Обзорная статья

УДК 004.032.26.056.5

ББК 32.818.1

Ч 25

DOI: 10.53598/2410-3225-2021-3-286-98-106

Обзор актуальных проблем основных методов фильтрации спама и анализ их эффективности (Рецензирована)

Вера Аркадьевна Частикова¹, Константин Валерьевич Козачёк²

^{1, 2} Кубанский государственный технологический университет, Краснодар,
Россия.

¹ *chastikova_va@mail.ru*

² *Kozachek.Work-mail@yandex.ru*

Аннотация. Представлен анализ основных проблем фильтрации почтового спама, современных методов фильтрации нежелательных писем и способов обхода систем защиты. Вводится понятие «легитимного спама» – новой проблемы, с которой сталкиваются пользователи электронной почты. Рассмотрены методы представления текста: *bag-of-words* и *Embedding-пространство*, а также методы классификации: искусственные нейронные сети, метод опорных векторов, наивный байесовский классификатор. В работе определены эффективные методы, построенные на анализе текста, для решения задач обнаружения различных видов спама: типичного (известного системе), составленного при помощи методов обхода систем детекции спама, и легитимного.

Ключевые слова: байесовский классификатор, искусственная нейронная сеть, спам, обработка естественного языка, метод опорных векторов

Review article

Overview of topical issues in basic spam filtering methods and analysis of their effectiveness

Vera A. Chastikova¹, Konstantin V. Kozachek²

^{1, 2} *Kuban State University of Technology, Krasnodar, Russia.*

¹ *chastikova_va@mail.ru*

² *Kozachek.Work-mail@yandex.ru*

Abstract. An analysis of the main problems of filtering mail spam, modern methods of filtering unwanted letters and methods of bypassing security systems is presented. The concept of “legitimate spam” is being introduced – a new problem that email users face. Methods of text presentation are considered: *bag-of-words* and *Embedding-space*, as well as classification methods: artificial neural networks, the method of reference vectors, naive Bayesian classifier. The work identifies effective methods based on text analysis, for solving the problems of detecting various types of spam: a typical (known to system), compiled using methods of bypassing spam detection systems, and legitimate.

Keywords: Bayesian classifier, artificial neural network, spam, natural language processing, reference vectors method

Введение. Проблема спама в 2021 году все еще актуальна. Согласно исследованию IPwarmup.com [1], в 2018 году с ростом числа аффилированных компаний и мошенничества с электронной почтой во всем мире около 85% глобального почтового

трафика составляет спам (согласно Talos Intelligence – также 85% [2], согласно Statista [3], – лишь 60%), при этом около 20% спама обнаруживается на 5–15 день после его рассылки лишь после маркирования несколькими пользователями данных писем как спам [1]. В исследовании калифорнийского университета сообщается, что около 12,5% пользователей переходили по ссылке спамовых сообщений [4], в России этот показатель равен 9,69% [5].

Из итогового отчета по спаму и фишингу за 2020 год от аналитиков «Лаборатории Касперского» [5] видно, что доля спама в почтовом трафике составляет 50,37%, при этом злоумышленники находят множество новых способов обойти антиспам системы [5, 6] (см. также рис. 1). Например, пользователю пишется безобидное письмо с темой, которая, возможно, близка пользователю, и после того, как с ним завязывается «мирная» переписка, прикрепляется ссылка на фишинговый сайт. Таким образом, настройка антиспам систем должна быть более тонкой, чтобы обнаружить возможную угрозу. В среднем фишинговые ссылки были заблокированы лишь у 13% пользователей продуктов «Лаборатории Касперского», а в некоторых странах фишинг не дошел до 20% пользователей [7]. В ответ на различные способы обхода антиспам систем и по прочим причинам, указанным в статье [8], в 2020 году компания Microsoft решила внести глобальные изменения в систему обнаружения спама [8, 9].

Рис. 1. Доля спама в почтовом трафике Рунета за 2020 г.

