УДК 004.056, 004.042, 004.852 DOI:10.31854/1813-324X-2020-6-4-80-90

Предобработка текстов электронных писем в задаче обнаружения спама

С.В. Корелов[1*], А.М. Петров[1], Л.Ю. Ротков[2], А.А. Горбунов[2]

[1]Национальный координационный центр по компьютерным инцидентам,
Москва, 107031, Российская Федерация
[2]Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского
Нижний Новгород, 603950, Российская Федерация
[*]Адрес для переписки: korelovsv@cert.gov.ru

Информация о статье
Поступила в редакцию 31.08.2020
Принята к публикации 23.11.2020.

Ссылка для цитирования: 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

Аннотация: *Функционирование практически любой организации в той или иной степени зависит от того, насколько надежно защищены ее информационные ресурсы от различных угроз безопасности информации, одной из которых является спам. При этом было совершено множество попыток раз и навсегда решить проблему его обнаружения. В данной предметной области постоянно ведутся исследования. По их результатам предлагаются и реализуются на практике различные подходы. Ранее авторами предложена модель электронных писем, учитывающая содержание электронных писем, которое зачастую меняется в зависимости от выполняемых пользователями задач и меняющихся их информационных потребностей.*
В настоящей статье обсуждается вопрос предобработки текстов электронных писем в задаче обнаружения спама с использованием модели электронных писем, полученной на основе генетического подхода к формированию математических моделей текстов, зарекомендовавшего себя для решения различных задач.

Ключевые слова: *информационная безопасность, спам, обнаружение, модель электронного письма, генетический подход, генетическая модель, электронная почта, электронные почтовые сообщения, электронные письма, предобработка текста.*

Введение

В условиях интенсивного развития различных сфер деятельности государства и общества использование передовых информационных технологий становится одним из наиболее важных, а часто и решающим фактором, определяющим эффективность всех уровней управления.

Наряду с заметным в последние годы ростом популярности использования таких форм онлайн-коммуникаций, как мессенджеры и социальные сети, электронная почта широко применяется для деловой переписки, а также является обязательным требованием для использования различных электронных услуг и сервисов. По оценкам The Radicati Group, Inc. [1], в 2020 г. электронной почтой пользуется половина населения Земли; при этом их количество до конца 2020 г. превысит отметку в 4 млрд. с прогнозом более 4,4 млрд. в 2024 г.

Однако столь высокая популярность электронных писем сопровождается рядом проблем. Одним из ставших классическими бизнес-рисков, связанным с использованием электронных почтовых сообщений, является спам. На сегодняшний день спам является настоящей проблемой для мирового потока электронных писем в частности и трафика в общем. Средняя доля спама в почтовом трафике в 2018 г. составила 52,48 % [2], а в 2019 – 56,51 % [3]. Спам является причиной различных негативных последствий для его получателей, а также серьезного негативного эффекта для мировой экономики [4–10].

Представляется очевидным, что обнаружение спама является не просто желательной, а остронеобходимой и неотъемлемой частью общей системы безопасности информационных систем. Необходимо отметить, что невозможно сформировать универсальное описание спамовых писем, поскольку возможна ситуация, когда в зависимости от интересов конкретного пользователя электронное письмо может быть отнесено к спамовым или легальным. Несмотря на это, ученые и специалисты по всему миру продолжают исследования в области обнаружения спама в поисках на 100 %

эффективного решения [8]. Однако из-за своей комплексности и сложности задача обнаружения спама не имеет единственно верного и универсального решения [4, 5].

Целью данной работы является оценка возможности повышения эффективности применения предложенной в [11] модели электронных писем в задаче обнаружения спама.

Краткий обзор современных исследований в области обнаружения спама

Электронные письма состоят из различных полей и частей, на основании которых можно выделить следующие два наиболее распространенных подхода для обнаружения спама [6]: базирующиеся на заголовках писем и их содержимом. Первый из них обладает недостатками, позволяющими отправителям спама достаточно легко обходить разработанные на его основе механизмы фильтрации спамовых сообщений [6, 12]. Успех методов машинного обучения в классификации текстов обратил внимание исследователей на обучающие алгоритмы для решения задачи обнаружения спама [12], в основе которых лежит второй подход. Его применение представляется более эффективным [6, 8].

Многие исследования последних лет в области обнаружения спама основываются именно на подходе анализа содержимого электронных писем и посвящены вопросу оценки эффективности применения в различных условиях методов машинного обучения в задаче обнаружения спама [например, 4, 6, 7, 10, 13], а также вопросу отбора признаков электронных писем.

Среди методов машинного обучения имеется большое количество разнообразных эффективных алгоритмов и их модификаций, используемых в задаче обнаружения спама. Они включают такие распространенные методы, как, например, наивный Байесовский классификатор [например, 6, 8, 9, 13–16], дерево решений [например, 17, 18], опорных векторов [например, 13, 14, 16, 19], *k*-ближайших соседей [16, 20], искусственные иммунные системы [5, 21, 22], искусственные нейронные сети [23–28] и другие. В качестве базовых признаков электронных писем в задаче их классификации в основном используются слова (и/или их сочетания) и рассчитываемые различными способами их веса [например, 5–9, 24, 25].

Методы машинного обучения позволяют автоматически строить списки слов с их весами на основе знания спамовых и легальных писем. Неправильная же классификация легальных писем как спамовых (ложноположительный результат) и неправильная классификация спамовых писем как легальных (ложноотрицательный результат) ведет к издержкам [4]. При этом отправители спама всячески стараются снизить вероятность обнаружения их писем путем использования слов, присущих легальным письмам.

