\bar{C}} \right|}{\left| EL^{tr}_{t_j} \right|} \right)^2, \quad (3.42)$$

где $\left| EL^{tr}_{\bar{C}} \right|$ – количество электронных писем обучающего набора, не принадлежащих классам C;

$\left| EL^{tr}_{t_j,\bar{C}} \right|$ – количество электронных писем обучающего набора, содержащих терм t_j и не принадлежащих классам C.

Смысловое значение формулы (3.42) (3.41) обратно смысловому значению формул (3.39) (3.40): чем больше значение $GI(t_j)$, тем сильнее разделяющая способность терма.

На основе вышеизложенного правило принятия решения в процедуре сокращения размерности признакового пространства (термов) электронных писем ϕ_T^{DR} будет иметь следующий вид:

$$\phi_T^{DR}: GI(t_j) \geq P. \quad (3.43)$$

Необходимо отметить, что в диссертационном исследовании не рассматривается вопрос обоснования выбора процедур сокращения признакового

пространства. При этом их использование в разработанном методе классификации электронных писем обусловлено только демонстрацией его работоспособности в целом и модели электронного почтового сообщения в частности. В связи с этим, а также на основе анализа источников [например, 75, 97, 126, 127, 140, 142, 174-192] в настоящем исследовании для оценки применения разработанной модели (2.14) и метода классификации будет использован индекс Джини. Он является универсальным применительно к решаемой задаче, независимым от алгоритмов классификации, а также обладает масштабируемостью, простотой применения и наименьшей вычислительной сложностью.

3.4 Правила классификации электронных писем для решения задачи обнаружения спама

Результаты анализа современных исследований, посвященных изучению методов машинного обучения в контексте решения задачи обнаружения спама показывает, что среди наиболее часто и эффективно используемых для ее решения методов машинного обучения, а также наиболее часто цитируемых можно выделить следующие: наивный Байесовский классификатор [например, 11, 19, 37-39, 40-43], опорных векторов [например, 47-50], нейросети [например, 13, 48, 55-61], k-ближайших соседей [например, 51, 52].

В наивном Байесовском классификаторе используется вероятностная модель определения принадлежности текстов к заданным классам по заданным признакам. Главным недостатком этого подхода является предположение о независимости слов текста (это упрощает вычисления и снижает затраты на них) [142].

Метод опорных векторов (SVM, *аббр. от англ.* Support Vector Machines) – один из популярных методов классификации, также применяемый для обнаружения спама. В основе данного метода лежит задача нахождения разделителя (гиперплоскости) в искомом пространстве, разделяющего наилучшим образом заданные классы (гиперплоскость находится на максимальном расстоянии до любых точек заданных классов).

В основу моделирования искусственных нейронных сетей для обработки информации положены известные принципы функционирования биологических нейронных сетей [142]. Они представляют собой наборы соединенных между собой узлов, имеющих вход, выход и функцию их активации, и способны обучаться на тренировочных наборах данных. Высокая помехоустойчивость, параллельная обработка данных, а также способность классификации линейно неразделимых классов относятся к основным достоинствам искусственных нейронных сетей [142]. Сложность выбора структуры сети и настройки ее параметров, т. е. сложность обучения, относят к их недостаткам [142].

В основе метода k-ближайших соседей лежит принцип сравнения исследуемого текстов писем с текстами писем обучающей выборки с целью нахождения наиболее близких по содержанию. Тем самым происходит определение класса, релевантного исследуемому тексту.

В основу метод k-ближайших соседей в простейшем виде положено правило [193, 194] присвоения классифицируемому объекту того же класса, что и класс ближайшего объекта классифицированной обучающей выборки. В таком виде алгоритм может иметь большую погрешность (например, при наличии выбросов в обучающих выборках). В общем виде метод k-ближайших соседей работает со следующим правилом: классифицируемому письму присваивают тот же класс, что и класс большинства из k-ближайших соседей обучающей выборки. При этом число соседей принимают нечетным в случае решения задачи с двумя классами. Это позволяет избежать ситуаций неоднозначности в случае принадлежности разным классам одинакового числа соседей [193]. В качестве меры расстояния могут быть применены различные функции, при этом для сравнения текстов часто используется косинусная мера.

В основе метода k-ближайших соседей заложена необходимость хранения в исходном виде текстов всех писем обучающей выборки, большой размер которой обуславливает обязательность решения следующих проблем технического характера: необходимость хранения большого объема данных с одновременной реализацией быстрого поиска для произвольного признака его k-ближайших

соседей [194]. Также необходимо учитывать проблему проклятия размерности, проявляющуюся при вычислении расстояния как суммы отклонений по отдельным признакам, количество которых велико. Согласно закону больших чисел значения данных сумм с большой вероятностью будут очень близки. Таким образом, высокая размерность признакового пространства приводит к практически равноудаленности признаков друг от друга, что делает результат применения метода k-ближайших соседей практически произвольным. Решение указанных проблем требует выделения набора признаков, информативных применительно к письмам обучающей выборки, включая сокращение размерности признакового пространства [194].

Таким образом, большие затраты вычислительных ресурсов и длительное время классификации относятся к недостаткам данного метода. Среди же его преимуществ можно выделить [142] простоту обучения малые ошибки классификации и простоту программной реализации.

Результаты анализа современных исследований, посвященных изучению методов машинного обучения в контексте решения задачи обнаружения спама, позволяют сделать следующие выводы:

1. Не существует универсального алгоритма, сочетающего достоинства малой вычислительной сложности и малой ошибки классификации при любых начальных условиях.

2. Малая ошибка отдельных методов достигается за счет существенного увеличения времени обучения и применения слабо формализуемых подходов к настройке их параметров.

3. В зависимости от состава обучающей выборки возможно получение противоречивых результатов в случае применения одного и того же метода.

4. В зависимости от состава обучающей выборки возможно получение соизмеримых ошибок при использовании разных методов. При этом время классификации может заметно отличаться.

Необходимо отметить, что в диссертационном исследовании не рассматривается вопрос обоснования выбора процедур классификации

электронных писем. При этом их использование в разработанном методе классификации электронных писем обусловлено только демонстрацией его работоспособности в целом и модели электронного почтового сообщения в частности. В связи с этим, а также с учетом изложенного в главе 1 настоящего диссертационного исследования и с учетом простоты реализации и обучения, интерпретируемости и устойчивости результатов, а также малой ошибки классификации, в настоящем исследовании для оценки применения разработанной модели (2.14) и метода классификации электронных писем для обнаружения спама и идентификации легальных электронных писем будут использованы:

- простое решающее правило;
- косинусная мера.

3.5 Разработка подхода к оценке эффективности (качества) метода классификации электронных писем для обнаружения спама и идентификации легальных электронных писем

Под эффективностью (качеством) метода классификации электронных писем для обнаружения спама и идентификации легальных электронных писем в настоящем диссертационном исследовании будем понимать комплексное операционное свойство (качество) обнаружения спама, характеризующее его приспособленность к достижению цели (выполнению задачи) [96].

В качестве критерия эффективности примем обобщенный показатель эффективности и правило выбора лучшего решения. В качестве частных показателей эффективности выступают характеристики, отражающие целевую направленность [96] метода.

Показатели можно разделить на частные показатели эффективности y, которые отражают существенные свойства метода классификации электронных писем и обобщенный показатель его эффективности [96]:

$$Y = \langle y_1, \dots, y_k \rangle, \tag{3.44}$$

отражающий совокупность всех свойств метода классификации электронных писем в целом. Показатель эффективности характеризует метод (алгоритм) и эффект от функционирования системы.

Для оценки эффективности (качества) разработанного метода классификации электронных писем в качестве частных показателей в настоящей работе используются показатели, определенные в разделе 2.3.1 главы 2 настоящего диссертационного исследования.

Таким образом, обобщенный показатель эффективности метода классификации электронных писем, обеспечивающего точность и полноту обнаружения спама, а также достоверность идентификации легальных электронных почтовых сообщений, в общем случае примет следующий вид:

$$Y = \langle R, P, F \rangle. \quad (3.45)$$

Необходимо отметить, что для обеспечения максимальной эффективности классификации электронных писем данный обобщенный показатель позволяет осуществить настройку метода классификации электронных писем с использованием пороговых значений (задаваемых пользователем) входящих в него частных показателей. При этом, основываясь на результатах экспериментальных исследований, приведенных в главе 2, правилами выбора лучших решений по обнаружению спама и достоверности идентификации легальных электронных почтовых сообщений целесообразно выбрать следующие:

- полнота обнаружения спама и F-мера их обнаружения;

- полнота обнаружения легальных писем.

Необходимо отметить, что для любого пользователя требуется максимальная полнота обнаружения легальных писем, что выдвигает требования к максимальному значению точности обнаружения спама. Также условием, безусловно, является максимальность F-меры обнаружения спама, демонстрирующая в таких условиях наилучший показатель обнаружения спама при одновременно максимальном значении обнаружения легальных писем. Следовательно, наилучший показатель обнаружения спама и легальных писем одновременно будет достигаться при их максимальных значениях F-меры.

Таким образом, обобщенный показатель эффективности метода классификации электронных писем, обеспечивающего точность и полноту обнаружения спама и достоверности идентификации легальных электронных почтовых сообщений, будет иметь следующий вид:

$$Y = \langle F^{Legal}, F^{Spam} \rangle. \quad (3.46)$$

Поскольку возможна ситуация, когда максимальное значение F-меры для спама будет достигаться при использовании одного правила классификатора, а для легальных писем – для другого, то целесообразно оценивать эффективность по единому значению, полученному на основе обоих значений F-мер, например, по их среднему гармоническому:

$$Y = HM(F^{Legal}, F^{Spam}). \quad (3.47)$$

Предложенный подход к оценке эффективности (качества) метода классификации электронных писем позволяет пользователю экспериментальным путем самостоятельно проводить оценку требуемого уровня эффективности работы классификатора путем настройки пороговых значений (критерия эффективности). Таким образом, настраивая работу классификатора под личные предпочтения, нужды, ожидания пользователя путем выбора критерия эффективности по точности и полноте обнаружения спама и достоверности идентификации легальных электронных почтовых сообщений. Т. е. он сам для себя задает, что для него значит «лучшая эффективность», корректируя пороговые значения (критерий), пока он не даст «лучший для пользователя результат». Данный подход позволяет сделать обнаружение спама персонализированным.

Необходимо отметить, что при решении задач классификации обычной является ситуация, когда имеется некий обучающий набор текстов, класс которых известен априорно. В таком случае на ней и производят обучение классификатора. При этом необходимо получить некоторые оценки качества классификации, которые можно будет использовать для сравнения различных правил классификации и оптимизации их параметров. Необходимо отметить, что обучающую выборку недопустимо использовать одновременно для обучения и для оценки эффективности классификатора. В противном случае есть высокая

вероятность получения недостоверных (завышенных) оценок качества классификации.

Обычно для оценки качества выборку классифицированных текстов разбивают на две: обучающую и тестовую. Классификатор обучают на первой, а затем применяют ко второй, на которой вычисляют показатели эффективности классификации. Очевидно, что эффективность классификации будет зависеть от качества разбиения исходной выборки. При этом необходимо отметить следующее [147]:

- чем больше по размеру обучающая выборка писем, тем лучше можно обучить классификатор, в то время как при ее малом размере оценки эффективности качества могут быть слишком грубыми;

- специально подобранное разбиение классифицированных текстов писем может сильно повлиять на результаты и привести к повышению или, наоборот, понижению оценок эффективности.

Поэтому, как правило, разбиение на обучающую и тестовую выборки выполняют случайно либо по некоторому признаку, не зависящему от содержания документа. Для сравнения эффективности различных классификаторов разбиение фиксируют.

Кроме сравнения на фиксированном разбиении часто используется метод усреднения показателей эффективности по различным разбиениям, который называется методом перекрестной проверки или кросс-валидацией [195]. Как было отмечено выше, побочным эффектом уменьшения размерности пространства признаков является переобучение. Для того, чтобы избежать этого, применяют различные методы, к которым, в том числе относится и метод кросс-валидации [195]. Таким образом, метод кросс-валидации служит для оценки метода классификации и его поведения на независимых данных, а также позволяет избежать переобучения.

При оценке метода имеющаяся в наличии выборка классифицированных текстов разбивается на m частей (m – параметр кросс-валидации). Затем на $m-1$ частях писем (обучающий набор) производится обучение, а оставшаяся часть

используется для тестирования (тестовый набор). Данная процедура повторяется m раз, в результате чего каждая из m частей текстов будет использована для тестирования. Полученные в результате m процедур показатели эффективности классификации усредняются. В результате получаются значения показателей эффективности выбранной модели с наиболее равномерным использованием имеющихся текстов писем.

Очевидно, описанный подход, являющийся, по сути, стандартной методикой тестирования и сравнения правил классификации, позволит получить более точную оценку эффективности разработанного метода классификации электронных писем.

3.6 Алгоритм классификации электронных писем для обнаружения спама и идентификации легальных электронных писем

Учитывая изложенное в главах 1-3 настоящего диссертационного исследования, становится возможным разработать алгоритм классификации электронных писем, который в укрупненном виде представлен на рисунке 3.2 [141].

В 1-3 блоках происходит установка параметров модели электронных писем и метода классификации. На следующих шагах осуществляется загрузка в память термов из базы данных, предварительно выделенных из электронных писем обучающей выборки, а также письма, подлежащего классификации. На шагах 6 и 7 происходит выделение термов классифицируемого письма в соответствии с моделью (2.14), а на шаге 8 – его классификация в соответствии с разработанным методом.

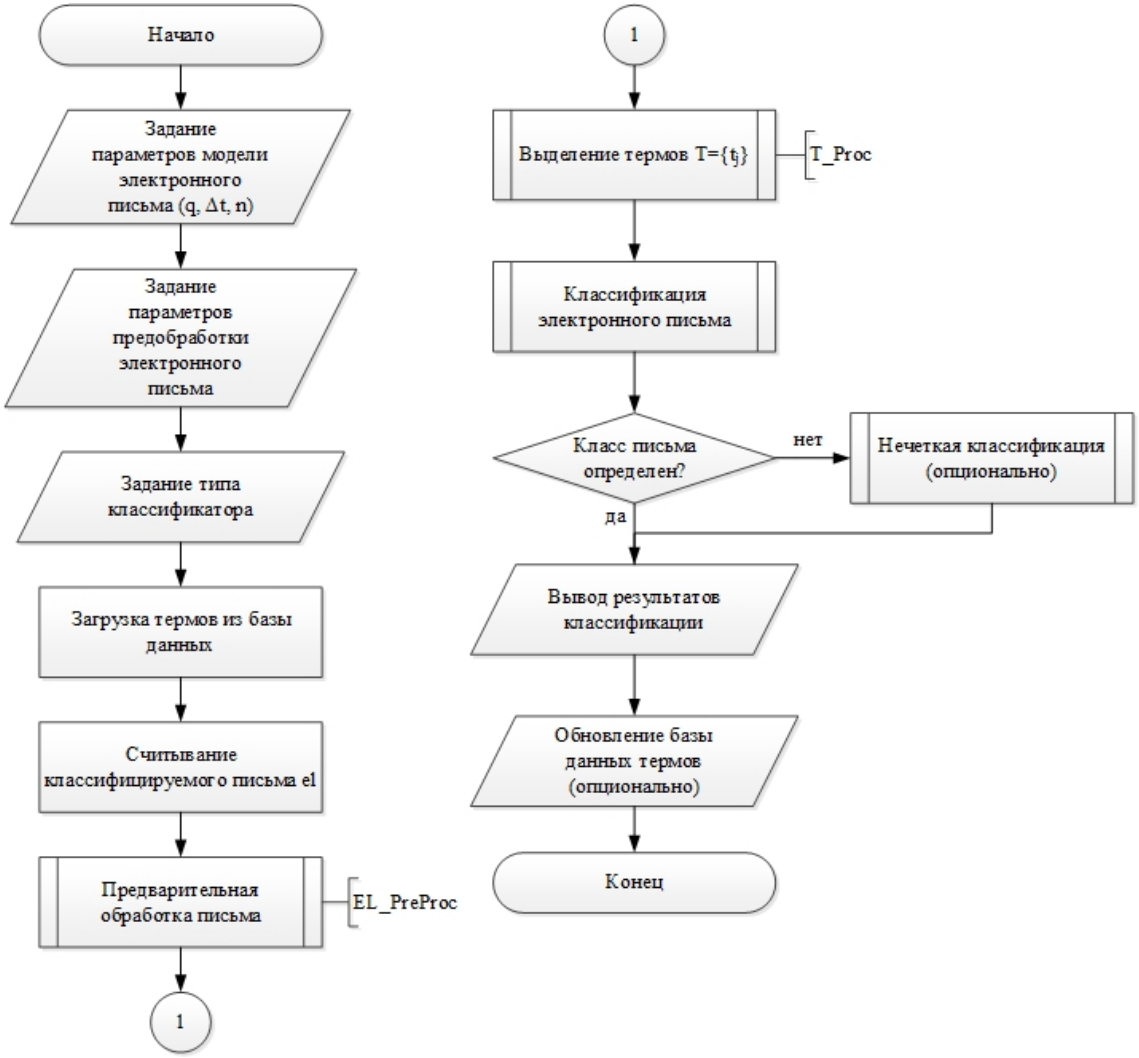

Рисунок 3.2 – Укрупненный алгоритм классификации электронных писем

В случае, если классификатор не позволил определить класс электронного письма, то для его определения может быть задействована процедура нечеткой классификации, использующая функцию расстояния Левенштейна. В ходе ее выполнения осуществляется сравнение термов классифицируемого письма с термами в базе данных. Из базы данных осуществляется выбор термов классов спама и легальных писем, наиболее близких к каждому классифицируемому терму. Далее по принципу их наибольшей близости к терму соответствующего класса принимается решение о принадлежности терма к классу. При этом если значение близости терма к термам спама и легальным термам равно, то ему присваивается неопределенный класс. В завершении осуществляется расчет общего количество термов каждого класса, по большему количеству которых принимается решение о

принадлежности письма к классу спама или легальных писем. При этом в случае равенства термов обоих классов письмо считается неклассифицированным.

Применение процедуры нечеткой классификации позволяет повысить эффективность обнаружения спама и достоверность идентификации легальных электронных писем, а также снизить количество неклассифицированных писем [141].

На завершающих шагах алгоритма осуществляется отображение результатов анализа и добавление термов анализируемого письма в базу данных (в случае необходимости).

Предложенный алгоритм является линейным, в нем отсутствуют критические участки, поэтому он выполняются за конечное время. Поскольку погрешность вводимых исходных данных на несколько порядков выше точности, обеспечиваемой применяемыми для расчетов типами данных, погрешностью вычислений в алгоритме можно пренебречь. Таким образом, алгоритм обнаружения спама является корректным.

Выводы по 3 главе

Основными результатами рассуждений, представленных в данной главе, являются:

1. Разработан метод классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, отличающийся использованием разработанной модели электронных писем.

2. Предложен подход к оценке эффективности (качества) метода классификации электронных писем для обнаружения спама и идентификации легальных электронных писем.

3. Разработан алгоритм классификации электронных писем, отличающийся наличием дополнительной процедуры определения «схожести» термов на основе расстояния Левенштейна, обеспечивающей вычисление мер принадлежности классифицируемого электронного письма к классам спама и легальных для повышения достоверности идентификации писем.

4. Разработанные метод и алгоритм позволяют решать следующие задачи: выбор способов предобработки, параметров модели и способа классификации; обучение на наборах писем; классификация писем; ведение базы данных термов; сохранение результатов в файл. Применение разработанного метода позволяет повысить эффективность обнаружения спама и достоверность идентификации легальных электронных писем, а также снизить количество неклассифицированных писем.

Предложенная последовательность анализируемых научных позиций стала обоснованием перехода к исследовательским материалам третьей главы, которая называется «Разработка архитектуры подсистемы классификации электронных писем для обнаружения спама» и посвящается:

1. Разработке архитектуры подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, реализующей метод и алгоритм, предложенные в настоящем диссертационном исследовании, применение которых позволяет повысить достоверность идентификации легальных электронных писем.

2. Разработке программных модулей исследовательского прототипа подсистемы классификации электронных писем.

3. Постановке эксперимента и экспериментальному исследованию разработанных модели и метода, а также оценки их эффективности для обнаружения спама и идентификации легальных электронных писем.

Глава 4 Разработка архитектуры подсистемы классификации электронных писем для обнаружения спама

4.1 Архитектура подсистемы классификации электронных писем

Под архитектурой (подсистемы[7]) понимают [196] ее основные понятия или свойства в окружающей среде, которые проявляются в ее элементах, отношениях между ними и конкретных принципах ее проектирования и жизненного цикла. Проецируя данное понимание архитектуры на предметную область настоящего диссертационного исследования, под архитектурой подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем можно понимать программную реализацию алгоритма, предложенного в главе 3, в виде взаимодействующих между собой программных модулей, обеспечивающих применение разработанных модели (2.14) и метода. Одновременно под модулем можно понимать [197] самостоятельный фрагмент кода, представляющий собой функционально завершенную часть программы с обеспечением возможности ее вызова из любого другого модуля.

При создании программных систем модульное программирование дает следующие преимущества [197], важные в нашем случае:

- упрощается их разработка и реализация;

- появляется возможность параллельной работы разработчиков различных подсистем в рамках одной системы;

- упрощается их настройка и модификация;

- можно создавать программные библиотеки;

- облегчается чтение и понимание программы;

- обеспечивается более полное тестирование;

- становится проще процедура загрузки в оперативную память большой программы.

[7] Здесь и далее применительно к настоящему диссертационному исследованию термины «система» и «подсистема» употребляются как синонимы.

В связи с изложенным, а также с учетом того, что, как было справедливо отмечено в главе 1 настоящего диссертационного исследования, обнаружение спама является неотъемлемой частью общей системы обеспечения безопасности информационных систем, представляется очевидным, что подсистему классификации электронных писем для обнаружения спама и идентификации легальных электронных писем целесообразно реализовывать в виде совокупности программных модулей, обеспечивающих выполнение алгоритма классификации электронных писем, представленного в третьей главе настоящего диссертационного исследования.

Программная реализация подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем была осуществлена с применением метода нисходящего проектирования [197]. Его применение подразумевает вначале построение модульной структуры программы с поочередным проектированием модулей, начиная с модуля самого верхнего уровня. После программной реализации всех модулей программы в том же нисходящем порядке поочередно осуществляется их тестирование и отладка. Описанный метод нисходящего проектирования также называют функциональной декомпозицией [197, 198].

На рисунке 4.1. представлена архитектура исследовательского прототипа подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем. Она включает в свой состав три взаимодействующих между собой основных модуля [199], реализующие разраотанные модель электронных писем (2.14) и метод классификации электронных писем, а также другие основные функции. Их вызов на исполнение осуществляется из основного модуля путем выбора определенного пункта меню.

