УДК 004.93275+519.764

Ю.Б. САВВА, Т.Ю. САВВА

Yu.B. SAVVA, T.Yu. SAVVA

**ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ВЫЯВЛЕНИЯ ОБФУСЦИРОВАННЫХ СООБЩЕНИЙ УЧАСТНИКОВ ВИРТУАЛЬНЫХ СОЦИАЛЬНЫХ СЕТЕЙ ПРИ КОМПЬЮТЕРНОМ АНАЛИЗЕ РУССКОЯЗЫЧНЫХ ТЕКСТОВ**

**INFORMATION TECHNOLOGY OF MACHINE LEARNING FOR DETECTING OBFUSCATED MESSAGES OF PARTICIPANTS OF ONLINE SOCIAL NETWORKS AT THE COMPUTER ANALYSIS OF TEXTS WRITTEN IN RUSSIAN**

*Описана информационная технология машинного обучения для выявления обфусцированных текстов, которыми обмениваются участники виртуальных социальных сетей при ведении ими противоправной и деструктивной деятельности. Приведены алгоритм классификации и оценка эффективности использования рассматриваемой технологии.*

*Ключевые слова: виртуальные социальные сети, информационные технологии, машинное обучение, компьютерный анализ текстов*

*Information technology of machine learning is described for revealing obfuscated texts exchanged by participants of the online social networks when they conduct unlawful and destructive activities. An algorithm for classifying and evaluating the effectiveness of the technology used is presented.*

*Keywords: online social networks, information technologies, machine learning, computer analysis of texts*

В настоящее время виртуальные социальные сети (ВСС) стали играть важную роль в жизни общества. К сожалению, в этих сетях, как и в реальной жизни, нашли свое место как отдельные их участники, так организованные сообщества, ведущие противоправную и деструктивную деятельность, заключающуюся в пропаганде терроризма, экстремизма, ксенофобии, употребления наркотических и психотропных веществ, а также вовлечении других участников ВСС в различные секты и террористические организации.

Очевидно, что подобные злоумышленники находят разнообразные средства, чтобы их преступные действия оставались незаметными для правоохранительных органов. В частности, тексты сообщений противозаконного содержания намеренно искажаются, обфусцируются, таким образом, чтобы они были понятны для людей, но оказались непригодными для распознавания компьютерными текстовыми анализаторами. Проблема вскрытия обфусцированных текстов сообщений участников ВСС рассмотрена в работе [1]. Для обеспечения возможности обнаружения подобных сообщений реализована информационная технология анализа текстов сообщений, способная с определенной долей вероятности выявлять случаи обфускации текста и декодировать его, которая использована в «Интеллектуальной системе поддержке принятия решений при выявлении противоправной и деструктивной деятельности в виртуальных социальных сетях «Невод-ВСС». Одним из критериев по определению участника сети как подозреваемого в ведении противоправной и деструктивной деятельности является наличие и достаточно частое использование в текстах его сообщений терминов определённой противоправной категории, например, такой, как наркомания, терроризм, мошенничество.

В современных поисковых и почтовых системах, а также популярных интернет-сервисах, как правило, присутствуют инструменты анализа контента. Примером подобного средства будет являться спам-фильтр почтовых служб. Так как зачастую алгоритм работы большинства из них основан на поиске ключевых слов в анализируемом тексте, обойти данную защиту злоумышленникам не составляет труда, им достаточно лишь изменить написание подозрительного слова в тексте сообщения, обфусцировав его. Кроме того, данным способом можно обойти и более продвинутые средства, например, основанные на семантическом анализе. Становится понятно: какой бы эффективной не была работа системы компьютерного анализа текста, её действия бесполезны в случае, если часть слов не может быть обработана должным образом. Отсюда следует, что для решения проблемы обнаружения и вскрытия обфусцированных тестов сообщений необходимо разработать метод, позволяющий производить первоначальный анализ текста с целью выявления в нём обфусцированных фрагментов и его декодирования для последующих их исследований.