Основываясь на изложенном, авторы приходят к выводу, что в исследованиях последних лет, посвященных решению задачи обнаружения спама, как правило, используются одни и те же методы классификации или предлагаются их модификации. В то же время большое внимание уделяется вопросу выделения и отбора признаков электронных писем и оценке эффективности обнаружения спама с применением известных методов классификации, но с использованием различных признаков (и/или их сочетаний).

Также авторы приходят к выводу о том, что «борьба» исследователей не останавливается [29] и идет буквально за каждые 0,01 % точности и полноты обнаружения спама. При этом продолжает оставаться актуальным вопрос выбора эффективных (с точки зрения качества обнаружения спама) признаков электронных почтовых сообщений для процесса классификации, что требует разработки новых подходов к определению и выделению признаков электронных писем и оценки эффективности их применения.

В связи с актуальностью и важностью данного направления исследований в задаче обнаружения спама в [11] авторами для обнаружения спама предложена, обоснована и описана генетическая модель электронных писем, позволяющая специфическим способом выделять текстовые отрезки электронных писем, являющиеся отражением их отличительных признаков:

$$\Psi_{el} = \langle gens, Gen\_Code \rangle. \quad (1)$$

Ключевой особенностью данной модели является то, что она оперирует с преобразованными в числовой вектор данными, полученными из исходных текстов электронных писем.

В качестве параметров модели электронных писем, оказывающих влияние на выделение текстовых отрезков писем, являющихся отражением их отличительных признаков, авторами в [30] обоснованы:

q – количество числовых кодов, сопоставляемых символам текста, в функции преобразования писем в числовой вектор;

Δt – шаг выборки символов текста в функции преобразования писем в числовой вектор;

n – длина «генератора» (последовательность, порождающая «ген»).

Там же продемонстрированы корректность и практическая применимость данной модели для обнаружения спама (классификации электронных писем на спамовые и легальные), а также обоснован выбор численного значения параметра n модели электронных писем.

Для методов машинного обучения важным является предварительная обработка данных [4, 7–9]. В связи с этим настоящая статья посвящена исследованию вопроса применения в задаче обнаружения спама предложенной модели (1) совместно с предварительной обработкой текстов электронных писем.

Краткий анализ предметной области

В машинном обучении важную роль играют непосредственно сами данные, а точнее – их подготовка [4, 7–9, 31, 32]. На практике электронные письма поступают от разных отправителей и составляются ими с использованием различных почтовых клиентов в различных форматах. При этом они могут содержать различного рода «шумы»: ошибки и искажения различной природы, незначимые с точки зрения содержания (смысла) слова и символы, а также неинформативные элементы, например, спецсимволы HTML, скрипты, рекламные вставки и т. п. [7, 9, 33–35]. Такие «шумы» негативно влияют на качество непосредственно самих данных для анализа и могут снизить полноту и точность классификации. Поэтому первым значимым этапом в задачах интеллектуального анализа текстов является их предобработка [7–9, 24, 33–35], представляющая собой процесс их очистки и подготовки к классификации [32].

Процесс предобработки текста можно разбить на отдельные операции; при этом выполняемые в ходе них действия обрабатывают текст различными способами. В качестве основных способов предобработки текстовых данных можно выделить следующие [31, 32, 34–37]:

1) удаление неинформативных с точки зрения содержания (смысла) 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 [36]. В [32, 36, 37] экспериментально продемонстрировано, что подобранные применительно к конкретным текстовым данным и решаемой задаче способы предобработки могут улучшить качество классификации.

Обобщая изложенное, авторы приходят к выводу, что использование предобработки электронных писем в модели электронных писем (3) также может повысить полноту и точность классификации в задаче обнаружения спама. И поскольку в модели применяется специфический способ выделения значимых характеристик, среди указанных актуальными в ее контексте являются 2–5 и их сочетания.

При этом необходимо отметить, что, по большому счету, ни один из них не оказывает существенного влияния на содержание (смысл) писем (кроме, вероятно, стоп-слов). Так, переносы строк необходимы для выделения мыслей и придания логической структуры тексту. Знаки препинания в основном предназначены для формирования логической связи. А приведение всех букв к одному регистру позволяет избежать различий при написании одних и тех же слов, стоящих первыми в предложениях и в других позициях.

Выбор же конкретных способов предобработки в задаче обнаружения спама с использованием модели электронных писем (1) целесообразно осуществить при проведении экспериментальных исследований [37].

Таким образом, проведение исследования вопроса применения в задаче обнаружения спама предложенной модели (1) с включением в ее состав процедуры предварительной обработкой текстов электронных писем, а также оценка получаемых при этом результатов обнаружения в сравнении с результатами без применения предварительной обработки является актуальным.

Экспериментальная часть

В результате правительственного расследования по факту банкротства компании Enron в начале 2000-х годов в открытом доступе стали доступны более 600 тысяч электронных писем ее сотрудников [15]. Ценность этого массива заключается в том, что все письма написаны людьми и представляют собой реальное человеческое общение на различные темы. На протяжении последних лет эти письма в том или ином объеме использовались исследователями в области обнаружения спама.

