Компьютерные инструменты в образовании, 2018
№ 3: 49–64
УДК: 159.9.07:316.454.52 + 004.8
http://ipo.spb.ru/journal
doi:10.32603/2071-2340-3-49-64

компьютер в учебном процессе

ИССЛЕДОВАТЕЛЬСКИЙ ПРОЕКТ КАК ИНСТРУМЕНТ ОБУЧЕНИЯ МЕТОДАМ АНАЛИЗА ТЕКСТА: ПРЕДСКАЗАНИЕ КЛАССА ПОСТА В СОЦИАЛЬНОЙ СЕТИ*

Суворова А. В.[1,2], Смирнова К. Р.[3], Будин Е. А.[3], Тулупьева Т. В.[1,3,4], Тулупьев А. Л.[1,4], Абрамов М. В.[1,4]

[1]Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, Россия
[2]Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург, Россия
[3]Северо-Западный институт управления — филиал РАНХиГС, Санкт-Петербург, Россия
[4]Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Аннотация

В статье описывается студенческий исследовательский проект по предсказанию класса поста в социальной сети на основе его текстового содержания. Обсуждаются особенности проекта как составной части траектории обучения методам анализа данных, в том числе методам и инструментам анализа текста, часто не включаемым в курсы по машинному обучению. Описана постановка задачи, этапы ее решения, последовательность рассмотрения новых методов как способов решения возникающих у студентов проблем, а также используемый инструментарий среды R. Приведены возможности расширения задачи и ее модификации в зависимости от уровня подготовки студентов.

Ключевые слова: *проблемно-ориентированное обучение, социальные сети, машинное обучение, анализ текста, классификация, автоматизация исследований, язык R.*

Цитирование: Суворова А. В., Смирнова К. Р., Будин Е. А., Тулупьева Т. В., Тулупьев А. Л., Абрамов М. В. Исследовательский проект как инструмент обучения методам анализа текста: предсказание класса поста в социальной сети // Компьютерные инструменты в образовании. 2018. № 3. С. 49–64.

1. ВВЕДЕНИЕ

Широкое развитие науки и техники, высокодинамичный современный мир, изменения характера выполнения многих профессиональных функций обусловливают необходимость модернизации учебного процесса, применяемых методов обучения. На первый план выходит задача формирования личности, способной творчески мыслить, принимать решения, свободно ориентироваться в потоке быстро меняющейся информации. Эти вызовы современности диктуют необходимость более широкого применения не репродуктивных, а проблемно-ориентированных, проблемно-поисковых

*Работа выполнена в рамках проекта по государственному заданию СПИИРАН № 0073-2018-0001, при частичной финансовой поддержке гранта РФФИ, проект № 16-31-60063-мол_а_дк, № 18-01-00626, № 18-37-00323.

методов обучения, которые, при опоре на интересы студента, обеспечивают глубокое понимание изучаемого материала, развивают аналитическое и креативное мышление. Многие исследования показывают, что проблемно-ориентированная и проектная деятельность стимулирует изучение новых методов [24, 35] и даже повышает удовлетворенность студентов процессом обучения [22]. Вовлеченность в решение интересной практической задачи позволяет сохранять интерес, даже если ответ не находится с первой попытки [37], а возникающие в процессе поиска решения проблемы позволяют плавно и логично вводить в обучение новые методы [28].

Проектный и проблемно-ориентированный подход широко применяется при обучении в различных сферах: в программировании [5], инженерном образовании [34], дискретной математике [14], статистике [18], медицине [26] и многих других [24].

В данной статье рассматривается пример применения указанного подхода для изучения методов анализа текста и машинного обучения в контексте следующей междисциплинарной задачи: необходимо предсказать класс (один из заранее определенных) поста в социальной сети на основании характеристик этого поста, включая текст.