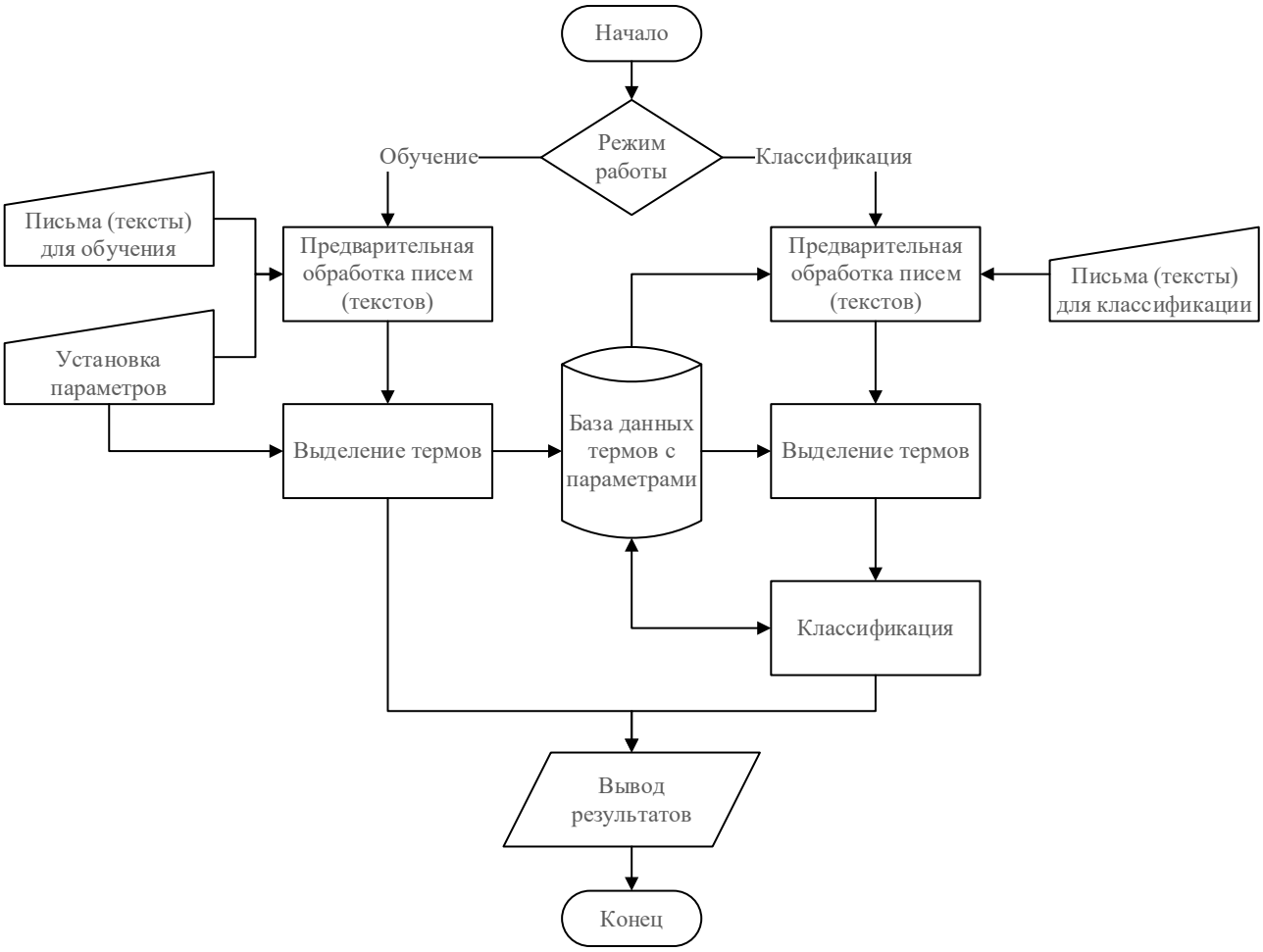

Рисунок 4.1 – Архитектура подсистемы классификации электронных писем

Предусмотрены два режима работы исследовательского прототипа подсистемы классификации электронных писем [199], входными данными для работы которого выступают тексты электронных почтовых сообщений:

1. Режим обучения. Применяется для построения признакового пространства (выделения термов) в виде базы данных термов на основе заранее известных писем спама и легальных писем. Позволяет пользователю выбрать и подготовить письма для обучения. В данном режиме задействуются только модули предварительной обработки электронных писем и выделения термов, а также база данных.

2. Режим классификации. Предназначен для определения классов неизвестных писем, для чего используются термы спама и легальных писем из базы данных, выделенных в режиме обучения. Позволяет пользователю выбрать базу данных известных термов, выбрать и подготовить письма для классификации. В данном режиме задействуются все разработанные модули исследовательского

прототипа подсистемы классификации электронных писем, представленные на рисунке 4.1.

Практическое внедрение предложенной архитектуры подсистемы может быть реализовано:

- путем интеграции в существующее продуктовое антиспам-решение с позиции его разработчика в виде отдельного, задействуемого для электронного почтового ящика конкретного пользователя модуля;

- в виде дополнения для существующих продуктовых антиспам-решений, поддерживающих подключение дополнений сторонних разработчиков (например, SpamAssasin);

- в виде дополнения для почтовых клиентов.

При этом предложенная подсистема может стать дополнительным рубежом для обнаружения спама с дополнительным набором семантических признаков. Ее использование для почтового сервиса организаций может позволить пользователю простыми манипуляциями именно со своим почтовым ящиком:

- создавать персональную базу данных термов спамовых и легальных писем;

- стать дополнительным рубежом классификации писем, не классифицированных существующими решениями, в том числе ориентированных на персональные (пользовательские) особенности (с т. з. информационных потребностей) применительно к конкретным пользователям.

4.2 Описание исследовательского прототипа подсистемы классификации электронных писем

Обработка поданных на вход исследовательского прототипа подсистемы классификации электронных писем начинается с их предобработки. Реализующий ее модуль состоит из последовательности опциональных процедур, применяемых к исходному тексту электронного письма. Если в настройках программы какая-либо предобработка выключена, то текст письма данной процедуре не подвергается, а передается дальше. На рисунке 4.2 представлена общая схема

модуля предварительной обработки, а в таблице 4.1 – перечень и описание его состав процедур.

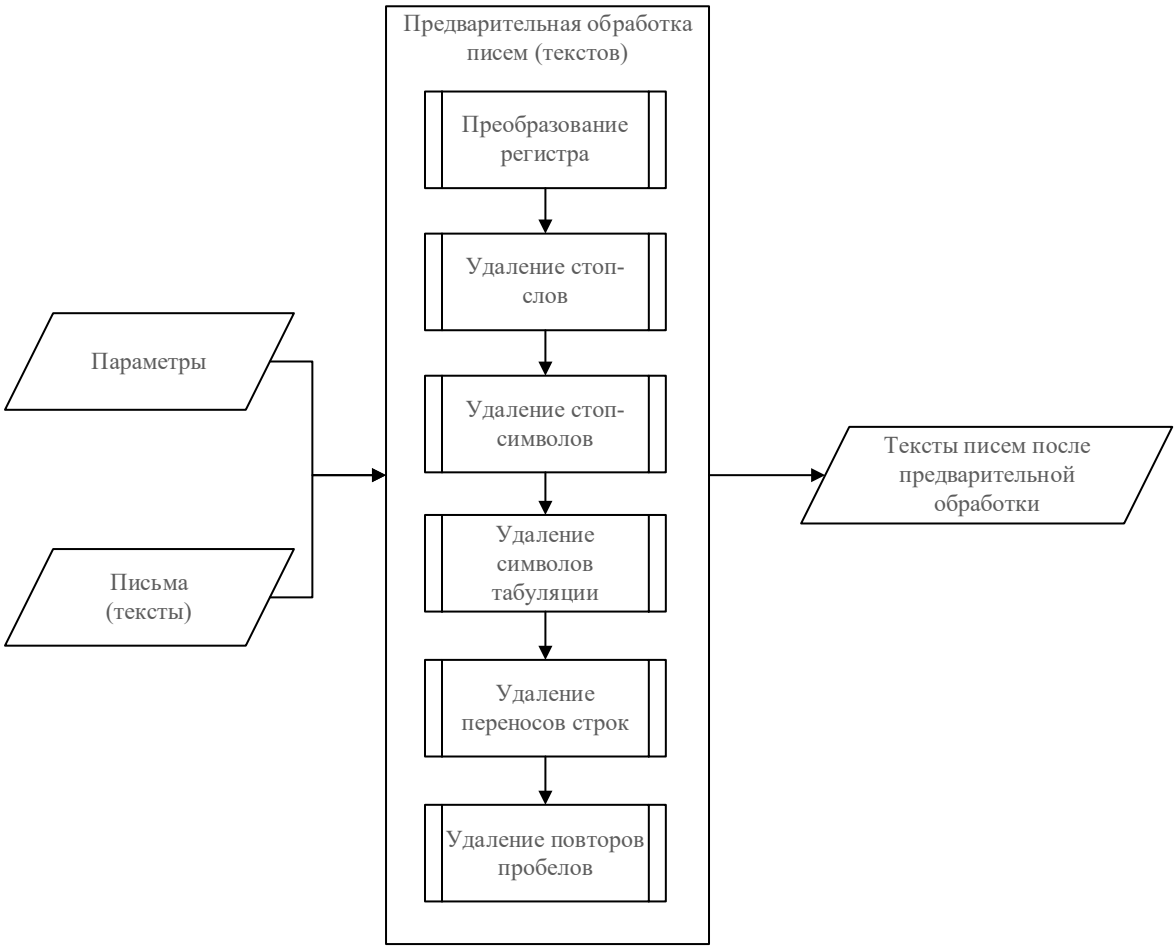

Рисунок 4.2 – Общая схема модуля предварительной обработки электронных писем

Таблица 4.1 – Спецификация процедур модуля предварительной обработки электронных писем

Условное название и назначение	Режимы
Преобразователь регистра (перевод текста в верхний регистр)	Вкл./Откл.
Удаление стоп-слов. Хранятся в заданном файле ProgramCatalog/stoplist/words.txt	Оставить/Удалить
Удаление стоп-символов. Хранятся в заданном файле ProgramCatalog/stoplist/syms.txt	Оставить/Удалить
Удаление символов табуляции. Например, “/t/t/t/t/t” преобразуется в “/t/” в режиме «Удалить повторения» или производится удаление всех символов “/t” табуляции в режиме «Удалить полностью»	Оставить/Удалить повторения/Удалить полностью

Условное название и назначение	Режимы
Удаление переносов строк. Например, текст: «abc bcde» преобразуется в «abc bcde» в режиме «Удалить повторения» или производится замена символов переноса строк на один символ пробела в режиме «Удалить полностью»	Оставить/Удалить повторения/Удалить полностью
Удаление повторов пробелов. Наример, текст «a b» преобразуется в «a b» в режиме «Удалить повторения» или производится удаление всех символов пробелов в режиме «Удалить полностью»	Оставить/Удалить повторения/Удалить полностью

Результатом работы модуля являются предварительно обработанные тексты электронных писем.

После выполнения процедур модуля предварительной обработки электронных писем осуществляется выделение из них термов в соответствии с моделью (2.14), реализованное также в виде модуля. Он состоит из трех процедур: непосредственно процедура выделения термов, процедура отбора «популярных» термов (опциональная) и процедуры снижения размерности признакового пространства (множества термов, подлежащих дальнейшему использованию в методе классификации и хранении в базе данных, опциональная). Эти процедуры применяются к предварительно обработанному тексту электронного письма. Общая схема модуля выделения термов представлена на рисунке 4.3, а спецификация входящих в его состав процедур в таблице 4.2.

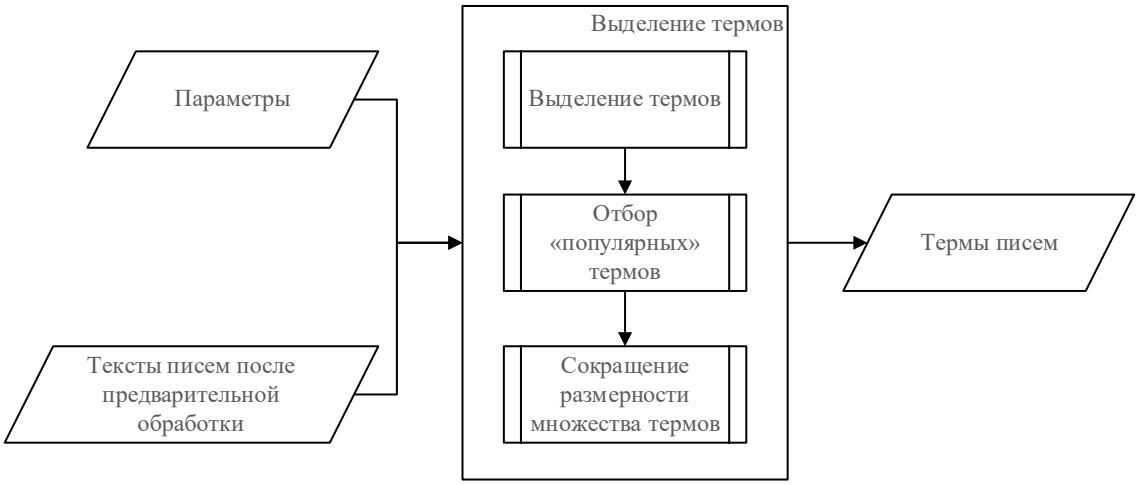

Рисунок 4.3 – Общая схема модуля выделения термов

Таблица 4.2 – Спецификация процедур модуля выделения термов

Условное название	Назначение
Выделение термов	Производится выделение термов в соответствии с моделью (2.14), а также производятся дополнительные «подготовительные» расчеты для последующего использования полученных результатов в процедуре расчета весов термов модуля классификации с целью определения значимости термов.
Отбор «популярных» термов	Опциональная процедура. Реализована через задание минимального процентного порога частотности, при котором термы со значением ниже его не включаются в базу данных.
Сокращение размерности множества термов	Опциональная процедура. Сокращение размерности базы данных возможно с использованием описанных в главе 3 настоящего диссертационного исследования: - прироста информации [174, 183]; - индекса Джини [192]; - взаимной информативности признаков [174, 177, 183]; - критерия χ^2 [178, 184, 188, 191].

Результатом работы модуля является описание термов писем в виде заданной структуры.

Процедуры модуля предварительной обработки, а также процедура выделения термов соответствующего модуля являются общими (одинаковыми) для режимов обучения и классификации.

Метод классификации реализован в виде модулей: выделения термов и классификации. Модуль классификации включает четыре процедуры:

- расчет весов термов;

- определение класса письма с использованием простейшего решающего правила по принципу простого большинства термов (весов термов) соответствующего класса;

- определение класса письма с использованием косинусной меры;

- определение класса письма с применением нечеткой классификации (опционально).

В работе перечисленных процедур используются термы из базы данных и термы классифицируемого электронного письма.

На рисунке 4.4 представлена общая схема модуля классификации.

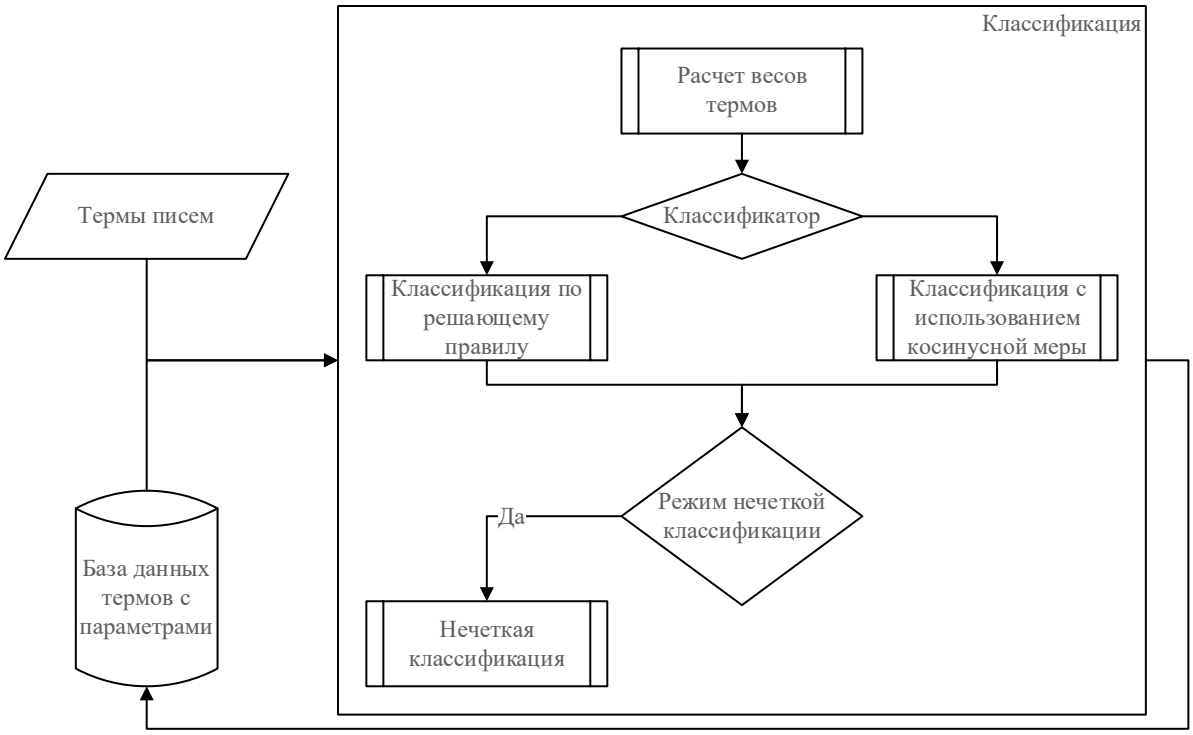

Рисунок 4.4 – Общая схема модуля классификации

Для дополнительной обработки неклассифицированных с применением простейшего решающего правила или с использованием косинусной меры писем может быть задействована процедура нечеткой классификации (в случае ее активации), включающая следующие этапы:

1. Осуществляется сравнение термов классифицируемого письма с термами в базе данных.

2. Из базы данных осуществляется выбор термов классов спама и легальных писем, наиболее близких к каждому классифицируемому терму.

3. Принимается решение о принадлежности терма к классу по принципу их наибольшей близости к терму соответствующего класса. Вместе с тем классифицируемый терм будет считаться неклассифицированным, если их значения для классов спама и легальных писем равны. Для определения близости в исследовательском прототипе подсистемы классификации электронных писем реализовано расстояние Левенштейна [200].

4. Подсчитываются суммарные значения количества термов каждого класса. Осуществляется присвоение класса письму по принципу большинства термов

соответствующего класса. При равном количестве термов обоих классов письмо считается неклассифицированным.

По результатам работы модуля классифицируемое письмо относится к классу спама или легального. При этом возможна ситуация, при которой письма может быть не классифицировано.

Таким образом, предложен исследовательский прототип подсистемы классификации электронных писем со следующими основными функциональными возможностями [199]: выбор способов предобработки, параметров модели и способа классификации; обучение на наборах писем; классификация писем; ведение базы данных термов; сохранение результатов в файл.

4.3 План проведения экспериментальных исследований

Предмет исследования позволяет провести эксперимент на реальных письмах спама и легальных электронных письмах. Для достижения целей настоящего диссертационного исследования для организации и проведения экспериментальных исследований среди их этапов, приводимых в литературе [например, 198, 201, 202], автор считает возможным ограничиться следующими:

- составление плана эксперимента с сокращением числа рассматриваемых факторов (внешних воздействий, условий) с целью уменьшения объема проводимого эксперимента;

- проведение и контроль хода эксперимента;

- анализ и интерпретация полученных результатов.

Исходя из цели настоящего диссертационного исследования целесообразно проводить активные экспериментальные исследования, подразумевающие их планирование [202]: выбор числа опытов и условий их проведения, необходимых и достаточных для достижения поставленных цели и задач. При этом в ходе эксперимента должен быть осуществлен поиск такого сочетания условий, при которых будет достигнуты наилучшие значения обобщенного показателя эффективности метода классификации электронных писем. Также в ходе

экспериментальных исследований должна быть подтверждена корректность и применимость разработанной модели (2.14).

При проведении экспериментальных исследований использован метод кросс-валидации [195], описанный в главе 3 настоящего диссертационного исследования.

Следует отметить, что такой подход, а также случайное изменение значений параметров (тем более их исчерпывающий набор сочетаний) модели (2.14) и метода требует больших временных затрат при проведении экспериментальных исследований. В связи с этим, с учетом изложенного в главах 2 и 3 настоящего диссертационного исследования и для достижения цели настоящей диссертационной работы автором проведение экспериментальных исследований запланировано и осуществлено по следующей схеме в два этапа с сокращением числа рассматриваемых значений параметров (уменьшения объема проводимого эксперимента) модели (2.14) (определены в 12, 118-120, 122 и описаны в главе 2) и метода.

Из групп писем набора Enron, описанных в главе 2, сформированы четыре поднабора, состоящие из 13 195 легальных писем и 13 577 писем спама, группы и количественный состав которых приведен в таблице 4.3. Необходимо отметить, что до настоящего времени набор Enron не утратил своей актуальности и продолжает оставаться одним из самых востребованных и наиболее распространенным в исследованиях в области обнаружения спама. На нем продолжают проводить исследования и эксперименты разные исследователи и исследовательские группы.

Таблица 4.3 – Набор электронных писем для проведения эксперимента[8]

	Поднабор 14		Поднабор 25		Поднабор 36		Поднабор 53	
	legal1	spam4	legal2	spam5	legal3	spam6	legal5	spam3
Кол-во писем	3 618	4 237	4 189	3 551	3 980	4 337	1 408	1 452

Необходимо отметить, что все поднаборы содержат несбалансированные классы электронных писем (с различным количеством писем и содержащихся в них общего количества термов разных классов), что соответствует реальной практике решения задач классификации.

[8] legal – легальные письма, spam – спам.

Письма в каждом из поднаборов спама и легальных писем были предварительно случайным образом пронумерованы и последовательно разбиты на 10 условно[9] равных частей, также пронумерованных от 1 до 10. На каждом поднаборе (14, 25, 36, 53) было проведено 10 последовательных испытаний, которым присвоен также номер от 1 до 10. При этом в соответствии с методом кросс-валидации одна из 10 частей (одинаковый номер для спама и легальных частей, соответствующий номеру эксперименту) принималась за тестовую выборку, остальные 9 – за обучающую.

На первом этапе экспериментальных исследований проведены испытания на поднаборах 14, 25, 36 с использованием классификаторов на основе решающего правила и косинусной меры с различными весами термов. Это позволило выбрать веса (для разного типа классификаторов), показавшие наилучшие результаты обнаружения.

На втором этапе экспериментальных исследований проведены испытания на поднаборе 53, состоящем из примерно одинакового количества термов спама и легальных термов, с использованием выбранных на первом этапе весов и с использованием снижения размера признакового пространства на основе индекса Джини. Оно осуществлялось путем уменьшения количества уникальных термов по задаваемому порогу в виде значения индекса Джини. В результате произведено уменьшение количество уникальных термов примерно до 20% от исходного количества (в среднем с 135 415 до 23 423 при значении индекса Джини равным 0,000003). Пороговое значение выбрано по правилу Парето.

Для составления и унификации описания схемы классификации применим кодификацию вариантов экспериментов, для чего введем следующие коды:

- поднабор писем: 14 – поднабор 14, 25 – поднабор 25, 36 – поднабор 36, 53 – поднабор 53;

[9] По причине того, что общее количество писем не кратно 10, в нескольких случайно выбранных частях количество писем составило на одно меньше, чем в остальных.

- классификатор: 11 – решающее правило, 12 – решающее правило с последующей нечеткой классификацией, 21 – косинусная мера, 22 – косинусная мера с последующей нечеткой классификацией;

- веса: 01 – единичный, 02 – веса (3.11, 3.13, 3.15, 3.18, 3.19, 3.21).

Тогда, принимая в качестве кода эксперимента значение вида «этап.поднабор писем.классификатор.вес», схему проведения экспериментальных исследований можно представить в виде последовательности испытаний, представленных в таблице 4.4. В результате каждого запуска (испытания) исследовательского прототипа подсистемы классификации электронных писем с целью оценки обобщенного показателя эффективности предложенного метода, обеспечивающего точность и полноту обнаружения спама и достоверности идентификации легальных электронных почтовых сообщений, фиксировались значения полноты, точности и F-меры.

Таблица 4.4 – Схема проведения экспериментальных исследований

№ испытания	Код	№ испытания	Код	№ испытания	Код
1 этап					
1	14.11.01	9	25.11.01	17	36.11.01
2	14.11.02	10	25.11.02	18	36.11.02
3	14.12.01	11	25.12.01	19	36.12.01
4	14.12.02	12	25.12.02	20	36.12.02
5	14.21.01	13	25.21.01	21	36.21.01
6	14.21.02	14	25.21.02	22	36.21.02
7	14.22.01	15	25.22.01	23	36.22.01
8	14.22.02	16	25.22.02	24	36.22.02
2 этап					
25	53.12.01	27	53.22.01		
26	53.12.02	28	53.22.02		

Обобщенная схема отдельно взятого испытания представлена на рисунке 4.5.

Таким образом, предложенная схема проведения экспериментальных исследований позволяет в целом оценить эффективность разработанного метода с применением предложенной модели (2.14) в различных условиях.

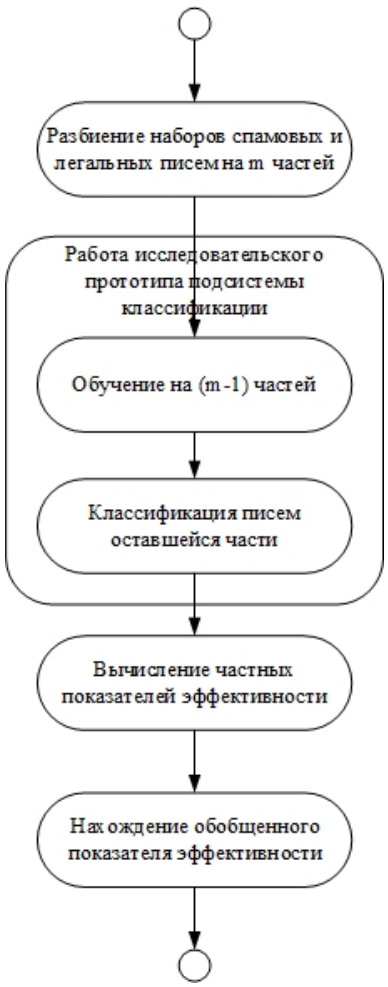

Рисунок 4.5 – Общая схема отдельного испытания

4.4 Экспериментальные исследования

4.4.1 Результаты экспериментальных исследований

В таблицах 4.5-4.16 представлены частные показатели эффективности классификации, полученные по результатам первого этапа экспериментальных исследований [169].