Для проведения эксперимента был использован основывающийся на этих письмах набор [38], сформированный и описанный в [15] с дополнительными изменениями в соответствии с [30]). Легальные письма в нем представлены упорядоченными по имени файла электронными письмами шести сотрудников компании Enron, почтовые ящики которых содержали достаточно большое количество электронных писем. Спамовые письма представляют собой письма из корпуса SpamAssassin, проекта HoneyPot, из коллекции Bruce Guenter, а также были собраны Georgios Paliouras [15]. Дубликаты среди них не удалялись, поскольку они являлись частью

естественного потока всех электронных писем (легальных и спамовых) на почтовый ящик конкретного отдельно взятого пользователя. Также авторами [15] осуществлена предварительная обработка писем – удалены в следующем объеме:

– сообщения, отправленные владельцем почтового ящика самому себе;

– все html-теги и html-заголовки (сохранены только темы писем и их содержание);

– спамовые сообщения, содержащие символы нелатинского набора.

Таким образом, для проведения эксперимента сформирован набор англоязычных писем, состоящий из 6 групп легальных писем общим количеством 16100 писем и 6 групп спамовых писем общим количеством 16420 писем. Их тексты содержат строчные и прописные буквы, цифры, знаки препинания и другие символы. Для целей настоящего эксперимента дополнительно из всех писем были удалены строки с их темами, после чего из набора были удалены письма с нулевой длиной (т. е. изначально содержащие только тему). Дополнительно проведенный анализ их содержимого показал следующее:

– практически все буквы переведены в нижний регистр (за исключением некоторых спамовых писем, в которых содержатся единожды встречающаяся бука «B» в слове «Binary»);

– письма не содержат знаков табуляции.

Также для эксперимента был сформирован набор русскоязычных электронных писем, состоящих из 3 групп рассылок порталов securitylab.ru, security.nnov.ru и xakep.ru (за период с 28 апреля 2009 г. по 4 марта 2011 г.) общим количеством 1242 письма и 2 группы спамовых писем, поступивших на индивидуальные почтовые адреса двух различных адресатов одного почтового сервера в зоне .ru, общим количеством 3215 писем. Из писем удалены все html-теги и html-заголовки (сохранено только их содержание). Их тексты содержат строчные и прописные буквы, цифры, знаки препинания и другие символы. Дополнительно проведенный анализ их содержимого показал следующее:

– письма содержат повторяющиеся пробелы, в особенности, их большое число содержится в письмах информационной рассылки портала security.nnov.ru;

– письма содержат небольшое число знаков табуляции, включая повторяющиеся;

– письма содержат строчные и прописные кириллические буквы, в отдельных письмах присутствуют латинские буквы.

С учетом изложенного в разделе «Краткий анализ предметной области» для эксперимента определены способы предобработки и их сочетания, которые перечислены ниже попунктно.

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

п. 2. Удаление:

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

b) 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);

c) всех символов табуляции;

d) всех переносов строк с заменой на пробел;

e) всех пробелов.

п. 3. Перевод всех буквенных символов в верхний регистр:

a) тождественно п. 3;

b) п. 3 + удаление стоп-символов;

c) п. 3 + удаление всех переносов строк с заменой на пробел;

d) п. 3 + удаление всех пробелов;

e) п. 3 + удаление стоп-символов и всех переносов строк с заменой на пробел;

f) п. 3 + удаление стоп-символов и всех пробелов;

g) п. 3 + удаление всех переносов строк с заменой на пробел и все пробелы;

h) п. 3 + удаление стоп-символов и всех переносов строк с заменой на пробел, а также всех пробелов.

В качестве значений параметров модели электронных писем заданы следующие:

$q = 256$;

$\Delta t = 1$ – шаг дискретизации равен одному символу;

$n = 1 \ldots 2$ [30].

Таким образом, с учетом изложенного, модель (1) примет следующий вид:

$$\Psi_{el} = \langle Prepr, gens, Gen\_Code \rangle, \qquad (2)$$

где $Prepr$ – процедура предобработки.

Эксперимент и оценка его результатов проводились аналогично описанным в [30]. Для каждой категории (класса) писем (легальные и спамовые) каждой группы писем были рассчитаны наборы «генов» и определен коэффициент принадлежности каждого письма к легальным или спамовым письмам, за который принято суммарное количе-

ство содержащихся в письме «генов», встретившихся в соответствующих категориях всех групп.

Решение о принадлежности письма к спамовым или легальным принималось с использованием простейшего решающего правила – по принципу большего суммарного количества «генов» соответствующей категории. При этом для классифицируемого письма расчет набора «генов» его группы велся только для писем, стоящих перед ним в списке, что позволило частично имитировать процесс получения писем адресатом.

В качестве мер оценки результатов эксперимента использованы полнота, точность и *F*-мера обнаружения (классификации) [39–43].

Под полнотой R обнаружения спамовых и легальных писем будем понимать соотношение числа всех верно классифицированных электронных писем к числу электронных писем, которые должны были быть отнесены к тому или иному классу:

$$R = \frac{N_{corr\_a}}{N_{corr\_a} + N_{incorr\_r}}, \quad (3)$$

где N_{corr_a} – количество электронных писем, корректно отнесенных к заданной категории (истинно положительные результаты или TP, *аббр. от англ.* True Positive); N_{incorr_r} – количество электронных писем, некорректно признанных не принадлежащими заданной категории (ложноотрицательные результаты или FN, *аббр. от англ.* False Negative).

Иначе, полнота характеризует потери процесса классификации электронных писем. Как следует из представленной формулы, чем выше значение полноты, тем меньше потери правильных классификаций. Таким образом, R определяет способность процесса классификации электронных писем обнаруживать заданный класс вообще.

