УДК 004.8

Қ.С. Оразбек*, Б.З. Ернияз
магистр, КазНИТУ им.К. Сатпаева, Алматы, Казахстан
магистр, КазНИТУ им.К. Сатпаева, Алматы, Казахстан
***Автор для корреспонденции:** kadyrzhan111@gmail.com

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ПРИ ОБНАРУЖЕНИИ СПАМ СООБЩЕНИЙ НА КАЗАХСКОМ ЯЗЫКЕ

Аннотация

В статье представлен сравнительный анализ производительности различных алгоритмов машинного обучения для обнаружения спама, с особым акцентом на их применении к казахскому языку. Рассматриваемые методы включают байесовскую фильтрацию спама (MNB), k-ближайших соседей (KNN), опорные векторные машины (SVM) и деревья решений (DT). Традиционно спам использовался для продвижения продуктов и услуг потенциальным клиентам. Однако он превратился в инструмент для взлома и распространения вирусов. Для решения этой проблемы учеными и исследователями были предложены различные методы обнаружения и фильтрации спама. Ниже приведены различные категории методов фильтрации спама: Методы фильтрации спама на основе случаев; Методы фильтрации на основе контента; Методы фильтрации на основе списков; Методы эвристической или основанной на правилах фильтрации спама; Методы адаптивной фильтрации спама. Разработка и оценка различных подходов машинного обучения для обнаружения спама на казахском языке может стать потенциальной исследовательской проблемой. Наша цель в решении этих исследовательских проблем заключается в обогащении текущей литературы путем предложения моделей машинного обучения, которые могут обнаруживать спам-сообщения на казахском языке.

Ключевые слова: машинное обучение, спам, обнаружение спама, методы фильтрации спама, казахский язык, алгоритмы машинного обучения, эффективность.

Введение

В эту эпоху мгновенного подключения понимание ключевой роли, которую играет электронная почта, необходимо для навигации в сложной сети современных коммуникаций. Согласно исследованиям, ожидается, что число пользователей электронной почты во всем мире вырастет примерно до 4,37 млрд в 2023 году. В 2024 году эта цифра составит около 4,48 млрд, а в 2025 году эта цифра, как ожидается, достигнет около 4,59 млрд [1].

По мере увеличения использования электронной почты увеличивается и количество спам-сообщений. Эти нежелательные сообщения могут исходить из любой точки мира, если есть доступ к Интернету. Примечательно, что, несмотря на достижения в области антиспамовых решений и технологий, количество спам-сообщений продолжает расти с тревожной скоростью. Спам — это массовая отправка нежелательных сообщений нескольким получателям. Традиционно спам использовался для продвижения продуктов и услуг потенциальным клиентам. Однако он превратился в инструмент для взлома и распространения вирусов. Для решения этой проблемы учеными и исследователями были предложены различные методы обнаружения и фильтрации спама. Ниже приведены различные категории методов фильтрации спама: Методы фильтрации спама на основе случаев; Методы фильтрации на основе контента; Методы фильтрации на основе списков; Методы эвристической или основанной на правилах фильтрации спама; Методы адаптивной фильтрации спама [2].

Особенности языка: лингвистические особенности казахского языка могут отличаться от особенностей других языков, и существующие методы извлечения признаков могут не быть эффективными при идентификации спам-сообщений на казахском языке. Разработка методов извлечения признаков, специфичных для языка, для казахского языка может стать

исследовательской проблемой для решения.

Ограниченное использование машинного обучения: разработка и оценка различных подходов машинного обучения для обнаружения спама на казахском языке может стать потенциальной исследовательской проблемой. Наша цель в решении этих исследовательских проблем заключается в обогащении текущей литературы путем предложения моделей машинного обучения, которые могут обнаруживать спам-сообщения на казахском языке.

Методы и материалы

Раздел обзора литературы в этой статье содержит всесторонний обзор существующих исследований методов фильтрации спама с использованием алгоритмов машинного обучения. В литературе были предложены различные алгоритмы и методы классификации.