Fig. 1. The share of spam in Runet mail traffic for 2020

Помимо этого, в последнее время возникла новая проблема – легитимный спам. Если пользователь подписывается на рассылку с какого-то сайта, надеясь получить интересный контент, но на деле интересный контент составляет малую долю писем, остальные же письма пользователю не нужны, их можно считать легитимным спамом. Зачастую пользователь тратит время на просмотр ненужных ему писем в поисках полезного. Почтовые сервисы сейчас не блокируют подобные письма, считая, что раз пользователь разрешил рассылку, то хочет видеть данное письмо. При этом для двух разных пользователей одно и то же письмо может являться полезным и быть легитимным спамом.

Также для полноты картины спама стоит сказать и про прочие сервисы. Количество sms-спама значительно снизилось с 2014 года, в том числе в Российской Федерации в связи с Федеральным Законом «О связи» [10]. Например, в Южной Корее (точной актуальной статистики по другим странам нет) пользователи получают спам-

сообщение раз в 9 дней, что в три раза реже, чем спам-письма в этой же стране [11].

Актуальной стала проблема спам-звонков. В РФ доля спам-звонков в первом квартале 2021 года составила 70% [12], и, хотя данные звонки являются нарушением ФЗ от 13.03.2006 N 38 «О рекламе», доказать данное нарушение непросто.

Спам в социальных сетях постоянно использует новые способы обхода антиспам систем. Так, несколько лет назад проводилась большая рассылка изображений, содержащих рекламу или незаконные темы, однако многие соцсети, например, Facebook, стали использовать нейросети для распознавания текста в изображениях, и доля такого спама снизилась. Сейчас злоумышленники записывают голосовые сообщения, но в ближайшее время и подобные методы распространения спама будут пресекаться.

Также большой проблемой в 2021 году является поисковый спам или спамдексинг – преднамеренное манипулирование индексами поисковых систем [13].

Основные методы борьбы со спамом, их плюсы и минусы

Существуют следующие группы методов борьбы со спамом:

1. Списки:

- черные списки – в них заносятся адреса, которые рассылали спам. Их преимущество в том, что они отсеивают большое количество спама. Недостаток: до того, как адрес попадает в черный список, он может разослать большое количество спама;

- белые списки – в них заносятся адреса доверенных пользователей. Плюсом является устранение проблемы ложных срабатываний. Минус заключается в том, что злоумышленник может получить доступ к почте доверенного пользователя (например, через фишинг или троян) и рассылать спам;

- черные списки в реальном времени (RBL) – черный список, управляемый сторонними службами. Достоинство – актуальность; недостаток – невозможность контроля;

- серые списки – создают временный отказ в приеме сообщения, провоцирующий почтовую систему отправителя отправить письмо еще раз. Программное обеспечение спамеров зачастую не делает повторной отправки. Главный недостаток – задержка в отправке равна 30 минут, что может быть критично для легитимного отправителя [14].

2. Основанные на контенте:

- фильтры на основе правил. Используют набор правил для слов, включенных в сообщение, чтобы узнать, является ли сообщение спамом. Хорошо отсеивают явные спамовые сообщения. Могут быть обмануты письмами, не содержащими спамовых слов, но содержащими картинки/ссылки;

- отдельным разделом можно выделить фильтры на основе статистических методов семантического анализа, например, на основе коэффициентов корреляции Спирмена и Пирсона. Такие методы находят широкое применение в отдельных областях применения обработки текста, однако в общем случае не могут конкурировать с более современными методами обработки текста [15];

- фильтры на основе байесовского классификатора основаны на теореме Байеса. Работают с частотами включений слова в «спам», «не спам», текст. Плюсом является простота реализации работы метода, скорость работы. По этой причине большинство антиспам систем используют именно этот метод. Недостатком является отсутствие понимания контекста, которое может быть нужно в некоторых случаях;

- метод опорных векторов (SVM) – метод классификации множеств. Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора, поэтому метод также известен как метод классификатора с максимальным зазором. Основной недостаток метода состоит в том,

что для классификации используется не все множество образцов, а лишь их небольшая часть, которая находится на границах. Помимо этого, SVM хуже работает в многомерных пространствах, а сообщения зачастую представляются именно таким образом;