Таким образом, выделяются три цели статьи: методическая — показать, как проблемы, возникающие у студентов в процессе работы над проектом, позволяют преподавателю вводить новые методы анализа (включая обработку текста) и акцентировать внимание на важности тех или иных показателей; дидактическая — продемонстрировать работу набора инструментов среды R для анализа текста и построение моделей классификации; и научная — построить модели для классификации постов в социальной сети по их текстам, что применяется, в том числе, для анализа психологических особенностей пользователей.

2. ОБОСНОВАНИЕ ТЕМАТИКИ ПРОЕКТА

Первое, что выделяет поставленную в проекте задачу — актуальность и ее привлекательность для студентов. На сегодняшний день социальные сети стали для многих неотъемлемой частью повседневной жизни, что особенно заметно среди молодежи [17]. Являясь источниками информации как для пользователей, так и о пользователях, они дают широкие возможности для исследований различных аспектов деятельности: этнических особенностей [16], миграционных потоков [39], принципов формирования сообществ [19], протестной деятельности [36], отражения личностных особенностей пользователей [31] и многих других. Знакомство студентов с примерами подобных исследований позволяет им по-новому взглянуть, в том числе, на их поведение в социальных сетях и на то, о чем может говорить та или иная информация, публикуемая на страницах.

Вторая особенность постановки подобной задачи — включенность ее в более общий проект, что позволяет ответить на вопросы, зачем нужно решение этой задачи и где оно потом будет применяться. Задача является необходимой частью реального исследования, а не формулируется только для изучения конкретного метода анализа. В данном случае контекстом задачи является проводимое исследование, направленное на изучение отражения психологических особенностей пользователей в контенте, публикуемом ими в их аккаунтах в социальной сети [9]. В проведенных ранее этапах исследования была рассмотрена взаимосвязь между постами, публикуемыми пользователями на своих страницах в социальной сети «ВКонтакте», и психологическими характеристиками этих пользователей (например личностными особенностями согласно опроснику Р. Б. Кеттелла, такими как мечтательность, подозрительность, эмоциональная стабильность, дипломатичность и т. д. — подробнее можно узнать в [9]). Для выявления характеристик

была разработана система классификации размещаемых постов [10, 11], причем распределение по классам проводилось экспертами вручную, что затрудняет работу с большими объемами данных. Поэтому в качестве исследовательского проекта была сформулирована задача автоматической классификации публикуемого контента.

В дальнейшем выводы о выраженности психологических характеристик пользователей позволят проводить экспресс-диагностику для различных приложений. Например, одной из развиваемых областей применения является тематика анализа социоинженерных атак — атак на информационную систему посредством манипулятивных воздействий на пользователя этой системы [2]. Полученная из социальных сетей оценка дает возможность оперативного построения или модификации профиля уязвимостей (оценка степени проявления которых основана на выраженности психологических особенностей) пользователя к тем или иным воздействиям, что в дальнейшем позволяет, в том числе, вырабатывать меры по повышению уровня защищенности всей информационной системы [1]. Еще одним подобным примером является диагностика студентов во время обучения, что позволяет более эффективно распределять студентов на проектные группы [4] или настраивать персонализированную программу обучения [32], стимулировать учащихся на продолжение обучения, что особенно актуально в массовых открытых онлайн курсах (MOOC) [25].

В качестве базового инструментария был выбран язык R [33] как продолжение предыдущих курсов по анализу данных и основам машинного обучения. На момент постановки задачи исследовательского проекта у студентов уже были навыки работы с языком R и основами статистического моделирования, в частности, был изучен такой метод как логистическая регрессия и ее применение в задачах классификации, что позволило сфокусировать задачу на методах работы с текстом. Вынесение методов анализа текста в исследовательский проект позволяет студентам познакомиться с широким классом задач и инструментарием для их решения, часто не включаемым в курсы по анализу данных. Проблема классификации текста в последние годы лишь набирает популярность, ей посвящено множество исследований [3, 6, 12], включая разработку методов работы с короткими текстами, что особенно актуально при анализе тестов из различных социальных сетей [7, 8]. В них можно найти результаты использования разнообразных алгоритмов классификации: метод k-средних [6], деревья решений [12], нейронные сети [8] и др., также можно найти статьи, направленные на сравнение перечисленных методов. Однако в большинстве статей исследователи обращают внимание на сами алгоритмы классификации текста без применения результатов для анализа особенностей авторов текстов. Именно в этом ключе будет интересно данное исследование, так как классы текста могут в дальнейшем использоваться для оценки психологических характеристик пользователей.