Очевидно, что общий поток сообщений, генерируемых участниками различных ВСС, составляет десятки миллионов в сутки. В связи с этим для компьютерного анализа текстов этого множества сообщений ключевой задачей является определение вероятности наличия в исследуемом тексте обфусцированных фрагментов по количественным характеристикам текста, что позволит с определённой долей достоверности отсеивать тексты, не нуждающиеся в приведении к виду, в котором их анализ осуществим и эффективен. Для решения этой задачи могут быть использованы следующие методы:

а) поиск ключевых слов, требующий хранения самых распространённых способов написания слов с искажениями с последующей проверкой текста на их наличие. Данный метод является тривиальным и достаточно громоздким, так как необходимо хранить очень большой набор слов, который необходимо постоянно пополнять. Если какой-либо способ обфускации применяется впервые, данный метод не позволит его обнаружить;

б) регулярные выражения. При использовании этого метода требуется задать набор правил все возможных обфускаций каждого слова, которое в дальнейшем проверяется на соответствие данного набора, и если оно совпадает, то делается вывод о том, что сообщение содержит искажение. Данный метод имеет более высокий уровень абстракции по сравнению с предыдущим, но также является объёмным и малопригодным для реализации, так как регулярные выражения также должны учитывать все возможные случаи обфускации, набор правил должен постоянно пополняться человеком, знающим язык регулярных выражений. Проблема вновь используемых способов обфускации также присутствует;

г) машинное обучение. Основная цель данного метода – вывести некоторые факты и знания о конкретном объекте на основе исходных данных с помощью некоторого алгоритма. Для этого составляется база знаний об объектах исследования, в которой для каждого из них указываются значения параметров, или признаков, характерных для всех из них. Далее, в соответствии с целями и задачами, поставленными перед разработчиком, выбирается алгоритм машинного обучения.

Преимущества методов машинного обучения в данном случае очевидны: не требуется хранить большие объёмы данных сложной структуры. Зная общие закономерности и способы искажения текстов, можно делать выводы об их возможной обфускации, не оперируя большой коллекцией частных случаев. При этом от исследователя в данной ситуации требуется правильное построение обучающей выборки и выявление необходимых признаков объектов, что позволит свести минимуму вероятность получения ошибочного решения. Поэтому именно метод машинного обучения был положен нами в основу рассматриваемой информационной технологии выявления обфусцированных сообщений участников ВСС при компьютерном анализе русскоязычных текстов.

Анализ предметной области, для которой рассматриваемой информационной технологии позволил сделать вывод о том, что случай выявления обфускации в текстах сообщений участников ВСС относится к классу задач бинарной классификации, т.е. в соответствии выявленными особенностям текста по некоторым количественным признакам сообщению можно присвоить класс «обфусцировано» и «не обфусцировано».

Использование рассматриваемой информационной технологии обеспечивает решение с использованием статистических методов при построении обучающих выборок следующих задач:

* + 1. Классификация поступающих в систему «Невод-ВСС» текстов сообщений с целью выявления наличия в них обфусцированных фрагментов. Это действие необходимо для отсеивания доли текстов сообщений, подлежащих анализу без декодирования и, следовательно, оптимизации работы алгоритма вскрытия обфускации и всей системы в целом. Данная задача решается путём реализации алгоритма машинного обучения, наиболее эффективно работающего со случаями бинарной классификации, а также сбора и анализа признаков обучающей выборки текстов сообщений. По результатам предварительных исследований в качестве метода классификации на данном этапе был выбран алгоритм логистической регрессии.

Для достижения достаточной эффективности работы метода классификации текстов сообщений, в рассматриваемой информационной технологии осуществляется проведение двух этапов классификации текста.

На первом этапе сообщение исследуется на наличие в тексте символов, не принадлежащих кириллице, для выявления случаев обфускации текста методом замены или вставки посторонних символов. Для построения данного классификатора использован метод логистической регрессии, так как на этапе предварительных исследований данный метод показал наилучший показатель эффективности. Первый этап позволяет сразу классифицировать тексты, содержащие недопустимое количество посторонних символов.

Если после прохождения первого этапа классификации текст был определён как необфусцированный, к сообщению применяется второй этап классификации, проводящий обнаружение искажения текста по дополнительным признакам. Данное дополнение исследует текст по количественным характеристикам сочетаемости последовательности букв каждого слова. Доказано, что при достаточно большой длине текста характеристики сочетаемости гласных и согласных букв являются устойчивыми [2]. Применение второго этапа классификации позволит с некоторой долей достоверности выявлять по вышеописанным количественным признакам случаи пропуска, замены и перестановки символов кириллицы в текстах, не прибегая при этом к более трудо- и ресурсоёмким средствам семантического анализа. В качестве классификатора на данном этапе применяется метод k ближайших соседей.

На рисунке 1 представлен алгоритм классификации текстов сообщений участников ВСС в виде диаграммы деятельности на языке UML, объясняющей логику его работы.

Как видно из диаграммы, процесс классификации текстов всегда должен начинаться с чтения файла, хранящего данные обучающей выборки. Активность «Настроить параметры модели» заключается отделении от признаков выборки значений целевой переменной. Затем выполняется действие по обучению модели.