– методы на основе искусственных нейронных сетей (ИНС), а также искусственных иммунных систем [16] – мейнстримное направление в области искусственного интеллекта. ИНС состоит из нейронов, которые складываются в слои, обрабатывают большие объемы данных, могут принимать на вход разные данные, что может быть полезно при обработке сложных данных с большим числом параметров [2, 17]. Например, в одной нейронной сети можно обрабатывать дату, время отправки сообщения, содержание (текст и изображения), тему, отправителя. Однако так как ИНС не могут обрабатывать естественный, человеческий язык, необходимо предварительно как-то интерпретировать его. В данной статье рассмотрены два метода обработки естественного языка: модель Bag of Word (BoW, «мешок слов») и слой Embedding. Первый из них прост в реализации, но не учитывает контекст применения слов. Суть его заключается в том, что каждое слово словаря представляется многомерным вектором, в котором единица ставится на месте этого слова в словаре, а остальные координаты равны 0. Вектор предложения или письма равен сумме векторов, соответствующих словам, включенным в предложение. Метод представления через слой Embedding преобразует каждое слово в n -мерный вектор. После обучения ИНС на наборе данных векторы слов, употребляемых в одинаковом контексте, будут близки. Недостатком метода ИНС можно считать необходимость дополнительных вычислений, а преимуществами – возможность автоматически обучаться при появлении новых типов спам-писем и, при определенной конфигурации, понимание смысла текстов.

3. Основанные на методе контроля массовых рассылок. Технология предполагает выявление в потоке почты массовых сообщений, которые абсолютно идентичны или различаются незначительно. Для обычных пользователей данный метод неприменим, так как многие отправители рассылают одинаковые и полезные письма (например, магазины). Подобная технология применима только для компаний. Недостаток – данный метод можно обойти при помощи программ, генерирующих разные сообщения с одинаковым смыслом.

4. Отдельной категорией можно выделить методы, анализирующие вложения. Данные методы анализируют документы, ссылки, изображения. Как правило, либо применяют ИНС (например, OCR), либо антивирусное программное обеспечение [5].

На основании вышеизложенного можно сделать вывод: следует применять совокупность решений. В данной работе была предложена комбинация методов, которая будет наиболее полно обеспечивать защищенность почты от спама (табл. 1).

Таблица 1

Совокупность решений по борьбе со спамом

Table 1. Set of anti-spam solutions

Метод	На решение какой проблемы направлен
Черный список	Отсекает все известные спамовые письма
Белый список	Предотвращает ложные срабатывания
ИНС, основанная на NLP	По содержанию письма распознает спам, даже если злоумышленник пытается имитировать легитимную почту
ИНС с OCR	Обнаруживает спам на изображениях

В статье основное внимание уделяется методам, основанным на анализе текста, так как множество способов обхода антиспам систем за 2020 г. основано именно на содержании письма [5]. Также применение подобных методов позволит снизить количество легитимного спама.

Сравнительный анализ существующего программного обеспечения по обнаружению спама

Spamoorona – программа от Яндекса, выпущена 22 января 2004 года, сейчас встроена в «Яндекс.Почта 360». Использует нейросетевые методы, применимые также к обработке естественных языков. Ранее использовала Байесовский классификатор. Помимо этого, использует OCR для распознавания текста на изображениях. В 2020 году обработала $78,3 \cdot 10^9$ писем, треть из которых спам. Из этого факта, в комбинации с информацией о том, что в 2020 году спам составил 50% всех писем [5], можно сделать вывод, что либо антиспам система Яндекса работает плохо, либо пользователям @ya.ru отправляют меньше спама, чем остальным пользователям.

SpamAssassin – популярная программа, работающая на Unix-подобных операционных системах. Выпущена 20 апреля 2001 года, последнее обновление выпущено в январе 2020 года, использует Байесовский классификатор.

Kaspersky Anti-Spam – использует лингвистические эвристики, сигнатурный анализ, OCR для распознавания текста на изображениях. Применяется для защиты пользователей корпоративных почтовых систем.

Система обнаружения спама от Google – применяется в Gmail, основана на методах обработки естественного языка, использует Байесовский классификатор, OCR для распознавания текста на изображениях, также применяется Natural Language API. В большей части построена на технологиях искусственного интеллекта и машинного обучения, например, 91,7% спама обнаруживается за счет линейной ML классификации. Gmail предотвращает 3,5% всех фишинговых и спам-атак новых типов, прогнозируя тенденции.

