УДК 004.9:316.472.4

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ КИБЕРЗАПУГИВАНИЯ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ

Зоткина А.А., Мартышкин А.И.

ФГБОУ ВО «Пензенский государственный технологический университет», Пенза,
e-mail: Alena.zotkina.97@mail.ru, Alexey314@yandex.ru

Социальная сеть VKontakte – одна из самых влиятельных платформ для обмена информацией в XXI в., популярность которой растет в геометрической прогрессии. Но, к сожалению, развитие социальной сети оказывает и негативное влияние на пользователей. В частности, к негативным последствиям относятся: киберзапугивание, киберпреступность, онлайн-троллинг и т.д. Киберзапугивание приводит к частым психическим и физическим расстройствам, особенно для учащихся учебных заведений, а иногда даже вынуждает их к попытке самоубийства. Следовательно, идентификация информации в социальной сети является актуальной. Цель данного исследования – прогнозирование и разработка эффективной техники обнаружения запугивающих и оскорбительных сообщений путем обработки естественного языка при помощи методов машинного обучения. В исследовании приведены примеры существующих работ по обнаружению киберзапугивания на основе машинного обучения. Обработка данных на естественном языке происходит в два этапа: с использованием алгоритма «Набор слов» и алгоритма «Частота, обратная частоте документа». Данные характеристики используются для анализа уровня точности четырех различных алгоритмов машинного обучения: дерево решений, наивный Байес, случайный лес, метод опорных векторов. В конце статьи сделаны соответствующие выводы.

Ключевые слова: социальная сеть, дерево решений, наивный Байес, случайный лес, машина опорных векторов

APPLICATION OF MACHINE LEARNING METHODS IN THE TASK OF PREDICTING CYBERBULLYING OF SOCIAL NETWORK USERS

Zotkina A.A., Martyshkin A.I.

Penza State Technological University, Penza,
e-mail: Alena.zotkina.97@mail.ru, Alexey314@yandex.ru

VKontakte social network is one of the most influential platforms for information exchange in the 21st century, the popularity of which is growing exponentially. But, unfortunately, the development of the social network also has a negative impact on users. In particular, the negative consequences include: cyberbullying, cybercrime, online trolling, etc. Cyberbullying leads to frequent mental and physical disorders, especially for students of educational institutions, and sometimes even forces them to attempt suicide. Therefore, the identification of information in the social network is relevant. The purpose of this study is to predict and develop an effective technique for detecting intimidating and abusive messages by processing natural language using machine learning methods. The study provides examples of existing work on the detection of cyberbullying based on machine learning. Data processing in natural language takes place in two stages: using the "Set of words" algorithm and the "Frequency inverse to the frequency of the document" algorithm. These characteristics are used to analyze the accuracy level of four different machine learning algorithms: decision tree, naive Bayes, random forest, support vector machine. At the end of the article, the relevant conclusions are drawn.

Keywords: social network, decision tree, naive Bayes, random forest, support vector machine

Социальные сети – популярный способ для общения, обмена информацией, создания социальных отношений между людьми. Пользователи проводят значительное количество времени в популярных социальных сетях, храня и обмениваясь большим количеством личной информации. В социальных сетях пользователи не только размещают письменный и мультимедийный контент, но и выражают свои чувства, эмоции и настроения. Именно поэтому в настоящее время социальные сети используются в различных секторах, таких как образование, бизнес и т.д., не только для обмена информацией, но и для проведения различных маркетинговых и социальных исследований. Социальные сети укрепляют мировую экономику, создавая множество новых рабочих мест. Хотя социальные сети имеют много преимуществ, у них также есть некоторые недостатки. Используя это средство массовой информации, злонамеренные пользователи совершают неэтичные и мошеннические действия, чтобы задеть чувства других и нанести ущерб их репутации. В последнее время киберзапугивание стало одной из основных проблем социальных сетей. По мере роста цифровой сферы и развития технологий киберзапугивание стало относительно распространенным явлением, особенно среди учащихся образовательных учреждений.

Примерно 76% пользователей социальных сетей в 2020–2021 гг. стали жертвами киберпреступлений. Киберпреступление оказывает физическое и психическое воз-

действие на жертву. Жертвы выбирают саморазрушительные действия, такие как самоубийство, из-за травмы от киберзапугивания, которую трудно пережить. Таким образом, выявление и предотвращение киберзапугивания важно для защиты учащихся.