Байесовская фильтрация спама (MNB)

Метод байесовской фильтрации спама включает применение теоремы Байеса. На этапе обучения фильтр вычисляет и запоминает важность каждого слова, найденного в тексте. Позже, когда сообщение получено, оно классифицируется как «спам» или «не спам» на основе того, превышает ли важность его слов заданный предел. Формула ниже иллюстрирует, как вычисляется вероятность того, что сообщение содержит определенное спамовое слово [3]:

$$P(sp \mid w) = \frac{P(sp) * P(w \mid sp)}{P(sp) * P(w \mid sp) + P(nsp) * P(w \mid nsp)}$$

При обнаружении определенного слова вероятность того, что сообщение будет считаться спамом, выражается как P(sp/w). P(sp) количественно определяет вероятность того, что сообщение будет классифицировано как спам, тогда как P(w/sp) обозначает вероятность того, что определенное слово будет встречаться в сообщении со спамом. P(nsp) обозначает общую вероятность того, что сообщение будет отнесено к категории не спам, тогда как P(w/nsp) указывает вероятность того, что определенное слово будет встречено в сообщении, не являющемся спамом.

Метод k ближайших соседей (KNN)

Обнаружение спама методом K-ближайшего соседа (KNN) — это подход, используемый для идентификации спам-сообщений. Алгоритм KNN анализирует характеристики нового сообщения и сравнивает их с характеристиками известных спам-сообщений и не-спам-сообщений. Рассматривая классификации KNN, алгоритм относит новое сообщение к классу (спам или не-спам), который наиболее распространен среди его ближайших соседей. В контексте фильтрации спама это означает, что сообщения, демонстрирующие схожие характеристики с известными спам-сообщениями, также, вероятно, являются спамом [4]. Основным фактором, используемым для идентификации схожих образцов, является мера расстояния. Обычно для определения того, насколько близки образцы друг к другу, используется евклидово расстояние. Когда получено новое сообщение без метки, алгоритм KNN оценивает его сходство с помеченными обучающими примерами, вычисляя евклидово расстояние. Для вычисления расстояния между двумя признаками, xi и xj, и поиска соседних образцов, определяется векторный признак < a1 (x), a2 (x), ..., an (x) > [5]. Расчет расстояния выполняется с использованием уравнения ниже:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r \left(x_i - \left(a_r \left(x_j \right) \right)^2 \right) \right)}$$

Метод опорных векторов (SVM)

Алгоритм Support Vector Machine (SVM) обычно используется для категоризации сообщений как спам или не спам (двоичные классы), хотя его также можно настроить для обработки нескольких классов. Он работает, определяя наиболее подходящую границу, известную как гиперплоскость, для различения положительных и отрицательных случаев. Оценивая эту границу, алгоритм определяет, относится ли новое сообщение к категории спама или не спама [6].

Дерево принятия решений (DT)

Дерево решений (DT) — это тип организованной древовидной структуры, используемой для идентификации спам-сообщений. Это DT использует множество признаков и шаблонов внутри содержимого сообщения, чтобы анализировать их способом, похожим на мыслительный процесс человека, чтобы вынести обоснованное суждение. Чтобы определить, является ли текстовое сообщение спамом или нет, подход DT использует входные слова (F), их частоты (V) и метки (C). Признаки, учитываемые в процессе принятия решения, представлены входными словами (F). Эти связанные со словами частоты (V) информируют нас о том, как часто они появляются в сообщении. Классификация сообщения как спама или нет указывается метками (C). Анализируя эти входные данные, алгоритм DT может точно классифицировать текстовые сообщения на основе их вероятности быть спамом [7].

Результаты

В этом разделе оценивается производительность алгоритмов посредством измерений точности(accuracy), точность(precision), полноты(recall) и F1-оценки. Чтобы оценить точность(accuracy) и точность(precision) классификатора, мы сравниваем прогнозы с фактическими правильными метками. Точность(accuracy) в контексте спам-сообщений вычисляется путем деления количества правильно предсказанных случаев на общее количество предсказаний, сделанных классификатором, в то время как точность(precision) измеряет способность классификатора точно идентифицировать не спам-сообщения. Полнота рассчитывается как количество истинно положительных прогнозов спам-сообщений, деленное на общее количество фактических спам- сообщений, присутствующих в наборе данных. F1-оценка представляет собой комбинированную метрику, которая уравновешивает точность и полноту, беря среднее их гармоническое среднее. В таблице ниже показан результат метрик для каждого метода классификации на тестовом наборе данных:

Таблица 1. Результаты тестирования алгоритмов машинного обучения

Название	Точность (accuracy)	Точность (precision)	Полнота(recall)	F1-оценка
SVM	0.975758	0.944444	0.708333	0.708333
KNN	0.963636	0.875000	0.583333	0.700000
MNB	0.954545	1.000000	0.375000	0.545455
DT	0.939394	0.590909	0.541667	0.565217

На основе полученных метрик можно заметить следующее: SVM имеет самую высокую точность и относительно высокую оценку f1, что уравновешивает точность и полноту. MNB имеет высокую точность, но относительно низкую полноту, что означает, что они хорошо избегают ложных срабатываний (все положительные прогнозы, сделанные этими моделями, верны), но не так хороши в обнаружении всех спам-сообщений. DT имеет относительно низкую точность и оценку f1 по сравнению с другими алгоритмами, что указывает на то, что он не так хорошо работает. SVM и KNN имеют наилучшую общую производительность для обнаружения спама на основе этих метрик.

Выводы

Результаты показывают, что и SVM, и KNN показали хорошие результаты по сравнению

с другими моделями, такими как MNB и DT, на основе всех четырех показателей. Результаты данного исследования поспособствует выбору наиболее эффективного выбора алгоритма машинного обучения при создании системы для обнаружения спам сообщений на казахском языке. В целом, существует значительный потенциал для дальнейшей работы в данной сфере.

Список литературы

1. Julie M. A Beginner's Guide to Successful Economic Marketing in 2023. 2023. Доступно на: https://webnus.net/ru/successful-email-marketing-on-2021/ (от 15 января 2025 г.).
2. Emmanuel G.D., Joseph S.B., Haruna Ch., Shafi'i M.A., Adebayo O.A., Opeyemi EA. Machine learning for email spam filtering: review, approaches and open research problems. 2019.
3. Ганиев С.К., Хамидов Ш.Ж., Олимов И.С. Анализ методов машинного обучения для фильтрации спам-сообщений в почтовых сервисах // In International Conference on Information Science and Communications Technologies (ICISCT), 2020.
4. M. RAZA, N.D. Jayasinghe, and M. M. A. Muslam. A comprehensive review on email spam classification using machine learning algorithms. 2021 International Conference on Information Networking (ICOIN), pages 327–332, Jan. 2021. URL https://doi.org/10.1109/ICOIN50884.2021.9334020.
5. Ş. Seral, P. Kemal, K. Halife, G. Salih. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis // Computers in Biology and Medicine, 2007, vol. 37(3), P. 415–423. https://doi.org/10.1016/j.compbiomed.2006.05.003
6. Niken L.O., Eko H.R., Christy A.S., DRIM Setiadi. Comparison of Multinomial Naïve Bayes Classifier, Support Vector Machine, and Recurrent Neural Network to Classify Email Spams // In International Seminar on Application for Technology of Information and Communication (iSemantic), September 19-20, 2020.
7. A. Wijaya and A Bisri. Hybrid decision tree and logistic regression classifier for email spam detection. 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), pages 1–4, Oct. 2016. URL https://doi.org/10.1109/ICITEED.2016.7863267.

References

1. Julie M. A Beginner's Guide to Successful Economic Marketing in 2023. 2023. Dostupno na: https://webnus.net/ru/successful-email-marketing-on-2021/ (ot 15 janvarja 2025 g.).
2. Emmanuel G.D., Joseph S.B., Haruna Ch., Shafi'i M.A., Adebayo O.A., Opeyemi EA. Machine learning for email spam filtering: review, approaches and open research problems. 2019.
3. Ganiev S.K., Hamidov Sh.Zh., Olimov I.S. Analiz metodov mashinnogo obuchenija dlja fil'tracii spam-soobshhenij v pochtovyh servisah // In International Conference on Information Science and Communications Technologies (ICISCT), 2020.
4. M. RAZA, N.D. Jayasinghe, and M. M. A. Muslam. A comprehensive review on email spam classification using machine learning algorithms. 2021 International Conference on Information Networking (ICOIN), pages 327–332, Jan. 2021. URL https://doi.org/10.1109/ICOIN50884.2021.9334020.
5. Ş. Seral, P. Kemal, K. Halife, G. Salih. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis // Computers in Biology and Medicine, 2007, vol. 37(3), P. 415–423. https://doi.org/10.1016/j.compbiomed.2006.05.003
6. Niken L.O., Eko H.R., Christy A.S., DRIM Setiadi. Comparison of Multinomial Naïve Bayes Classifier, Support Vector Machine, and Recurrent Neural Network to Classify Email Spams // In International Seminar on Application for Technology of Information and Communication (iSemantic), September 19-20, 2020.