Таблица 4.5 – Результаты испытаний №№ 1, 2

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n = 1$	$n = 2$	$n = 1$	$n = 2$	$n = 1$	$n = 2$
Легальные письма						
1 (ед.)	99,50	97,46	94,58	88,31	**96,98**	92,65
2 (3.11)	89,63	90,01	68,74	68,19	77,79	77,56
2 (3.13)	89,81	90,12	69,04	68,49	78,05	77,80
2 (3.15)	94,25	88,30	91,79	88,89	93,00	88,59
2 (3.18)	96,83	95,64	94,39	89,06	95,59	92,22
2 (3.19)	96,83	95,70	94,39	89,06	95,59	92,25
2 (3.21)	98,12	93,97	94,80	89,14	***96,43***	91,49

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Спам						
1 (ед.)	96,93	97,86	97,90	82,89	**97,41**	89,74
2 (3.11)	77,75	78,77	92,97	83,88	84,68	81,23
2 (3.13)	77,94	78,97	93,06	83,93	84,83	81,36
2 (3.15)	93,20	94,18	94,97	80,29	94,07	86,67
2 (3.18)	95,37	96,58	96,77	83,86	96,05	89,76
2 (3.19)	95,37	96,58	96,77	83,86	96,06	89,76
2 (3.21)	95,75	94,73	98,21	85,46	***96,96***	89,85
Y						
1 (ед.)					**97,19**	91,18
2 (3.11)					81,09	79,35
2 (3.13)					81,30	79,54
2 (3.15)					93,53	87,62
2 (3.18)					95,82	90,98
2 (3.19)					95,82	90,99
2 (3.21)					***96,69***	90,66

Таблица 4.6 – Результаты испытаний №№ 3, 4

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
3 (ед.)	99,39	97,05	95,19	96,10	**97,24**	96,57
4 (3.11)	89,64	90,06	68,79	72,69	77,83	80,42
4 (3.13)	89,82	90,15	69,10	73,00	78,09	80,65
4 (3.15)	94,25	88,41	91,85	93,39	93,03	90,83
4 (3.18)	96,81	95,34	94,44	95,19	95,60	95,26
4 (3.19)	96,81	95,39	94,44	95,19	95,60	95,29
4 (3.21)	98,12	93,81	94,86	93,64	***96,46***	93,72
Спам						
3 (ед.)	96,51	97,45	98,94	96,51	**97,71**	96,98
4 (3.11)	77,79	80,13	93,18	92,42	84,79	85,83
4 (3.13)	77,98	80,32	93,27	92,47	84,94	85,96
4 (3.15)	93,21	94,40	95,19	88,84	94,18	91,53
4 (3.18)	95,38	96,53	97,07	95,14	96,21	95,83
4 (3.19)	95,38	96,53	97,07	95,19	96,21	95,85
4 (3.21)	95,76	94,89	98,42	94,01	***97,07***	94,44
Y						
3 (ед.)					**97,47**	96,77
4 (3.11)					81,16	83,04
4 (3.13)					81,37	83,22
4 (3.15)					93,60	91,18
4 (3.18)					95,91	95,54
4 (3.19)					95,91	95,57
4 (3.21)					***96,76***	94,08

Таблица 4.7 – Результаты испытаний №№ 5, 6

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
5 (ед.)	48,81	87,77	99,92	93,62	65,58	**90,59**
6 (3.11)	71,13	87,54	97,84	68,21	82,36	76,65
6 (3.13)	71,32	87,66	97,82	70,15	82,49	77,90
6 (3.15)	77,95	80,18	72,58	83,80	75,13	81,93
6 (3.18)	71,13	87,54	97,84	68,21	82,36	76,65
6 (3.19)	71,32	87,66	97,82	70,15	82,49	77,90
6 (3.21)	48,81	87,77	99,92	93,62	65,58	**90,59**
Спам						
5 (ед.)	1,00	98,82	10,27	79,18	18,59	***87,79***
6 (3.11)	97,39	78,40	65,80	82,04	78,52	80,17
6 (3.13)	97,36	79,64	66,13	81,90	78,74	80,74
6 (3.15)	77,89	88,69	82,16	72,65	79,95	79,84
6 (3.18)	97,39	78,40	65,80	82,04	78,52	80,17
6 (3.19)	97,36	79,64	66,13	81,90	78,74	80,74
6 (3.21)	1,00	98,82	10,27	79,18	18,59	**87,91**
Y						
5 (ед.)					28,97	**89,23**
6 (3.11)					80,39	78,37
6 (3.13)					80,57	79,29
6 (3.15)					77,47	80,87
6 (3.18)					80,39	78,37
6 (3.19)					80,57	79,29
6 (3.21)					28,97	**89,23**

Таблица 4.8 – Результаты испытаний №№ 7, 8

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
7 (ед.)	48,82	87,90	99,97	98,12	65,60	**92,72**
8 (3.11)	71,14	87,73	97,90	72,72	82,39	79,50
8 (3.13)	71,33	87,84	97,87	74,65	82,51	80,69
8 (3.15)	77,97	80,65	72,64	88,31	75,17	84,29
8 (3.18)	71,14	87,73	97,90	72,72	82,39	79,50
8 (3.19)	71,33	87,84	97,87	74,65	82,51	80,69
8 (3.21)	48,82	87,90	99,97	98,12	65,60	**92,72**
Спам						
7 (ед.)	1,00	98,59	10,48	87,73	18,94	**92,83**
8 (3.11)	97,40	79,81	66,01	90,58	78,67	84,85
8 (3.13)	97,37	80,98	66,35	90,44	78,89	85,43
8 (3.15)	77,94	89,46	82,37	81,19	80,07	85,10
8 (3.18)	97,40	79,81	66,01	90,58	78,67	84,85
8 (3.19)	97,37	80,98	66,35	90,44	78,89	85,43
8 (3.21)	1,00	98,59	10,48	87,73	18,94	**92,83**

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Y						
7 (ед.)					29,39	**92,78**
8 (3.11)					80,49	82,09
8 (3.13)					80,66	82,99
8 (3.15)					77,54	84,69
8 (3.18)					80,49	82,09
8 (3.19)					80,66	82,99
8 (3.21)					29,39	**92,78**

Таблица 4.9 – Результаты испытаний №№ 9, 10

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
9 (ед.)	97,76	97,86	98,95	92,98	**98,35**	95,35
10 (3.11)	82,65	86,19	72,81	74,34	77,39	79,81
10 (3.13)	82,75	86,26	73,00	74,77	77,54	80,09
10 (3.15)	86,19	91,83	97,04	90,62	91,28	91,22
10 (3.18)	94,96	96,79	96,85	93,67	95,89	95,21
10 (3.19)	95,00	96,82	96,85	93,67	95,91	95,22
10 (3.21)	96,63	94,86	99,12	95,20	***97,86***	95,03
Спам						
9 (ед.)	99,56	98,72	95,33	89,19	**97,39**	93,71
10 (3.11)	71,97	74,89	81,92	82,26	76,60	78,39
10 (3.13)	72,14	75,22	82,00	82,26	76,73	78,57
10 (3.15)	95,96	91,13	81,55	86,79	88,15	88,91
10 (3.18)	96,36	96,85	93,78	90,09	95,04	93,34
10 (3.19)	96,36	96,82	93,83	90,09	95,07	93,33
10 (3.21)	98,98	96,75	95,86	90,23	**97,39**	93,37
Y						
9 (ед.)					**97,87**	94,52
10 (3.11)					76,99	79,09
10 (3.13)					77,14	79,32
10 (3.15)					89,69	90,05
10 (3.18)					95,46	94,27
10 (3.19)					95,49	94,27
10 (3.21)					***97,62***	94,19

Таблица 4.10 – Результаты испытаний №№ 11, 12

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
11 (ед.)	97,40	96,18	99,26	98,54	**98,32**	97,35
12 (3.11)	82,65	85,57	72,81	76,22	77,39	80,61
12 (3.13)	82,75	85,65	73,00	76,65	77,54	80,89
12 (3.15)	86,19	91,14	97,04	92,50	91,28	91,81

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
12 (3.18)	94,96	95,47	96,90	97,09	95,91	96,27
12 (3.19)	95,00	95,49	96,90	97,06	95,94	96,27
12 (3.21)	96,63	94,10	99,12	97,09	***97,86***	95,57
Спам						
11 (ед.)	99,45	98,64	96,31	94,34	**97,85**	96,44
12 (3.11)	71,97	75,29	81,92	84,31	76,60	79,53
12 (3.13)	72,14	75,62	82,00	84,31	76,73	79,72
12 (3.15)	95,96	91,24	81,55	88,85	88,15	90,03
12 (3.18)	96,36	96,85	93,78	93,69	95,04	95,24
12 (3.19)	96,36	96,82	93,83	93,69	95,07	95,23
12 (3.21)	98,98	96,73	95,86	92,28	***97,39***	94,45
Y						
11 (ед.)					**98,09**	96,89
12 (3.11)					76,99	80,07
12 (3.13)					77,14	80,30
12 (3.15)					89,69	90,91
12 (3.18)					95,48	95,76
12 (3.19)					95,50	95,75
12 (3.21)					***97,62***	95,01

Таблица 4.11 – Результаты испытаний №№ 13, 14

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
13 (ед.)	99,81	98,82	77,99	90,05	87,54	**94,23**
14 (3.11)	83,30	88,11	82,64	66,72	82,95	75,93
14 (3.13)	82,37	87,52	86,77	67,75	84,51	76,36
14 (3.15)	66,69	81,46	96,90	91,65	79,00	86,24
14 (3.18)	83,30	88,11	82,64	66,72	82,95	75,93
14 (3.19)	82,37	87,52	86,77	67,75	84,51	76,36
14 (3.21)	99,81	98,82	77,99	90,05	87,54	**94,23**
Спам						
13 (ед.)	79,43	91,26	99,77	95,04	88,44	**93,11**
14 (3.11)	79,78	70,08	80,34	85,69	80,03	77,09
14 (3.13)	83,40	70,57	78,01	84,88	80,60	77,05
14 (3.15)	92,30	90,83	42,83	71,67	58,47	80,10
14 (3.18)	79,78	70,08	80,34	85,69	80,03	77,09
14 (3.19)	83,40	70,57	78,01	84,88	80,60	77,05
14 (3.21)	79,43	91,26	99,77	95,04	88,44	**93,11**
Y						
13 (ед.)					87,99	**93,66**
14 (3.11)					81,46	76,51
14 (3.13)					82,51	76,70
14 (3.15)					67,21	83,06
14 (3.18)					81,46	76,51
14 (3.19)					82,51	76,70
14 (3.21)					87,99	**93,66**

Таблица 4.12 – Результаты испытаний №№ 15, 16

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
15 (ед.)	99,81	97,87	77,99	91,93	87,54	**94,80**
16 (3.11)	83,30	87,34	82,64	68,61	82,95	76,84
16 (3.13)	82,37	86,79	86,77	69,63	84,51	77,25
16 (3.15)	66,69	81,10	96,90	93,53	79,00	86,87
16 (3.18)	83,30	87,34	82,64	68,61	82,95	76,84
16 (3.19)	82,37	86,79	86,77	69,63	84,51	77,25
16 (3.21)	99,81	97,87	77,99	91,93	87,54	**94,80**
Спам						
15 (ед.)	79,43	91,36	99,77	97,10	88,44	**94,14**
16 (3.11)	79,78	70,52	80,34	87,75	80,03	78,19
16 (3.13)	83,40	71,01	78,01	86,93	80,60	78,16
16 (3.15)	92,30	90,97	42,83	73,72	58,47	81,43
16 (3.18)	79,78	70,52	80,34	87,75	80,03	78,19
16 (3.19)	83,40	71,01	78,01	86,93	80,60	78,16
16 (3.21)	79,43	91,36	99,77	97,10	88,44	**94,14**
Y						
15 (ед.)					87,99	**94,47**
16 (3.11)					81,46	77,51
16 (3.13)					82,51	77,70
16 (3.15)					67,21	84,06
16 (3.18)					81,46	77,51
16 (3.19)					82,51	77,70
16 (3.21)					87,99	**94,47**

Таблица 4.13 – Результаты испытаний №№ 17, 18

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
17 (ед.)	92,20	95,33	98,72	94,10	**95,34**	94,71
18 (3.11)	90,91	91,48	65,40	84,37	75,96	87,77
18 (3.13)	90,94	91,49	65,44	84,50	76,00	87,85
18 (3.15)	64,40	83,66	99,05	95,03	78,05	88,97
18 (3.18)	90,84	93,80	96,48	93,82	93,56	93,80
18 (3.19)	90,89	93,73	96,53	93,84	93,61	93,78
18 (3.21)	89,34	89,41	98,82	95,88	***93,83***	92,53
Спам						
17 (ед.)	99,66	99,46	88,38	80,82	**93,66**	89,15
18 (3.11)	74,83	87,69	93,84	84,64	83,23	86,13
18 (3.13)	74,85	87,80	93,87	84,64	83,25	86,18
18 (3.15)	98,36	97,27	49,64	74,82	65,96	84,57
18 (3.18)	96,65	97,93	90,71	82,52	93,57	89,54
18 (3.19)	96,70	97,95	90,75	82,45	93,61	89,51
18 (3.21)	98,85	98,41	89,02	81,44	**93,66**	89,11

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n = 1$	$n = 2$	$n = 1$	$n = 2$	$n = 1$	$n = 2$
Y						
17 (ед.)					**94,49**	91,85
18 (3.11)					79,43	86,94
18 (3.13)					79,46	87,01
18 (3.15)					71,50	86,71
18 (3.18)					93,56	91,62
18 (3.19)					93,61	91,60
18 (3.21)					***93,74***	90,78

Таблица 4.14 – Результаты испытаний №№ 19, 20

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n = 1$	$n = 2$	$n = 1$	$n = 2$	$n = 1$	$n = 2$
Легальные письма						
19 (ед.)	91,12	89,25	99,10	99,02	**94,93**	93,87
20 (3.11)	90,91	87,95	65,40	86,76	75,96	87,34
20 (3.13)	90,94	87,96	65,43	86,88	75,99	87,41
20 (3.15)	64,40	81,17	99,05	97,41	78,05	88,54
20 (3.18)	90,82	89,05	96,51	97,69	93,56	93,16
20 (3.19)	90,87	89,01	96,56	97,71	93,61	93,15
20 (3.21)	89,34	86,46	98,82	98,27	***93,83***	91,98
Спам						
19 (ед.)	99,67	99,32	90,32	87,16	**94,75**	92,83
20 (3.11)	74,84	88,00	93,87	87,76	83,24	87,87
20 (3.13)	74,86	88,11	93,89	87,76	83,27	87,92
20 (3.15)	98,36	97,27	49,67	77,93	65,98	86,52
20 (3.18)	96,65	97,89	90,92	87,25	93,68	92,25
20 (3.19)	96,71	97,91	90,96	87,18	***93,73***	92,22
20 (3.21)	98,85	98,36	89,05	84,55	93,67	90,93
Y						
19 (ед.)					**94,84**	93,35
20 (3.11)					79,44	87,60
20 (3.13)					79,46	87,66
20 (3.15)					71,51	87,52
20 (3.18)					93,62	92,70
20 (3.19)					93,67	93,68
20 (3.21)					***93,75***	91,45

Таблица 4.15 – Результаты испытаний №№ 21, 22

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n = 1$	$n = 2$	$n = 1$	$n = 2$	$n = 1$	$n = 2$
Легальные письма						
21 (ед.)	99,58	98,01	62,34	77,51	76,60	**86,56**
22 (3.11)	81,62	88,39	60,55	60,50	69,49	71,80
22 (3.13)	79,97	88,77	67,31	62,06	73,07	73,02
22 (3.15)	80,17	83,20	72,46	86,48	76,09	84,80

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
22 (3.18)	81,62	88,39	60,55	60,50	69,49	71,80
22 (3.19)	79,97	88,77	67,31	62,06	73,07	73,02
22 (3.21)	99,58	98,01	62,34	77,51	76,60	**86,56**
Спам						
21 (ед.)	74,34	83,27	99,68	90,43	85,15	**86,70**
22 (3.11)	70,75	71,47	87,34	84,55	78,16	77,45
22 (3.13)	73,84	72,37	84,41	84,64	78,76	78,01
22 (3.15)	76,81	88,43	83,42	75,84	79,96	81,63
22 (3.18)	70,75	71,47	87,34	84,55	78,16	77,45
22 (3.19)	73,84	72,37	84,41	84,64	78,76	78,01
22 (3.21)	74,34	83,27	99,68	90,43	85,15	**86,70**
Y						
21 (ед.)					80,65	**86,63**
22 (3.11)					73,57	74,52
22 (3.13)					75,81	75,43
22 (3.15)					77,98	83,19
22 (3.18)					73,57	74,52
22 (3.19)					75,81	75,43
22 (3.21)					80,65	**86,63**

Таблица 4.16 – Результаты испытаний №№ 23, 24

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
23 (ед.)	99,58	93,43	62,34	79,90	76,60	**86,13**
24 (3.11)	81,62	83,99	60,55	62,89	69,49	71,89
24 (3.13)	79,97	84,43	67,31	64,45	73,07	73,06
24 (3.15)	80,17	80,51	72,46	88,87	76,09	84,48
24 (3.18)	81,62	83,99	60,55	62,89	69,49	71,89
24 (3.19)	79,97	84,43	67,31	64,45	73,07	73,06
24 (3.21)	99,58	93,43	62,34	79,90	76,60	**86,13**
Спам						
23 (ед.)	74,34	83,67	99,70	93,54	85,16	**88,33**
24 (3.11)	70,75	72,15	87,36	87,66	78,17	79,14
24 (3.13)	73,84	73,03	84,44	87,76	78,77	79,71
24 (3.15)	76,82	88,74	83,44	78,95	79,97	83,55
24 (3.18)	70,75	72,15	87,36	87,66	78,17	79,14
24 (3.19)	73,84	73,03	84,44	87,76	78,77	79,71
24 (3.21)	74,34	83,67	99,70	93,54	85,16	**88,33**
Y						
23 (ед.)					80,65	**87,22**
24 (3.11)					73,57	75,34
24 (3.13)					75,81	76,24
24 (3.15)					77,98	84,01
24 (3.18)					73,57	75,34
24 (3.19)					75,81	76,24
24 (3.21)					80,65	**87,22**

По результатам первого этапа экспериментальных исследований можно сделать следующие промежуточные выводы:

- применение процедуры нечеткой классификации продемонстрировало повышение эффективности обнаружения спама и легальных писем;

- в составленной схеме эксперимента для различных наборов электронных писем с использованием решающего правила и косинусной меры наилучшие результаты достигаются при использовании единичного веса и веса (3.21) [169], в связи с чем эти веса будут использованы на этапе 2 экспериментальных исследований.

В таблицах 4.17-4.18 представлены частные показатели эффективности классификации, полученные по результатам второго этапа экспериментальных исследований.

Таблица 4.17 – Результаты испытаний №№ 25, 26

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
25 (ед.)	0,945	0,844	0,974	0,959	0,959	0,897
26 (3.21)	0,907	0,824	0,968	0,932	0,937	0,874
Спам						
25 (ед.)	0,987	0,966	0,924	0,789	0,954	0,868
26 (3.21)	0,967	0,928	0,902	0,781	0,933	0,848
Y						
25 (ед.)					0,957	0,883
26 (3.21)					0,935	0,861

Таблица 4.18 – Результаты испытаний №№ 27, 28

№ испытания (вес)	P, %		R, %		F-мера, %	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
27 (ед.)	0,861	0,885	0,996	0,852	0,923	0,867
28 (3.21)	0,861	0,885	0,996	0,852	0,923	0,867
Спам						
27 (ед.)	0,996	0,863	0,842	0,866	0,912	0,864
28 (3.21)	0,996	0,863	0,842	0,866	0,912	0,864
Y						
27 (ед.)					0,918	0,866
28 (3.21)					0,918	0,866

Анализ результатов экспериментальных исследований в целом позволяет сделать следующие выводы:

1. Разработанные модель электронного почтового сообщения и метод классификации электронных писем позволяют эффективно обнаруживать спам и идентифицировать легальные электронные письма.

2. Разработанная модель электронных писем является универсальной с т. з. содержания (текстов) электронных писем. Параметры модели не зависят от обучающего набора электронных писем $EL^{tr} \subseteq EL$ и процедур ϕ_T^W, ϕ_T^{DR}, $\tilde{\phi}_{\Psi_{el}}$.

3. Разработанный метод классификации электронных писем позволяет использовать не только знания о спаме, но и о легальных письмах. Это позволяет повысить не только эффективность обнаружения спама, но и снизить вероятность ошибочной классификации легальных писем. Т. е. метод позволяет «задавать» необходимые показатели обнаружения спама с учетом показателей идентификации легальных писем.

4. Целесообразно использование весов термов. Это, в том числе позволяет исключить фактор случайности в предложенном методе, что обусловлено достижением возможного максимума разнообразия уникальных термов спама и легальных писем при бесконечном увеличении числа обучающих писем [141].

В составленной схеме эксперимента наиболее эффективной процедурой классификации явилось простое решающее правило с единичным весом или мерой $TF - IDF$ в формулировке поисковой системы $INQUERY$ (3.21) [141].

Применение процедуры нечеткой классификации продемонстрировало повышение эффективности обнаружения спама и достоверность идентификации легальных электронных писем, а также снижение количества неклассифицированных писем.

На первом этапе получены следующие наилучшие значения (при наилучшем показателе эффективности) с единичным весом и весом в формулировке поисковой системы $INQUERY$ (3.21):

- для простого решающего правила: точность классификации легальных писем и спама – 0,974/0,966 и 0,995/0,990, полноты классификации легальных писем и спама – 0,993/0,991 и 0,963/0,959 соответственно;

- для косинусной меры: точность классификации легальных писем и спама – 0,979/0,979 и 0,914/0,914, полноты классификации легальных писем и спама – 0,919/0,919 и 0,971/0,971 соответственно.

На втором этапе получены следующие наилучшие значения с единичным весом и весом в формулировке поисковой системы $INQUERY$:

- для простого решающего правила: точность классификации легальных писем и спама – 0,945/0,907 и 0,987/0,967, полноты классификации легальных писем и спама – 0,974/0,968 и 0,924/0,902 соответственно;

- для косинусной меры: точность классификации легальных писем и спама – 0,861/0,861 и 0,996/0,996, полноты классификации легальных писем и спама – 0,996/0,996 и 0,842/0,842 соответственно.

Результаты второго этапа в случае простого решающего правила демонстрируют незначительное ухудшение значений частных показателей качества при примерно пятикратном уменьшении количества термов.

5. Несбалансированность классов обучающей выборки электронных писем влияет на эффективность классификации, что подтверждает тезис о том, что она должна быть сбалансирована, т. е. в выборке должно быть присутствовать одинаковое количество объектов каждого класса [198].

Таким образом, результаты экспериментальных исследований подтверждают достижение поставленной цели диссертационного исследования и свидетельствуют о применимости и эффективности метода с использованием разработанной модели (2.14) (как степени приспособленности метода для обнаружения спама и идентификации легальных электронных писем как такового).

В дальнейшем целесообразно провести исследование эффективности метода классификации в зависимости от различных процедур сокращения размерности признакового пространства, правил классификации электронного письма, а также

зависимости выбираемых процедур метода классификации электронных писем от соотношения количества известных термов спама и легальных термов.

В связи с тем, что в реальной практике превалируют несбалансированные обучающие выборки легальных писем и спама, целесообразно провести исследование по выбору техники искусственной модификации наборов и их параметров для выравнивания соотношения классов электронных писем.

4.4.2 Сравнение результатов эксперимента на исследовательском прототипе с результатами аналогичных исследований

Проведенный в главе 1 анализ современного состояния исследований в области обнаружения спама позволяет провести сравнение результатов проведенного эксперимента с результатами обнаружения спама с применением различных классификационных пайплайнов на наборе писем Enron без постановки практического эксперимента.

Для этого были отобраны следующие отдельные исследования, содержащие результаты экспериментов на наборе Enron и охватывающие разнообразные классификаторы:

- Barushka A. etc. [10] – содержит результаты по предложенному авторами подходу в сравнении с результатами других опубликованных исследований. Поскольку результаты предложенного авторами подхода дали лучший результат при сопоставлении с иными выбранными для сравнения различными классификаторами (более 10), для целей настоящего раздела был выбран именно он. Авторами предложен модифицированный алгоритм на основе глубоких нейронных сетей (DBB-RDNN-ReL, *аббр. от англ.* distribution-based balancing algorithm and a regularized deep multi-layer perceptron NN model with rectified linear units);

- Sharaff A. etc. [22] – эксперименты с 4 классификаторами;

- Mohammad A. etc. [203] – авторами предложены алгоритмы извлечения признаков: адаптированные алгоритм пчелиной колонии (AABC, *аббр. от англ.* adapted artificial bee colony), муравьиный алгоритм (AACO, *аббр. от англ.* adapted ant colony optimization) и роя частиц (APSO, *аббр. от англ.* Adapted particle swarm

optimization). Проведены экспериментальные исследования с классификатором на основе случайного леса с использованием предложенных подходов.