Рассматриваемый в этой статье пример (кейс) исследовательского проекта и его результаты, таким образом, интересны как с точки зрения исследований, связанных с выявлением психологических особенностей пользователей социальных сетей, так и с образовательной точки зрения, иллюстрируя последовательный процесс обучения студентов основам обработки текста и алгоритмам машинного обучения.

3. ЭТАПЫ ПРОЕКТА

Первый этап включал знакомство с предметной областью, разработанной ранее классификацией и предварительный описательный анализ данных. Студентам была

предоставлена собранная на предыдущих этапах исследования выборка из более 2000 постов студентов из социальной сети «ВКонтакте» и проведена ручная (неавтоматизированная) классификация по системе, предложенной в [9]. При этом эксперты проводили классификацию, ориентируясь только на тексты постов, в предоставленном им наборе не было картинок, что «уравнивает» исходные данные для экспертов и алгоритмов, так как алгоритмы в качестве входных параметров используют именно текст, игнорируя прикрепленные картинки, а эксперты, наоборот, в первую очередь обращают внимание именно на картинку, если она есть.

Разработанная классификация предполагает характеристику постов по трем категориям, каждая из которых, в свою очередь, делится на непересекающиеся классы (рис. 1). Так, категория «эмоции» предполагает три класса (позитивный, негативный, нейтральный / неэмоциональный), по критерию «действие» посты разделены на три класса (побудительный к действию, благотворительный, продающий, не побудительный), а категория «информация» дает наибольшее число классов (формальный, событийный, личный, цитатный, ссылочный, кулинарный, неинформационный). Каждый пост отнесен к одному из классов в каждой категории, например, <личный позитивный не побудительный> или <нейтральный кулинарный продающий>. По окончании данного этапа было выявлено, что «кулинарных» и «продающих» постов недостаточно для дальнейшей классификации автоматизированными средствами (менее 5 постов обоих типов).