Исходя из вышеописанного, целью данного исследования является прогнозирование и разработка эффективной техники обнаружения запугивающих и оскорбительных сообщений путем обработки естественного языка при помощи методов машинного обучения.

В связи с этим предложена модель обнаружения киберзапугивания, которая основана на машинном обучении. Модель может определить, относится ли текст к киберзапугиванию или нет. Было исследовано несколько алгоритмов машинного обучения, включая наивный байесовский классификатор, метод опорных векторов, дерево решений и случайный лес. В предлагаемой модели обнаружения киберзапугивания проведены эксперименты с двумя наборами данных из комментариев и постов социальной сети VKontakte.

Материалы и методы исследования

Существует несколько работ по обнаружению киберзапугивания на основе машинного обучения. Д. Инь, З. Сюэ, Л. Хонг, Б. Дэвисон, А. Контостатис и Л. Эдвардс предложили контролируемый алгоритм машинного обучения, использующий подход «мешок слов» для определения настроения и контекстуальных особенностей предложения. Этот алгоритм показывает 61,9% точности. К. Рейнольдс, А. Контостатис использовали метод машины опорных векторов для обнаружения киберзапугивания комментариев на YouTube. Результат проекта был улучшен до 66,7% точности для применения вероятностного моделирования. Чтобы улучшить обнаружение киберзапугивания, автор статьи использовал в качестве характеристики личности, эмоции и сентименты [1].

Также было введено несколько моделей, основанных на глубоком обучении, для обнаружения киберзапугивания. Модель на основе глубокой нейронной сети применяется для обнаружения киберзапугивания с использованием реальных данных [2]. Авторы сначала систематически анализируют киберзапугивание, а затем используют трансфертное обучение для выполнения задачи обнаружения. Баджатия и др. [3] представили метод, использующий архитектуры глубоких нейронных сетей для обнаружения ненавистнических высказываний. Для обнаружения киберзапугивания была предложена модель на основе сверточной нейронной сети [4]. Авторы использовали встраивание слов там, где похожие слова имеют аналогичное встраивание. За последние несколько десятилетий многие работы по киберзапугиванию были сосредоточены на анализе текста. Разнообразие данных об издевательствах на социальных платформах не может быть удовлетворено обычными методами текстового анализа.

Система обнаружения киберзапугивания состоит из двух основных частей: обработка естественного языка (*NLP – Natural Language Processing*) и машинное обучение (*ML – machine learning*).

На первом этапе наборы данных, содержащие оскорбительные тексты, сообщения и публикации, собираются и подготавливаются для алгоритмов машинного обучения с использованием обработки естественного языка [5]. Далее обработанные наборы данных используются для обучения алгоритмов машинного обучения.

Обработка на естественном языке: сообщения или текст в реальном мире содержат различные ненужные символы или текст. Например, цифры или знаки препинания не имеют отношения к выявлению издевательств. Прежде чем применять алгоритмы машинного обучения к комментариям, нам нужно очистить и подготовить их к фазе обнаружения. На этом этапе выполняются различные задачи обработки, включая удаление всех нерелевантных символов, таких как стоп-слова («если», «но», «а» и т.д.), знаки препинания и цифры, токенизация (разделение строк на более мелкие части, называемые токенами, например разбивка на основе пробелов, знаков препинания) и т.д. [6]. Данные шаги необходимы для уменьшения шума, который наблюдается у любого текста, а также для повышения точности результатов классификатора.

После предварительной обработки информация разбивается на части по двум подходам:

1. «Мешок слов» (*Bag-of-Word)*. Прежде чем применять алгоритмы машинного обучения, мы должны преобразовать текст в векторы или числа, так как алгоритмы не могут работать с необработанным текстом. После обработанные данные преобразуются в набор слов для следующего этапа. Это называется «мешком» слов, потому что всякая информация о порядке или структуре слов в документе отбрасывается.

2. Частота, обратная частоте документа – это статистический показатель, который позволяет оценить, насколько релевантно слово для документа в коллекции документов. В отличие от подхода «Мешок

слов», здесь словам, которые встречаются в тексте чаще, придают большее значение (поскольку они полезны для классификации), чем в пункте 1, когда каждому слову присваивается одинаковое значение.