Все исследования содержат результаты по полноте, точности и F-мере, а также по таким мерам как ошибки первого и второго рода, для которых были произведены дополнительные вычисления.

Ошибка первого рода рассчитывается как:

$$FP\_rate = \frac{N_{incorr\_a}}{N_{incorr\_a}+N_{corr\_r}}, \quad (4.1)$$

где N_{corr_r} – количество писем, корректно признанных не принадлежащими заданной категории (ложноотрицательные результаты или TN, *аббр. от англ.* True Negative),

а ошибка второго рода – по формуле:

$$FN\_rate = \frac{N_{incorr\_r}}{N_{incorr\_r}+N_{corr\_a}}. \quad (4.2)$$

Результаты сравнения представлены в таблице 4.19.

Таблица 4.19 – Сравнение результатов эксперимента с результатами аналогичных исследований

Классификатор	P	R	FP_rate	FN_rate	F-мера
DBB-RDNN-ReL [10]	n/a	**0,9983**[10]	0,0212	**0,0017**	n/a
J48[11] [22]	0,93	0,933	0,284	0,067[12]	0,93
Опорные вектора [22]	0,898	0,884	0,671	0,116[13]	0,85
Байесовский [22]	0,951	0,931	0,022	0,069[14]	0,936
LazyIBK[15] [22]	0,924	0,892	0,095	0,108[16]	0,901
Случайный лес + AABC [203]	0,762	0,711	n/a	0,289[17]	0,7356
Случайный лес + AACO [203]	0,8427	0,7676	n/a	0,2324[18]	0,8034
Случайный лес + APSO [203]	0,8831	0,8554	n/a	0,1446[19]	0,8690
Решающее правило (ед. вес)	0,995	0,963	**0,005**	0,037	0,979
Решающее правило (вес 3.21)	0,990	0,959	0,01	0,041	0,974
Косинусная мера (ед. вес)	0,914	0,971	0,079	0,029	0,941
Косинусная мера (вес 3.21)	0,914	0,971	0,079	0,029	0,941

[10] Получено расчетным способом.
[11] Реализация на языке Java алгоритма для построения дерева решений C4.5.
[12] Получено расчетным способом.
[13] Получено расчетным способом.
[14] Получено расчетным способом.
[15] Реализация в Weka (свободное программное обеспечение для анализа данных и машинного обучения) k-ближайших соседей.
[16] Получено расчетным способом.
[17] Получено расчетным способом.
[18] Получено расчетным способом.
[19] Получено расчетным способом.

Анализ результатов сравнения показывает, что среди сравниваемых предложенные модель и метод по своим показателям в целом по результатам уступают только модифицированному алгоритму на основе глубоких нейронных сетей [10].

Выводы по 4 главе

Основными результатами рассуждений, представленных в данной главе, являются:

1. Разработана архитектура подсистемы классификации писем, реализующая предложенные в работе метод и алгоритм, применение которых позволяет повысить достоверность идентификации легальных писем с учетом меняющихся информационных потребностей конкретного пользователя (персонализации).

2. На различных наборах электронных писем фиксированных размеров проведены экспериментальные исследования по их классификации для обнаружения спама. Показана эффективность использования разработанных модели электронных писем (2.14) и метода с применением различных весов и методов машинного обучения. Это делает их универсальными для обнаружения спама и позволяет сделать этот процесс персонализированным.

3. В ходе экспериментов исследовано влияние весов термов и методов машинного обучения на обобщенный показатель эффективности предложенного метода, обеспечивающего точность и полноту обнаружения спама и достоверности идентификации легальных электронных почтовых сообщений.

4. Экспериментально продемонстрирована обоснованность разработанных в диссертационном исследовании модели электронного почтового сообщения и метода классификации электронных писем, а также их эффективность.

Заключение

Диссертационная работа посвящена повышению эффективности обнаружения спама и достоверности идентификации легальных писем на основе классификации их содержания за счет создания модели электронного письма, учитывающей содержание электронных писем конкретного пользователя (персонализацию). Основные результаты диссертационной работы сводятся к следующим положениям:

1. Проведен анализ современного состояния исследований в области обнаружения спама, результаты которого позволили выделить наиболее значимые отличительные особенности спама, определить группы и перечень наиболее важных информативных признаков, которые позволяют отнести то или иное электронное письмо к классу спама или легальных, обосновать целесообразность использования при создании модели электронных писем метода выделения термов, позволяющего усилить смысловое содержание термов за счет применения метода «генетических карт».

2. Разработана модель электронного почтового сообщения для классификации электронных писем на основе метода «генетических карт», отличающаяся от известных моделей методом выделения значимых последовательностей символов текста (признаков электронных писем на основе их содержания, термов), позволяющим усилить смысловое содержание термов.

3. Разработан метод классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, отличающийся от известных использованием разработанной модели электронных писем, применение которого позволяет повысить эффективность обнаружения спама и достоверность идентификации легальных электронных писем с учетом меняющихся информационных потребностей конкретного пользователя (персонализации), а также снизить количество неклассифицированных писем.

4. Разработан алгоритм классификации электронных писем, отличающийся от известных наличием дополнительной процедуры определения «схожести»

термов на основе расстояния Левенштейна, обеспечивающей вычисление мер принадлежности классифицируемого электронного письма к классам спама и легальных, применение которого позволяет повысить достоверность идентификации легальных электронных писем с учетом меняющихся информационных потребностей конкретного пользователя (персонализации), а также уменьшить количество неклассифицируемых писем.

5. Разработана архитектура подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, отличающаяся от известных блоком выделения термов и блоком нечеткой классификации, реализующая предложенные в работе метод и алгоритм, применение которых позволяет повысить достоверность идентификации легальных электронных писем с учетом меняющихся информационных потребностей конкретного пользователя (персонализации).

6. Разработаны программные модули исследовательского прототипа подсистемы классификации электронных писем и экспериментально продемонстрирована обоснованность разработанных в диссертационном исследовании модели электронного почтового сообщения и метода классификации электронных писем, а также их эффективность со следующими наилучшими результатами (при наилучшем показателе эффективности): точность классификации легальных писем и спама – 0,974/0,966 и 0,995/0,990, полнота классификации легальных писем и спама – 0,992/0,991 и 0,963/0,959 соответственно.

Совокупность результатов диссертационного исследования дает основание утверждать, что достигнута цель работы – повышена эффективность обнаружения спама и достоверность идентификации легальных электронных почтовых сообщений на основе классификации их содержания.

Объем проведенной работы позволяет автору утверждать, что полученные в ее ходе эмпирические и статистические данные достаточно объективны, валидны и репрезентативны, лишены личной предубежденности, основаны на научных и практических позициях российских и зарубежных исследователей и ученых.

Материалы настоящего исследования могут быть использованы в дальнейших научных разработках по схожим тематикам, в практической деятельности разработчиков средств для обнаружения спама, а также в программах лекционных и семинарских занятий по соответствующим дисциплинам, предлагаемым слушателям высших учебных заведений.

Сложность и многогранность естественного языка, а также многообразие и специфика содержимого спама оставляют открытым большое количество вопросов для научной проработки.

В дальнейшем целесообразно провести исследование эффективности метода классификации в зависимости от различных процедур сокращения размерности признакового пространства, правил классификации электронного письма, а также зависимости выбираемых процедур метода классификации электронных писем от соотношения количества известных термов спама и легальных термов.

В связи с тем, что в реальной практике превалируют несбалансированные обучающие выборки легальных писем и спама, целесообразно провести исследование по выбору техники искусственной модификации наборов и их параметров для выравнивания соотношения классов электронных писем.

Список литературы

1. Что такое спам [электронный ресурс]. АО «Лаборатория Касперского». Режим доступа: https://encyclopedia.kaspersky.ru/knowledge/what-is-spam (дата обращения: 19.01.2021).

2. Вергелис М., Щербакова Т., Сидорина Т. Спам и фишинг в 2018 году [электронный ресурс]. Securelist. – 2019. Режим доступа: https://securelist.ru/spamand-phishing-in-2018/93453 (дата обращения 19.01.2021).

3. Вергелис М., Щербакова Т., Сидорина Т., Куликова Т. Спам и фишинг в 2019 году [электронный ресурс]. Securelist. – 2020. Режим доступа: https://securelist.ru/spam-report-2019/95727 (дата обращения 19.01.2021).

4. Куликова Т., Щербакова Т., Сидорина Т. Спам и фишинг в 2020 году [электронный ресурс]. Securelist. – 2021. Режим доступа: https://securelist.ru/spam-and-phishing-in-2020/100408/ (дата обращения 21.04.2021).

5. Куликова Т., Щербакова Т. Спам и фишинг в 2021 году [электронный ресурс]. Securelist. – 2022. Режим доступа: https://securelist.ru/spam-and-phishing-in-2021/104407/ (дата обращения 31.08.2023).

6. Куликова Т., Деденок Р., Свистунова О., Ковтун А., Шимко И. Спам и фишинг в 2022 году [электронный ресурс]. Securelist. – 2023. Режим доступа: https://securelist.ru/spam-phishing-scam-report-2022/106719/ (дата обращения 31.08.2023).

7. Bibi A., Latif R., Khalid S., Ahmed W., Shabir R. A., Ansari M., et al. Spam Mail Scanning Using Machine Learning Algorithm // Journal of Computers. 2020. Vol. 15. No. 2. PP. 73-84. DOI:10.17706/jcp.15.2.73-84.

8. Androutsopoulos I., Paliouras G., Michelakis E. Learning to Filter Unsolicited Commercial E-Mail // NCSR «Demokritos». Tech. Report number: 2004/2. 2004.

9. Radhakrishnan A., Vaidhehi V. Email Classification Using Machine Learning Algorithms // International Journal of Engineering and Technology (IJET). 2017. Vol. 9. No. 2. PP. 335-340. DOI:10.21817/ijet/2017/v9i1/170902310.

10. Barushka A., Hajek P. Spam Filtering Using Integrated Distribution-Based Balancing Approach and Regularized Deep Neural Networks // Applied Intelligence. 2018. Vol. 48. PP. 3538-3556. DOI:10.1007/s10489-018-1161-y.

11. Shen H., Li Z., Leveraging Social Networks for Effective Spam Filtering // IEEE Transactions on Computers. 2014. Vol. 63. No. 11. PP. 2743-2759. DOI:10.1109/TC.2013.152.

12. Корелов С. В., Петров А. М., Ротков Л. Ю., Горбунов А. А. Предобработка текстов электронных писем в задаче обнаружения спама // Труды учебных заведений связи. 2020. Т. 6. № 4. С. 80-90. DOI:10.31854/1813-324X-2020-6-4-80-90.

13. Чернопрудова Е. Н. Защита почтовых сервисов от несанкционированных рассылок на основе контентной фильтрации электронных сообщений: автореф. дис. … канд. техн. наук: 05.13.19/Чернопрудова Елена Николаевна. – Уфа, 2013. – 16 с.

14. Bhattacharya P., Singh A. E-mail Spam Filtering using Genetic Algorithm based on Probabilistic Weights and Words Count // International Journal of Integrated Engineering. 2020. Vol. 12. No. 1. PP. 40-49. DOI:10.30880/ijie. 2020.12.01.004.

15. Кирьянов К. Г. Генетический код и тексты: динамические и информационные модели сложных систем /Ред. Л. Ю. Ротков, А. В. Якимов. – Нижний Новгород: ТАЛАМ, 2002. – 100 с.

16. Кирьянов К. Г. Выбор оптимальных базовых параметров источников экспериментальных данных при их идентификации // Идентификация систем и задачи управления SICPRO’04: тр. III Междунар. конф. – М.: Изд-во ИПУ РАН, 2004. – С. 187–208.

17. Email Statistics Report, 2023–2027 [электронный ресурс]. The Radicati Group, Inc. – 2023. Режим доступа: https://www.radicati.com/?p=18089 (дата обращения: 06.11.2023).

18. Abdulhamid Sh. M., Shuaib M., Osho O., Ismaila I., Alhassan J. K. Comparative Analysis of Classification Algorithms for Email Spam Detection //

International Journal of Computer Network and Information Security (IJCNIS). 2018. Vol. 10. No. 1. PP. 60-67. DOI:10.5815/ijcnis.2018.01.07.

19. Rusland N., Wahid N., Kasim Sh., Hafit H. Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets // Proceedings of International Research and Innovation Summit (IRIS2017, Melaka, Malaysia, 6-7 May 2017). IOP Conference Series: Materials Science and Engineering. Bristol: IOP Publishing, 2017. Vol. 226. DOI:10.1088/1757-899X/226/1/012091.

20. Verma T., Gill N. S. Email Spams via Text Mining using Machine Learning Techniques // International Journal of Innovative Technology and Exploring Engineering (IJITEE). 2020. Vol. 9. No. 4. PP. 2535-2539. DOI:10.35940/ijitee.D1915.029420.

21. Alguliyev R., Aliguliyev R., Saadat A. Classification of Textual E-Mail Spam Using Data Mining Techniques // Applied Computational Intelligence and Soft Computing. 2011. Vol. 2011. Article ID 416308, 8 pages. DOI:10.1155/2011/416308.

22. Sharaff A., Nagwani N., Dhadse A. Comparative Study of Classification Algorithms for Spam Email Detection // Shetty N., Prasad N., Nalini N. (eds) Emerging Research in Computing, Information, Communication and Applications. New Delhi: Springer, 2016. PP. 237-244. DOI:10.1007/978-81-322-2553-9_23.

23. Yasin A. Spam Reduction by using E-mail History and Authentication (SREHA) // International Journal of Computer Network and Information Security (IJCNIS). 2016. Vol. 8. No. 7. PP. 17-22, 2016. DOI:10.5815/ijcnis.2016.07.03

24. Корелов С. В., Ротков Л. Ю., Рябов А. А. Вероятностный метод идентификации спама // Доклады Томского государственного университета систем управления и радиоэлектроники. 2010. № 1 (21), часть 1. С. 150-152.

25. Ущерб от рассылки спама в России составляет 55 млн. долларов в год [электронный ресурс]. Positive Technologies. – 2004. Режим доступа: https://www.securitylab.ru/news/213700.php (дата обращения: 19.01.2021).

26. Николаев И. А., Титова М. В. Спам: экономические потери [электронный ресурс]. ФБК Grant Thornton. – 2009. Режим доступа: http://www.fbk.ru/upload/images/economic_losses-final.pdf (дата обращения: 19.01.2021).

27. РАЭК выпустила первое в России масштабное исследование по спаму по итогам 2009 года [электронный ресурс]. НП «РАЭК». – 2010. Режим доступа: http://2010.raec.ru/news/meeting100203/ (дата обращения: 19.01.2021).

28. How much does spam cost the world? [электронный ресурс]. Fastnet SA Blog. – 2017. Режим доступа: https://www.mailcleaner.net/blog/spam-world-news/how-much-does-spam-cost-the-world (дата обращения: 20.01.2021).

29. Пользователи по всему миру столкнулись с огромной волной спама [электронный ресурс]. Positive Technologies. – 2020. Режим доступа: https://www.securitylab.ru/news/514827.php (дата обращения: 19.01.2021).

30. «Код Безопасности» ранжировал ИБ-инциденты 3 и 4 квартала 2010 г. и проанализировал тенденции серверной виртуализации [электронный ресурс]. ООО «Код Безопасности». – 2011. Режим доступа: https://www.securitycode.ru/company/news/kod_bezopasnosti_ranzhiroval_ib_intsidenty_3_i_4_kvartala_2010_g_i_proanaliziroval_tendentsii_server/ (дата обращения: 24.07.2018).

31. Sattler J. Why Spam is On the Rise – Again [электронный ресурс]. F-Secure Blog. – 2018. Режим доступа: https://blog.f-secure.com/why-spam-is-on-the-rise-again (дата обращения: 19.01.2021).

32. Состав технических параметров компьютерного инцидента, указываемых при представлении информации в ГосСОПКА, и форматы представления информации о компьютерных инцидентах [электронный ресурс]. НКЦКИ – 2019. Режим доступа: https://safe-surf.ru/specialists/article/5252/638030/ (дата обращения: 06.09.2021).

33. Стандарт Банка России СТО БР ИББС-1.0-2014 Обеспечение информационной безопасности организаций банковской системы Российской Федерации. Общие положения. – М.: Банк России, 2014.

34. Стандарт Банка России СТО БР ИББС-1.3-2016 Обеспечение информационной безопасности организаций банковской системы Российской Федерации. Сбор и анализ технических данных при реагировании на инциденты

информационной безопасности при осуществлении переводов денежных средств. – М.: Банк России, 2016.

35. Стандарт Банка России СТО БР БФБО-1.5-2023 Безопасность финансовых (банковских) операций управление инцидентами, связанными с реализацией информационных угроз, и инцидентами операционной надежности о формах и сроках взаимодействия банка России с кредитными организациями, некредитными финансовыми организациями и субъектами национальной платежной системы при выявлении инцидентов, связанных с реализацией информационных угроз, и инцидентов операционной надежности. – М.: Банк России, 2023.

36. Временный регламент передачи данных участников информационного обмена в Центр мониторинга и реагирования на компьютерные атаки в кредитнофинансовой сфере Банка России (Версия 2.3) [электронный ресурс]. Банк России. Режим доступа: https://cbr.ru/StaticHtml/File/14408/inforegl_23.pdf (дата обращения: 19.01.2021).

37. Androutsopoulos I., Koutsias J., Chandrinos K., Spyropoulos C. An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-Mail Messages // Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR’00, Athens, Greece, 24–28 July 2000). New York: Association for Computing Machinery, 2000. PP. 160-167. DOI:10.1145/345508.345569.

38. Metsis V., Androutsopoulos I., Paliouras G. Spam Filtering with Naive Bayes – Which Naive Bayes? // Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006, Mountain View, USA, 27-28 July 2006). 2006. PP. 28-69.

39. Мезенцева Е. М. Исследование и разработка статистических алгоритмов фильтрации сообщений в интерактивных ресурсах инфокоммуникационных сетей: автореф. дис. … канд. техн. наук: 05.12.13/Мезенцева Екатерина Михайловна. – Самара, 2013. – 16 с.

40. Sahami M., Dumais S., Heckerman D., Horvitz E. A Bayesian Approach to Filtering Junk E-Mail // Proceedings of 1998 AAAI Workshop on Learning for Text Categorization. AAAI Technical Report WS-98-05. AAAI, 1998. PP. 55-62.

41. Graham P. A Plan for Spam [электронный ресурс] // Graham P. – 2002. Режим доступа: http://www.paulgraham.com/spam.html (дата обращения: 21.01.2021).

42. Graham P. Better Bayesian filtering [электронный ресурс] // Graham P. – 2003. Режим доступа: http://www.paulgraham.com/better.html (дата обращения: 21.01.2021).

43. Robinson G. A Statistical Approach to the Spam Problem // Linux Journal, Iss. 107. 2003.

44. Carreras X., Marquez L. Boosting Trees for Anti-Spam Email Filtering // Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing (RANLP, 5-7 September 2001). 2001. PP. 58-64.

45. Sheu JJ., Chen YK., Chu KT., Tang JH., Yang WP. An Intelligent Three-Phase Spam Filtering Method Based on Decision Tree Data Mining // Security and Communication Networks. 2016. Vol. 9. No. 17. PP. 4013-4026. DOI:10.1002/sec.1584.

46. Павлов А. С. Исследование и разработка методов построения программных средств обнаружения текстового спама: автореф. дис. … канд. физ.-мат. наук: 05.13.11/Павлов Антон Сергеевич. – М., 2011. – 15 с.

47. Drucker H., Wu D., Vapnik V. Support Vector Machine for Spam Categorization // IEEE Transactions on Neural Networks. 1999. Vol. 10. No. 5. PP. 1048-1054. DOI:10.1109/72.788645.

48. Мироненко А. Н. Алгоритм контентной фильтрации спама на базе совмещения метода опорных векторов и нейронных сетей: автореф. дис. … канд. техн. наук: 05.13.19/Мироненко Антон Николаевич. – СПб., 2012. – 18 с.

49. Блинов С. Ю. Методы и алгоритмы классификации информации для защиты от спама: автореф. дис. … канд. техн. наук: 05.13.19/Блинов Станислав Юрьевич. – СПб., 2013. – 22 с.

50. Розинкин А. Н. Система защиты от массовых несанкционированных рассылок электронной почты на основе методов Data Mining: автореф. дис. … канд. физ.-мат. наук: 05.13.11/Розинкин Андрей Николаевич. – М., 2006. – 16 с.

51. Jiang S., Pang G., Wu M., Kuang L. An Improved k-Nearest-Neighbor Algorithm for Text Categorization // Expert System with Applications. 2012. Vol. 39. No. 1. PP. 1503–1509. DOI:10.1016/j.eswa.2011.08.040.

52. Sakkis G., Androutsopoulos I., Paliouras G., Karkaletsis V., Spyropoulos C. D., Stamatopoulos P. A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists // Information Retrieval. 2003. Vol. 6. PP. 49-73. DOI:10.1023/A:1022948414856.

53. Yue X., Abraham A., Chi ZX., Hao YY., Mo H. Artificial Immune System Inspired Behavior-Based Anti-Spam Filter // Soft Computing. 2007. Vol. 11. PP. 729–740. DOI:10.1007/s00500-006-0116-0.

54. Малыхина М. П., Частикова В. А., Биктимиров А. А. Методика обнаружения спама на основе искусственных иммунных систем // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2018. № 3. С. 38–48. DOI:10.24143/2072-9502-2018-3-38-48.

55. Clark J., Koprinska I., Poon J. A Neural Network Based Approach to Automated Email Classification // Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI 2003, Halifax, Canada, 13–17 October 2003). IEEE, 2003. PP. 702–705. DOI:10.1109/WI.2003.1241300.

56. Катасёв А. С., Катасёва Д. В., Кирпичников А. П. Нейросетевая технология классификации электронных почтовых сообщений // Вестник технологического университета. 2015. Т. 18. № 5. С. 180–183.

57. Катасёв А. С., Катасёва Д. В., Кирпичников А. П., Семёнов Я. Е. Спам-фильтрация электронных почтовых сообщений на основе нейросетевой и нейронечеткой моделей // Вестник технологического университета. 2015. Т. 18. № 15. С. 217–221.

58. Катасёв А. С., Катасёва Д. В. Разработка нейросетевой системы классификации электронных почтовых сообщений // Вестник Казанского государственного энергетического университета. 2015. № 1 (25). С. 68–78.

59. Ларионова А. В., Хорев П. Б. Метод фильтрации спама на основе искусственной нейронной сети // Науковедение. 2016. Т. 8. № 3. URL: http://naukovedenie.ru/PDF/04TVN316.pdf (дата обращения 26.11.2020)

60. Ларионова А. В., Хорев П. Б. Оценка эффективности метода фильтрации спама на основе искусственной нейронной сети // Науковедение. 2016. Т. 8. № 2. DOI:10.15862/134TVN216.

61. Никитин А. П. Многоуровневая многоагентная система фильтрации спама в организации: автореф. дис. … канд. техн. наук: 05.13.19/Никитин Андрей Павлович. – Уфа, 2009. – 16 с.

62. Корелов С. В., Крюков А. К., Ротков Л. Ю. Методы цифрового анализа текстовых сообщений для идентификации спама // Труды (Десятой) Научной конференции по радиофизике, посвященная 90-летию ННГУ и 100-летию со дня рождения Г. С. Горелика (Нижний Новгород, Российская Федерация, 5-25 мая 2006). Нижний Новгород: ННГУ, 2006. URL: http://old.rf.unn.ru/rus/sci/books/06/doc/11InfSys06.doc (дата обращения 21.01.2021).

63. Агаджанов В. В., Корелов С. В., Ротков Л. Ю. Обнаружение спама при помощи аппарата wavelet-преобразований // Труды XII научной конференции по радиофизике, посвященной 90-летию со дня рождения М. М. Кобрина (Нижний Новгород, 7 мая 2008 г.) /Под ред. А. В. Якимова, С. М. Грача. Нижний Новгород: Изд-во ТАЛАМ, 2008. С. 276-277.

64. Корелов С. В., Грачева О. К. Идентификация спама на классах сообщений // Труды XIV научной конференции по радиофизике, посвященной 80-й годовщине со дня рождения Ю. Н. Бабанова (Нижний Новгород, 7 мая 2010 г.) /Под ред. С. М. Грача, А. В. Якимова. Нижний Новгород: ННГУ, 2010. С. 288-289.

65. Семенова М. А. Модель и метод градуированной фильтрации «спама»: автореф. дис. … канд. техн. наук: 05.13.19/Семенова Мария Александровна. – СПб., 2009. – 20 с.