Под точностью P обнаружения спамовых и легальных писем будем понимать соотношение числа верно классифицированных электронных писем к числу всех классифицированных электронных писем как принадлежащих к тому или иному классу:

$$P = \frac{N_{corr\_a}}{N_{corr\_a} + N_{incorr\_a}}, \quad (4)$$

где N_{incorr_a} – количество электронных писем, некорректно признанных принадлежащими заданной категории (ложноположительные результаты или FP (*аббр. от англ.* False Positive).

Иначе, точность можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными. Таким образом, P определяет способность процесса классификации электронных писем правильно обнаруживать заданный класс (долю правильных классификаций), т. е. чем лучше выстроен процесс классификации, тем меньше будет неверно классифицированных электронных писем как принадлежащих заданному классу.

Безусловно, чем выше полнота и точность, тем лучше. Но очевидно, что достичь максимальной полноты и точности одновременно невозможно. В связи с этим для выбора лучшего варианта классификации использована сбалансированная *F*-мера обнаружения (классификации) [40–44] спамовых и легальных писем, которая позволяет объединить полноту и точность в агрегированную величину для оценки, представляющую собой их среднее гармоническое:

$$F = \frac{2 \cdot P \cdot R}{P + R}. \quad (5)$$

Из (5) следует, что *F*-мера достигает максимума при полноте и точности, равными единице, и близка к нулю, если один из аргументов близок к нулю. Таким образом, *F*-мера позволяет определить наилучший процесс классификации электронных писем с учетом одновременно полноты и точности, т. е. чем лучше выстроен процесс классификации, тем больше значение *F*-меры.

Результаты эксперимента на англоязычных и русскоязычных письмах округлены до сотых долей процента по правилам простого математического округления и представлены в таблице 1, а значения *F*-меры в виде гистограмм приведены на рисунках 1a и 1b, соответственно.

В связи с наличием повторяющихся пробелов в письмах информационной рассылки портала security.nnov.ru эксперимент был дополнен исследованиями с предварительным удалением повторений пробелов. Их результаты также приведены в таблице 1 и на рисунке 1c.

Как видно из таблицы (где в желтых клетках – примерно равно значению «без предобработки», в зеленых – превышение значения «без предобработки»), удаление повторений пробелов приводит к улучшению результатов обнаружения при использовании не всех способов предобработки. Однако основываясь на описании модели [11] и полученных результатах, можно предположить, что повторения символов пробелов, табуляции и переносов строк могут ухудшать результаты обнаружения, поскольку создают короткие («мусорные», с точки зрения модели) «гены».

Также повторения таких знаков могут быть созданы умышленно с целью возможного подстраивания спамовых писем под легальные (особенно, форматированные письма различных рассылок). Это дает основания утверждать целесообразность осуществлять предварительную очистку (предобработку) писем от повторений указанных символов (знаков) с целью обеспечения независимости процесса классификации с применением модели (1) от наборов писем (даже с учетом возможного небольшого снижения результатов обнаружения).

ТАБЛИЦА 1. Результаты эксперимента на наборе электронных писем, %

TABLE 1. Experimental Result on the Electronic Letters Set, %

№ п.	*R*		*P*		*F*-мера	
	$n = 1$	$n = 2$	$n = 1$	$n = 2$	$n = 1$	$n = 2$
англоязычных						
1	**95,92**	90,94	**96,63**	96,44	**96,27**	93,61
2a	95,57	91,46	96,38	96,50	95,97	93,91
2b	**95,89**	87,44	**96,66**	96,56	**96,28**	91,77
2c	**95,92**	90,94	**96,63**	96,44	**96,27**	93,61
2d	95,78	91,08	96,48	96,31	96,13	93,62
2e	95,47	85,38	96,40	95,72	95,93	90,26
3a	**95,92**	90,94	**96,63**	96,44	**96,27**	93,61
3	95,57	91,46	96,38	96,50	95,97	93,91
3c	95,78	91,08	96,48	96,31	96,13	93,62
3d	95,47	85,38	96,40	96,18	95,93	90,46
3e	95,16	92,11	96,24	96,46	95,70	94,23
3f	94,97	86,04	95,94	95,93	95,45	90,71
3g	95,26	83,57	96,36	95,55	95,81	89,16
3h	94,68	84,07	95,80	95,16	95,24	89,28
русскоязычных						
1)	**93,04**	85,62	**93,74**	91,45	**93,39**	88,44
2a)	90,04	85,78	90,87	91,31	90,45	88,45
2b)	92,73	85,44	93,34	91,56	93,03	88,39
2c)	**93,02**	85,62	**93,74**	91,47	**93,38**	88,45
2d)	86,29	81,62	87,39	90,79	86,84	85,96
2e)	92,53	90,89	93,54	**93,82**	93,03	92,33
3a)	92,95	87,03	**93,71**	92,03	**93,33**	89,46
3b)	89,68	87,37	90,57	91,86	90,12	89,56
3c)	85,48	83,35	86,37	91,32	85,93	87,16
3d)	92,37	92,37	93,48	**99,32**	92,92	**95,72**
3e)	80,50	89,68	81,36	**97,94**	80,93	**93,63**
3f)	92,42	92,87	93,32	**99,11**	92,86	**95,89**
3g)	86,31	82,10	87,45	98,07	86,88	89,37
3h)	86,09	83,47	87,66	98,77	86,87	90,48
с предварительным удалением повторений пробелов						
1)	**92,64**	91,56	93,48	**98,74**	93,06	**95,02**
2a)	92,53	91,95	93,28	**98,70**	92,90	**95,20**
2b)	**92,87**	91,65	93,56	**99,22**	93,21	**95,29**
2c)	**92,69**	91,52	93,42	**98,77**	93,05	**95,00**
2d)	90,85	83,67	91,61	97,08	91,22	89,88
2e)	92,53	90,89	93,54	93,82	93,03	92,33
3a)	**92,78**	**92,78**	93,51	**99,04**	93,14	**95,81**
3b)	92,33	**93,25**	93,08	**98,76**	92,70	**95,93**
3c)	89,88	85,98	90,92	97,68	90,40	91,46
3d)	92,37	92,37	93,48	**99,32**	92,92	**95,72**
3e)	88,67	86,87	89,68	97,51	89,17	91,88
3f)	92,42	**92,87**	93,32	**99,11**	92,86	**95,89**
3g)	86,31	82,10	87,45	98,07	86,88	89,37
3h)	86,09	83,47	87,66	**98,77**	86,87	90,48

