ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

УДК 004.852

ПРОФИЛИРОВАНИЕ АВТОРА ЭЛЕКТРОННОГО СООБЩЕНИЯ

В. В. Бондаренко, Д. А. Кривов

В данной работе авторами рассматривается проблема определения пола автора электронного сообщения. Для ее решения разработана компьютерная программа, которая на основе статистических методов способна определять гендерную принадлежность автора текста. Для формирования обучающей выборки корпуса текста на русском языке предложен метод сбора и анализа большого количества электронных сообщений. На его основе проведен ряд тестов с целью нахождения наиболее результативного подхода к обучению программ по определению требуемых сведений об авторе сообщения. Помимо этого, в работе представлены и разобраны результаты проведенных экспериментов, а также проведен их анализ для дальнейшего повышения точности работы подобных программ.

Ключевые слова: текст; пол; гендер; метод опорных векторов; научный текст; сообщение; моделирование личности по тексту; математические методы в лингвистике.

В современном мире одной из наиболее серьезных проблем является мошенничество в сети интернет [1]. Благодаря особенностям данной сети можно без особых трудностей скрыть всю информацию о своей личности. Именно этим пользуются злоумышленники и часто выдают себя за совершенно другого человека. Так, например, для получения доверия со стороны потенциальной жертвы преступник может притвориться красивой девушкой с целью дальнейшего получения денежных средств.

Однако у каждой категории людей существуют свои речевые особенности [2]. Например, женщины употребляют для общения некоторые слова намного чаще мужчин. Поэтому, если проанализировать большое количество сообщений от представителей разных полов, можно с достаточно высокой вероятностью определить гендерную принадлежность автора электронного сообщения. Это может быть полезно для определения потенциального преступника. Именно поэтому, целями данной работы является:

- **разработка** компьютерной программы для определения пола автора текста;
- **обучение** раннее созданной утилиты;
- **разработка** метода для ускорения обучения приложения;
- **сравнение** и **анализ** полученных результатов.

Разработка программы для определения пола автора электронного сообщения

Необходимо отметить, что подобные разработки проводились для работы с иностранными языками. Так в работе [3] был представлен инструмент для турецкого языка, который позволял определить пол автора текста по вводу короткого электронного сообщения. Количество правильных ответов составляло 90 процентов от общего числа ответов, что на сегодняшний день является одним из наилучших результатов. Также в работе [4], основанной на корпусе англоязычных электронных писем, с помощью метода

Бондаренко Владимир Владимирович (bondarenko.vv@ssau.ru),
доцент кафедры безопасности информационных систем;
Кривов Даниил Андреевич (krylov_danechka@list.ru),
студент IV курса механико-математического факультета Самарского университета,
443086, Россия, г. Самара, Московское шоссе, 34.

опорных векторов была достигнута точность идентификации пола в 82 процента по функциональным словам и характеристикам уровня символов.

Однако для русского языка данная проблема до сих пор остается крайне актуальной, ибо подобные исследования ранее не производились. В работе [5] был продемонстрирован подход для создания потенциально мощного инструмента для определения данных автора текста, однако данное приложение так и не было разработано.

Именно поэтому было решено создать подобную программу. Идея работы приложения следующая: на вход подается какое-либо сообщение, далее идет обработка исходного текста сообщения и уже после этого используется расстановка весов для каждого слова при помощи метода опорных векторов (support vector machine). Схема предварительного преобразования сообщения представлена на рисунке 1.

Как становится видно из рисунка 1, изначально производится токенизация текста по словам, а именно для дальнейшей корректной работы необходимо создать массив, элементами которого будут являться слова из поданного на вход электронного сообщения.