66. Junejo K., Yousaf M., Karim A. A Two-Pass Statistical Approach for Automatic Personalized Spam Filtering // Proceedings The Discovery Challenge Workshop of 17th European Conference on Machine Learning (ECML) and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD) (ECML-PKDD 2006, Berlin, Germany, 18-22 September 2006). 2006. PP. 16-27.

67. Junejo K., Karim A. PSSF: A Novel Statistical Approach for Personalized Service-side Spam Filtering // Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'07, Fremont, California, USA, 2-5 November 2007). IEEE, 2007. PP. 228-234, DOI:10.1109/WI.2007.47.

68. Cohen W. Learning Rules that Classify E-Mail // Proceedings of 1996 AAAI Spring Symposium on Machine Learning in Information Access (Stanford, 25-27 March 1996). AAAI Technical Report SS-96-05. 1996. PP. 18-25.

69. Cohen W., Singer Y. Context-sensitive learning methods for text categorization // ACM Transactions on Information Systems. 1999. Vol. 17. No. 2. PP. 141-173. DOI:10.1145/306686.306688.

70. Delany S., Cunningham P., Coyle L. An Assessment of Case-Based Reasoning for Spam Filtering // Artificial Intelligence Review. 2005. Vol. 24. PP. 359-378, DOI:10.1007/s10462-005-9006-6.

71. Gee K. Using Latent Semantic Indexing to Filter Spam // Proceedings of the 2003 ACM Symposium on Applied computing (SAC'03, Melbourne, Florida, USA, 9-12 March, 2003). New York: Association for Computing Machinery, 2003. PP. 460-464. DOI: 10.1145/952532.952623.

72. Visani Ch., Jadeja N., Modi M. A Study on Different Machine Learning Techniques for Spam Review Detection // Proceedings of the International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS, Chennai, India, 1-2 August 2017). IEEE, 2017. PP. 676-679. DOI:10.1109/ICECDS.2017.8389522.

73. Hussain N., Turab Mirza H., Rasool G., Hussain I., Kaleem M. Spam Review Detection Techniques: A Systematic Literature Review // Applied Sciences. 2019. Vol. 9. No. 5. PP. 10-26. DOI:10.3390/app9050987.

74. Как работает фильтр Spamtest [электронный ресурс]. Режим доступа: https://securelist.ru/kak-rabotaet-fil-tr-spamtest/110/, свободный (дата обращения: 17.10.2018).

75. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys. 2002. Vol. 34, No. 1, 2002, PP. 1-47, DOI: 10.1145/505282.505283.

76. Терейковский И. А. Применение семантического анализа содержимого электронных писем в системах распознавания спама [электронный ресурс]. Режим доступа: https://refdb.ru/look/1498468.html, свободный (дата обращения: 06.08.2018).

77. Ермаков А. Е. Эксплицирование элементов смысла текста средствами синтаксического анализа-синтеза // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2003. 2003. С. 136-140.

78. Сегалович И., Тейблюм Д., Дилевский А. Принципы и технические методы работы с незапрашиваемой корреспонденцией [электронный ресурс]. Режим доступа: https://cache-mskm908.cdn.yandex.net/download.yandex.ru/company/spamooborona-latest.pdf, свободный (дата обращения: 07.08.2018).

79. Тутубалин А. Распределенные методы обнаружения спама: обзор существующих решений, анализ перспектив [электронный ресурс]. Режим доступа: http://www.lexa.ru/articles/distributed-antispam-1.html, http://www.lexa.ru/articles/distributed-antispam-2.html, свободный (дата обращения: 07.08.2018).

80. Зайцев О. Технологии рассылки спам сообщений и методы защиты от него. // Компьютер пресс, № 2, 2007 г. [Электронный ресурс]. Режим доступа: http://www.compress.ru/article.aspx?id=17269&iid=799, свободный (дата обращения: 07.08.2018).

81. Федотов Н. Н. Спам обречен? // Мир ПК, № 11, 2003 г.

82. Федеральный закон от 13 марта 2006 года № 38-ФЗ «О рекламе».

83. Наумов В. Спам: юридический анализ явления [электронный ресурс]. Режим доступа: http://www.russianlaw.net/law/media/spam/a25/, свободный (дата обращения: 07.08.2018).

84. Нормы пользования Сетью (OFIPS-008) [электронный ресурс]. Режим доступа: http://www.ofisp.org/documents/ofisp-008.html, свободный (дата обращения: 07.08.2018).

85. Постановление Правительства Российской Федерации от 10 сентября 2007 года № 575 «Об утверждении Правил оказания телематических услуг связи».

86. Enron-Spam datasets. URL: http://www2.aueb.gr/users/ion/data/enron-spam (дата обращения 26.11.2020).

87. Спам и фишинг [электронный ресурс]. АО «Лаборатория Касперского». Режим доступа: https://securelist.ru/threat-category/spam-i-fishing/, свободный (дата обращения: 17.01.2022).

88. Анти-спам решения и безопасность [электронный ресурс]. Режим доступа: http://www.brain-work.ru/en/articles/102-anti-spam-solutions-and-security, свободный (дата обращения: 22.05.2014).

89. Афонин А. И. Что такое спам? // Наука и образование, № 02, 2013 г.

90. Технические средства фильтрации спама [электронный ресурс]. АО «Лаборатория Касперского». Режим доступа: https://securelist.ru/tehnicheskie-sredstva-fil-tratsii-spa/72/, свободный (дата обращения: 08.08.2018).

91. Ашманов И. Борьба со спамом техническими средствами // BYTE Россия. 2004. № 1 (65). URL: https://www.bytemag.ru/articles/detail.php?ID=8952 (дата обращения 14.07.2020).

92. Слепов О. Борьба со спамом // Jet Info Информационный бюллетень. 2004. № 9 (136). – 20 с.

93. Anti-spam techniques. [электронный ресурс] // ВикипедиЯ. Режим доступа: http://en.wikipedia.org/wiki/Anti-spam_techniques, свободный (дата обращения: 05.10.2018).

94. Леонтьева Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы: учеб. пособие для студ. лингв. фак. вузов / Нина Николаевна Леонтьевна. – М.: Издательский центр «Академия», 2006. – 306 с.

95. Теория множеств. [электронный ресурс] // ВикипедиЯ. Режим доступа: https://ru.wikipedia.org/wiki/Теория_множеств (дата обращения: 17.01.2022).

96. Анфилатов В. С., Емельянов А. А., Кукушкин А. А. Системный анализ в управлении: Учеб. пособие – М.: Финансы и статистика, 2020. – 368 с.

97. Рамеев О. А., Коваленко А. П. Методы анализа многомерных данных // Учебно-методические материалы, Москва, 1988. – 621 с.

98. Теория принятия решений. [электронный ресурс] // ВикипедиЯ. Режим доступа: https://ru.wikipedia.org/wiki/Теория_принятия_решений (дата обращения: 17.01.2022).

99. Генетическая информация. [электронный ресурс] // ВикипедиЯ. Режим доступа: https://ru.wikipedia.org/wiki/Генетическая_информация (дата обращения: 17.12.2018).

100. Введение в генетику. Лекция [электронный ресурс]. Режим доступа: https://studfiles.net/preview/5283790/ (дата обращения: 23.01.2019).

101. Корелов С. В., Петров А. М., Ротков Л. Ю., Горбунов А. А. Модель электронных писем в задаче обнаружения спама // Вестник Поволжского государственного технологического университета. Сер.: Радиотехнические и инфокоммуникационные системы. 2020. № 2 (46). С. 44-54. DOI:10.25686/2306-2819.2020.2.44.

102. Кирьянов К. Г. Анализ и диагностирование последовательностей данных с помощью «генетических карт» // В сб. «Техническая диагностика». IV Всесоюзное совещание. Тезисы докладов. Москва, 1987. С. 4.

103. Кирьянов К. Г. Диагностирование последовательностей данных по их генетическим картам // В сб. «Комбинаторно-алгебраические методы и их применение». Межвузовский сборник. – Горький: ГГУ, 1987. С. 40-46.

104. Петрунин Д. Н., Ротков Л. Ю. Анализ трафиков каналов и сетей связи с помощью генетических карт // Труды (пятой) научной конференции по

радиофизике, посвященной 100-летию со дня рождения А. А. Андронова. 7 мая 2001 г. /Ред. А. В. Якимов. – Нижний Новгород: ТАЛАМ, 2001. С. 349-350.

105. Кирьянов К. Г., Рязанов В. М., Сахаров Б. А. Анализ спектров по генетическим картам процессов // Сб. «Развитие и внедрение новой техники радиоприемных устройств и обработки сигналов». – М.: Радио и Связь. 1989. С. 108.

106. Кирьянов К. Г. Идентификация сложных нестационарных объектов и процессов по их генетическим картам // Тезисы докладов 3-й конференции «Нелинейные колебания механических систем» – Н. Новгород: НИИПМК при ННГУ, 1993.

107. Кирьянов К. Г. Исследование сложных объектов и процессов по их генетическим картам. Синергетические измерения (ч. 1). // Техника средств связи. Серия РТ, М.: ЭКОС, вып. 4, 1991, С.45-78.

108. Кирьянов К. Г. Измерение динамической сложности процессов по их генетическим картам // Сб. «Развитие и внедрение новой техники радиоприемных устройств и обработки сигналов» – М.: Радио и Связь, 1989. С. 100.

109. Raidl G. Evolutionary Computation: An Overview and Recent Trends [электронный ресурс]. Режим доступа: https://www.ac.tuwien.ac.at/files/pub/raidl-05c.pdf, свободный (дата обращения: 17.12.2018)

110. Батищев Д. И. Генетические алгоритмы решения экстремальных задач: учеб. пособ. / под ред. акад. Я. Е. Львовича. – Воронеж: Воронеж. гос. техн. ун-т: Нижегород. гос. ун-т, 1995.

111. Батищев Д. И., Неймарк Е. А., Старостин Н. В. Применение генетических алгоритмов к решению задач дискретной оптимизации. // Учебно-методические материалы по программе повышения квалификации «Информационные технологии и компьютерное моделирование в прикладной математике». – Нижний Новгород: ННГУ им. Н. И. Лобачевского, 2007.

112. Горбунов А. А. Алгоритмы структурной идентификации математических моделей криптосистем на основе определения базовых параметров // Доклады ТУСУРа. 2009. № 1 (19). Ч. 2. С. 21-23.

113. Корелов С. В., Ротков Л. Ю. Метод генетических карт в задаче идентификации спама // Информационно-измерительные и управляющие системы. 2011. № 3. Т. 9. С. 72-75.

114. Корелов С. В., Крюков А. К., Ротков Л. Ю. Применение метода построения генетической карты текста для идентификации спама // Труды XII научной конференции по радиофизике, посвященной 90-летию со дня рождения М. М. Кобрина (Нижний Новгород, 7 мая 2008 г.) /Под ред. А. В. Якимова, С. М. Грача. Нижний Новгород: Изд-во ТАЛАМ, 2008. С. 277-278.

115. Корелов С. В. Обнаружение текстового спама методом генетических карт // Труды XV научной конференции по радиофизике, посвященной 110-й годовщине со дня рождения А. А. Андронова (Нижний Новгород, 1-13 мая 2011 г.) /Под ред. С. М. Грача, А. В. Якимова. Нижний Новгород: ННГУ, 2011. С. 265-267.

116. Корелов С. В., Ротков Л. Ю. Идентификация текстового спама методом генетических карт // Вестник Нижегородского университета им. Н.И. Лобачевского. 2012. № 4 (1). С. 101-104.

117. Kanaris I, Kanaris K, and Stamatatos E. Spam Detection Using Character N-Grams. In SETN. Lecture Notes in Computer Science. Springer. 2006. Vol. 3955. PP. 95-104. DOI:10.1007/11752912_12.

118. Корелов С. В., Петров А. М., Ротков Л. Ю., Горбунов А. А. Определение длины выборки в модели электронных писем // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. 2020. № 4 (36). С. 31-47. DOI:10.15593/2224-9397/2020.4.02.

119. Корелов С. В., Петров А. М., Ротков Л. Ю., Горбунов А. А. К вопросу об определении численного значения параметра в модели электронных писем // Труды XXIV научной конференции по радиофизике, посвященной 75-летию радиофизического факультета (Нижний Новгород, 13 – 31 мая 2020 г.). Нижний Новгород: ННГУ, 2020. С. 471-474.

120. Корелов С. В., Петров А. М., Ротков Л. Ю., Горбунов А. А. К вопросу об определении численного значения параметра модели электронных писем //

Материалы всероссийской научно-технической конференции «Автоматизированные системы управления и информационные технологии» (г. Пермь, 9–11 июня 2020 г.). 2020. Т.2. С. 519-525.

121. Корелов С. В., Петров А. М., Ротков Л. Ю., Горбунов А. А. Комбинирование значений параметра модели электронных писем // Материалы XII Международной интернет-конференции молодых ученых, аспирантов, студентов «Инновационные технологии: теория, инструменты, практика» (г. Пермь, 16 ноября – 31 декабря 2020 г.). 2020. С. 448-455.

122. Корелов С. В., Петров А. М., Сидоркина И. Г., Ротков Л. Ю., Горбунов А. А. Выбор размера кодовой таблицы в модели электронных писем // Вестник Поволжского государственного технологического университета. Сер.: Радиотехнические и инфокоммуникационные системы. 2021. № 3 (51). С. 49-62. DOI:10.25686/2306-2819.2021.3.49.

123. Uysal A., Gunal S. The Impact of Preprocessing on Text Classification // Information Processing & Management. 2014. Vol. 50. No. 1. PP. 104-112. DOI:10.1016/j.ipm.2013.08.006.

124. Enron Corpus. [электронный ресурс] // ВикипедиЯ. Режим доступа: https://en.wikipedia.org/wiki/Enron_Corpus (дата обращения: 30.03.2022).

125. Sebastiani F. Text Categorization // Zanasi A. (ed.). Text Mining and its Applications. Southampton: WIT Press, 2005. PP. 109-129.

126. Aas K., Eikvil L. Text Categorisation: A Survey // Norwegian Computing Center. Tech. Report number: 941, 1999.

127. Manning C., Raghavan P., Shütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. DOI:10.1017/CBO9780511809071.

128. Sokolova M., Lapalme G. A Systematic Analysis of Performance Measures for Classification Tasks // Information Processing & Management. 2009. Vol. 45. Iss. 4. PP. 427–437. DOI:10.1016/j.ipm.2009.03.002.

129. Шаров С. А. Частотный словарь. РосНИИ ИИ. [электронный ресурс]. Режим доступа: http://www.artint.ru/projects/frqlist.php (дата обращения: 30.03.2022).

130. Бойков В. В., Жукова Н. А., Романова Л. А. Распределение длины слов в русских, английских и немецких текстах. [электронный ресурс]. Режим доступа: http://tverlingua.ru/archive/001/01_1-006.htm (дата обращения: 30.03.2022).

131. Климов Д. В. Предобработка текстовых сообщений для метрического классификатора // Символ науки. 2017. № 12. С. 25–32.

132. Haddi E., Liu X., Shi Y. The Role of Text Pre-processing in Sentiment Analysis // Procedia Computer Science. 2013. Vol. 17. PP. 26–32. DOI:10.1016/j.procs.2013.05.005.

133. Devaraj S., Krishnakumar A. Effective Search Engine Spam Classification // International Journal of Recent Technology and Engineering (IJRTE). 2019. Vol. 8. No. 2S8. PP. 1541–1545. DOI:10.35940/ijrte.B1100.0882S819.

134. HaCohen-Kerner Y., Miller D., Yigal Y. The Influence of Preprocessing on Text Classification Using a Bag-of-Words Representation // PLoS ONE. 2020. Vol. 15 (5): e0232525. DOI:10.1371/journal.pone.0232525.

135. Vijayarani S., Ilamathi J., Nithya M. Preprocessing Techniques for Text Mining – An Overview // International Journal of Computer Science & Communication Networks. 2015. Vol. 5. No. 1. PP. 7–16.

136. Weng J. NLP Text Preprocessing: A Practical Guide and Template. URL: https://towardsdatascience.com/nlp-textpreprocessing-a-practical-guide-and-template-d80874676e79 (дата обращения 14.07.2020).

137. Корелов С. В., Петров А. М., Сидоркина И. Г., Горбунов А. А. Анализ результатов реализации подхода к выделению термов в модели электронных писем на случайность // Труды XXV научной конференции по радиофизике, (Нижний Новгород, 14 – 26 мая 2021 г.). Нижний Новгород: ННГУ, 2021. С. 498-502.

138. Бессмертный И. А., Нугуманова А. Б., Платонов А. В. Интеллектуальные системы: учебник и практикум для вузов – М.: Издательство Юрайт, 2019. – 243 с.

139. Солодухин А. С. Классификация текстов на основе приближенных оценок вероятностей классов // Вестник Воронежского государственного

университета. Сер.: Системный анализ и информационные технологии. 2008. № 1. С. 86-91.

140. Епрев А. С. Автоматическая классификация текстовых документов // Математические структуры и моделирование. 2010. Вып. 21. С. 65-81.

141. Корелов С. В., Петров А. М., Сидоркина И. Г., Ротков Л. Ю. Модель процесса классификации электронных писем и алгоритм его реализации в задаче обнаружения спама // Труды Международного научно-технического конгресса «Интеллектуальные системы и информационные технологии - 2023» («ИС & ИТ-2023», «IS&IT'23»). Научное издание в 2-х т. Т. 2. – Таганрог: Изд-во Ступина С.А., 2023. С. 3-9.

142. Некрасов И. В. Разработка и исследование метода классификации библиографической текстовой информации: дис. … канд. техн. наук: 05.13.01/Некрасов Иван Валериевич. – Москва, 2005. – 152 с.

143. Корелов С. В., Петров А. М., Сидоркина И. Г., Ротков Л. Ю. Применение весов термов в задаче обнаружения спама с использованием модели электронных писем // Труды XXVI научной конференции по радиофизике, посвященной 120-летию М. Т. Греховой, (Нижний Новгород, 12 – 27 мая 2022 г.). Нижний Новгород: ННГУ, 2022. С. 522-526.

144. Nokel M. A., Bolshakova E. I., Loukachevitch N. V. Combining Multiple Features for Single-Word Term Extraction // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). Т. 2: Доклады специальных секций. Вып. 11. – М.: РГГУ. 2012. PP. 490–501.

145. Нокель М. А., Лукашевич Н. В. Использование тематических моделей в извлечении однословных терминов // Selected Papers of the 15th All-Russian Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» (Yaroslavl, Russia, October 14-17). CEUR Workshop Proceedings. 2013. Vol. 1108. С. 52-60.

146. Salton G., Buckley C. Term-Weighting Approaches in Automatic Text Retrieval // Information Processing & Management. 1988. Vol. 24. Iss. 5. PP. 513-523. DOI:10.1016/0306-4573(88)90021-0.

147. Агеев М. С. Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов: дис. … канд. физ.-мат. наук: 05.13.11/Агеев Михаил Сергеевич. – Москва, 2004. – 136 с.

148. Church K., Gale W. Poisson mixtures // Natural Language Engineering. 1995. Vol. 1, Iss. 2. PP. 163-190. DOI:10.1017/S1351324900000139.

149. Yamamoto M., Church K. Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus // Computational Linguistics. 2001. Vol. 27, Iss. 1. PP. 1-30. DOI: 10.1162/089120101300346787.

150. Church K., Gale W. Inverse Document Frequency (IDF): A Measure of Deviations from Poisson // Natural Language Processing Using Very Large Corpora. 1999. Vol. 11. PP. 283-295. DOI:10.1007/978-94-017-2390-9_18.

151. Лукашевич Н. Г., Логачев Ю. М. Использование методов машинного обучения для извлечения слов-терминов // Труды XII национальной конференции по искусственному интеллекту с международным участием КИИ-2010 (20-24 сентября 2010 г., Тверь, Россия). Т. 1. – М.: Физматлит. 2010. С. 292–299.

152. Лукашевич Н. В. Модели и методы автоматической обработки неструктурированной информации на основе базы знаний онтологического типа: дис. … докт. техн. наук: 05.25.05/Лукашевич Наталья Валентиновна. – Москва, 2014. – 312 с.

153. Клышинский Э. С., Кочеткова Н. А. Метод извлечения технических терминов с использованием меры странности // Новые информационные технологии в автоматизированных системах: материалы семнадцатого научно-практического семинара. – М.: ИПМ им. М. В. Келдыша. 2014. № 17. С. 365-370.

154. Kurz D. and Xu F. Text Mining for the Extraction of Domain Retrieval Terms and Term Collocations // Proceedings of the International Workshop on Computational Approaches to Collocations (Vienna, Austria, July 22-23). 2002. URL:

https://www.coli.uni-saarland.de/publikationen/softcopies/Kurz:2002:TME.pdf (дата обращения 11.04.2022).

155. Xu F., Kurz D., Piskorski J. and Schmeier S. Term Extraction and Mining of Term Relations from Unrestricted Texts in the Financial Domain // In: 5th International Conference on Business Information Systems (Poznan, Poland). 2002. URL: https://www.coli.uni-saarland.de/publikationen/softcopies/Xu:2002:TEM.pdf (дата обращения 11.04.2022).

156. Петров А. С., Шульга Т. Э. Математическая модель русскоязычного текстового документа для решения задачи автоматического извлечения терминов из текста // Вестник Воронежского государственного университета. Сер.: Системный анализ и информационные технологии. 2017. № 3. С. 195-203.

157. Frantzi K., Ananiadou S., Mima H. Automatic recognition of multi-word terms: The C-value/NC-value method // International Journal on Digital Libraries. 2000. Vol. 3. Iss. 2. PP. 115-130. DOI:10.1007/3-540-49653-X_35.

158. Frantzi K., Ananiadou S. Automatic Term Recognition using Contextual Cues // Proceedings of Third DELOS Workshop: Cross Language Information Retrieval (Zurich, Switzerland, March 5-7). 1997. URL: https://www.ercim.eu/publication/ws-proceedings/DELOS3/Frantzi.pdf (дата обращения 11.04.2022).

159. Liu M., Yang J. An Improvement of TFIDF Weighting in Text Categorization // 2012 International Conference on Computer Technology and Science (ICCTS 2012). 2012. Vol. 47. PP. 44-47. DOI:10.7763/IPCSIT.2012.V47.9.

160. Чернопрудова Е. Н., Соловьев Н. А., Юркевская Л. А. Фильтрация несанкционированных сообщений в почтовых электронных сервисах // Моделирование, оптимизация и информационные технологии. Научный журнал. 2017. № 5 (4). URL: https://moitvivt.ru/ru/journal/pdf?id=403 (дата обращения 11.04.2022).

161. Агеев М. С., Добров Б. В., Лукашевич Н. В., Сидоров А. В., Штернов С. В. «Отправная точка» для дорожки по поиску в РОМИП (предварительный анализ). // Труды первого российского семинара по оценке методов информационного поиска. – СПб.: НИИ Химии СПбГУ, 2003. С. 87–109.

162. Агеев М. С., Добров Б. В., Лукашевич Н. В., Сидоров А. В. Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line» // Труды второго российского семинара по оценке методов информационного поиска. – СПб: НИИ Химии СПбГУ, 2004. С. 62–89.

163. Zhang Z., Brewster C., Ciravegna F. A Comparative Evaluation of Term Recognition Algorithms // Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08). 2008. PP. 2108-2113.

164. Браславский П. И., Соколов Е. А. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.). – М.: Изд-во РГГУ, 2006. С. 88-94.

165. Браславский П., Соколов Е. Сравнение пяти методов извлечения терминов произвольной длины // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4-8 июня 2008 г.). Вып. 7 (14). – М.: Изд-во РГГУ, 2008. С. 67-74

166. Nokel M., Loukachevitch N. An Experimental Study of Term Extraction for Real Information-Retrieval Thesauri // Proceedings of 10th International Conference on Terminology and Artificial Intelligence. 2013. PP. 69-76.

167. Bolshakova E., Loukachevitch N., Nokel M. Topic Models Can Improve Domain Term Extraction // Proceedings of ECIR 2013. 2013. Vol. 7814. PP. 684-687. DOI:10.1007/978-3-642-36973-5_60.

168. Loukachevitch N. Automatic Term Recognition Needs Multiple Evidence // Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). 2012. PP. 2401-2407.

169. Корелов С. В., Петров А. М., Сидоркина И. Г., Ротков Л. Ю. Применение весов термов в задаче обнаружения спама // Труды XXVII научной конференции по радиофизике (Нижний Новгород, 15 – 25 мая 2023 г.). Нижний Новгород: ННГУ, 2023. С. 516-521.

170. Allan J., Ballesteros L., Callan J. P., Croft W. B., Lu Z. Recent Experiments with INQUERY // Proceedings of the Fourth Text REtrieval Conference (TREC-4). Gaithersburg, MD: NIST Special Publication 500-236, 1996. PP. 49-63.