Машинное обучение: этот модуль включает в себя применение различных подходов к машинному обучению, таких как дерево решений (*DT*), случайный лес (*RF*), машина опорных векторов (*SVM*), наивный Байес (*NB*) для обнаружения оскорбительного сообщения и текста. Классификатор с наивысшей точностью обнаруживается для конкретного общедоступного набора данных о киберзапугивании.

1. Дерево решений: классификатор дерева решений может использоваться как для классификации, так и для регрессии [7]. Это может помочь представить решение, а также принять решение. Дерево решений – это древовидная структура, где каждый внутренний узел представляет условие, а каждый конечный узел представляет решение. Дерево классификации возвращает класс, к которому относится цель. Дерево регрессии выдает прогнозируемое значение для адресованного входного сигнала.

2. Наивный Байес – эффективный алгоритм машинного обучения, основанный на теореме Байеса [8]. Исходя из названия, следует, что все переменные в наборе данных «наивные», т.е. не коррелируют друг с другом. Использование теоремы Байеса с сильным предположением о независимости между признаками является основой наивной байесовской классификации. Алгоритм предсказывает в зависимости от вероятности объекта. Проблемы бинарной и многоклассовой классификации могут быть быстро решены с помощью этого метода. Основываясь на теореме Байеса, находится вероятность наступления события с учетом вероятности другого события, которое уже произошло.

3. Случайный лес: классификатор случайного леса состоит из нескольких классификаторов дерева решений [9]. Каждое дерево дает прогноз класса индивидуально. Максимальное количество прогнозируемого класса – это наш конечный результат. Этот классификатор представляет собой модель контролируемого обучения, которая обеспечивает точный результат, поскольку для получения результата объединяются несколько деревьев решений. Вместо того чтобы полагаться на одно дерево решений, случайный лес берет прогноз из каждого сгенерированного дерева и на основе большинства голосов за прогнозы определяет окончательный результат.

4. Метод опорных векторов: *SVM* – это контролируемый алгоритм машинного обучения, который может применяться как для классификации, так и для регрессии дерева решений. Он может однозначно различать классы в n-мерном пространстве [10]. Таким образом, *SVM* выдает более точный результат, чем другие алгоритмы, за меньшее время. На практике *SVM* создает множество гиперплоскостей в бесконечномерном пространстве, а *SVM* реализуется с помощью ядра, которое преобразует пространство входных данных в требуемую форму.

Результаты исследования и их обсуждение

В данной работе было использовано четыре алгоритма машинного обучения, чтобы классифицировать комментарии как издевательства или не издевательства. На данном этапе были собраны комментарии пользователей социальной сети *VKontakte* из разных постов. Социальная сеть снабжена методами для взаимодействия и извлечения информации при помощи VK_API.

Чтобы обратиться к методу API ВКонтакте, необходимо выполнить POST или GET запрос следующего вида: requests.get (‘https://api.vk.com/method/wall.get’). Данный запрос состоит из нескольких частей, таких как params – входные параметры, в состав которых входят token – ключ доступа, v – используемая версия API. Ниже представлена часть кода обращения к методу API:

token = ‘b6e60a65b6e60a65b6e60a65ffb69e88f8bb6e6b6e60a65d603965d2127f9cdc8a7beb8’
version = 5.131
domain = ‘___’
count = 100
offset = 0

Тексты или комментарии были разделены на два типа следующим образом:

− текст без издевательств;
− текст с издевательством.

Алгоритмы обнаружения издевательств реализованы с использованием пакета машинного обучения на языке программирования *Python* [11]. *Python* – язык программирования, считается высокоуровневым, поддерживает динамическую строгую типизацию, т.е. переменная начинает работать с типом в момент ее присваивания, что обозначает, что одна и та же переменная может принимать различные типы данных. Еще одним преимуществом использования *Python* является свойство автоматического управления памятью.