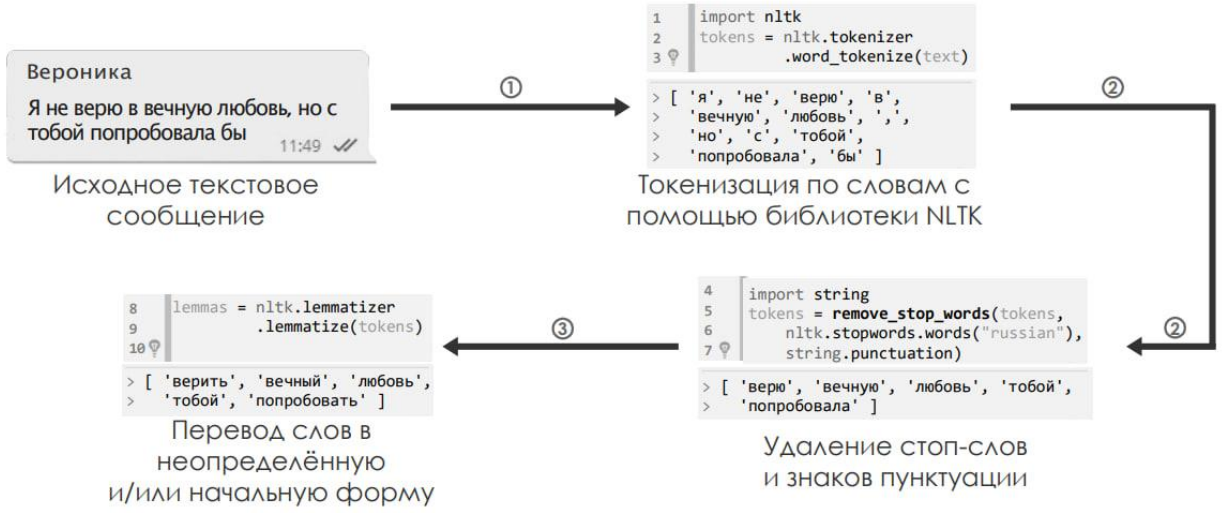

Рис. 1. Схема преобразования текста электронного сообщения

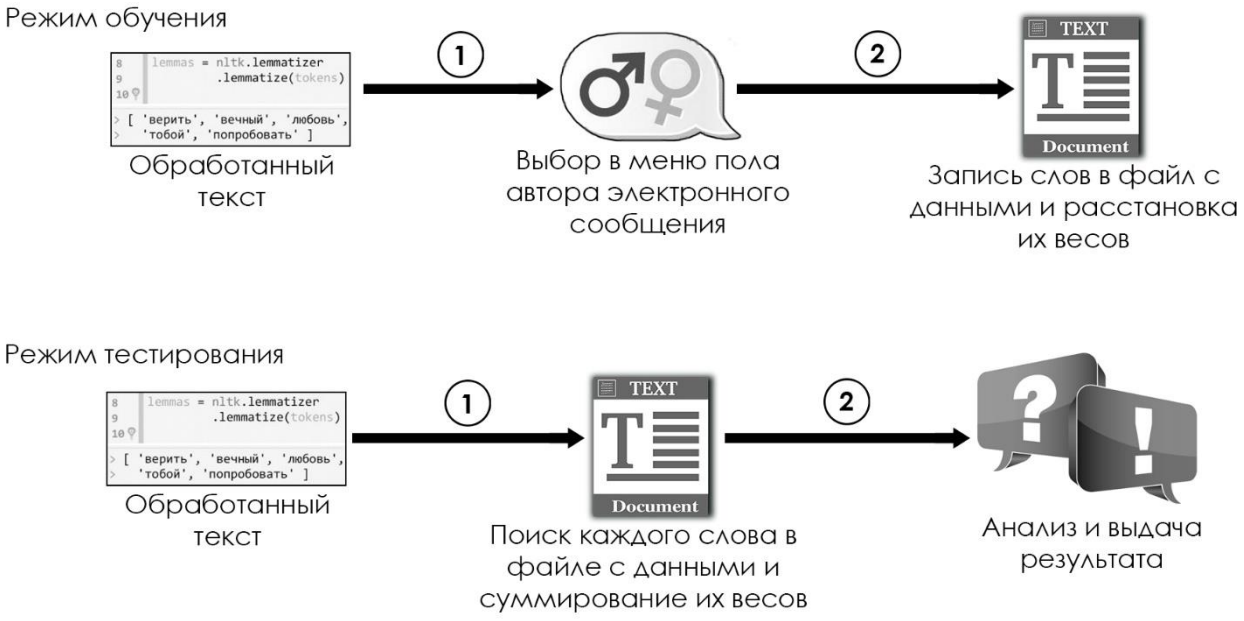

Рис. 2. Режимы работы приложения профилирования автора электронного сообщения

После этого происходит удаление всех стоп-слов из заготовленного заранее списка, а также знаков препинания. Если пропустить данный шаг, то точность определения пола автора сообщения алгоритмом становится в разы ниже. После проведения данных манипуляций с текстом остается наиболее важная часть преобразования – лемматизация. На данном этапе производится перевод всех глаголов в неопределенную форму, а также удаление всех окончаний у существительных. Все вышеперечисленные действия реализуются при помощи библиотеки NLTK для алгоритмического языка Python.