В работе используются результаты исследования [18] (см. табл. 2). Однако в исследовании нет результатов тестирования системы обнаружения спама от Google, поэтому точность данной системы будет считаться равной 99,9%.

Таблица 2

Сравнительный анализ точности работы антиспам систем

Table 2. System antispam accuracy benchmarking

Название антиспам системы	Ошибки первого рода	Ошибки второго рода
Spamoorona	0,45%	0,002%
SpamAssassin	12,93%	0%
Kaspersky Anti-Spam	1,3%	0,002%

Как видно, наилучшие результаты показывают системы, использующие большее количество интеллектуальных методов. Также можно сказать, что это результаты, на которые следует ориентироваться в своих исследованиях.

Практическая реализация

Для проведения экспериментов использовались следующие датасеты:

1. Переведенный на русский язык датасет спам-писем, собранный в 2018 году и состоящий из 5500 писем [19].
2. Тестовый набор данных, состоящий из 100 писем, 50 из которых – спам, составленный по аналогии с письмами из отчета Лаборатории Касперского за 2020 год (в том числе легитимная почта).
3. Датасет, размером 600 сообщений, состоящий из писем реальной почты, предназначен для обучения и тестирования моделей на обнаружение легитимного спама.

Все данные были предварительно обработаны: удалены знаки препинания, тексты приведены к нижнему регистру, составлены словари употребляемых слов.

Был совершен предварительный анализ данных. Так, например, в первом датасете спам-письма были либо слишком короткими (содержали просто ссылку), либо слишком длинными (в попытке запутать антиспам систему).

В Google Collabs с использованием Keras был проведен ряд экспериментов:

1. Для первого датасета применены следующие методы:

– модель ИНС с bag of words, состоящая из 4-х скрытых слоев, максимальная достигнутая точность – 98,2% (рис. 2);

Рис. 2. Результаты работы ИНС с bag of words

Fig. 2. Results of artificial neural network operation with bag of words

– модель ИНС с Embedding-слоем в наилучшей конфигурации достигла точности в 94,9% (рис. 3);

Рис. 3. Результаты работы ИНС с Embedding-слоем

Fig. 3. Results of artificial neural network operation with Embedding-layer

– модель, основанная на методе опорных векторов. Была реализована с помощью библиотеки Sklearn. Точность составила 94,8%.

– модель на основе наивного байесовского классификатора. В библиотеке Sklearn есть готовая модель классификатора на основе «Наивного байесовского алгоритма», которая и применялась. Слова были переведены в BoW-представление. Точность достигла 96,8%.

2. Второй набор данных применяется для тестирования моделей, обученных на первом наборе. Были получены результаты, представленные в таблице 3.

Таблица 3

Результаты тестирования моделей

Table 3. Model test results

Модель	Точность, %
ИНС с BoW	72
ИНС с Embedding-слоем	71
Метод опорных векторов	67
Наивный байесовский классификатор	68

3. Третий набор данных применяется для тестирования к моделям, обученным на первом датасете. Достаточно плохие результаты предсказуемы, ведь модели для решения проблемы легитимного спама должны обучаться на индивидуальных для каждого пользователя наборах данных. При использовании третьего датасета для обучения и тестирования получились результаты, представленные в таблице 4.

Таблица 4

Результаты обучения и тестирования моделей

Table 4. Model training and testing results

Модель	Точность, %
ИНС с BoW	90,5
ИНС с Embedding-слоем	93
Метод опорных векторов	89,5
Наивный байесовский классификатор	91

На основании проведенных экспериментов сделаны следующие выводы:

1. Для переведенного датасета из 5500 писем наибольшую эффективность показала ИНС с методом представления BoW. Максимальная достигнутая точность равна 98,6%, стабильная точность – 98,2%. Более слабый результат показала модель на основе наивного байесовского классификатора, который и используется в большинстве современных антиспам системах.

2. Однако для новых типов писем результаты изменились в пользу ИНС: 72% – для BoW и 71% – для ИНС с Embedding-слоем. Байесовский классификатор показал точность 68%, определив верно в спам лишь 19 писем из 50.