171. Broglio J., Callan J. P., Croft W. B., Nachbar D. W. Document Retrieval and Routing Using the INQUERY System // Proceedings of Third Text Retrieval Conference (TREC-3). Gaithersburg, MD: NIST Special Publication 500-225. 1999. PP. 29-38.

172. Callan J. P., Croft W. B., Harding S. M. The INQUERY Retrieval System // Database and Expert Systems Applications. 1992. PP. 78-83. DOI:10.1007/978-3-7091-7557-6_14.

173. Переобучение. [электронный ресурс] // ВикипедиЯ. Режим доступа: https://ru.wikipedia.org/wiki/Переобучение (дата обращения: 07.10.2019).

174. Yang Y., Pedersen J. A Comparative Study on Feature Selection in Text Categorization // Proceedings of the Fourteenth International Conference on Machine Learning (ICML ’97). 1997. PP. 412-420.

175. Forman G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification // Journal of Machine Learning Research. 2003. Vol. 3. PP. 1289-1305. DOI:10.1162/153244303322753670.

176. Forman G. Choose Your Words Carefully: An Empirical Study of Feature Selection Metrics for Text Classification // 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2002). Lecture Notes in Computer Science. Vol. 2431. PP. 150-162. DOI:10.1007/3-540-45681-3_13.

177. Simeon M., Hilderman R. Categorical Proportional Difference: A Feature Selection Method for Text Categorization // Proceedings of the Seventh Australasian Data Mining Conference (AusDM 2008). 2008. V. 87. PP. 201-208.

178. Zheng Zh., Wu X., Srihari R. Feature Selection for Text Categorization on Imbalanced Data // ACM Sigkdd Explorations Newsletter. 2004. Vol. 6. Iss. 1. PP. 80-89.

179. Yang Y., Liu X. A Re-Examination of Text Categorization Methods // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 2003. PP. 42-49. DOI:10.1145/312624.312647.

180. Gabrilovich E., Markovitch S. Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5 // Proceedings of the Twenty-First International Conference on Machine Learning, (ICML 2004). 2004. PP. 41. DOI:10.1145/1015330.1015388.

181. Nicolosi N. Feature Selection Methods for Text Classification // Department of Computer Science, Rochester Institute of Technology, Tech. Rep. 2008.

182. Dasgupta A., Drineas P., Harb B., Josifovski V., Mahoney M. Feature Selection Methods for Text Classification // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2007. PP. 230-239. DOI:10.1145/1281192.1281220.

183. Li Sh., Xia R., Zong Ch., Huang Ch. A Framework of Feature Selection Methods for Text Categorization // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009. Vol. 2. PP. 692-700. DOI:10.3115/1690219.1690243.

184. Zareapoor M., Seeja K. R. Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection // International Journal of Information Engineering and Electronic Business. 2015. Vol. 7. No. 2. PP. 60-65. DOI:10.5815/ijieeb.2015.02.08.

185. Ramya M., Pinakas J. Alwin Different Type of Feature Selection for Text Classification // International Journal of Computer Trends and Technology. 2014. Vol. 10. No. 2. PP. 102-107. DOI:10.14445/22312803/IJCTT-V10P118.

186. Uysal A. K. An improved global feature selection scheme for text classification // Expert Systems With Applications. 2016. Vol. 43. PP. 82-92. DOI:10.1016/j.eswa.2015.08.050.

187. Villacampa O. Feature Selection and Classification Methods for Decision Making: A Comparative Analysis. Diss. Nova Southeastern University, 2015.

188. Meng J., Lin H., Yu Y. A two-stage feature selection method for text categorization // Computers & Mathematics with Applications. 2011. Vol. 62. Iss. 7. PP. 2793-2800. DOI: 10.1016/j.camwa.2011.07.045.

189. Rogati M., Yang Y. High-Performing Feature Selection for Text Classification // Proceedings of the eleventh international conference on Information and knowledge management. 2002. PP. 659-661. DOI:10.1145/584792.584911.

190. Uysal A. K. Comparative Analysis of Recent Feature Selection Methods for Sentiment Classification // Anadolu University Journal of Science and Technology A-Applied Sciences and Engineering. 2018. Vol. 19 No. 3. PP. 645-659. DOI:10.18038/aubtda.412532.

191. Sahin D., Kilic E. Two new feature selection metrics for text classification // Automatika. 2019. Vol. 60. Iss. 2. PP. 162-171. DOI:10.1080/00051144.2019.1602293.

192. Shang W., Huang H., Zhu H., Lin Y., Qu Y., Wang Z. A novel feature selection algorithm for text categorization // Expert Systems with Applications, 2007. Vol. 33. No 1. PP. 1-5. DOI:10.1016/j.eswa.2006.04.001.

193. Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин) [электронный ресурс]. Режим доступа: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения: 20.12.2019).

194. Метод ближайших соседей [электронный ресурс]. // MachineLearning.ru – Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Метод_ближайших_соседей, свободный (дата обращения: 20.12.2019).

195. Перекрёстная проверка. [электронный ресурс]. // ВикипедиЯ. Режим доступа: https://ru.wikipedia.org/wiki/Перекрёстная_проверка, свободный (дата обращения: 07.10.2019).

196. ГОСТ Р 57100-2016. Системная и программная инженерия. Описание архитектуры. М., 2019. 31 с.

197. Назаров С. В. Архитектура и проектирование программных систем: монография /С. В. Назаров. – М.: ИНФРА-М, 2013. – 350 с.

198. Соловьев Н. А., Чернопрудова Е. Н., Тишина Н. А., Юркевская Л. А. Программное обеспечение защиты почтовых сервисов от несанкционированных рассылок на основе контентной фильтрации электронных сообщений: монография/ Н.А. Соловьев, Е.Н. Чернопрудова, Н.А. Тишина, Л.А. Юркевская. Оренбург, 2016. – 128 с.

199. Корелов С.В., Петров А. М., Сидоркина И. Г., Ротков Л. Ю. Подсистема классификации электронных писем на основе модели электронных писем // Информационные технологии обеспечения комплексной безопасности в цифровом обществе: сборник материалов VI Всероссийской молодежной научно-практической конференции с международным участием (г.Уфа, 19-20 мая 2023 года) / отв. ред. Д. С. Юнусова – Уфа: РИЦ УУНиТ, 2023. – С. 31-35.

200. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. 1965. – Т. 163. № 4. С. 845-848.

201. Реброва И. А. Теория планирования эксперимента [Электронный ресурс]: учебное пособие/ И. А. Реброва. – Омск: СибАДИ, 2016.

202. Юдин Ю. В., Майсурадзе М. В., Водолазский Ф. В. Организация и математическое планирование эксперимента: учебное пособие – Екатеринбург: Изд-во Урал. ун-та, 2018. – 124 с.

203. Mohammad A. H., Alwada’n T. Email Filtering Using Hybrid Feature Selection Model // CMES-Computer Modeling in Engineering & Sciences, 2022. Vol. 132. No 2. PP. 435-450. DOI:10.32604/cmes.2022.020088.

Приложение А. Результаты эксперимента по выбору значений длины выборки в модели

Таблица А.1 – Результаты эксперимента на англоязычных письмах

n	Легальные письма			Спам			Набор в целом		
	R	P	F-мера	R	P	F-мера	R	P	F-мера
1	0,947	0,977	0,962	0,972	0,956	0,964	0,959	0,966	0,963
2	0,915	0,963	0,939	0,904	0,966	0,934	0,909	0,964	0,936
3	0,689	0,980	0,809	0,805	0,980	0,884	0,748	0,980	0,848
4	0,532	0,991	0,692	0,709	0,986	0,825	0,621	0,989	0,763
5	0,466	0,995	0,634	0,647	0,993	0,783	0,557	0,994	0,714
6	0,389	0,992	0,559	0,607	0,975	0,748	0,499	0,982	0,662
7	0,382	0,995	0,552	0,577	0,987	0,728	0,480	0,990	0,647
8	0,377	0,993	0,546	0,553	0,989	0,709	0,466	0,991	0,634
9	0,363	0,993	0,532	0,534	0,990	0,693	0,449	0,991	0,618
10	0,325	0,993	0,490	0,518	0,990	0,680	0,422	0,991	0,592
11	0,300	0,993	0,460	0,497	0,989	0,661	0,399	0,990	0,569
12	0,281	0,994	0,438	0,479	0,985	0,644	0,381	0,988	0,550
13	0,270	0,988	0,424	0,462	0,989	0,630	0,367	0,989	0,535
14	0,257	0,992	0,408	0,445	0,986	0,613	0,352	0,988	0,519
15	0,245	0,992	0,393	0,431	0,992	0,601	0,339	0,992	0,505
16	0,214	0,994	0,352	0,425	0,962	0,589	0,320	0,972	0,482
17	0,209	0,992	0,345	0,418	0,988	0,587	0,314	0,989	0,477
18	0,199	0,993	0,332	0,410	0,990	0,579	0,305	0,991	0,467
19	0,188	0,994	0,317	0,407	98,99	0,576	0,299	0,991	0,459
20	0,184	0,993	0,310	0,396	0,994	0,566	0,291	0,994	0,450

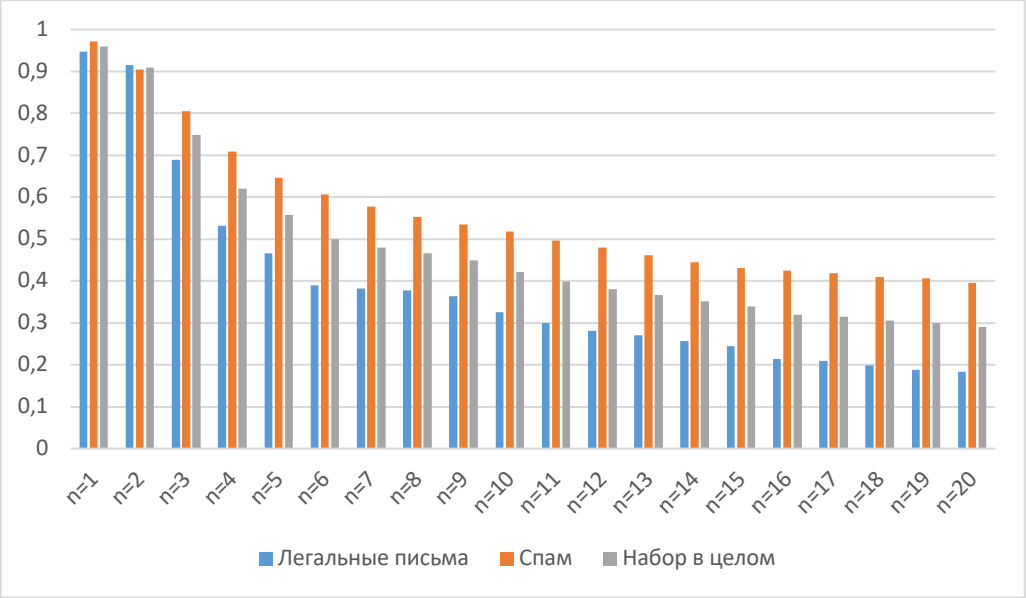

Рисунок А.1 – Полнота обнаружения R на англоязычных письмах

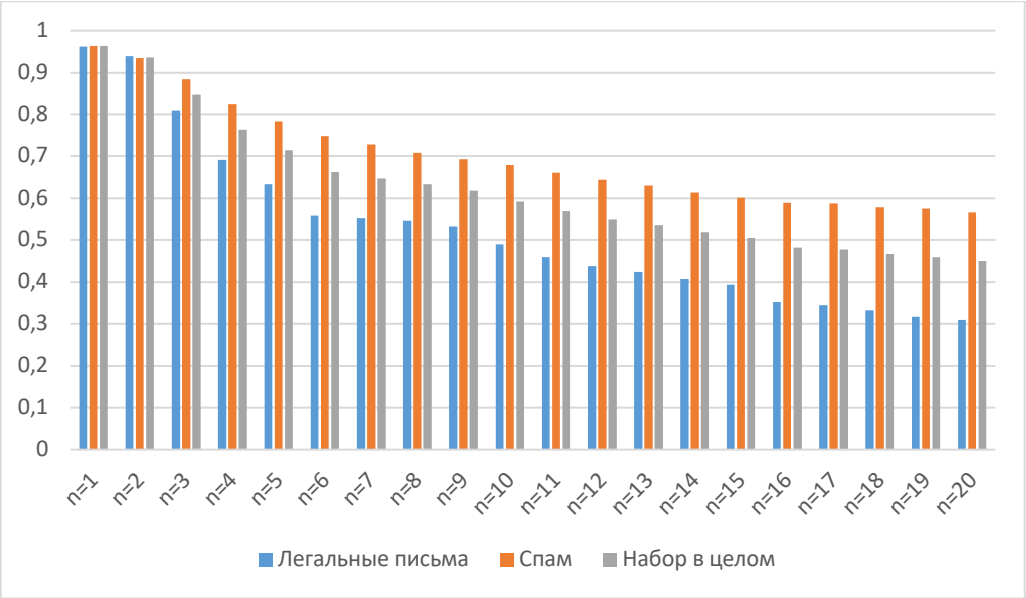

Рисунок А.2 – Значения F-меры на англоязычных письмах

Таблица А.2 – Результаты эксперимента на русскоязычных письмах

n	Легальные письма			Спам			Набор в целом		
	R	P	F-мера	R	P	F-мера	R	P	F-мера
1	0,998	0,819	0,900	0,905	0,999	0,949	0,930	0,937	0,934
2	0,740	0,960	0,836	0,901	0,901	0,901	0,856	0,915	0,884
3	0,718	0,999	0,836	0,682	0,865	0,763	0,692	0,900	0,782
4	0,668	1	0,801	0,587	0,862	0,698	0,610	0,900	0,727
5	0,786	0,999	0,880	0,515	0,940	0,665	0,591	0,961	0,732
6	0,738	1	0,850	0,477	0,970	0,640	0,550	0,981	0,705
7	0,774	1	0,873	0,456	1	0,626	0,544	1	0,705
8	0,763	0,999	0,865	0,441	1	0,612	0,531	1	0,693
9	0,410	1	0,581	0,425	0,999	0,597	0,421	1	0,592
10	0,335	0,998	0,502	0,408	0,999	0,579	0,388	0,999	0,558
11	0,317	1	0,482	0,381	1	0,552	0,363	1	0,533
12	0,311	1	0,474	0,367	1	0,537	0,352	1	0,520
13	0,316	1	0,480	0,361	1	0,531	0,348	1	0,517
14	0,306	1	0,469	0,357	1	0,526	0,342	1	0,510
15	0,316	1	0,481	0,352	1	0,521	0,342	1	0,510
16	0,268	1	0,423	0,345	1	0,513	0,324	1	0,489
17	0,227	1	0,370	0,339	1	0,507	0,308	1	0,471
18	0,231	1	0,375	0,333	1	0,500	0,305	1	0,467
19	0,223	1	0,365	0,329	1	0,495	0,299	1	0,461
20	0,209	1	0,345	0,317	1	0,481	0,287	1	0,445

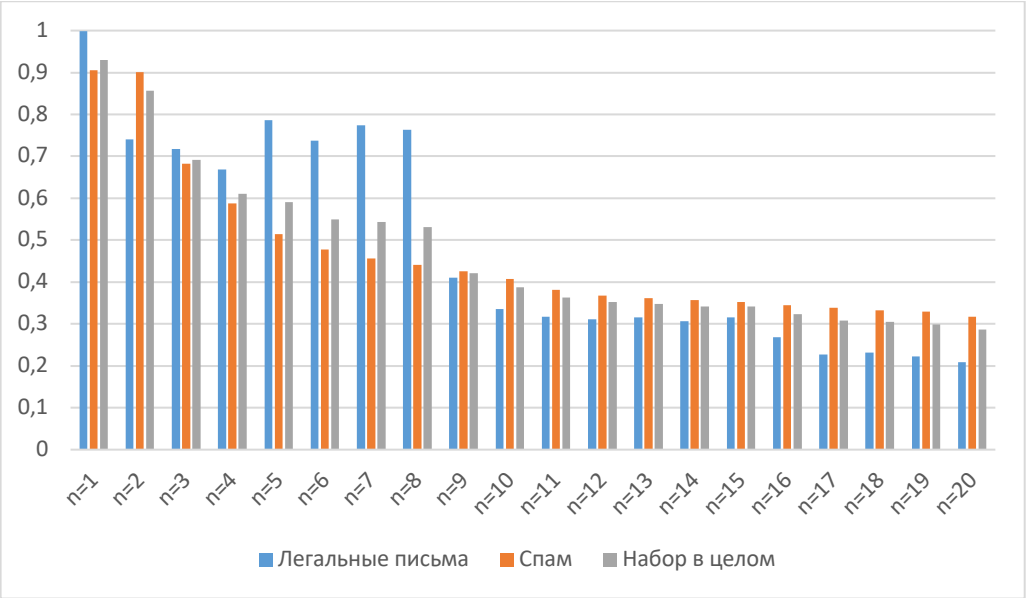

Рисунок А.3 – Полнота обнаружения R на русскоязычных письмах

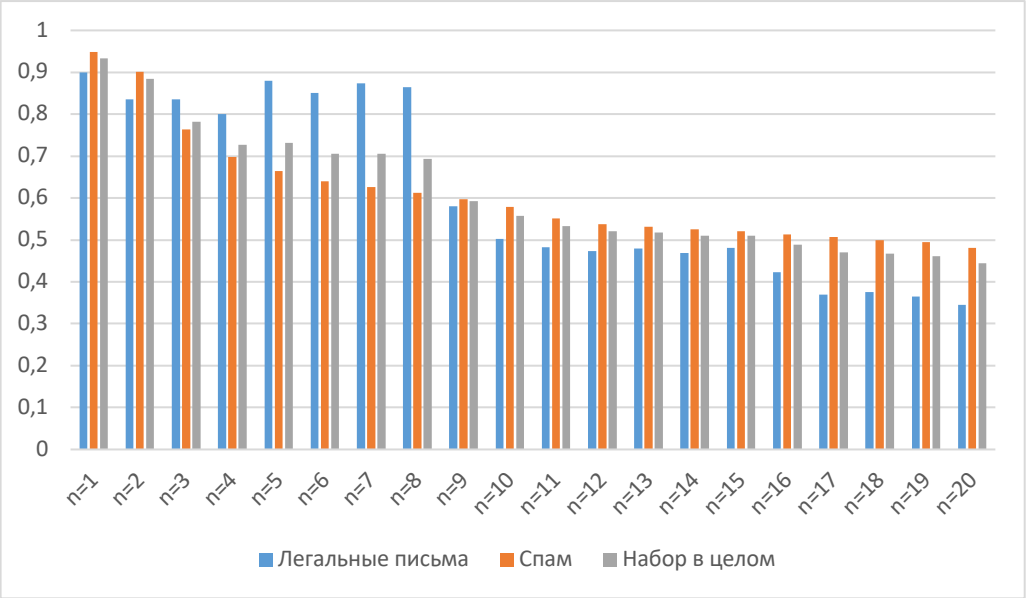

Рисунок А.4 – Значения F-меры на русскоязычных письмах

Приложение Б. Результаты эксперимента по выбору размера кодовой таблицы модели

Таблица Б.1 – Результаты эксперимента на англоязычных письмах

q	полнота R		точность P		F-мера	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
256	0,947	0,915	0,977	0,963	0,962	0,939
224	0,943	0,923	0,978	0,962	0,960	0,942
192	0,917	0,933	0,979	0,958	0,947	0,945
160	0,918	0,938	0,977	0,960	0,947	0,949
128	0,900	0,941	0,978	0,958	0,938	0,950
96	0,831	0,948	0,975	0,957	0,897	0,952
64	0,660	0,947	0,955	0,956	0,781	0,951
32	0,214	0,898	0,704	0,952	0,329	0,924
Спам						
256	0,972	0,904	0,956	0,966	0,964	0,934
224	0,972	0,907	0,953	0,967	0,963	0,936
192	0,975	0,915	0,932	0,967	0,953	0,940
160	0,974	0,920	0,932	0,969	0,953	0,944
128	0,975	0,920	0,920	0,969	0,946	0,944
96	0,969	0,928	0,867	0,968	0,915	0,947
64	0,954	0,938	0,766	0,964	0,849	0,951
32	0,573	0,945	0,631	0,917	0,601	0,931
Набор в целом						
256	0,959	0,909	0,966	0,964	0,963	0,936
224	0,958	0,915	0,965	0,964	0,961	0,939
192	0,946	0,924	0,954	0,962	0,950	0,943
160	0,946	0,929	0,953	0,964	0,950	0,946
128	0,938	0,930	0,947	0,964	0,942	0,947
96	0,901	0,938	0,913	0,962	0,907	0,950
64	0,809	0,942	0,832	0,960	0,820	0,951
32	0,395	0,922	0,649	0,934	0,491	0,928

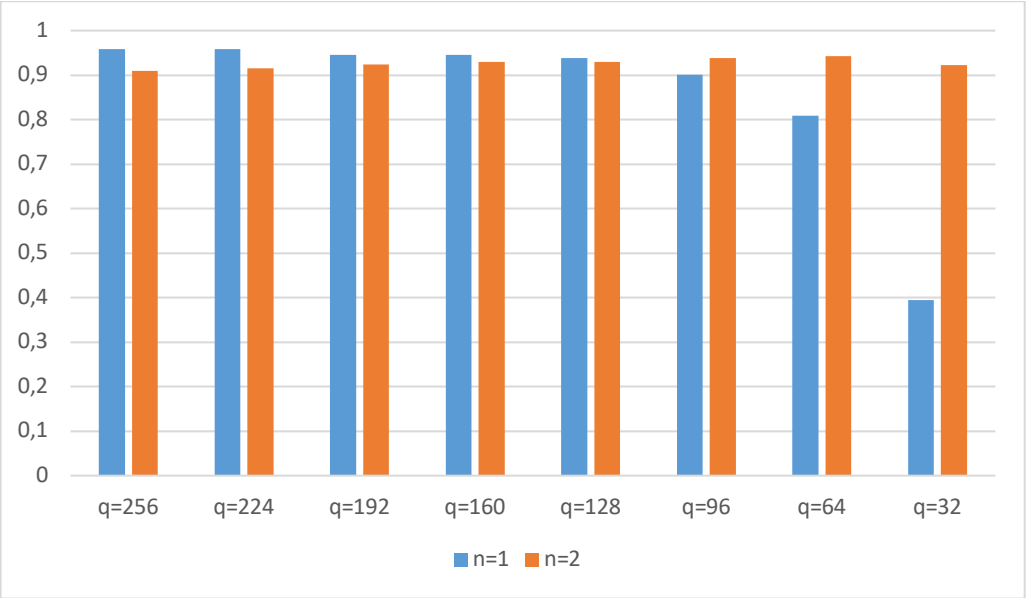

Рисунок Б.1 – Полнота обнаружения R на англоязычных письмах (в целом)

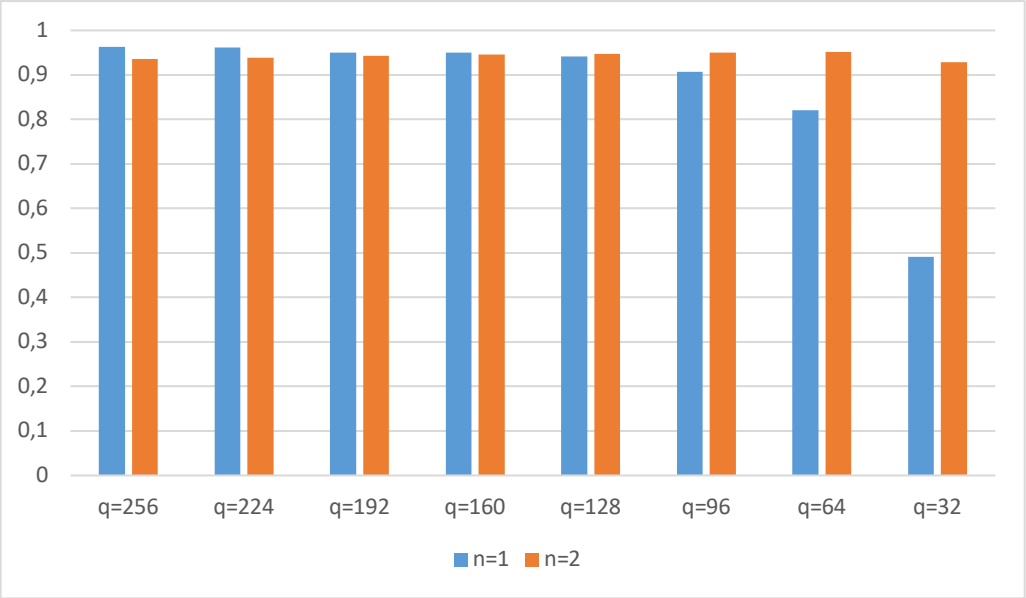

Рисунок Б.2 – Значения F-меры на англоязычных письмах (в целом)