Непосредственная классификация представляет собой циклический процесс, выполняемый для каждой строки текстов сообщений, поступающих в систему из текстового файла, и заключается в вычислении параметров строки – определении признаков, по которым проводится классификация, и вычисления оценки принадлежности к классам – применение алгоритма классификации по обученной ранее модели к признакам строки текста.

Разделение процесса классификации на два этапа в данном случае оправдано тем, что для разных способов обфускации текста необходимо выделять признаки, составляющие разные по своей структуре пространства состояний объектов. Так, для случаев вставки и замены с использованием латинских символов, цифр или специальных знаков, пространство состояний представляется линейно разделимым. Тогда как в случае анализа последовательностей кириллических символов на обнаружение вставки, замены или пропуска можно говорить о том, что пространство состояний будет определяться объектами, принадлежащими некоторой n-мерной области, определяющей некоторые значения признаков, характерные для одного класса, и объектами, выходящими за её пределы. Таким образом, совместное использование и последовательное применение вышеописанных методов классификации позволяет выявлять практически все возможные основные случаи обфускации текста сообщения.



Рисунок 1 – Алгоритм классификации текстов сообщений

* + 1. Сбор, обработка, оценка и интерпретация статистических данных русского языка для реализации и эффективного функционирования статистической модели языка в рамках алгоритма вскрытия обфускации текстов сообщений. Поскольку для реализации алгоритма вскрытия обфускации в качестве модели представления данных в системе «Невод\_ВСС», как описано в работе [1], используется скрытая марковская модель (СММ), то необходимо рассчитать её параметры, в частности, такие как вероятностное распределение наблюдаемых состояний. Данная задача решается с использованием таблицы возможных обфускаций (наблюдаемых состояний модели) соответствующих символам русского алфавита (скрытых состояний модели), представленной в работе [1], Далее проводится оптимизация начальных предположений о параметрах модели на основе размеченной обучающей выборки. В качестве метода оптимизации используется EM-алгоритм оценки максимального правдоподобия.
    2. Классификация участников социальной сети по данным из их сообщений. Данная задача решается путём построения тематической модели сообщений участников ВСС и программы [3]. Это позволяет проводить классификацию текстов сообщений (формально, определять категорию смоделированных тем), давать рекомендации по добавлению в словарь новых терминов с указанием возможной их категории, а также проводить классификацию участников виртуальной социальной сети на основе категорий их сообщений. На основе текстового корпуса сообщений участников сети, методами тематического моделирования формируется тематическая модель, определяющая вероятностное распределение тем в анализируемых сообщениях и вероятностное распределение слов в темах. В качестве метода тематического моделирования используется алгоритм латентного распределения Дирихле (LDA, latent Dirichlet allocation). Далее к распределениям слов каждой темы применяется метод классификации, для определения темы к конкретной категории. Для этого необходима обучающая выборка, фактически являющаяся словарём, где каждое слово помечено категорией. Пример одной из категорий, а именно «Жаргон в сфере незаконного оборота наркотических средств и психотропных веществ», представлен в [4] После этого принимается решение о принадлежности каждого участника ВСС к определённому заранее классу, характеризующему подозрительность содержимого сообщений участника сети. Для двух вышеперечисленных действий используется наивный байесовский классификатор в качестве метода реализации. Также одним из результатов тематического моделирования является выявление новых слов, не найденных в базах данных системы «Невод-ВСС» и рекомендация пользователям этой системы о присвоении им возможной категории.

Для повышения эффективности использования тематической модели программным способом реализуется следующее преобразование над текстом: перед применением к сообщениям методов тематического моделирования, из каждого текста происходит удаление так называемых стоп-слов – конструкций языка, не представляющих интерес с точки зрения понимания темы документа. Это, например, предлоги, частицы, союзы, местоимения, междометия, вводные слова, наречия места, времени, и т.д. Для удаления таких слов был составлен список стоп-слов русского языка, который помещён в словарь системы «Невод-ВСС», в котором каждое слово этого списка было помечено соответствующей категорией, имеющей название «stopword».

Для обеспечения функционирования рассматриваемой информационной технологии необходим следующий набор входных данных:

а) тексты сообщений участников ВСС, полученные из открытых источников (сообщения со «стены» в социальной сети собираются специальной программой обходчиком графа сети – т.н. «краулером», которая входит в состав системы «Невод-ВСС») в текстовом файле;

б) обучающая выборка признаков сообщений, хранящаяся в табличной форме в svg-файле;

в) декодированные тексты сообщений для построения тематической модели;

г) словари русского языка с размеченной категорией каждого слова, представленные в системе «Невод-ВСС» в форме специализированных по определенным темам баз данных.