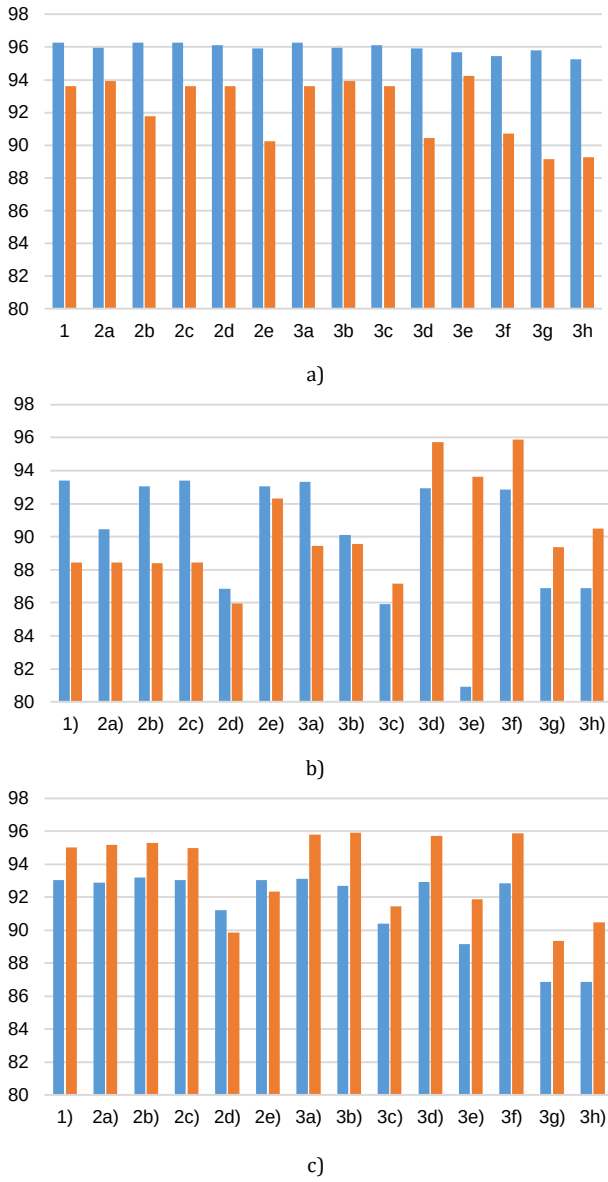

Рис. 1. Значения *F*-меры на наборе электронных писем, %: a) англоязычных, b) русскоязычных и c) русскоязычных с предварительным удалением повторений пробелов

Fig. 1. F-Measure Values on the Electronic Letters Set, %: a) of the English-Language, b) of the Russian-Language & c) of the Russian-Language with Pre-Deleting Repetitions of Spaces

Обобщая полученные результаты эксперимента на англоязычных и русскоязычных письмах, авторы приходят к следующим выводам.

Во-первых, применение предобработок текстов англоязычных электронных писем в задаче обнаружения спама с применением модели (1) в целом не приводит к существенному изменению результатов обнаружения в сравнении с результатами без предобработок.

Во-вторых, применение «атомарных» способов 2a–3a предобработки текстов русскоязычных электронных писем в задаче обнаружения спама с применением модели (1) также в целом не приводит к

существенному изменению результатов обнаружения в сравнении с результатами без предобработок. Вместе с тем, следует отметить существенное улучшение точности обнаружения (более 99 %) при использовании способов предобработки 3d и 3f (совместное использование «атомарных»). Также точность обнаружения превышает 97 % практически при использовании любых предобработок при условии предварительного удаления повторений пробелов.

В-третьих, имеется зависимость результатов обнаружения от конкретных применяемых способов предобработки в совокупности со значениями ее параметров, что коррелирует с опубликованными в [19, 32, 37] выводами.

Также целесообразно отметить, что предложенная авторами модель электронных писем с использованием описанного решающего правила позволяют достичь неплохих результатов, в целом не сильно уступающих результатам, полученным в некоторых аналогичных исследованиях, а иногда и немного превосходящих их.