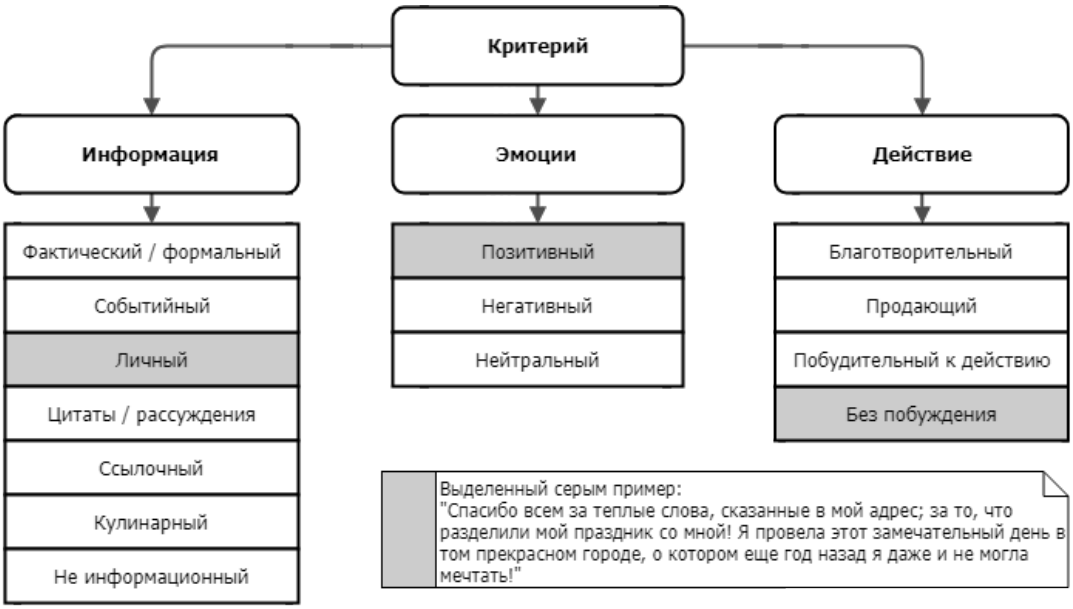


Рис. 1. Структура классификации и пример поста

Второй этап посвящен базовым навыкам предобработки текстов и знакомству с соответствующими инструментами. Со студентами были разобраны функции пакета `stringr` [43] языка R, позволяющего работать со строками (например как с помощью простых регулярных выражений можно удалить все цифры и знаки пунктуации — листинг 1), и пакета `tidytext` [38], реализующего различные методы анализа текста, включая используемые на этом этапе функции разбиения текста на токены (в данной задаче в качестве токенов рассматривались отдельные слова, но во время разбора методов обсуждались и другие варианты).

Затем с помощью изученных в рамках предыдущих курсов инструментов анализа, в частности функций пакета dplyr [42], предназначенных для удобной работы с таблицами данных, был проведен первичный частотный анализ слов: выявлены наиболее частые слова, проведен анализ частот слов в зависимости от типа поста.

```
# подключение библиотеки
library(stringr)

# удаление знаков пунктуации, цифр и переходов на следующую строку
data$text = str_replace_all(data$text, fixed("\n"), " ")
data$text = str_replace_all(data$text, "[[:punct:]]", "")
data$text = str_replace_all(data$text, "[[0-9]]+", "")
```

Листинг 1. Примеры работы с текстом с помощью функций пакета stringr

Полученные к этому моменту результаты начали вызывать дополнительные вопросы у студентов, в частности из-за того, что разные формы одного и того же слова считаются отдельно, а наиболее частыми оказываются служебные слова, не несущие смысловой нагрузки и не отличающиеся для разных классов (например «и», «не», «в»). Такие вопросы позволяют перейти к следующему этапу предобработки и рассмотреть понятия «стоп-слов», а также инструменты для преобразования слов к нормальной форме (лемматизации). Для работы с текстами на русском языке была рассмотрена специализированная консольная программа MyStem, разработанная сотрудниками Яндекса [30], которая позволяет не только проводить лемматизацию, но и определять грамматическую форму слова, что тоже может служить признаком в дальнейшей классификации.

Таким образом, на этапе предобработки были удалены числа, знаки препинания, латинские символы, пустые строки, стоп – слова, проведена лемматизация слов и базовый частотный анализ (код представлен в листинге 2).

```
# подключение библиотек
library(tidytext)
library(dplyr)

# разбиение на слова: output - название нового столбца со словами,
#       input - название исходного столбца, в котором содержится
#       анализируемый текст (уже лемматизированный)
text.tidy = unnest_tokens(tbl = data, output = words, input = text,
                          token = ""words, to_lower = TRUE)

# создание списка стоп-слов (стандартный и некоторые дополнительные)
rustopwords <- data.frame(words=c(stopwords::stopwords("ru"),
                                  "который", "весь", "это", "г"),
                          stringsAsFactors=FALSE)
# удаление стоп-слов
text.tidy = text.tidy %>% anti_join(rustopwords)

# подсчет частоты встречаемости слов на всех текстах
countText = text.tidy %>% count(words)

# подсчет частоты встречаемости слов отдельно по каждому классу информационного критерия
countTextInformation = text.tidy %>% count(words, Information)
```