В результате компьютерной обработки и анализа текстов из потока сообщений участников ВСС методом машинного обучения в системе «Невод-ВСС» формируются следующие данные:

а) данные о принадлежности сообщений участника ВСС к классу обфусцированных;

б) категории тем сообщений участника ВСС;

в) проценты сообщений участника ВСС по каждой из выявленных категорий, относящихся к противоправным, в которых обнаружено употребление терминов каждой категории;

г) данные о принадлежности участников ВСС к определённому классу опасности (классы опасности предварительно задаются пользователем системы «Невод-ВСС»);

д) новые часто используемые в текстах сообщений термины, не включённые в словари, хранящихся в базах данных системы «Невод-ВСС».

При разработке информационной технологии машинного обучения для выявления обфусцированных текстов сообщений как системы «Невод-ВСС» учитывалось, что использовать ее в практической работе будут оперативные работники компетентных органов не знакомые с методами машинного обучения и компьютерного анализа текстов на естественных и искусственных языках. Поэтому от пользователей системы при работе с ней требуется только первоначальное определение категорий противоправной и деструктивной деятельности и последующая их корректировка, а также оценка новых терминов на предмет их отнесения к той или иной категории. Кроссплатформенность разработанной информационной технологии как системы «Невод-ВСС» обеспечивается использованием языка программирования Python и системы управления базами данных PostgeSQL.

Для оценки эффективности использования разработанной информационной технологии классификации текстов сообщений участников ВСС в рамках системы «Невод-ВСС», реализованный метод классификации был применён к ста тысячам строк текстов из сети «ВКонтакте», находящихся в открытом доступе (статьи, комментарии пользователей). В данном контексте положительным классом является класс, включающий обфусцированные сообщения. В класс Истинный\_Отрицательный попали сообщения, не содержащие элементы обфускации текста. По результатам проверки работы метода классификации, доля истинных положительных классификации составляет 41%, доля истинных отрицательных – 42%, доля ложных положительных (ошибок первого рода) – 12%, ложных отрицательных (ошибок второго рода) – 5% (см. рисунок 2).

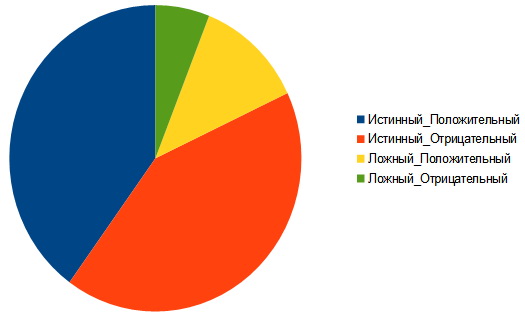


Рисунок 2 – Диаграмма оценки эффективности классификации текстов

СПИСОК ЛИТЕРАТУРЫ

1. Никольская А.Г., Савва Ю.Б. О проблеме вскрытия обфусцированных русскоязычных текстов участников виртуальных социальных сетей // Информационные системы и технологии. № 6 (98). 2016. С.44-55.
2. Варфоломеев, А.А. Информационная безопасность. Математические основы криптологии. Учебное пособие, ч.1.[Текст]/ А.А. Варфоломеев, В.М. Фомичев. – М.: МИФИ, 1995. – 114 с.
3. Савва, Ю.Б. Программа кластеризации и шкалирования нечетких данных: свидетельство об офиц. регистрации программы для ЭВМ № 2015612445 Рос. Федерация/ Ю.Б. Савва; заявитель и правообладатель ФГБОУ ВПО «Госуниверситет – УНПК» (RU). – № 2014663471; заявл. 23.12.2014; зарегистрировано в реестре программ для ЭВМ 18.02.2015. – [1] с.
4. Савва, Ю.Б. Жаргон в сфере незаконного оборота наркотических средств и психотропных веществ: свидетельство о государственной регистрации базы данных № 2016620197 Рос. Федерация/ Ю.Б. Савва, Ю.В. Давыдова, В.Т. Еременко; заявитель и правообладатель ФГБОУ ВО «Приокский государственный университет» (RU). – № 2015621008; заявл. 28.07.2015; зарегистрировано в реестре баз данных 10.02.2016. – [1] с.

**Савва Юрий Болеславович**

Орловский государственный университет имени И.С. Тургенева, г. Орел

К.т.н., доцент кафедры информационной безопасности

Тел. +7-910-207-61-85

E-mail: su\_fio@mail.ru

**Савва Татьяна Юрьевна**

Орловский государственный университет имени И.С. Тургенева, г. Орел

К.т.н., доцент кафедры информационных систем

Тел. +7-953-620-14-80

E-mail: t.savva@mail.ru