Так, в [15] результаты проведенных экспериментов с использованием Байесовского классификатора с различными параметрами на наборе писем [38] демонстрируют полноту обнаружения в среднем от около 93 % до 97 %. В [44] на наборе из 3196 писем (740 легальных и 2456 спамовых) эксперименты с разработанными алгоритмами выявили лучшую полноту обнаружения спама около 81 % при точности свыше 98 %. В [45] на наборе из 908 писем (424 легальных и 484 спамовых) эксперименты с разработанными алгоритмами продемонстрировали лучшую полноту обнаружения около 90 % при точности около 96 %. Также полученные результаты показывают лучшие значения полноты, точности и *F*-меры, чем некоторые из полученных в [5–7] результаты.

Таким образом, основываясь на полученных результатах генетическую модель электронных писем (1) целесообразно дополнить функцией предобработки электронных писем:

$$\Psi_{el} = \langle Prepr, gens, Gen\_Code \rangle, \quad (6)$$

где

$$Prepr = \{ws\_reps\_del, tabs\_reps\_del, lb\_reps\_del, up\_case\}, \quad (7)$$

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.

Заключение

Результаты проведенного эксперимента подтверждают сделанные в [30] выводы о корректности и применимости разработанной авторами и описанной в [11] модели электронных писем (1) для обнаружения спама также в условиях применения различных способов предобработки текстов. При этом результаты обнаружения напрямую зависят от выбора способов предобработки в совокупности со значениями параметров модели, что подтверждается опубликованными в [19, 32, 37] выводами.

С целью снижения зависимости процесса классификации с применением модели (1) от конкретных писем среди способов предобработки целесообразно рассматривать только предварительную очистку (предобработку) писем от повторений символов пробелов, табуляции и переносов строк и перевод всех букв в верхний (нижний) регистр. Применение каких-либо иных способов предобработки в модели (1) целесообразно только при условии их предварительной экспериментальной оценки и постоянной периодической корректировки с течением времени для адаптации процесса классификации применительно к индивидуальным особенностям написания электронных писем их автором.

Список используемых источников

1. Email Statistics Report, 2016–2020 // The Radicati Group. URL: https://www.radicati.com/?p=13546(дата обращения 25.11.2020)

2. Вергелис М., Щербакова Т., Сидорина Т. Спам и фишинг в 2018 году // Securelist. URL: https://securelist.ru/spam-and-phishing-in-2018/93453 (дата обращения 17.09.2019)

3. Вергелис М., Щербакова Т., Сидорина Т., Куликова Т. Спам и фишинг в 2019 году // Securelist. URL: https://securelist.ru/spam-report-2019/95727 (дата обращения 29.10.2020)

4. 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

5. 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

6. 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

7. 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

8. 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

9. 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

10. 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

11. Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. Модель электронных писем в задаче обнаружения спама // Вестник Поволжского государственного технологического университета. Серия: Радиотехнические и инфокоммуникационные системы. 2020. № 2(46). С. 44–54. DOI:10.25686/2306-2819.2020.2.44

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

13. 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

14. 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

15. 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.

16. 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

17. 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.

18. 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

19. 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

20. 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

21. 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

22. 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

23. 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

24. 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.

25. 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.

26. Катасёв А.С., Катасёва Д.В. Разработка нейросетевой системы классификации электронных почтовых сообщений // Вестник Казанского государственного энергетического университета. 2015. № 1(25). С. 68–78.

27. 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)

28. 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

29. 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. 1–26. DOI:10.3390/app9050987

30. 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, Российская Федерация, 13–31 мая 2020). Нижний Новгород: ННГУ, 2020. С. 471–474. URL: http://www.rf.unn.ru/wp-content/uploads/sites/21/2020/10/rf-conf-2020-book-1.pdf (дата обращения 26.11.2020)

31. Климов Д.В. Предобработка текстовых сообщений для метрического классификатора // Символ науки. 2017. № 12. С. 25–32.

32. 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

33. 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

34. 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

35. 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.

36. Weng J. NLP Text Preprocessing: A Practical Guide and Template. URL: https://towardsdatascience.com/nlp-text-preprocessing-a-practical-guide-and-template-d80874676e79 (дата обращения 14.07.2020)
37. 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
38. Enron-Spam datasets. URL: http://www2.aueb.gr/users/ion/data/enron-spam (дата обращения 26.11.2020)
39. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys. 2002. Vol. 34. No. 1. PP. 1–47. DOI:10.1145/505282.505283
40. Sebastiani F. Text Categorization // Zanasi A. (ed.). Text Mining and its Applications. Southampton: WIT Press, 2005. PP. 109–129.
41. Aas K., Eikvil L. Text Categorisation: A Survey // Norwegian Computing Center. Tech. Report number: 941, 1999.
42. Manning C., Raghavan P., Shütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. DOI:10.1017/CBO9780511809071
43. 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
44. Мироненко А.Н. Алгоритм контентной фильтрации спама на базе совмещения метода опорных векторов и нейронных сетей. Автореферат дис. … канд. техн. наук. Санкт-Петербург, 2012. 18 с.
45. Чернопрудова Е.Н. Защита почтовых сервисов от несанкционированных рассылок на основе контентной фильтрации электронных сообщений. Автореферат дис. … канд. техн. наук. Уфа, 2013. 16 с.