Дальнейшее функционирование программы зависит от выбранного в начале режима работы приложения. Режимы работы программы представлены на рисунке 2. Из рисунка 2 становится понятно следующее.

1. Если была поставлена задача обучения, то все представленные слова записываются в текстовый документ, после чего им расставляются веса в зависимости от выбранного пункта меню (мужчина/женщина). Также стоит упомянуть, что в случае, если данное слово уже присутствовало в файле с данными, то оно не будет снова записано, а лишь увеличится его вес.

2. Если же выбран режим тестирования программы, то в данной ситуации приложение начинает искать каждое из слов в файле с данными и суммировать все значения их весов. Затем полученные результаты сравниваются, и программа выдает заключение о том, кто написал поданное на вход сообщение, также при этом высчитывается вероятность правильного ответа. Вероятность высчитывается путем деления значения весов для одного из гендеров на сумму весов обоих гендеров и умножением на 100%. Она может принимать значения в промежутке от 50 до 100 процентов.

Основной проблемой для обучения программы становится сбор и анализ большого количества данных. Именно поэтому был разработан следующий метод.

Реализация метода быстрого обучения программы

За основу создания метода для эффективного обучения программы берется информация о том, что особенности употребления различных слов мужчинами и женщинами отображаются не только при отправке сообщений, но и при устном общении [6], то есть, женщины более склонны употреблять при разговоре определенные слова, нежели мужчины. Основываясь на этом, возникла идея обучения программы на основе аудиосообщений. Для возможности использования аудиосообщений необходимо было реализовать функцию преобразования речи в текст. При анализе данного вопроса было обнаружено, что в настоящее время в языке Python наиболее широкое распространение получили 3 библиотеки с подобным функционалом: speech_recognition, pocketsphinx, а также Vosk.

Для реализации функции преобразования была выбрана последняя из указанных библиотек, Vosk. Первая, а именно speech_recognition, не подошла из-за постоянной необходимости доступа к сети интернет. Вторая библиотека также не подошла, ибо качество распознавания текста было на среднем уровне и часто отображались некорректные слова, что в итоге могло привести к уменьшению точности программы по определению пола автора электронного сообщения. Совершенно иначе обстояли дела с пакетом Vosk, он позволяет работать в режиме оффлайн, качество распознавание достаточное для решения поставленной задачи.

Однако у библиотеки Vosk, есть свои особенности. Для корректного отображения всех слов необходимо выбрать и загрузить на свое устройство языковую модель. На момент написания статьи существовало 2 основных языковых модели для русского языка: «большая» и «маленькая». Преимуществом «большой» модели является точность перевода устной речи в текст, но при этом ее вес составляет почти 2 гигабайта. В то же время, «маленькая» модель занимает всего 45 мегабайт памяти на диске, однако и точность распознавания у нее в разы ниже. Для получения более точных результатов была выбрана «большая» языковая модель. После определения необходимого инструменты для работы с речью, данный пакет был встроен в программу профилирования автора электронного сообщения.

На данном этапе разработки было выявлено, что сбор обучающих данных стал в разы быстрее и теперь было необходимо определиться с тем, откуда брать необходимый датасет.

Результаты обучения программы

Основной проблемой обучения программы стало отсутствие большого количества аудиосообщений в открытом доступе. Учитывая данные обстоятельства, было принято решение попробовать обучить приложение при помощи видеороликов двух типов: новостных и в формате интервью. Перед тем, как «отдавать» программе выбранный материал, было необходимо его преобразовать. Схема преобразования видеороликов представлена на рисунке 3.