7. Wijaya and A Bisri. Hybrid decision tree and logistic regression classifier for email spam detection. 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), pages 1–4, Oct. 2016. URL https://doi.org/10.1109/ICITEED.2016.7863267.

Қ.Оразбек*, Б.З. Ернияз
магистр, Қ.И. Сәтбаев атындағы ҚазҰТЗУ, Алматы, Қазақстан
магистр, Қ.И. Сәтбаев атындағы ҚазҰТЗУ, Алматы, Қазақстан
***Корреспондент авторы:** kadyrzhan111@gmail.com

ҚАЗАҚ ТІЛІНДЕ СПАМ ХАБАРЛАМАЛАРДЫ АНЫҚТАҒАН КЕЗДЕ МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІНІҢ ТИІМДІЛІГІН САЛЫСТЫРУ

Түйін

Мақалада спамды анықтау үшін машиналық оқытудың әртүрлі алгоритмдерінің өнімділігін салыстырмалы талдау, оларды қазақ тіліне қолдануға ерекше назар аударылады. Қарастырылып отырған әдістерге Байес Спамын сүзу, k-жақын көршілер, тірек векторлық машиналар және шешім ағаштары жатады. Дәстүрлі түрде спам әлеуетті клиенттерге өнімдер мен қызметтерді жылжыту үшін пайдаланылды. Алайда ол вирустарды бұзу және тарату құралына айналды. Бұл мәселені шешу үшін ғалымдар мен зерттеушілер спамды анықтау мен сүзудің әртүрлі әдістерін ұсынды. Төменде спамды сүзу әдістерінің әртүрлі санаттары берілген: жағдайларға негізделген спамды сүзу әдістері; мазмұнға негізделген сүзу әдістері; тізімге негізделген сүзу әдістері; эвристикалық немесе спамды сүзу ережелеріне негізделген әдістер; адаптивті спамды сүзу әдістері. Қазақ тілінде спамды анықтау үшін машиналық оқытудың әртүрлі тәсілдерін әзірлеу және бағалау әлеуетті зерттеу проблемасына айналуы мүмкін. Осы зерттеу мәселелерін шешудегі біздің мақсатымыз қазақ тіліндегі спам-хабарламаларды анықтай алатын Машиналық оқыту модельдерін ұсыну арқылы ағымдағы әдебиеттерді байыту болып табылады.

Кілттік сөздер: машиналық оқыту, спам, спамды анықтау, спамды сүзу әдістері, қазақ тілі, машиналық оқыту алгоритмдері, тиімділік.

K.S. Orazbek*, B.Z. Yeryaz
Master's degree, K.I. Satbayev Kazakh NRTU, Almaty, Kazakhstan
Master's degree, K.I. Satbayev Kazakh NRTU, Almaty, Kazakhstan
***Corresponding author's email:** kadyrzhan111@gmail.com

COMPARISON OF THE EFFECTIVENESS OF MACHINE LEARNING ALGORITHMS IN DETECTING SPAM MESSAGES IN KAZAKH

Abstract

The article presents a comparative analysis of the performance of various machine learning algorithms for spam detection, with a special focus on their application to the Kazakh language. The methods under consideration include Bayesian spam filtering, k-nearest neighbors, support vector machines, and decision trees. Traditionally, spam has been used to promote products and services to potential customers. However, it has become a tool for hacking and spreading viruses. To solve this problem, scientists and researchers have proposed various methods for detecting and filtering spam. The following are the different categories of spam filtering methods: Case-based spam filtering methods; Content-based filtering methods; List-based filtering methods; Heuristic or rule-based spam filtering methods; Adaptive Spam filtering methods. The development and evaluation of various machine learning approaches for detecting spam in the Kazakh language may become a potential research problem. Our goal in solving these research problems is to enrich the current literature by offering machine learning models that can detect spam messages in Kazakh. Kazakh language, machine learning algorithms, efficiency.

Keywords: machine learning, spam, spam detection, spam filtering methods, Kazakh language, machine learning algorithms, efficiency.