* * *

Preprocessing of the Emails in the Spam Detection Task

S. Korelov[1], A. Petrov[1], L. Rotkov[2], A. Gorbunov[2]

[1]National Computer Incident Response & Coordination Center,
Moscow, 107031, Russian Federation
[2]National Research Lobachevsky State University of Nizhny Novgorod
Nizhny Novgorod, 603950, Russian Federation

Article info
DOI:10.31854/1813-324X-2020-6-4-80-90
Received 31st August 2020
Accepted 23rd November 2020

For citation: Korelov S., Petrov A., Rotkov L., Gorbunov A. Preprocessing of the Emails in the Spam Detection Task. *Proc. of Telecom. Universities.* 2020;6(4):80–90. (in Russ.) DOI:10.31854/1813-324X-2020-6-4-80-90

Abstract: *The functioning of almost any organization to one degree or another depends on how reliably its information resources are protected from various information security threats, one of which is spam. At the same time there have been many attempts to solve the problem of its detection once and for all. Research is ongoing in this subject area constantly. Based on its results, various approaches are proposed and implemented in practice. The authors previously proposed a model of e-mails that takes into account the content of e-mails, which often changes depending on the tasks performed by users and their changing information needs.*
This article discusses the issue of preprocessing e-mail texts in the problem of spam detection using a model of e-mails obtained on the basis of a genetic approach to the formation of mathematical models of texts, which has proven itself for solving various problems.

Keywords: *information security, spam, detection, electronic letter model, genetic approach, genetic model, email, e-mail messages, electronic letters, text preprocessing.*

References

1. The Radicati Group. *Email Statistics Report, 2016–2020.* Available from: https://www.radicati.com/?p=13546 [Accessed 25th November 2020]
2. Vergelis M., Shcherbakova T., Sidorina T. *Spam and Phishing in 2018*. (in Russ) Available from: https://securelist.ru/spam-and-phishing-in-2018/93453 [Accessed 17th September 2019]

3. Vergelis M., Shcherbakova T., Sidorina T., Kulikova T. *Spam and Phishing in 2019*. (in Russ) Available from: https://securelist.ru/spam-report-2019/95727 [Accessed 29th October 2020]

4. Barushka, A., Hajek, P. Spam Filtering Using Integrated Distribution-Based Balancing Approach and Regularized Deep Neural Networks. *Applied Intelligence*. 2018;48:3538–3556. DOI:10.1007/s10489-018-1161-y

5. Bhattacharya P., Singh A. E-mail Spam Filtering using Genetic Algorithm based on Probabilistic Weights and Words Count. *International Journal of Integrated Engineering*. 2020;12(1): 40–49. DOI:10.30880/ijie.2020.12.01.004

6. 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;15(2):73–84. DOI:10.17706/jcp.15.2.73-84

7. 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;10(1):60–67. DOI:10.5815/ijcnis.2018.01.07

8. Radhakrishnan A., Vaidhehi V. Email Classification Using Machine Learning Algorithms. *International Journal of Engineering and Technology (IJET)*. 2017; 9(2):335–340. DOI:10.21817/ijet/2017/v9i1/170902310

9. 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, 6–7 May 2017, Melaka, Malaysia. IOP Conference Series: Materials Science and Engineering*. Bristol: IOP Publishing; 2017. vol.226. DOI:10.1088/1757-899X/226/1/012091

10. Verma T., Gill N.S. Email Spams via Text Mining using Machine Learning Techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*. 2020;9(4):2535–2539. DOI:10.35940/ijitee.D1915.029420

11. Korelov S., Petrov A., Rotkov L.Yu., Gorbunov A.A. Model of Email Messages in the Problem of Detecting Spam. *Vestnik of Volga State University of Technology. Series "Radio Engineering and Infocommunication Systems"*. 2020;2(46):44–54. DOI:10.25686/2306-2819.2020.2.44

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

13. Sharaff A., Nagwani N., Dhadse A. Comparative Study of Classification Algorithms for Spam Email Detection. In: Shetty N., Prasad N., Nalini N. (eds) *Emerging Research in Computing, Information, Communication and Applications*. New Delhi: Springer; 2016. p.237–244. DOI:10.1007/978-81-322-2553-9_23

14. 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, 24–28 July 2000, Athens, Greece*. New York: Association for Computing Machinery; 2000. p.160–167. DOI:10.1145/345508.345569

15. 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, 27–28 July 2006, Mountain View, USA*. 2006. p.28–69.

16. 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, 1–2 August 2017, Chennai, India*. IEEE; 2017. p.676–679. DOI:10.1109/ICECDS.2017.8389522

17. 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. p.58–64.

18. 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;9(17):4013–4026. DOI:10.1002/sec.1584

19. Drucker H., Wu D., Vapnik V. Support Vector Machine for Spam Categorization. *IEEE Transactions on Neural Networks*. 1999;10(5):1048–1054. DOI:10.1109/72.788645

20. Jiang S., Pang G., Wu M., Kuang L. An Improved k-Nearest-Neighbor Algorithm for Text Categorization. *Expert System with Applications*. 2012;39(1):1503–1509. DOI:10.1016/j.eswa.2011.08.040

21. Yue X., Abraham A., Chi ZX., Hao YY., Mo H. Artificial Immune System Inspired Behavior-Based Anti-Spam Filter. *Soft Computing*. 2007;11:729–740. DOI:10.1007/s00500-006-0116-0

22. Malykhina M.P., Chastikova V.A., Biktimirov A.A. Method of Spam Detection Based on Artificial Immune Systems. *Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and Informatics*. 2018;3:38–48. (in Russ.) DOI:10.24143/2072-9502-2018-3-38-48

23. 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, 13–17 October 2003, Halifax, Canada*. IEEE; 2003. p.702–705. DOI:10.1109/WI.2003.1241300

24. Katasev A.S., Kataseva D.V., Kirpichnikov A.P. Neural Network Technology for Classifying Electronic Mail Messages. *Vestnik tekhnologicheskogo universiteta*. 2015;18(5):180–183. (in Russ.)