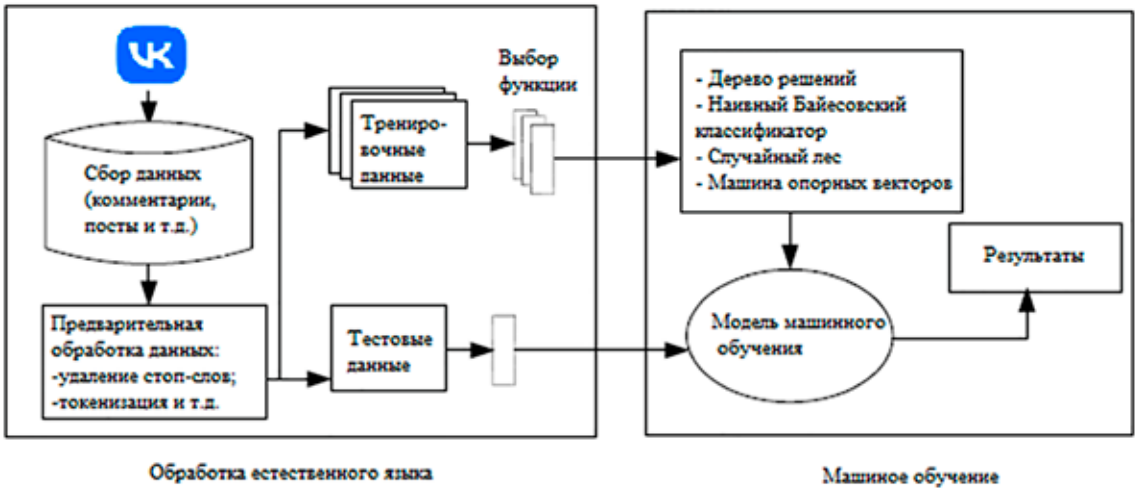

Рис. 1. Предлагаемая модель обнаружения киберзапугивания пользователей социальной сети

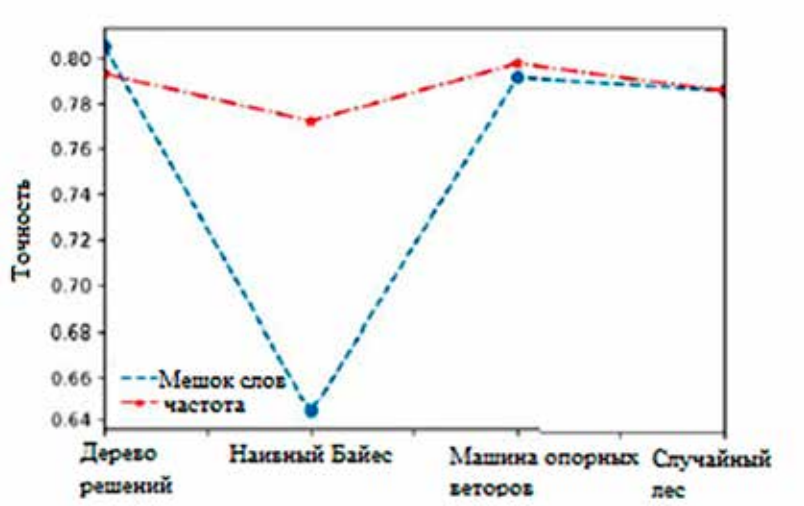

Рис. 2. Результаты точности двух описанных подходов

Существует множество ресурсов, которые облегчают разработку в машинном обучении с использованием *Python*, больше, чем для любого другого языка. Использование специальных инструментов, таких как пакеты *pandas* – библиотека с открытым исходным кодом, предоставляющая высокопроизводительные, простые в использовании структуры данных и инструменты анализа для языка программирования *Python*, и *numpy*, позволяет достичь высокой производительности в обработке данных.

На рис. 1 показана модель предлагаемого решения для обнаружения киберзапугивания пользователей социальной сети *Vkontakte*.

На рис. 2 показаны результаты точности машинного обучения. Следуя полученным результатам, видно, что *SVM* превосходит все алгоритмы. Результаты также показывают, что подход, основанный на частоте встречающихся слов, обеспечивает лучшую точность, чем «Мешок слов». Это связано с тем, что вместо того, чтобы разбивать почти все слова на векторы, второй подход использует наиболее часто встречающиеся слова и обеспечивает лучшую производительность.

На рис. 3 представлены кривые рабочих характеристик приемника для обеих функций. Для подхода «Мешок слов» и «Частота, обратная частоте документа» ясно, что *SVM* обеспечивает более высокую производительность, чем другие алгоритмы машинного обучения.