Листинг 2. Подсчет частот слов

На третьем этапе были предложены первые подходы для классификации постов на основе наиболее часто встречающихся слов в каждом классе. Наиболее часто встречающиеся слова в каждом классе были визуально представлены в виде облаков слов, построенных средствами библиотеки `wordcloud` [21] (пример для класса событийных постов приведен в листинге 3).

```
# подключение библиотеки
library(wordcloud)

# выбираем событийные посты (второй класс), а затем 50 наиболее частных слов
countEvent = countTextInformation %>% filter(Information == 2) %>% top_n(50)

# строим облако
wordcloud(words = countEvent$words, # слова
          freq = countEvent$n, # частоты
          scale=c(2,0.3)) # шкала размера текста - от 0.3 до 2
```

Листинг 3. Построение облака слов для событийных постов

Облако слов, построенное по всей выборке, представлено на рис. 2, примеры для других классов приведены на рис. 3 и 4. Важно отметить, что в «эмоциональных» постах было выявлено меньше десяти часто встречающихся слов, на основе которых проводить классификацию не имело бы смысла (слова: «спасибо», «день», «год», «хотеть» и т.д.).; следовательно, для данного типа постов необходим принципиально иной подход к классификации, например использование словаря тональностей.

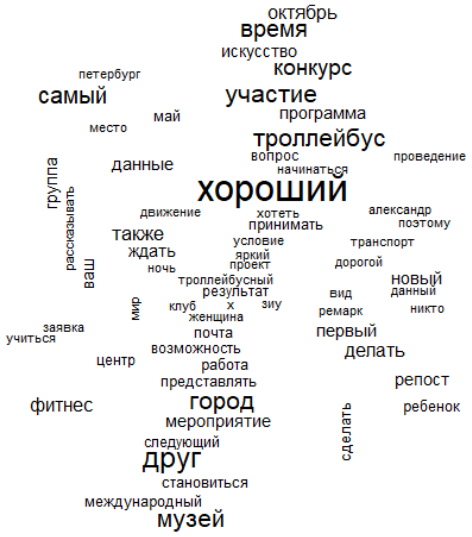
Рис. 2. Облако слов для всей выборки

По сформированным спискам наиболее частых слов в каждом классе постов студентами был предложен следующий принцип классификации: если в тексте встречалось слово из набора, соответствующего конкретному классу, то ему присваивался этот класс. Процент правильной классификации данным методом составил 50%, после чего был сделан вывод о бесполезности подобного классификатора, что позволило логично перейти к следующему этапу проекта.

(а) Событийные посты

(б) Цитаты / рассуждения

Рис. 3. Примеры облаков слов для событийных и цитатных постов



(а) Побудительные посты

(б) Благотворительные посты

Рис. 4. Примеры облаков слов для побудительных и благотворительных постов

На четвертом этапе проекта были рассмотрены возможности применения методов машинного обучения для создания моделей классификации. Для построения моделей были использованы изученные ранее методы — логистическая регрессия (функция `glm()`) и деревья решений (функция `rpart()` из библиотеки `rpart` [40]), в которых в качестве предикторов использовался тот факт, встретилось ли слово из соответствующего списка, относящегося к конкретному классу. Для оценивания результатов

классификации выборка была разделена на тестовую и обучающую части (20% и 80% выборки соответственно), построение модели проводилось на обучающей выборке, оценка качества предсказания — на тестовой. Каждый класс предсказывался отдельно, то есть, например, при предсказании цитатных постов все тексты разделялись на два класса — цитатные и не цитатные (все остальные)

Результаты тестирования показали, что логистическая регрессия дает не больше 50% точности предсказания, а показатели для деревьев решений лучше только для цитатных (70% точности предсказания) и ссылочных (92% точности предсказания) постов. После этого для улучшений результатов в качестве дополнительного предиктора была добавлена информация о репостах (то есть запись размещена именно пользователем страницы или это повторное размещение поста из сообщества/страницы другого пользователя), однако результат не изменился.

Так как большинство постов невозможно было классифицировать, используя ранее описанные модели, то следующим этапом стало использование тематического моделирования, в частности — метод латентного размещения Дирихле (latent Dirichlet allocation, LDA) [15], реализованный в R через функцию `LDA()` из библиотеки `topicmodels` [23].

