УДК 004.738.1-048.445

ГЕНДЕРНАЯ КЛАССИФИКАЦИЯ ПОЛЬЗОВАТЕЛЕЙ REDDIT

А.А. Усатов
Студент ФПМИ, БГУ,
исследователь BelPrime

Д.М. Прокурат
Магистр математики и
информационных технологий

ООО «БелПрайм», г. Минск, Республика Беларусь
Белорусский государственный университет, г. Минск, Республика Беларусь
E-mail: ausatov@icloud.com

А.А. Усатов
Студент факультета прикладной математики и информатики Белорусского государственного университета. Работает в BelPrime в должности исследователя.

Д.М. Прокурат
Магистр математики и информационных технологий.

Аннотация. В данной работе рассмотрена проблема гендерной классификации пользователей социально-новостного сайта Reddit. Рассмотрены следующие подходы для решения задачи: алгоритм Ахо-Корасика, метод опорных векторов (SVM), предложен алгоритм для поиска имени в начале строки.

Ключевые слова: классификация, имена, SVM, метод опорных векторов, n-grams, reddit, алгоритм Ахо-Корасика.

Введение. Бизнес и авторы социальных исследований регулярно сталкиваются с необходимостью определения гендерной принадлежности своей аудитории либо аудитории рекламных компаний. В случае с большим большим объёмом аудитории процесс нуждается в автоматизации с помощью модели, способной обрабатывать данные в реальном времени. Рассмотрим один из самых сложных случаев определения пола автора сообщения: пользователи социально-новостного сайта Reddit.

Причины, по которым мы считаем Reddit наиболее сложным случаем:

1. Пользователи, как правило, не используют свои настоящие фото, в основном их аватары – это картинки (причём, предлагаемые Reddit’ом по умолчанию).
2. Reddit не просит пользователей указывать их настоящие имена, ввиду чего ник пользователя зачастую является не имеющим к имени пользователя словом или же набором букв и символов.
3. Не представляется возможным получить список подписок пользователя, используя открытые исходные данные.
4. Нет дополнительных данных: email, ссылки на другие социальные сети и т.д.

Сбор и подготовка данных. Используя открытые источники, предоставляемые Reddit, удалось получить для каждого пользователя: сообщения, комментарии, аватар и ник. Аватар, как выяснилось в процессе, обычно является стандартной картинкой, предлагаемая Reddit'ом. Исходя из имеющихся данных мы первоначально сформировали несколько стратегий для

определения пола автора: с использованием сообщений, ника и их комбинации. В процессе подготовки данных выяснилось, что профили многих пользователей не поддаются разметке, так как нет никаких признаков, указывающих на пол пользователя. Пример такого профиля изображён на рисунке 1. Поэтому далее в работе будут рассмотрены только те профили, пол обладателя которых представляется возможным определить вручную.

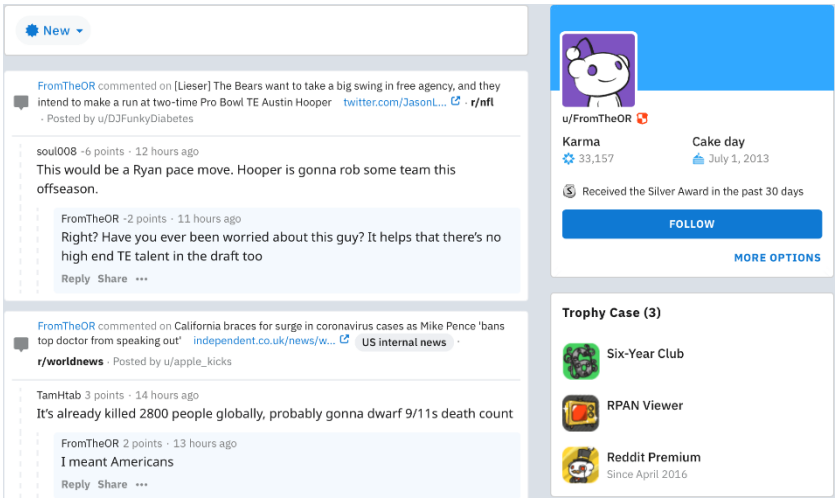

Рисунок *1*. – Пример профиля, не поддающегося классификации

Обзор работ. Рассмотрим уже описанные в научных работах подходы к определению пола пользователей социальных сетей. В работе [1] отмечается, что мужчины и женщины стоят свою речь примерно одинаково, из-за чего не представляется возможным использовать текст публикаций или комментариев для решения поставленной задачи. Кроме того, если предположить, что у женщин в сообщениях чаще будут встречаться некоторые определённые слова – например, слово "маникюр" – и взять это слово как указывающее на то, что автор сообщения – женщина, то, если мы захотим посмотреть гендерное распределение среди пользователей, которые упомянули это слово, мы получим заведомо недостоверный результат. Поэтому в статье будут рассмотрены подходы без анализа текста постов и комментариев. В работе [2] было показано, как можно использовать имя как признак для классификации пользователей Twitter. В работе [3] предложен вариант разбиения строки на два имени, используемый для китайских имён. К сожалению, для Reddit этот подход не подходит.

