УДК 004.42

АВТОМАТИЗИРОВАННОЕ ЦЕНЗУРИРОВАНИЕ ТЕКСТА

Шалаев Алексей Дмитриевич
г. Москва, Школа № 1537 "Информационные технологии", 11 класс
Научный руководитель: Минченко Михаил Михайлович, г. Москва, Школа № 1537, учитель информатики, канд. эконом. наук

Российское законодательство квалифицирует использование в общении нецензурных выражений как административное нарушение. С 1 февраля 2021 года на территории Российской Федерации вступила в силу новая редакция Федерального закона от 27 июля 2006 г. № 149-ФЗ "Об информации, информационных технологиях и защите информации" [1], которая предписывает обязательное удаление матерных постов из соцсетей.

Для достижения цели снижения употребления нецензурной лексики в речи молодежи нужно искать различные методы. Одним из таких методов является принятие закона, запрещающего использовать нецензурную лексику, при нарушении закона придется заплатить административный штраф. Автоматизированная система (АС), описанная в данной работе, так же может рассматриваться как одни из методов борьбы с нецензурной лексикой.

Создание представляемой АС ориентировано на поддержку функционала, обеспечивающего автоматизированный анализ текста на предмет выявления и последующего исключения нецензурной лексики. Анализу и корректировке можно подвергнуть текстовый контент из различных источников: социальных сетей, мессенджеров и т.д.

Методологическую основу разработанных и программно реализованных алгоритмов составляет метод нормализации слов [2-4]. Алгоритмическую и программную структуру автоматизированной системы можно представить в виде следующих этапов:

1. Токенизация – это самый первый шаг при обработке текста. Заключается в разбиении (разделении) длинных строк текста на слова (от пробела до пробела).
2. Приведение всех символов к нижнему регистру.
3. Замена букв на аналогично выглядящие латинские («п» – «n», «и» – «u»).
4. Удаление внутри каждого слова всех символов, не являющихся буквами русского и английского алфавитов.
5. Стемминг – процесс нахождения основы слова для заданного исходного слова (на основе реализации алгоритма Стемминга Портера).
6. Сравнение полученной “нормальной” формы слова с базой нецензурных слов и его замена при совпадении.

Токенизация, или, другими словами, сегментация текста, производится во время получения текста пользователя. Программа считывает текст построчно и разбивает каждую строку на отдельные слова. За счет встроенных в стандартные библиотеки C# функций алгоритмом токенизации является создание подстроки от начала слова до ближайшего вхождения пробела. Данный алгоритм несет в себе цель рассмотрения каждого слова по отдельности. Далее используется копия вектора слов, полученных на данном шаге.

На втором этапе программа заменяет все символы, являющиеся заглавными буквами русского и английского алфавитов, на строчные по кодировке ASCII. Это необходимо для упрощения реализации последующих этапов программы и сокращения времени ее работы. По-другому – это оптимизация процесса, которая позволяет АС сократить алфавит символов, используемых в программе.

Третий этап ориентирован на борьбу с наиболее популярными способами «маскировки» нецензурной лексики – такими, как: смешение кириллицы и транслита, замена букв на аналогично выглядящие латинские («п» — «n», «и» — «u»), замена согласных на парные созвучные им (например: «х» на «к»).

Следующий немаловажный этап программной обработки текста – удаление внутри слова символов, не являющихся буквами русского и английского алфавитов. Помимо “маскировки” символов, реализуемой на 3 этапе, в интернете появилась тенденция ”ложных” символов, появление которых не учитывается. АС определяет значение символа, и, если он не является буквой, то АС удаляет его, то есть символы – такие, как знаки препинания, подчеркивания, дефисы, косые черты и т.д.

Пятый этап программной обработки текста является основной функцией нормализации слов. Стемминг – это нахождение основы слова (стеммы). Термин стемминг образован от слова «stem» – ствол, стебель, основа. Алгоритм не использует баз основ слов, а лишь, применяя последовательно ряд правил, отсекает окончания и

суффиксы, основываясь на особенностях языка, в связи с чем работает быстро, но не всегда безошибочно. Для реализации алгоритма Стемминга Портера [5] была взята базовая библиотека C++ “regex” (библиотека регулярных выражений) [6].

Завершающий этап программы – цензурирование. После получения “нормальной” формы слова остается понять, принадлежит ли оно к нецензурной лексике. Для этого автором был создан текстовый файл-тезаурус, содержащий большое количество матерных и оскорбляющих слов, прошедших алгоритм Стемминга Портера [7]. Все исходные данные после обработки сверяются с массивом “плохих” слов. А также происходит особая проверка слова на случайное срабатывание цензурирования. Слово сверяется со списком исключений, которые не являются плохими словами, но содержат в своей основе матерное слово, например, глагол “оскорблять”. При совпадении исходное слово заменяется на нейтральный набор символов: “[censored]”.

Разработанная АС может найти практическое применение в СМИ и различных интернет-ресурсах. Программа также призвана способствовать уменьшению «загрязнения» русского языка и снижению использования нецензурной лексики в речи молодежи.

Первоначально предполагается использование данного продукта в личных целях. Люди, получающие данный продукт, смогут корректировать свои тексты. Пользователь будет взаимодействовать с интерфейсом, который предоставляет связь между АС и человеком. После усовершенствования продукта планируется интеграция с частными компаниями, которые будут использовать АС в индивидуальных целях. Более масштабным планом является сотрудничество с более крупными компаниями и проектами – такими, как ВКонтакте, Telegram и др.

Данный проект – полностью бесплатный, однако при добавлении мобильного приложения в Google play придется заплатить за аккаунт разработчика 25 долларов. В таком случае приложение будет платным, цена которого будет около 100 рублей, тем самым оно окупится спустя 20 покупок.

Список литературы:

1. Федеральный закон от 30.12.2020 № 530-ФЗ "О внесении изменений в Федеральный закон "Об информации, информационных технологиях и о защите информации". Режим доступа: http://publication.pravo.gov.ru/Document/View/0001202012300062 (дата обращения 15.06.2022 г.)
2. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, 2003. URL: https://cache-mskmar02.cdn.yandex.net/download.yandex.ru/company/iseg-las-vegas.pdf
3. Jurafsky D., Martin J. H. Speech and Language Processing. New Jersey: Prentice Hall, 2008. 1024 p.
4. Большакова Е.И., Воронцов К.В. и др. Автоматическая обработка текстов на естественном языке и анализ данных: Учеб. пособие. М.: Изд-во НИУ ВШЭ, 2017. 269 с.
5. Russian stemming algorithm. Режим доступа: http://snowball.tartarus.org/algorithms/russian/stemmer.html (дата обращения 15.06.2022 г.)
6. Regular expressions library. Режим доступа: https://en.cppreference.com/w/cpp/regex (дата обращения 15.06.2022 г.).
7. Список нецензурной лексики. Режим доступа: https://yadi.sk/d/McrzPrrcj3hd7 (дата обращения 15.06.2022 г.)