Тематическая модель позволяет выделить несколько не определенных заранее тем (в рассматриваемом проекте было выделено четыре темы), выявить, в какой степени каждый документ относится к той или иной теме, а также какие слова характеризуют каждую тему. Последнее дает возможность при необходимости интерпретировать каждую из полученных тем. Логика построения модели представлена в листинге 4.

```
# подключение библиотек
library(topicmodels)
library(tidyr)

# считаем частоты отдельно по каждому посту
word_counts <- text.tidy %>%
    count(post.id, words) %>%
    ungroup()

# строим Document-Term Matrix
post_dtm <- word_counts %>%
    cast_dtm(post.id, words, n)

# построение модели
post_lda <- LDA(post_dtm, k = 4, control = list(seed = 32654))

# получим вероятности того, что слово относится к той или иной теме
# (per-topic-per-word probabilities), обозначаемые beta
topics <- tidy(post_lda, matrix = "beta")

# топ-10 слов в каждой теме
topics_top_terms <- topics %>%
    group_by(topic) %>% # группируем по теме
    top_n(10, beta) %>% # находим топ-10 в каждой теме
    arrange(topic, -beta) # упорядочиваем по убыванию вероятности

# получим вероятности того, что документ относится к той или иной теме
# (per-document-per-topic probabilities), обозначаемые gamma
documents <- tidy(post_lda, matrix = "gamma")
```

Листинг 4. Построение тематической модели

На рис. 5 представлены наиболее часто встречающиеся в каждой теме слова.