Как становится видно из рисунка 3, изначально все видеоролики из формата .mp4 переводились в аудиоформат .wav. Делалось это в ручном режиме, ибо также был необходим ряд других правок, которые в то же время производились в программе для обработки видео VegasPro. Приложению нельзя было подавать на вход необработанную аудиодорожку, ибо в подобных видеороликах, как правило, присутствуют разговоры как и мужчин, так и женщин. При обучении приложения можно выбирать только один из вариантов пола и даже если в видео был хотя бы небольшой фрагмент разговора представителя другого пола, то это могло негативно сказаться на полученных результатах. Именно поэтому все видеоролики были обработаны таким образом, чтобы в них присутствовала речь либо мужчины, либо женщины, но никак не представителей обоих полов сразу.

Суммарно было обработано 60 роликов, из них 30 были новостными, а 30 в формате интервью. Отличия двух категорий в том, что, как правило, в новостных передачах ведущий рассказывает о новостях по заготовленному заранее сценарию, в то время как в интервью чаще оба собеседника беседуют в свободной форме. Проблема сценариев в том, что точно неизвестно, кто именно их писал. Ведущим может выступать женщина, а вот автором сценария может быть мужчина, что в итоге негативно сказывается на точности работы программы профилирования автора электронного сообщения. В результате, после обучения программы двумя разными способами в каждом из случаев в памяти приложения осталось примерно 25 тысяч слов. Стоит отметить, что такое количество получилось без учета местоимений, частиц, предлогов и союзов, так как они убирались на этапе удаления стоп-слов. При этом же, в первом случае получилось 5982 уникальных слов, а во втором 5602, что наглядно показывает различие между новостными и разговорными видеороликами.

Далее была разработана специальная программа, которая позволяет проводить тестирования основного приложения. Алгоритм работы следующий:

1) на вход программы по профилированию автора электронного сообщения подается заготовленное сообщение из документа;

2) фиксируется полученный от программы ответ (мужчина/женщина);

Рис. 3. Схема преобразования исходного видеоролика

3) сверяется с правильным ответом о поле автора сообщения, который также хранится в текстовом файле из пункта 1;

4) случае верного определения, значение счетчика увеличивается на 1, в ином случае значение счетчика не изменяется;

5) значение счётчика делится на общее количество сообщений в файле и умножается на 100 %.

Также стоит отметить, что количество сообщений для тестирования в документе равняется 500 штук. Данные сообщения взяты из различных источников, в том числе из личных переписок авторов работы. Темы сообщений тоже совершенно разнообразные. Так, например, были сообщения на нейтральные темы вроде погоды, но встречались и сообщения, в которых разговор шел об обстановке в мире. После проделывания данных манипуляций для каждого из двух случаев, получились результаты, представленные в таблице 1.

Как становится видно из таблицы 1, обучать подобного рода программы намного эффективнее при помощи записей «живого» разговора людей. Новостные передачи хоть и позволяют создать инструмент для профилирования личности автора сообщения, однако его точность будет существенно ниже, чем в случае с реальным разговором между людьми. Как было ранее сказано в данной работе, связано это с тем, что у каждой новостной передачи имеется четко прописанный сценарий и далеко не всегда получается так, что пол автора сценария совпадает с полом ведущего, отчего не всегда получается корректно обучить программу. Именно это и влияет на итоговые результаты.

Эксперименты по определению оптимального объема корпуса русского языка для определения пола автора сообщения с целью повышения точности работы алгоритма показали, что после 25000 тысяч слов точность работы программы по определению пола автора сообщения расти переставала в обоих случаях, и при обучении на новостных видеороликах, и на видео в формате интервью. Зависимость количества слов в памяти программы от точности работы приложения представлена на рисунке 4.

Разработка бота в мессенджере Telegram

После завершения обучения программы возник вопрос о том, как предоставить доступ всем желающим опробовать приложение по определению пола автора сообщения. Передавать программу при помощи физического носителя не является оптимальным методом, ибо в таком случае количество потенциальных пользователей является крайне малым. Именно поэтому было решено распространить приложение при помощи сети интернет. Для того, чтобы облегчить процесс установки, был реализован бот в мессенджере Telegram.