Таблица Б.2 – Результаты эксперимента на русскоязычных письмах

q	полнота R		точность P		F-мера	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
256	0,999	0,998	0,744	0,942	0,853	0,970
224	0,998	0,998	0,680	0,922	0,809	0,959
192	0,998	0,998	0,661	0,922	0,795	0,958
160	0,998	1	0,558	0,876	0,716	0,934
128	0,998	1	0,520	0,892	0,684	0,943
96	0,998	0,998	0,418	0,861	0,589	0,925

q	полнота R		точность P		F-мера	
	$n = 1$	$n = 2$	$n = 1$	$n = 2$	$n = 1$	$n = 2$
64	0,998	0,999	0,309	0,831	0,472	0,908
32	0,998	0,998	0,279	0,568	0,436	0,724
Спам						
256	0,853	0,807	1	1	0,921	0,893
224	0,803	0,823	0,999	1	0,890	0,903
192	0,785	0,840	0,999	0,999	0,879	0,916
160	0,677	0,843	0,999	1	0,807	0,915
128	0,627	0,856	0,999	1	0,771	0,921
96	0,449	0,865	0,999	1	0,620	0,928
64	0,126	0,877	0,995	1	0,223	0,934
32	0,004	0,681	0,800	0,999	0,007	0,810
Набор в целом						
256	0,894	0,860	0,903	0,980	0,898	0,916
224	0,857	0,872	0,867	0,973	0,862	0,920
192	0,844	0,884	0,855	0,974	0,850	0,926
160	0,767	0,887	0,776	0,958	0,772	0,921
128	0,731	0,894	0,740	0,964	0,735	0,928
96	0,602	0,902	0,608	0,952	0,605	0,927
64	0,369	0,911	0,373	0,941	0,371	0,926
32	0,281	0,769	0,281	0,784	0,281	0,777

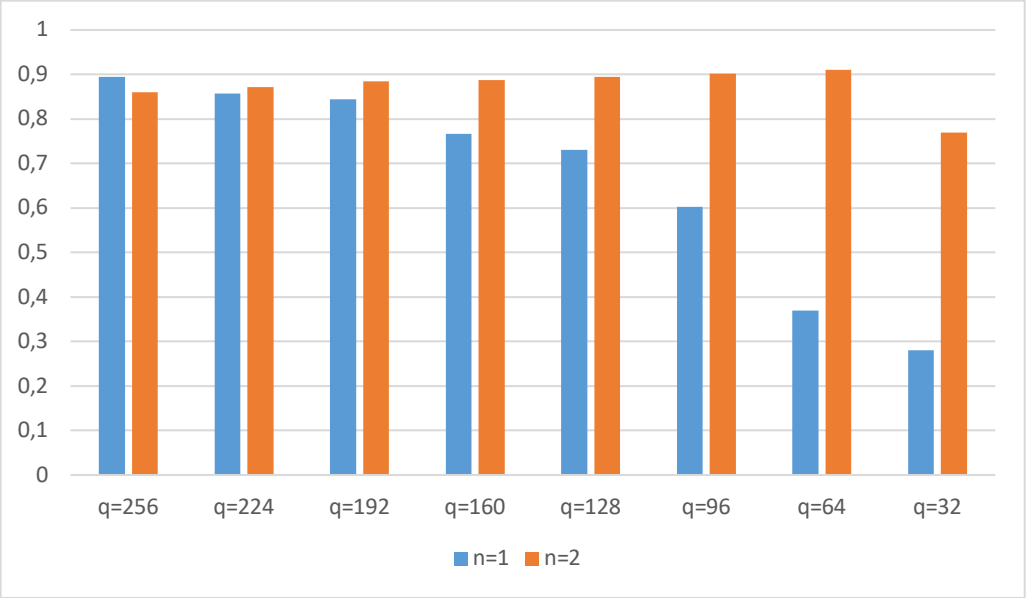

Рисунок Б.3 – Полнота обнаружения R на русскоязычных письмах (в целом)

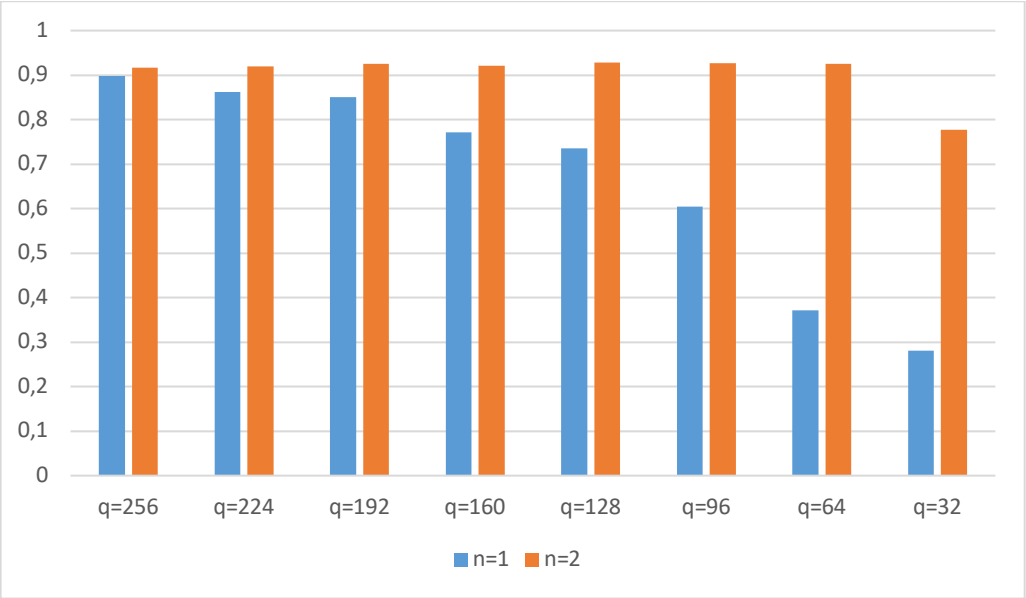

Рисунок Б.4 – Значения F-меры на русскоязычных письмах (в целом)

Таблица Б.3 – Дополнительные расчеты для англоязычных писем

q форм./факт	$n = 1$		$n = 2$	
	Общ. кол. термов	Доля дубл.	Общ. кол. термов	Доля дубл.
Легальные письма, 16 100 шт.				
256/70	5 759 020	0,942	1 552 107	0,495
224/62	5 882 742	0,945	1 610 497	0,504
192/53	6 177 100	0,955	1 762 040	0,527
160/45	6 213 628	0,958	1 821 655	0,536
128/37	6 387 774	0,963	1 899 117	0,551
96/29	7 080 916	0,978	2 120 066	0,582
64/21	8 203 543	0,994	2 601 061	0,670
32/12	10 444 625	0,999	3 617 634	0,871
Спам, 16 420 писем				
256/74	4 227 358	0,907	1 135 505	0,591
224/65	4 345 352	0,913	1 176 315	0,595
192/56	4 601 888	0,928	1 291 107	0,610
160/48	4 643 083	0,932	1 346 455	0,617
128/39	4 801 385	0,942	1 396 520	0,622
96/31	5 318 615	0,967	1 587 667	0,646
64/23	6 250 451	0,992	1 965 054	0,707
32/13	8 045 896	0,999	2 766 282	0,873
Набор в целом				
256/74	9 986 378	0,936	2 687 612	0,541
224/65	10 228 094	0,940	2 786 812	0,548
192/56	10 778 988	0,952	3 053 147	0,569
160/48	10 856 711	0,955	3 168 110	0,578
128/39	11 189 159	0,961	3 295 637	0,590
96/31	12 399 531	0,979	3 707 733	0,620
64/23	14 453 994	0,995	4 565 115	0,701
32/13	18 490 521	0,999	6 383 916	0,888

Таблица Б.4 – Дополнительные расчеты для русскоязычных писем

q форм./факт	$n = 1$		$n = 2$	
	Общ. кол. термов	Доля дубл.	Общ. кол. термов	Доля дубл.
Легальные письма, 16 100 шт.				
256/120	694 398	0,781	177 781	0,326
224/109	736 430	0,800	186 756	0,336
192/95	762 006	0,810	199 297	0,357
160/83	852 064	0,845	221 128	0,392
128/68	906 830	0,863	229 184	0,401
96/55	1 039 324	0,906	283 272	0,481
64/40	1 267 652	0,954	364 257	0,567
32/23	1 744 542	0,993	575 354	0,788
Спам, 16 420 писем				
256/132	1 024 814	0,848	294 868	0,622
224/120	1 098 135	0,863	312 487	0,629
192/105	1 115 810	0,866	325 949	0,633
160/92	1 256 805	0,891	353 311	0,640
128/77	1 340 844	0,904	367 377	0,646
96/61	1 505 785	0,929	434 722	0,674
64/43	1 797 348	0,962	545 806	0,716
32/24	2 458 700	0,994	836 033	0,849
Набор в целом				
256/134	1 719 212	0,833	472 649	0,512
224/122	1 834 565	0,849	499 243	0,522
192/107	1 877 816	0,855	525 246	0,530
160/94	2 108 869	0,883	574 439	0,548
128/78	2 247 674	0,897	596 561	0,556
96/62	2 545 109	0,928	717 994	0,603
64/43	3 065 000	0,964	910 063	0,665
32/24	4 203 242	0,995	1 411 387	0,837

Таблица Б.5 – Дополнительные расчеты для случайной последовательности англоязычных символов

q форм./факт	$n = 1$		$n = 2$	
	Общ. кол. термов	Доля дубл.	Общ. кол. термов	Доля дубл.
Последовательность символов длиной 26 Мб				
256/70	2 625 067	0,047	299 347	0,0001
224/62	2 794 167	0,058	337 450	0,0001
192/53	3 027 987	0,077	393 912	0,0002
160/45	3 296 772	0,107	462 536	0,0005
128/37	3 648 898	0,151	560 297	0,0009
96/29	4 143 789	0,226	708 978	0,0019
64/21	4 909 447	0,391	968 559	0,0066
32/12	6 608 901	0,809	1 643 876	0,0435

Таблица Б.6 – Дополнительные расчеты для случайной последовательности русскоязычных символов

q форм./факт		$n = 1$		$n = 2$	
		Общ. кол. термов	Доля дубл.	Общ. кол. термов	Доля дубл.
Последовательность символов длиной 4 Мб					
256/120		326 648	0,0109	28 961	0
224/109		343 524	0,0135	31 736	0
192/95		368 192	0,0173	36 423	0
160/83		394 568	0,0232	41 662	0
128/68		437 216	0,0349	50 755	0
96/55		487 857	0,0517	62 531	0,0001
64/40		575 549	0,0913	85 369	0,0003
32/23		768 208	0,255	145 875	0,0026

Приложение В. Результаты эксперимента по комбинированию значений параметра n модели

Таблица В.1 – Результаты эксперимента на англоязычных письмах

n	Легальные письма			Спам			Набор в целом		
	R	P	F-мера	R, %	P	F-мера	R	P	F-мера
1	0,947	0,977	0,962	0,972	0,956	0,964	0,959	0,966	0,963
1 ÷ 2	0,956	0,958	0,957	0,955	0,959	0,957	0,955	0,959	0,957
1 ÷ 3	0,952	0,958	0,955	0,955	0,956	0,956	0,954	0,957	0,956
1 ÷ 4	0,945	0,960	0,952	0,957	0,956	0,957	0,951	0,958	0,954
1 ÷ 5	0,953	0,960	0,957	0,958	0,956	0,957	0,955	0,958	0,957

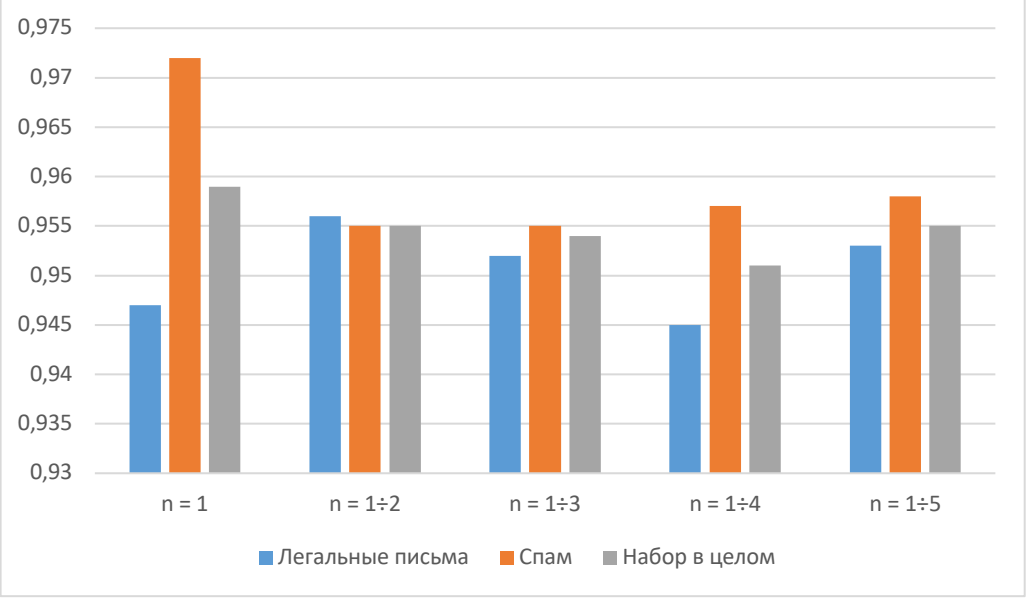

Рисунок В.1 – Полнота обнаружения R на англоязычных письмах

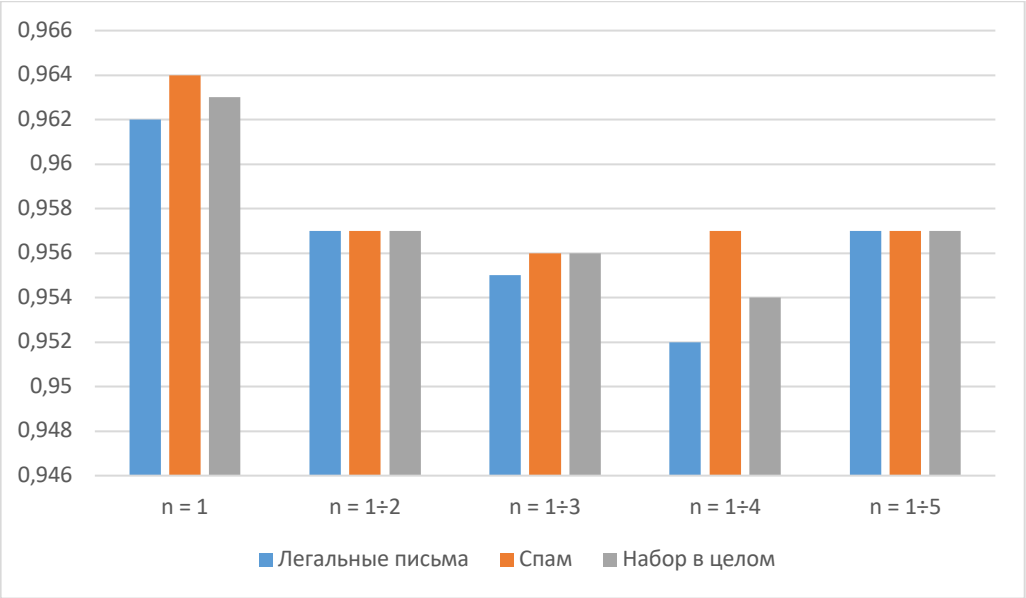

Рисунок В.2 – Значения F-меры на англоязычных письмах

Приложение Г. Результаты эксперимента по выбору способов предварительной обработки

Таблица Г.1 – Результаты эксперимента на англоязычных письмах

№ п/о	полнота R		точность P		F-мера	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
1	0,947	0,915	0,977	0,963	0,962	0,939
2а	0,941	0,921	0,978	0,962	0,959	0,941
2б	0,943	0,853	0,980	0,973	0,961	0,909
2в	0,947	0,915	0,977	0,963	0,962	0,939
2г	0,942	0,917	0,979	0,962	0,960	0,939
2д	0,949	0,834	0,969	0,968	0,959	0,896
3а	0,947	0,915	0,977	0,963	0,962	0,939
3б	0,941	0,921	0,978	0,962	0,959	0,941
3в	0,942	0,917	0,979	0,962	0,960	0,939
3г	0,949	0,834	0,969	0,968	0,959	0,896
3д	0,936	0,931	0,979	0,962	0,957	0,946
3е	0,943	0,843	0,965	0,965	0,954	0,900
3ж	0,946	0,809	0,969	0,963	0,958	0,879
3з	0,940	0,814	0,964	0,959	0,952	0,881
Спам						
1	0,972	0,904	0,956	0,966	0,964	0,934
2а	0,971	0,908	0,951	0,968	0,961	0,937
2б	0,974	0,895	0,954	0,959	0,964	0,926
2в	0,972	0,904	0,956	0,966	0,964	0,934
2г	0,973	0,905	0,952	0,964	0,962	0,934
2д	0,960	0,873	0,959	0,956	0,960	0,913
3а	0,972	0,904	0,956	0,966	0,964	0,934
3б	0,971	0,908	0,951	0,968	0,961	0,937
3в	0,973	0,905	0,952	0,964	0,962	0,934
3г	0,960	0,873	0,959	0,956	0,960	0,913
3д	0,967	0,912	0,947	0,968	0,957	0,939
3е	0,956	0,877	0,954	0,954	0,955	0,914
3ж	0,959	0,862	0,958	0,949	0,958	0,903
3з	0,953	0,867	0,953	0,945	0,953	0,904
Набор в целом						
1	0,959	0,909	0,966	0,964	0,963	0,936
2а	0,956	0,915	0,964	0,965	0,960	0,939
2б	0,959	0,874	0,967	0,966	0,963	0,918
2в	0,959	0,909	0,966	0,964	0,963	0,936
2г	0,958	0,911	0,965	0,963	0,961	0,936
2д	0,955	0,854	0,964	0,962	0,959	0,905
3а	0,959	0,909	0,966	0,964	0,963	0,936
3б	0,956	0,915	0,964	0,965	0,960	0,939
3в	0,958	0,911	0,965	0,963	0,961	0,936
3г	0,955	0,854	0,964	0,962	0,959	0,905
3д	0,952	0,921	0,962	0,965	0,957	0,942

№ п/о	полнота R		точность P		F-мера	
	$n = 1$	$n = 2$	$n = 1$	$n = 2$	$n = 1$	$n = 2$
3е	0,950	0,860	0,959	0,959	0,955	0,907
3ж	0,953	0,836	0,964	0,956	0,958	0,892
3з	0,947	0,841	0,958	0,952	0,952	0,893

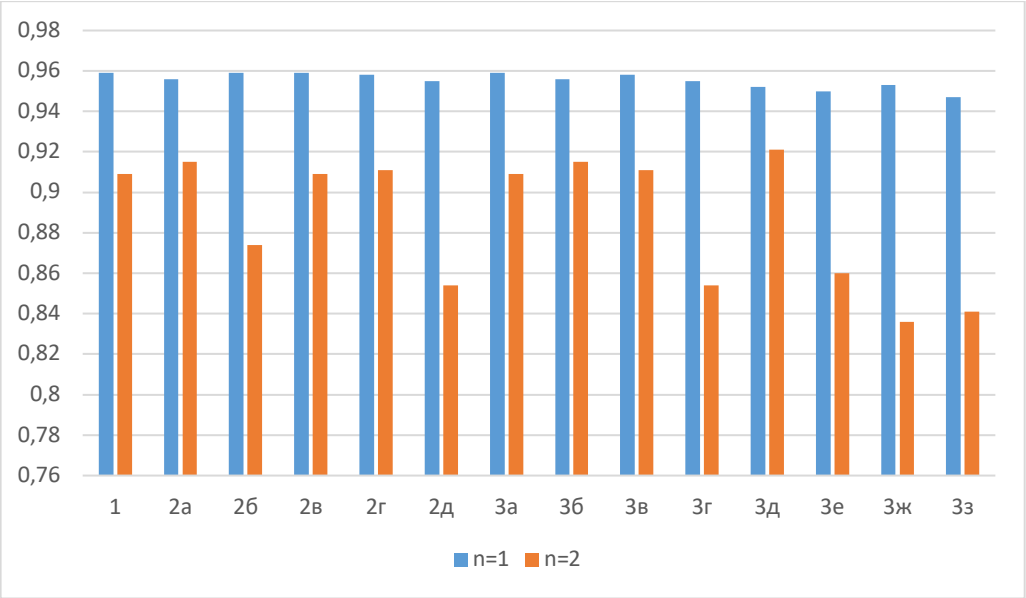

Рисунок Г.1 – Полнота обнаружения R на англоязычных письмах (в целом)

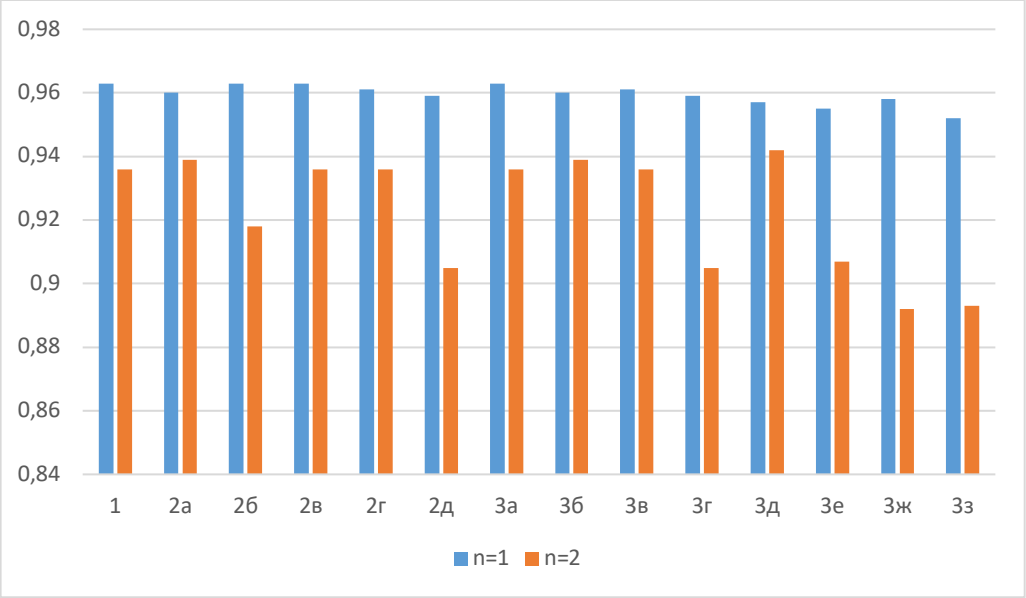

Рисунок Г.2 – Значения F-меры на англоязычных письмах (в целом)

Таблица Г.2 – Результаты эксперимента на русскоязычных письмах

№ п/о	полнота R		точность P		F-мера	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
1	0,998	0,740	0,819	0,960	0,900	0,834
2а	0,998	0,734	0,756	0,959	0,860	0,831
2б	0,998	0,734	0,809	0,972	0,894	0,836
2в	0,998	0,740	0,819	0,961	0,900	0,836
2г	0,999	0,750	0,691	0,939	0,817	0,834
2д	0,998	0,998	0,814	0,980	0,897	0,989
3а	0,998	0,754	0,818	0,965	0,899	0,846
3б	0,998	0,751	0,750	0,956	0,856	0,841
3в	0,999	0,759	0,674	0,943	0,805	0,841
3г	0,998	0,998	0,813	0,980	0,896	0,989
3д	0,999	0,998	0,602	0,938	0,751	0,967
3е	0,999	0,998	0,809	0,973	0,894	0,985
3ж	0,998	0,997	0,693	0,946	0,818	0,971
3з	0,999	0,996	0,698	0,966	0,821	0,981
Спам						
1	0,905	0,901	0,999	0,900	0,949	0,901
2а	0,863	0,906	0,999	0,900	0,926	0,903
2б	0,900	0,901	0,999	0,899	0,947	0,900
2в	0,904	0,901	0,999	0,900	0,949	0,901
2г	0,810	0,842	1	0,898	0,895	0,869
2д	0,897	0,878	0,999	0,999	0,945	0,933
3а	0,903	0,915	0,999	0,907	0,949	0,911
3б	0,858	0,921	0,999	0,907	0,923	0,914
3в	0,799	0,862	1	0,904	0,888	0,882
3г	0,895	0,895	0,999	0,999	0,944	0,944
3д	0,730	0,858	1	0,999	0,844	0,923
3е	0,895	0,902	1	0,999	0,945	0,948
3ж	0,811	0,753	0,999	1	0,895	0,859
3з	0,808	0,772	1	0,999	0,893	0,871
Набор в целом						
1	0,930	0,856	0,937	0,915	0,934	0,884
2а	0,900	0,858	0,909	0,913	0,905	0,885
2б	0,927	0,854	0,933	0,916	0,930	0,884
2в	0,930	0,856	0,937	0,915	0,934	0,885
2г	0,863	0,816	0,874	0,908	0,868	0,860
2д	0,925	0,909	0,935	0,993	0,930	0,949
3а	0,930	0,870	0,937	0,920	0,933	0,895
3б	0,897	0,874	0,906	0,919	0,901	0,896
3в	0,855	0,834	0,864	0,913	0,859	0,872
3г	0,924	0,924	0,935	0,993	0,929	0,957
3д	0,805	0,897	0,814	0,979	0,809	0,936
3е	0,924	0,929	0,933	0,991	0,929	0,959
3ж	0,863	0,821	0,875	0,981	0,869	0,894
3з	0,861	0,835	0,877	0,988	0,869	0,905