3. При обучении на наборе данных, предназначенном для того, чтобы обучить систему распознавать бесполезные для конкретного пользователя письма, наилучший результат показала ИНС с Embedding-слоем. Таким образом, можно сказать, что ИНС, понимающая смысл сообщений, точнее может определить, что нужно пользователю.

4. Во всех экспериментах модель, основанная на методе опорных векторов, дала наихудший результат.

Заключение

В данной работе проведен анализ методов фильтрации спама, выполнена программная реализация и исследование работы некоторых из них. Составлены датасеты, выявлены новые проблемы в задачах фильтрации спама, предложены возможные решения. Можно сказать, что из классических методов обнаружения спама искусственные нейронные сети являются наилучшим вариантом, однако для понимания интересов конкретного человека необходимо использовать методы обработки естественного языка.

Список литературы

1. Emailstatistics. URL: <https://ipwarmup.com/> (дата обращения: 10.06.2021).
2. Частикова В.А., Жерлицын С.А., Воля Я.И. Нейросетевой подход к решению задачи

построения фоторобота по словесному описанию // Известия Волгоградского государственного технического университета. 2018. № 8 (218). С. 63–67.

3. Global spam volume as percentage of total e-mail traffic from January 2014 to September 2020, by month. URL: <https://www.statista.com/statistics/420391/spam-email-traffic-share/> (дата обращения: 11.06.2021).

4. Rajput P.H.N. Phish Muzzle: This Fish Won't Bite. Los Angeles: University of California, 2017.

5. Спам и фишинг в 2020 году. URL: <https://securelist.ru/spam-and-phishing-in-2020/100408/> (дата обращения: 10.06.2021).

6. Спам и фишинг в I квартале 2021 года. URL: <https://securelist.ru/spam-and-phishing-in-q1-2021/101270/> (дата обращения: 11.06.2021).

7. Security Week 08: спам в 2020 году. URL: <https://habr.com/ru/company/kaspersky/blog/543690/> (дата обращения: 10.06.2021).

8. Changes in anti-spam system. URL: <https://admin.microsoft.com/AdminPortal/Home?ref=MessageCenter&id=MC226683> (дата обращения: 11.06.2021).

9. Microsoft makes anti-spam changes. URL: <https://www.gosolis.com/blog/microsoft-makes-anti-spam-changes/> (дата обращения: 10.06.2021).

10. Независимое тестирование различных АнтиСпам решений (коммерческие и свободные продукты). URL: <https://habr.com/ru/post/56779/> – сравнение антиспамов (дата обращения: 11.06.2021).

11. Number of spam messages received per person per day in South Korea from 1st half 2011 to 2nd half 2019. URL: <https://www.statista.com/statistics/647840/south-korea-spam/> (дата обращения: 10.07.2021).

12 «Лаборатория Касперского»: доля спам-звонков в первом квартале 2021 года составила 70%. URL: https://www.kaspersky.ru/about/press-releases/2021_laboratoriya-kasperskogo-dolya-spam-zvonkov-v-pervom-kvartale-2021-goda-sostavila-70 (дата обращения: 11.06.2021).

13 Spamdexing. URL: <https://inlinks.net/trends/topic?n=Spamdexing> (дата обращения: 10.07.2021).

14. Learn to Combat Spam with Greylisting. URL: <https://www.greylisting.org/> (дата обращения: 27.04.2021).

15. Тарасов Е.С. Разработка лингвосемантических методов обработки экспертной информации для ситуационных центров органов государственной власти: дис. ... канд. техн. наук / Кубанский гос. техн. ун-т. Краснодар, 2011. 198 с.

16. Малыхина М.П., Частикова В.А., Биктимиров А.А. Методика обнаружения спама на основе искусственных иммунных систем // Вестник Астраханского государственного технического университета. Сер.: Управление, вычислительная техника и информатика. 2018. № 3. С. 38–48.

17. Нейросетевая технология обнаружения аномального сетевого трафика / В.А. Частикова, С.А. Жерлицын, Я.И. Воля, В.В. Сотников // Прикаспийский журнал: управление и высокие технологии. 2020. № 1 (49). С. 20–32.