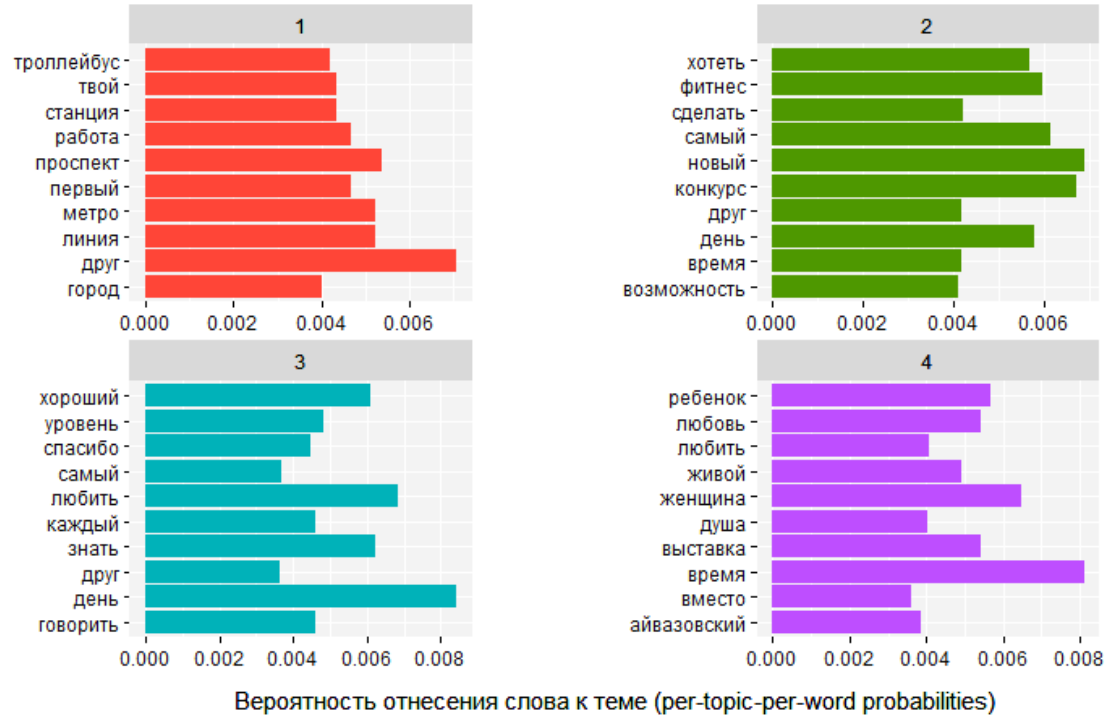


Рис. 5. Тематическое моделирование

Далее, используя вероятность отнесения к каждой из трех (так как общая сумма равна единице, то вероятность для четвертой темы не вносит дополнительной информации) тем в качестве предикторов, были заново построены логистические регрессии и деревья решений. Логистическая регрессия со стандартным порогом разбиения (0,5) так и не помогла различить классы постов, а вот деревья решений помогли различить фактические (77 % точность предсказания, точность (precision) 0,22, полнота (recall) 0,06) и цитатные (64 % точности, точность 0,56, полнота 0,47) посты. При изменении порога принятия решения в логистической регрессии, оптимальное значение которого определялось на обучающей выборке, появилась возможность различать фактические, но хуже, чем для дерева решений (75 %, порог равен 0,25, точность 0,26, полнота 0,15), событийные (74 %, порог равен 0,06, точность 0,44, полнота 0,10), личные (63 %, порог равен 0,16, точность 0,04, полнота 0,09), ссылочные (61 %, порог равен 0,05, точность 0,06, полнота 0,44) и побудительные (62 %, порог равен 0,1, точность 0,13, полнота 0,60) посты.

По результатам этого этапа студенты обратили внимание на еще одну особенность данных — сильное смещение классов в выборке. В частности, при изучении моделей на основе дерева решений для, например, побудительных к действию постов видно, что модель относит 100 % выборки к классу «не побудительных» постов, то есть построена модель вида «все посты не побуждают к действию». Более того, только 8 % постов из всей анализируемой выборки на самом деле являются побудительными к действию. Такая же ситуация складывается и с другими классами: чаще всего модели, показывающие хорошую точность предсказания, просто относят большую часть постов к противоположному классу. Такие выборки называются несбалансированными и достаточно часто встречаются при решении практических задач, одним из наиболее ярких примеров

является выявление аномалий или мошеннических действий — они происходят гораздо реже, чем регулярные и корректные действия [13]. Основная проблема при работе с подобными данными — тот факт, что алгоритмы, используемые при построении моделей, направлены на минимизацию ошибок, соответственно, может оказаться выгоднее отнести все объекты к наибольшему классу, не пытаясь как-то выделить объекты меньшего класса. То есть при работе с несбалансированными выборками классификаторы могут получаться очень плохие с точки зрения точности (precision) или полноты (recall).

Столкновение студентов с подобной проблемой в результате проектной деятельности, а не только в рамках одной из тем в курсе машинного обучения позволяет сформировать более полное понимание ситуации и вызванных ею последствий и в дальнейшем осознанно применять необходимые методы, а не просто следовать некоему алгоритму [37] (в частности, не ориентироваться только на метрику accuracy при оценивании модели).

Для решения сформулированной проблемы на шестом этапе проекта были рассмотрены различные методы работы с несбалансированными выборками [20]. Была использована функция `ROSE()` из пакета `ROSE` [29] для искусственного увеличения количества постов, представленных сравнительно небольшим количеством примеров в тренировочной выборке. В отличие от методов уменьшения количества наблюдений большего класса (down-sampling) или искусственного увеличения количества наблюдений меньшего класса (over-sampling) ROSE предполагает одновременную работу с двумя классами. В результате получилось лучше различить фактические (логистическая регрессия, 72% правильных предсказаний, точность 0,22, полнота 0,18) и распознать благотворительные (дерево, 77% правильных предсказаний, точность 0,05, полнота 0,98) посты. Однако событийные посты перестали различаться.