Для того, чтобы воспользоваться ботом необходимо в поисковой строке ввести «kda1_bot», после чего перейти в чат и программа начнет свою работу. Если отправить боту любое сообщение, то он в автоматическом режиме определит пол его автора. Стоит отметить, что точность определения на данный момент составляет 81%, так как данная версия была обучена на видеороликах в формате интервью. Интерфейс взаимодействия с чат-ботом представлен на рисунке 5.

Таблица 1

Результаты тестирования программы по определению пола автора сообщения

Способ обучения	Количество уникальных слов в памяти программы после обработки	Точность (в процентах)
Обучено на основе новостных передач	5982	64
Обучено на основе видео с интервью	5602	81

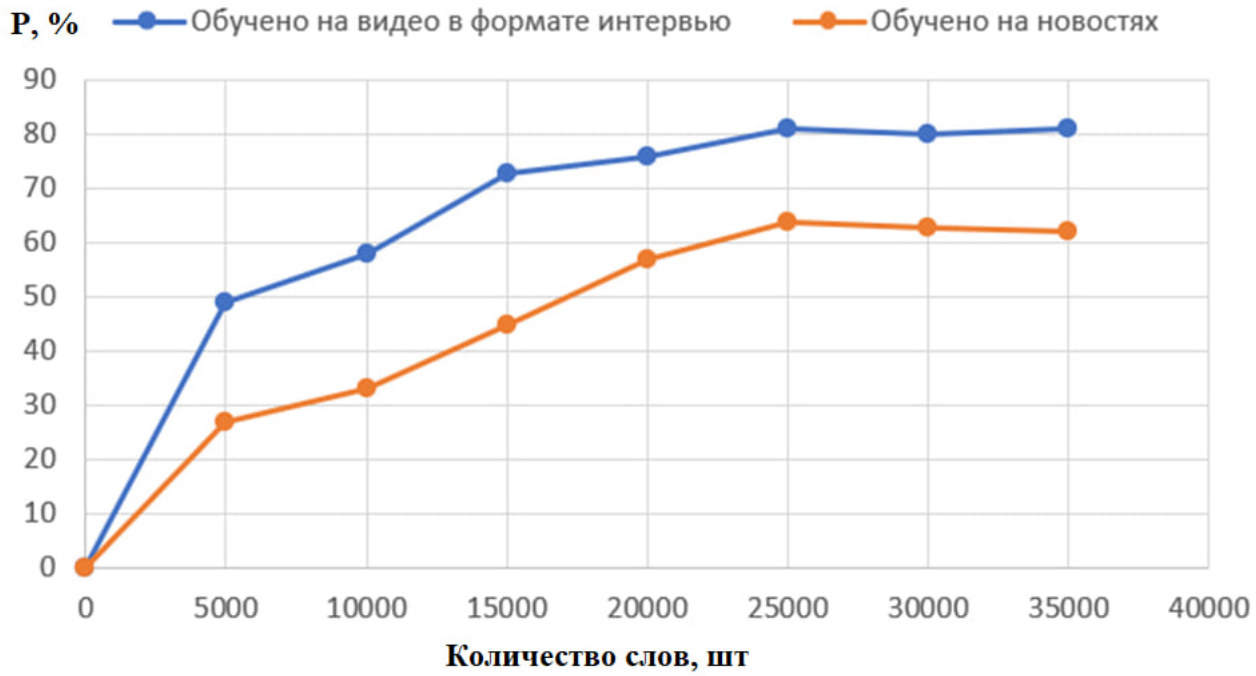

Рис. 4. График зависимости количества слов в памяти программы от точности её работы