25. Katasev A.S., Kataseva D.V., Kirpichnikov A.P., Semenov J.E. Spam Filtering of E-Mail Messages Based on Neural Network and Neural Fuzzy Models. *Vestnik tekhnologicheskogo universiteta*. 2015;18(15):217–221. (in Russ.)

26. Katasev A.S., Kataseva D.V. The Neural Network System Development for Classification of E-Mail Messages. *Vestnik Kazanskogo gosudarstvennogo yenergeticheskogo universiteta*. 2015;1(25):68–78. (in Russ.)

27. Larionova A.V., Khorev P.B. Spam Filtering Method Based on Artificial Neural Network. *Naukovedeniye*. 2016;8(3). (in Russ.) Available from: http://naukovedenie.ru/PDF/04TVN316.pdf [Accessed 26 November 2020]

28. Larionova A.V., Khorev P.B. Efficiency Evaluating of Spam Filtering Method Based on Artificial Neural Network. *Naukovedeniye*. 2016;8(2). (in Russ.) DOI:10.15862/134TVN216

29. Hussain N., Turab Mirza H., Rasool G., Hussain I., Kaleem M. Spam Review Detection Techniques: A Systematic Literature Review. *Applied Sciences*. 2019;9(5):1–26. DOI:10.3390/app9050987

30. Korelov S.V., Petrov A.M., Rotkov L.Yu., Gorbunov A.A. On the Question of Determining the Numerical Value of a Parameter in the Email Model. *Proceedings of the XXIV Scientific Conference on Radiophysics devoted to the 75th anniversary of the Radiophysics Faculty, 13–31 May 2020, Nizhny Novgorod, Russian Federation*. Nizhny Novgorod: National Research Lobachevsky

State University of Nizhny Novgorod Publ.; 2020. p.471–474. (in Russ.) Available from: http://www.rf.unn.ru/wp-content/uploads/sites/21/2020/10/rf-conf-2020-book-1.pdf [Accessed 26 November 2020]

31. Klimov D.V. Preprocessing Text Messages for Metric Classifier. *Simvol nauki*. 2017;12:25–32. (in Russ.)

32. Haddi E., Liu X., Shi Y. The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science*. 2013;17:26–32. DOI:10.1016/j.procs.2013.05.005

33. Devaraj S., Krishnakumar A. Effective Search Engine Spam Classification. *International Journal of Recent Technology and Engineering (IJRTE)*. 2019;8(2S8):1541–1545. DOI:10.35940/ijrte.B1100.0882S819

34. HaCohen-Kerner Y., Miller D., Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation. *PLoS ONE*. 2020;15(5):e0232525. DOI:10.1371/journal.pone.0232525

35. Vijayarani S., Ilamathi J., Nithya M. Preprocessing Techniques for Text Mining – An Overview. *International Journal of Computer Science & Communication Networks*. 2015;5(1):7–16.

36. Weng J. NLP Text Preprocessing: A Practical Guide and Template. Available from: https://towardsdatascience.com/nlp-text-preprocessing-a-practical-guide-and-template-d80874676e79 [Accessed 14th July 2020]

37. Uysal A., Gunal S. The Impact of Preprocessing on Text Classification. *Information Processing & Management*. 2014;50(1): 104–112. DOI:10.1016/j.ipm.2013.08.006

38. *Enron-Spam datasets*. Available from: http://www2.aueb.gr/users/ion/data/enron-spam [Accessed 26th November 2020]

39. Sebastiani F. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*. 2002;34(1):1–47. DOI:10.1145/505282.505283

40. Sebastiani F. Text Categorization. In: Zanasi A. (ed.). *Text Mining and its Applications*. Southampton: WIT Press; 2005. p.109–129.

41. Aas K., Eikvil L. *Text Categorisation: A Survey*. Norwegian Computing Center. Tech. Report number: 941, 1999.

42. Manning C., Raghavan P., Shütze H. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press; 2008. DOI:10.1017/CBO9780511809071

43. Sokolova M., Lapalme G. A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management.* 2009;45(4):427–437. DOI:10.1016/j.ipm.2009.03.002

44. Mironenko A.N. *Algorithm for Content Filtering of Spam Based on Combining Support Vector Machines and Neural Networks*. PHD Thesis. St. Petersburg; 2012. 18 p. (in Russ.)

45. Chernoprudova E.N. *Protection of Mail Services from Unauthorized Mailings Based on Content Filtering of Electronic Messages.* PHD Thesis. Ufa; 2013. 16 p. (in Russ.)

Сведения об авторах:

КОРЕЛОВ Сергей Викторович | сотрудник Национального координационного центра по компьютерным инцидентам (г. Москва), korelovsv@cert.gov.ru

ПЕТРОВ Артем Михайлович | сотрудник Национального координационного центра по компьютерным инцидентам (г. Москва), pam@cert.gov.ru

РОТКОВ Леонид Юрьевич | кандидат технических наук, доцент, начальник Управления информационной безопасности, заведующий кафедрой «Безопасность информационных систем» Национального исследовательского Нижегородского государственного университета им. Н.И. Лобачевского, rtv@rf.unn.ru

ГОРБУНОВ Александр Александрович | преподаватель кафедры «Безопасность информационных систем» Национального исследовательского Нижегородского государственного университета им. Н.И. Лобачевского, aagor@rf.unn.ru