Рассматриваемые варианты:

1. Вариант 1. Найти все вхождения имён в ник пользователя.
2. Вариант 2. Найти самое длинное имя в начале строки.
3. Вариант 3. SVM по n-граммам.

Описание идеи. Вариант 1. Предположим, что пользователи часто используют свои имена в никах. Соберём большой (около 100к) набор имён имён. Воспользуемся алгоритмом Ахо-Корасика для поиска всех поиска всех этих имён в никах, поступающих на вход классификатора.

Описание реализации. Вариант 1. Воспользуемся библиотекой pyahocorasick. Построение бора происходит один раз при запуске приложения и достаточно быстро – а именно, за линейное от суммарной длины строк – чтобы не беспокоиться об этом. Время построения автомата так же линейное. Время нахождения всех имён в строке, поступающей на вход – линейное от длины этой строки, так что алгоритм позволяет обрабатывать данные онлайн без распараллеливания.

Результаты. Обнаружилось, что в именах находятся подстроки, которые соответствуют другим именам. Причём, часто эти имена используются для противоположных полов Установить какую-либо взаимосвязь между найденными именами и полом пользователя не удалось.

Ниже приведены имена, которые нашлись в нике 'dreammanalishi'.

Таблица 1. – Найденные в 'dreammanalishi' имена

Найденное имя	**Доля мужчин**	**Доля женщин**
dream	0.72	0.28
amman	0.5	0.5
manal	0.0	1.0
manali	0.0	1.0
anali	0.0	1.0
analis	0.0	1.0

Чтобы этого избежать, было предложено применить различные эвристики:

– рассмотреть только непересекающиеся множества подстрок
– брать подстроку наибольшей длины
– брать первую попавшуюся строку

Вышеуказанные эвристики не позволили достичь приемлемой точности классификации.

Описание идеи и реализации. Вариант 2. Изложенная ниже идея появилась после неудовлетворительных результатов алгоритма Ахо-Корасика. Будем рассматривать строку без последнего символа до тех пор, пока и если не найдём подстроку в наборе имён. Например, для строки ‘alexanderusatov’ рассмотрим поочерёдно ‘alexanderusatov’, ‘alexanderusato’, ‘alexanderusat’ и т.д. Будем продолжать убирать последний символ до того момента, пока ‘alexander’ не найдётся в именах. Определим гендер по этому имени.

Результаты. Вышеописанный подход позволил получить довольной высокую точность (Accuracy около 0.8) в случаях, когда ник пользователя начинается с его реального имени. В противном же случае алгоритм не подходит для классификации.

Описание идеи и реализации. Вариант 3. Попробуем использовать SVM. Для этого воспользуемся:

1.Униграммами

Рисунок *2.* – Униграммы

Данные для обучения - набор имён, размером около 100 000. В этом случае Accuracy была 0.69 среди пользователей, которых удалось разметить вручную.

2. Биграммами

Рисунок *3*. – Биграммы

Данные для обучения - набор имён, размером около 100 000. В этом случае Accuracy была 0.78 среди пользователей, которых удалось разметить вручную.

3.Триграммами

Рисунок *4*. – Триграммы

В этом случае Recall для мужчин составил 1. Однако при этом абсолютно все пользователи были классифицированы как мужчина, и Accuracy была на уровне константного классификатора, который относит всех пользователей к мужчинам, что не несёт никакой пользы.

Вывод. Поскольку среди пользователей Reddit большинство не указывает никакой информации, которая позволяла бы определить пол, была рассмотрена только та часть пользователей, которая указала своё имя или нечто похожее на него. На этом множестве пользователей удалось получить Accuracy = 0.78.

Список литературы

[1]. Плюснина, А.В. Характеристики мужской и женской письменной речи в гендерном сознании коммуникантов / А.В. Плюснина // Ярославский педагогический вестник – 2012 – No 1 – Том I (Гуманитарные науки) – С.184-188.

[2]. Wendy Liu and Derek Ruthsю What’s in a Name? Using First Names as Features for Gender Inference in Twitter. Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium - P. 10-16.

[3]. Hua Zhao, Fairouz Kamareddine Recursion identify algorithm for Gender Prediction with Chinese names / А.В. Плюснина // Int'l Conf. Data Science | ICDATA'18 | P. 137-142.

GENDER CLASSIFICATION OF REDDIT USERS

A.A. Usatov
Student of FAMCS, BSU,
researcher at BelPrime

D.M. Prokurat
Master of Mathematics and
Information Technology

«BelPrime» Ltd., Republic of Belarus
Faculty of Applied Mathematics and Computer Science of BSU
Mechanics and Mathematics Faculty of BSU
Belarusian State University, Republic of Belarus
E-mail: ausatov@icloud.com

Abstract. This paper considers the problem of gender classification of users of the social news site Reddit. The following approaches to solving problems are considered: the Aho-Korasik algorithm, the support vector method (SVM), and the proposed algorithm for finding a name at the beginning of a line.

Keywords: classification, names, SVM, support-vector machines, n-grams, reddit, Aho-Corasick algorithm.