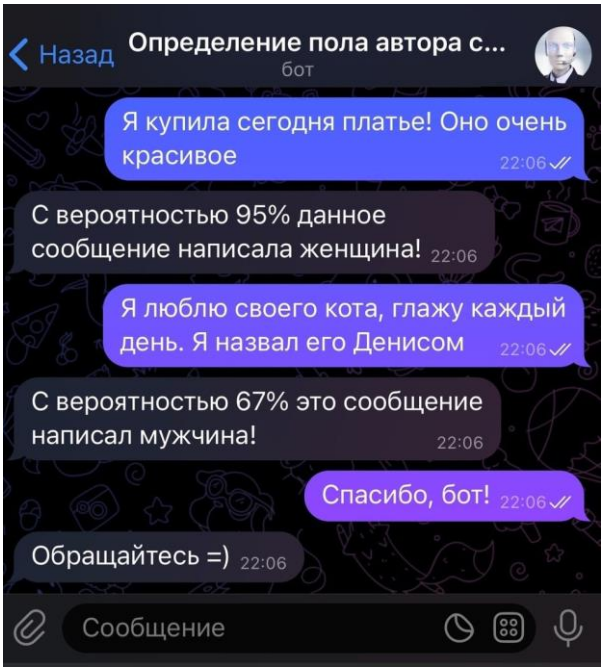

Рис. 5. Чат-бот Telegram по определению пола автора текста

Также важно отметить, что для реализации бота была использована библиотека Telebot, которая позволяет не только использовать Telegram API, но и обеспечить стабильную работу бота и быстрый отклик после ввода сообщения.

Заключение

Подводя итоги, можно сказать, что вопрос о профилировании автора сообщения до сих пор остается очень актуальным. Как показали результаты экспериментов, проведенных в рамках настоящей работы, определить пол автора сообщения на русском языке становится возможным, причем с довольно высокой точностью. Метод обучения подобного рода программ при помощи видеороликов с «живым» общением позволяет не только максимально быстро обучить приложение, но и

повысить точность результатов тестирования. Точность верных ответов в 81% может быть улучшена, если использовать предварительную фильтрацию текста не только удаляя стоп-слова, но и используя другие методы, а также изменяя значение весов наиболее часто употребляемых слов и у мужчин, и у женщин.

Необходимо отметить, что созданный механизм обучения и распознавания является базовой платформой для определения психофизических характеристик автора сообщения. Также в дальнейшем после соответствующей настройки и доработки предполагается применение вышеописанного метода для определения возрастной категории, психоэмоциональных характеристик, уровня образования и других параметров.

Литература

1. Белоножко Е. С., Чеджемов Г. А. Мошенничество в сети Интернет // Наука XXI века: актуальные направления развития. 2017. №. 1-1. С. 85-88.

2. Гендер и язык: есть ли разница между мужской и женской речью? // Theory&Practice. URL: https://theoryandpractice.ru/posts/9451-gender-language (дата обращения: 07.02.2023).

3. Doyle J., Keselj V. Automatic categorization of author gender via n-gram analysis // The 6th Symposium on Natural Language Processing, SNLP. 2005.

4. Köse C., Özyurt Ö., Amanmyradov G. Mining Chat Conversations for Sex Identification. Emerging Technologies in Knowledge Discovery and Data Mining Lecture Notes in Computer Science. 2007. Vol. 4819. P. 45–55.

5. Проблема диагностирования пола автора письменного текста: фактор жанра / Т. А. Литвинова [и др.] // Russian Journal of Education and Psychology. 2014. №. 1 (33). С. 4.

6. Сравнительный анализ гендерных моделей речевой деятельности в английском, немецком и русском языках / Е. Ю. Ковалева [и др.] // Балтийский гуманитарный журнал. 2017. Т. 6. №. 4 (21). С. 105-109.

PROFILING OF THE AUTHOR OF ELECTRONIC COMMUNICATION

V. V. Bondarenko, D. A. Krivov

In this paper, the authors consider the problem of determining the gender of the author of an electronic communication. In order to solve the problem, a computer programme had been developed that was able to determine the gender of the author of the text on the basis of a set of methods. A method for collecting and analyzing a large number of electronic messages has been proposed for the development of a learning body of text in Russian. Based on this, a number of tests have been conducted to find the most effective approach to training programs to determine the required information about the communicant. In addition, the paper presented and reviewed the results of the experiments carried out and analysed them in order to further improve the accuracy of the work of such programmes.

Key words: text; sex; gender; support vector method; scientific text; message; personality modeling by text; mathematical methods in linguistics.

Статья поступила в редакцию 27.05.2023 г.

Bondarenko Vladimir Vladimirovich (bondarenko.vv@ssau.ru),
associate professor of Information Systems Security department;
Krivov Daniil Andreyevich (krylov_danechka@list.ru),
IVth year student of the Faculty of Mechanics and Mathematics of Samara University,
443086, Russia, Samara, Moskovskoye shosse, 34.