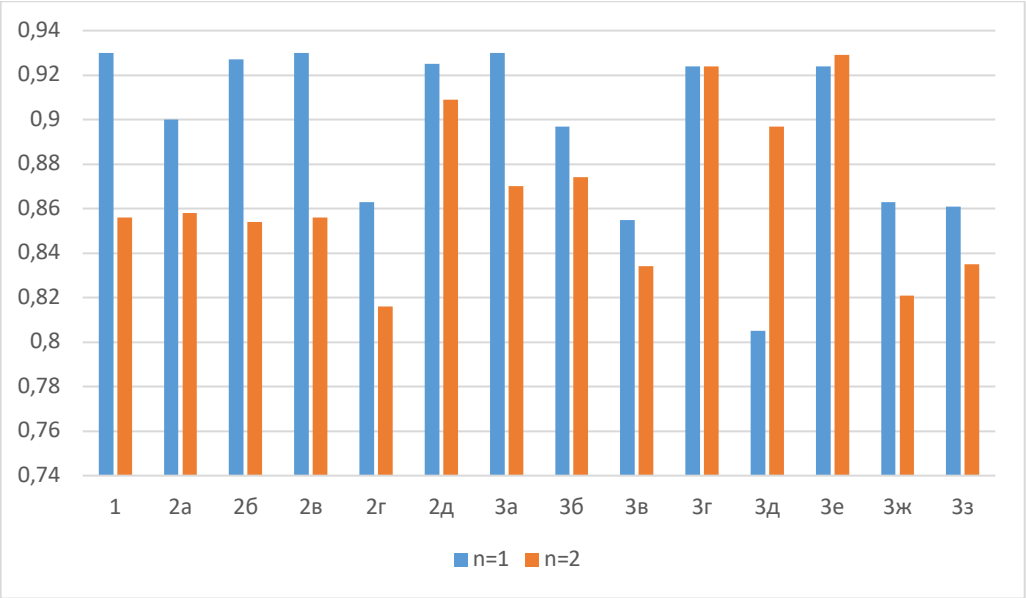

Рисунок Г.3 – Полнота обнаружения R на русскоязычных письмах (в целом)

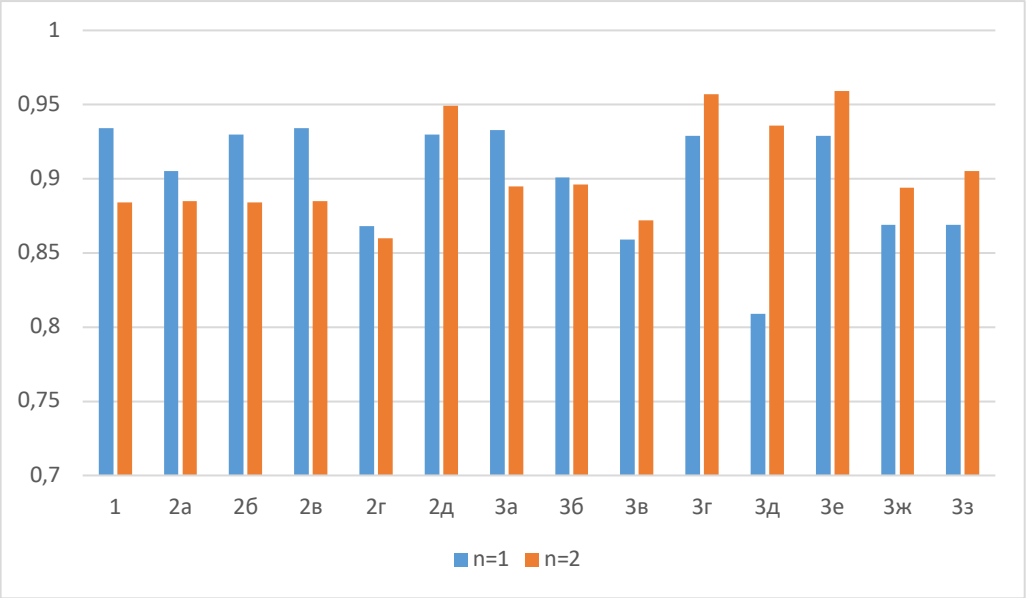

Рисунок Г.4– Значения F-меры на русскоязычных письмах (в целом)

Таблица Г.3 – Результаты эксперимента на русскоязычных письмах (с предварительным удалением повторений пробелов)

№ п/о	полнота R		точность P		F-мера	
	$n=1$	$n=2$	$n=1$	$n=2$	$n=1$	$n=2$
Легальные письма						
1	0,998	0,998	0,813	0,961	0,896	0,979
2а	0,998	0,997	0,808	0,960	0,893	0,978
2б	0,998	0,998	0,815	0,977	0,897	0,987
2в	0,998	0,998	0,811	0,962	0,895	0,980
2г	0,999	0,996	0,770	0,918	0,870	0,956
2д	0,998	0,998	0,814	0,980	0,897	0,989
3а	0,999	0,998	0,813	0,970	0,896	0,984
3б	0,999	0,998	0,803	0,961	0,890	0,979
3в	0,999	0,996	0,757	0,932	0,861	0,963
3г	0,998	0,998	0,813	0,980	0,896	0,989
3д	0,999	0,997	0,732	0,927	0,845	0,961
3е	0,999	0,998	0,809	0,973	0,894	0,985
3ж	0,998	0,997	0,693	0,946	0,818	0,971
3з	0,999	0,996	0,697	0,966	0,821	0,981
Спам						
1	0,899	0,884	0,999	0,999	0,946	0,938
2а	0,897	0,890	0,999	0,999	0,945	0,941
2б	0,902	0,885	0,999	0,999	0,948	0,939
2в	0,899	0,883	0,999	0,999	0,947	0,938
2г	0,873	0,775	1	0,999	0,932	0,873
2д	0,897	0,875	0,999	0,999	0,945	0,933
3а	0,900	0,901	1	0,999	0,947	0,947
3б	0,894	0,901	1	0,999	0,944	0,951
3в	0,860	0,807	1	1	0,925	0,893
3г	0,895	0,895	0,999	0,999	0,944	0,944
3д	0,843	0,819	1	0,999	0,915	0,900
3е	0,895	0,902	1	0,999	0,945	0,948
3ж	0,811	0,753	0,999	1	0,895	0,859
3з	0,808	0,772	1	0,999	0,893	0,871
Набор в целом						
1	0,926	0,916	0,935	0,987	0,931	0,950
2а	0,925	0,920	0,933	0,987	0,929	0,952
2б	0,929	0,917	0,936	0,992	0,932	0,953
2в	0,927	0,915	0,934	0,988	0,931	0,950
2г	0,909	0,837	0,916	0,971	0,912	0,899
2д	0,925	0,909	0,935	0,993	0,930	0,949
3а	0,928	0,928	0,935	0,990	0,931	0,958
3б	0,923	0,933	0,931	0,988	0,927	0,959
3в	0,899	0,860	0,909	0,977	0,904	0,915
3г	0,924	0,924	0,935	0,993	0,929	0,957
3д	0,887	0,869	0,897	0,975	0,892	0,919
3е	0,924	0,929	0,933	0,991	0,929	0,959
3ж	0,863	0,821	0,875	0,981	0,869	0,894
3з	0,861	0,835	0,877	0,988	0,869	0,905

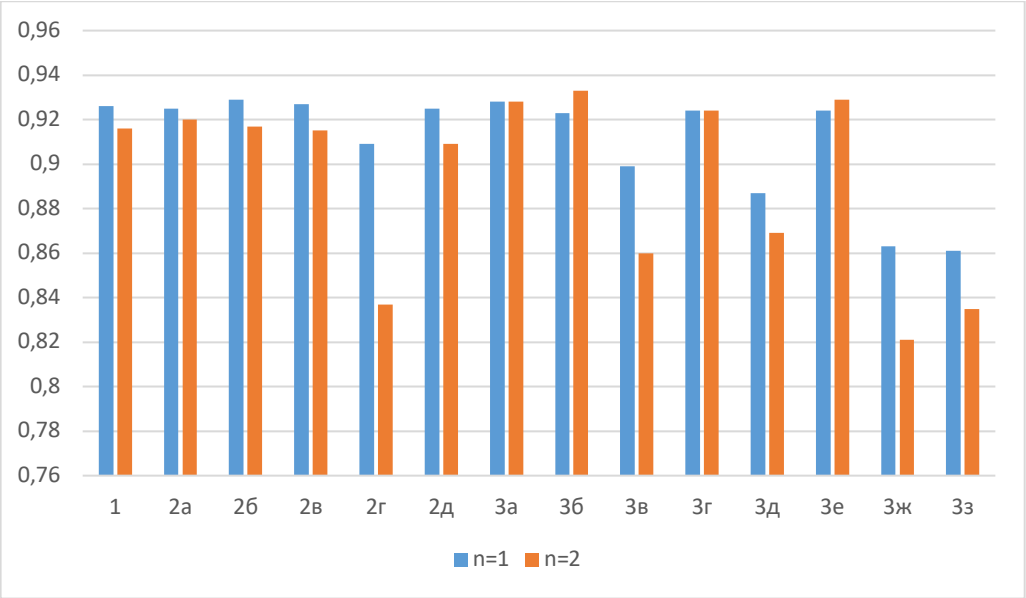

Рисунок Г.5 – Полнота обнаружения R на русскоязычных письмах (с предварительным удалением повторений пробелов, в целом)

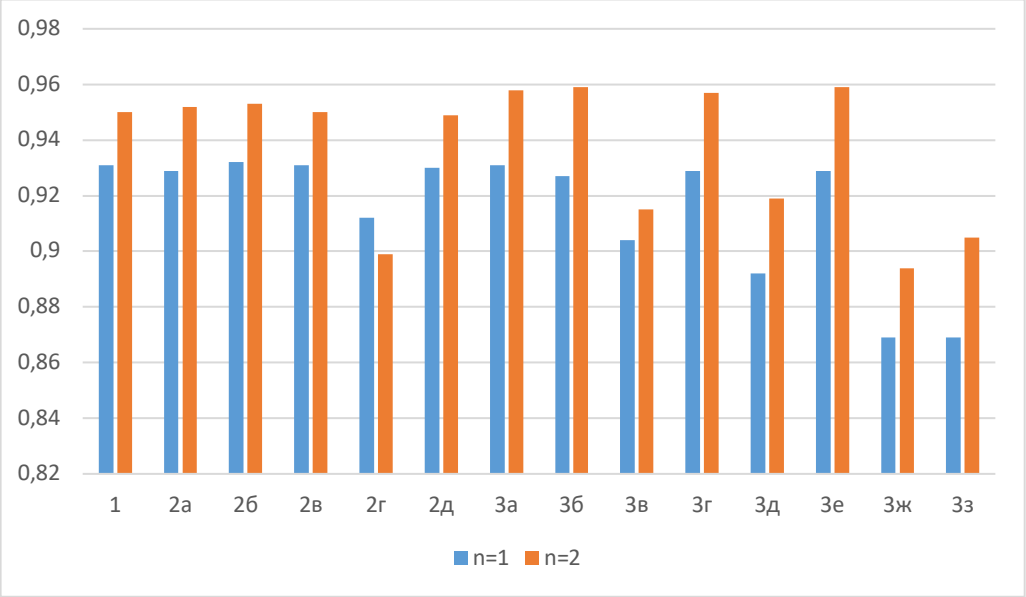

Рисунок Г.6 – Значения F-меры на русскоязычных письмах (с предварительным удалением повторений пробелов, в целом)

Приложение Д. Результаты эксперимента по обоснованию неслучайности результатов обнаружения спама с применением разработанной модели

Таблица Д.1 – Результаты на англоязычных письмах

n	полнота R		точность P		F-мера	
	модель	«псевдослуч.»	модель	«псевдослуч.»	модель	«псевдослуч.»
Легальные письма						
1	0,947	0,873	0,977	0,869	0,962	0,871
2	0,915	0,758	0,963	0,780	0,939	0,769
3	0,689	0,465	0,980	0,899	0,809	0,613
Спам						
1	0,972	0,854	0,956	0,882	0,964	0,868
2	0,904	0,551	0,966	0,867	0,934	0,674
3	0,805	0,379	0,980	0,917	0,884	0,536
Набор в целом						
1	0,959	0,864	0,966	0,875	0,963	0,869
2	0,909	0,654	0,964	0,814	0,936	0,725
3	0,748	0,421	0,980	0,907	0,848	0,575

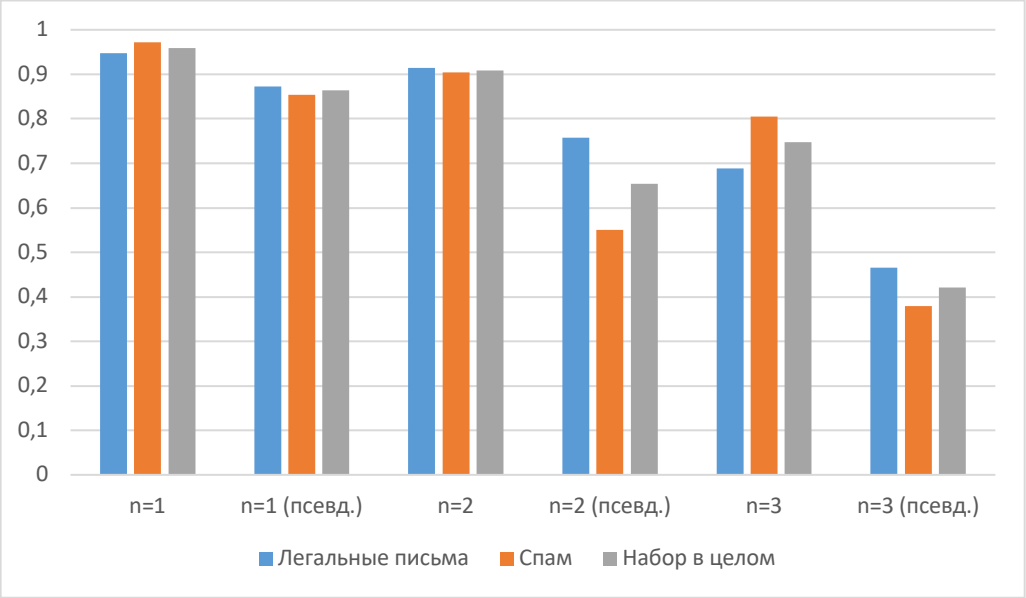

Рисунок Д.1 – Полнота обнаружения R на англоязычных письмах

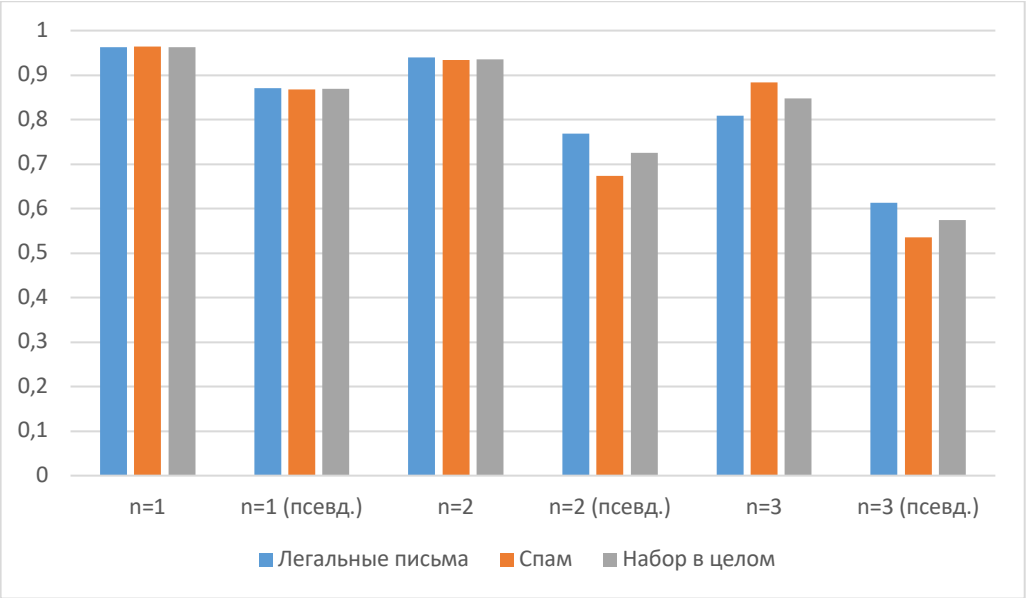

Рисунок Д.2 – Значения F-меры на англоязычных письмах

Приложение Е. Акты внедрения результатов работы

Общество с ограниченной ответственностью «Омега Софт»
ИНН 1215233374, КПП 121501001 ОКПО 45613208
424033, г. Йошкар-Ола, Воскресенский пр., д.17, этаж 2, тел. +7 906 139 57 77,
e-mail: alexey@omega-r.com.

"УТВЕРЖДАЮ"
Генеральный директор
ООО «Омега Софт»

Рыбаков Алексей Евгеньевич
подпись

«11» марта 2024 г.

АКТ
об использовании результатов
диссертационной работы
Корелова Сергея Викторовича
на соискание ученой степени кандидата технических наук

Комиссия в составе Генерального директора ООО"ОМЕГА СОФТ" Рыбакова Алексея Евгеньевича, Технического директора Якимова Сергея Александровича, Руководителя отдела мобильной разработки Князева Антона Викторовича, составила настоящий акт о том, что результаты диссертационной работы Корелова С.В. «МЕТОД И АЛГОРИТМ ОБНАРУЖЕНИЯ СПАМА НА ОСНОВЕ ВЫДЕЛЕНИЯ ПРИЗНАКОВ ЭЛЕКТРОННЫХ ПИСЕМ С ИСПОЛЬЗОВАНИЕМ КОНТЕНТНОЙ ФИЛЬТРАЦИИ», представленной на соискание ученой степени кандидата технических наук по специальности 2.3.6 – «**Методы и системы защиты информации, информационная безопасность**», используются в деятельности ООО «Омега Софт» (Omega), а именно, алгоритм классификации электронных писем на основе методов машинного обучения, отличающийся наличием дополнительной процедуры определения «схожести» термов на основе расстояния Левенштейна, обеспечивающей вычисление мер принадлежности классифицируемого электронного письма к классам спама и легальных для повышения достоверности идентификации электронных писем, позволяющий осуществить программную реализацию метода в виде плагина.

Настоящий акт подтверждает что практические результаты, представленные в работе Корелова С.В. «МЕТОД И АЛГОРИТМ ОБНАРУЖЕНИЯ СПАМА НА ОСНОВЕ ВЫДЕЛЕНИЯ ПРИЗНАКОВ ЭЛЕКТРОННЫХ ПИСЕМ С ИСПОЛЬЗОВАНИЕМ КОНТЕНТНОЙ ФИЛЬТРАЦИИ» в виде программного средства внедрены и используются на практике.

Члены комиссии:

Якимов С.А.
Князев А.В.

ООО «ТРЭВЕЛ ЛАЙН СИСТЕМС»
Россия, 424003, Республика Марий Эл,
г. Йошкар-Ола, Ленинский пр-т 56А
8 800-555-20-30 www.travelline.ru

ОГРН 1141215003214 ИНН 1215180595 КПП 121501001
Р/С 40702810910090007251 в Филиал «Центральный» Банка ВТБ (ПАО) в г. Москве
БИК 044525411 к/с 30101810145250000411

«УТВЕРЖДАЮ»

_____ *А.Н. Дудин*
подпись
«__» _____ 2024 г.

М.П.

АКТ
об использовании результатов
диссертационной работы
Корелова Сергея Викторовича
на соискание ученой степени кандидата технических наук

Комиссия в составе: председателя Герасимова А.В., членов комиссии – к.т.н. Лучинина З.С., Лежнина А.В, составили настоящий акт о том, что результаты диссертационной работы «МЕТОД И АЛГОРИТМ ОБНАРУЖЕНИЯ СПАМА НА ОСНОВЕ ВЫДЕЛЕНИЯ ПРИЗНАКОВ ЭЛЕКТРОННЫХ ПИСЕМ С ИСПОЛЬЗОВАНИЕМ КОНТЕНТНОЙ ФИЛЬТРАЦИИ», представленной на соискание ученой степени кандидата технических наук по специальности 2.3.6 – «**Методы и системы защиты информации, информационная безопасность**», активно используются в деятельности ООО «ТРЭВЕЛ ЛАЙН СИСТЕМС», а именно:

- Архитектура подсистемы классификации электронных писем для обнаружения спама и идентификации легальных электронных писем на основе алгоритма , отличающаяся от известных блоком выделения термов и блоком нечеткой классификации, реализующая метод классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, отличающийся использованием модели электронного почтового сообщения для классификации электронных писем на основе метода «генетических карт»;
- Программный модуль на основе предложенного алгоритма.

Использование указанных результатов позволило гарантировать корректное обновление данных, соответствующее правилам предметной области и целостности данных; сократить затраты на разработку новых программных продуктов, работающих с почтовыми сервисами.

Председатель комиссии:
_____ Герасимов А.В.

Члены комиссии:
_____ к.т.н. Лучинин З.С.
_____ Лежнин А.В.

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования «Поволжский государственный технологический университет» (ФГБОУ ВО «ПГТУ»)

пл. Ленина, д. 3, г.Йошкар-Ола, Республика Марий Эл, 424000
Телефон (8362) 68-68-70, факс (8362) 41-08-72
E-mail: info@volgatech.net,
http://www.volgatech.net./
ИНН/КПП 1215021281/121501001,
_____ № _____
На №_____ от _____

УТВЕРЖДАЮ:
Директор департамента образовательной деятельности ФГБОУ ВО «ПГТУ»
_____ / Т.Л.Кошохова___/
«19» 04 2024 г.

А К Т

об использовании результатов научных исследований
Корелова Сергея Викторовича
в учебном процессе ФГБОУ ВО ПГТУ

Научно-техническая комиссия в составе: председателя
А.А. Кречетов, к.т.н., доц., декан ФИиВТ
(Фамилия И.О уч.ст., уч.зв., должность.)
и членов комиссии: Васяевой Н.С., к.т.н, доц. каф. ИВС, Кубашевой Е.С. к.т.н., доц. каф. ИВС, Савинова А.Н.__ к.т.н., доц. каф. ИВС
(уч.ст., уч.зв., должность, Фамилия И.О.)
составила настоящий акт о том, что материалы и результаты научных исследований Корелова С.В. на тему «МЕТОД И АЛГОРИТМ ОБНАРУЖЕНИЯ СПАМА НА ОСНОВЕ ВЫДЕЛЕНИЯ ПРИЗНАКОВ ЭЛЕКТРОННЫХ ПИСЕМ С ИСПОЛЬЗОВАНИЕМ КОНТЕНТНОЙ ФИЛЬТРАЦИИ» использованы в учебном процессе подготовки обучающихся по направлению подготовки 10.04.01 Информационная безопасность
(код и наименование направления подготовки)
в следующих формах:

№	Результат исследования	Учебная дисциплина	Форма использования
1.	Модель электронного почтового сообщения для классификации электронных писем на основе метода «генетических карт»	"Проектная деятельность", "Управление информационной безопасностью", "Системная инженерия"	Внедрение в образовательный процесс с целью совершенствования процесса формирования проектной компетентности обучающихся
2.	Метод классификации электронных писем для обнаружения спама и идентификации легальных электронных писем, отличающийся использованием	"Проектная деятельность", "Управление информационной безопасностью"	Внедрение в образовательный процесс с целью совершенствования процесса формирования проектной

	модели электронного почтового сообщения для классификации электронных писем на основе метода «генетических карт»		компетентности обучающихся

Материалы обсуждены и одобрены на расширенном заседании кафедры
Информационной безопасности
(наименование кафедры)
Протокол № 8 « 6 » 02 2024 г.

Председатель комиссии: / А.А. Кречетов

Члены комиссии: / Н.С. Васяева

/ Е.С. Кубашева

/ А.Н.Савинов

Заведующий кафедрой ИБ: / И.Г. Сидоркина

Исп. Добрынина Т.В.
88362587842