18. Специалисты «Яндекса» назвали самые популярные уловки спамеров в 2020. URL: <https://4pda.to/2021/02/02/381414/> (дата обращения: 27.06.2021).

19. Spam Mails Dataset. URL: <https://www.kaggle.com/venky73/spam-mails-dataset> (дата обращения: 11.06.2021).

References

1. Email statistics. URL: <https://ipwarmup.com/> (access date: 10.06.2021).
2. Chastikova V.A., Zherlitsyn S.A., Volya Ya.I. Neural network approach to the solution of the problem of construction of an identikit using description // News of Volgograd State Technical University. 2018. No. 8 (218). P. 63–67.
3. Global spam volume as percentage of total e-mail traffic from January 2014 to September 2020, by month. URL: <https://www.statista.com/statistics/420391/spam-email-traffic-share/> (access date: 11.06.2021).

4. Rajput P.H.N. Phish Muzzle: This Fish Won't Bite. Los Angeles: University of California, 2017.
5. Spam and fishing in 2020. URL: <https://securelist.ru/spam-and-phishing-in-2020/100408/> (access date: 10.06.2021).
6. Spam i fishing in the first quarter of 2021. URL: <https://securelist.ru/spam-and-phishing-in-q1-2021/101270/> (access date: 11.06.2021).
7. Security Week 08: спам в 2020 году. URL: <https://habr.com/ru/company/kaspersky/blog/543690/> (access date: 10.06.2021).
8. Changes in anti-spam system. URL: <https://admin.microsoft.com/AdminPortal/Home?ref=MessageCenter&id=MC226683> (access date: 11.06.2021).
9. Microsoft makes anti-spam changes. URL: <https://www.gosolis.com/blog/microsoft-makes-anti-spam-changes/> (access date: 10.06.2021).
10. Independent testing of various AntiSpam solutions (commercial and free products). URL: <https://habr.com/ru/post/56779/> – comparison of antispams (access date: 11.06.2021).
11. Number of spam messages received per person per day in South Korea from 1st half 2011 to 2nd half 2019. URL: <https://www.statista.com/statistics/647840/south-korea-spam/> (access date: 10.07.2021).
12. Kaspersky Lab: The share of spam calls in the first quarter of 2021 was 70%. URL: https://www.kaspersky.ru/about/press-releases/2021_laboratoriya-kasperskogo-dolya-spam-zvonkov-v-pervom-kvartale-2021-goda-sostavila-70 (access date: 11.06.2021).
13. Spamdexing. URL: <https://inlinks.net/trends/topic?n=Spamdexing> (access date: 10.07.2021).
14. Learn to Combat Spam with Greylisting. URL: <https://www.greylisting.org/> (access date: 27.04.2021).
15. Tarasov E.S. Development of linguistic-semantic methods of processing expert information for situational centers of public authorities: Diss. for the Cand. of Technology degree / Kuban State Technological university. Krasnodar, 2011. 198 p.
16. Malykhina M.P., Chastikova V.A., Biktimirov A.A. Spam detection technique based on artificial immune systems // Bulletin of Astrakhan State Technical University. Ser.: Management, computer technology and informatics. 2018. No. 3. P. 38–48.
17. Neural network technology for detecting anomalous network traffic / V.A. Chastikova, S.A. Zherlitsyn, Ya.I. Volya, V.V. Sotnikov // Caspian Journal: Management and High Technologies. 2020. No. 1 (49). P. 20–32.
18. Yandex specialists named the most popular tricks of spammers in 2020. URL: <https://w3bsit3-dns.com.to/2021/02/02/381414/> (access date: 27.06.2021).
19. Spam Mails Dataset. URL: <https://www.kaggle.com/venky73/spam-mails-dataset> (access date: 11.06.2021).

Авторы заявляют об отсутствии конфликта интересов.

The authors declare no conflicts of interests.

Статья поступила в редакцию 29.07.2021; одобрена после рецензирования 26.08.2021; принята к публикации 27.08.2021.

The article was submitted 29.07.2021; approved after reviewing 26.08.2021; accepted for publication 27.08.2021.

© В.А. Частикова, К.В. Козачёк, 2021