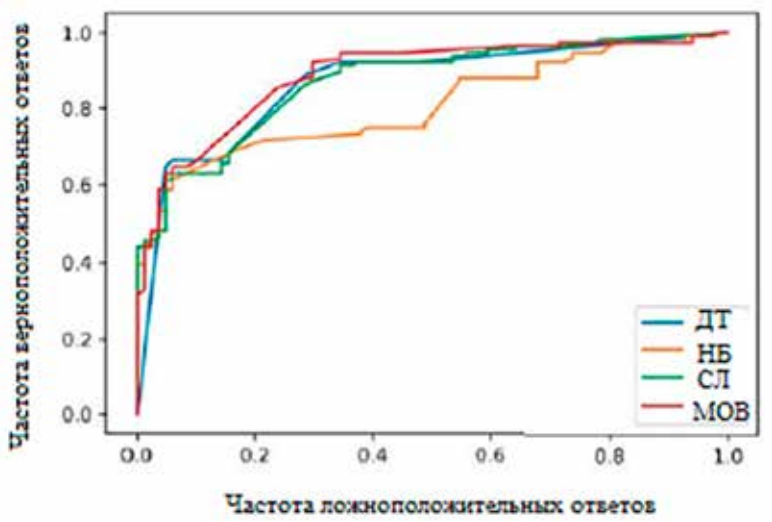

Рис. 3. Результаты рабочих характеристик приемника для четырех видов машинного обучения

Заключение

В настоящее время киберзапугивание стало распространенным явлением и начало вызывать серьезные социальные проблемы в связи с развитием социальных сетей и ростом их использования. Учитывая важность обнаружения киберзапугивания, в этом исследовании была исследована автоматическая идентификация сообщений в социальных сетях, связанных с киберзапугиванием, с учетом двух подходов «Мешок слов» и «Частота, обратная частоте документа». Для выявления запугивающего текста использовались четыре алгоритма машинного обучения: случайный лес, дерево решений, наивный Байес, метод опорных векторов, среди которых было выявлено, что *SVM* обеспечивает более высокую производительность, чем другие алгоритмы машинного обучения.

Список литературы

1. Balakrishnan V., Khan S., Arabnia H.R. Improving cyberbullying detection using twitter users' psychological features and machine learning. Computers & Security. 2020. Vol. 90. P. 101710.

2. Agrawal S., Awekar A. "Deep learning for detecting cyberbullying across multiple social media platforms" in European Conference on Information Retrieval. Springer. 2018. P. 141–153.

3. Badjatiya P., Gupta S., Gupta M., Varma V. "Deep learning for hate speech detection in tweets" in Proceedings of the 26th International Conference on World Wide Web Companion. 2017. P. 759–760.

4. Al-Ajlan M.A., Ykhlef M. "Deep learning algorithm for cyberbullying detection". International Journal of Advanced Computer Science and Applications. 2018. Vol. 9. No. 9.

5. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М.: Изд-во НИУ ВШЭ, 2017. 269 с.

6. Семантический анализ для автоматической обработки естественного языка. [Электронный ресурс]. URL: https://rdc.grfc.ru/2021/09/semantic_analysis/ (дата обращения: 12.10.2022).

7. Деревья решений в машинном обучении. [Электронный ресурс]. URL: https://biconsult.ru/products/derevya-resheniy-v-mashinnom-obuchenii (дата обращения: 12.10.2022).

8. Наивный байесовский классификатор. [Электронный ресурс]. URL: http://bazhenov.me/blog/2012/06/11/naive-bayes.html (дата обращения: 12.10.2022).

9. Случайный лес [Электронный ресурс]. URL: https://alexanderdyakonov.wordpress.com/2016/11/14/% D1%81%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D1%8B%D0%B9-%D0%BB%D0%B5%D1%81-random-forest/ (дата обращения: 12.10.2022).

10. Метод опорных векторов (Support Vector Machines). [Электронный ресурс]. URL: http://statsoft.ru/home/textbook/modules/stmachlearn.html#Support_Vector_Machines (дата обращения: 12.10.2022).

11. Маккинни У. Python и анализ данных / Пер. с англ. А.А. Слинкина. 2-е изд., испр. и доп. М.: ДМК Пресс, 2020. 540 с.