Конечным итогом исследования явились конкретные модели классификации для каждого типа постов, представленные в таблице 1. Студентами были выбраны те модели, которые показали наилучшее качество предсказания на тестовой выборке. Таким образом, был предложен первоначальный подход к быстрой классификации контента из социальных сетей для дальнейшего их изучения и применения в сфере психологии и социологии, который в то же время позволил изучить базовые понятия анализа текста и применить их для обучения классификационных моделей.

Таблица 1. Результаты классификации

Класс постов	**Метод классификации**	**Результат**
Фактические	Логистическая регрессия после тематического моделирования с пакетом `ROSE` (порог 0,5)	72%
Событийные	Логистическая регрессия после тематического моделирования (порог 0,06)	74%
Личные	Логистическая регрессия после тематического моделирования (порог 0,16)	63%
Цитатные	Дерево решений по частотным словам (порог 0,5)	70%
Ссылочные	Дерево решений по частотным словам (порог 0,5)	92%
Благотворительные	Дерево решений после тематического моделирования с `ROSE` (порог 0,5)	77%
Побудительные	Логистическая регрессия после тематического моделирования (порог 0,1)	62%

4. ИСПОЛЬЗОВАННЫЕ ИНСТРУМЕНТЫ

Как уже упоминалось, анализ текста проводился методами языка R и ряда пакетов. Для удобства дальнейшего использования перечислим использованный инструментарий еще раз с кратким описанием, для чего применяется тот или иной пакет, и ссылками на документацию или тьюториалы:

- `dplyr` — удобная обработка данных: фильтрация, обобщение и т. д. [42]; тьюториал https://dplyr.tidyverse.org/index.html;
- `tidyr` — преобразование данных – слияние, разделение, переформатирование и т. д. [41]; тьюториал https://tidyr.tidyverse.org/index.html;
- `stringr` — работа со строками – заменить, обрезать, найти, выделить часть и т. д. [43]; тьюториал https://stringr.tidyverse.org/index.html;
- `tidytext` — работа с текстами – разбиение на слова, подсчет частот, подсчет метрики tf-idf, Document-Term Matrix и т. д. [38]; тьюториал https://www.tidytextmining.com/;
- `topicmodels` — построение тематических моделей (Latent Dirichlet Allocation — LDA и Correlated Topic Models — CTM) [23];
- `wordcloud` — построение облаков слов [21];
- `rpart` — построение классификационных и регрессионных деревьев решений методом CART [40];
- `caret` —- упрощение работы с методами машинного обучения, включает методы-обертки для алгоритмов, реализованных в других пакетах, для единого формата их вызова, а также различные дополнительные функции —- вычисление метрик качества, кросс-валидация, предобработка, выбор признаков, значимость признаков и т. д. [27]; тьюториал https://topepo.github.io/caret/index.html;
- `ROSE` — реализация метода ROSE (Random Over-Sampling Examples) для корректировки несбалансированных выборок при бинарной классификации [29].

Кроме того, для лемматизации тестов на русском языке использовалась программа MyStem [30].

5. ЗАКЛЮЧЕНИЕ

Таким образом, в работе, во-первых, решена прикладная научная задача — построены модели для классификации постов, публикуемых в социальной сети ВКонтакте, на основе текста этих постов. Во-вторых, предложен подход, позволяющий включить изучение методов анализа текста, обычно не рассматриваемых в курсах по анализу данных и машинному обучению, в индивидуальную траекторию обучения студента за счет организации исследовательского проекта (индивидуального или в малых группах). В данной работе рассмотрен пример конкретной задачи (предсказание классов и потенциальная взаимосвязь с психологическими особенностями), но общий принцип применим для любой достаточно сложной задачи — это могут быть данные с отзывами на фильмы, игры, любые товары, тексты песен, описания приложений и т. д. Основных ограничений несколько:

1) для постановки задачи предсказания в данных должно быть разбиение на какие-то классы (жанры, исполнители, полярность и др.) для задачи классификации или какой-то метрический показатель (количество скачиваний, оценка, продажи и т. д.) для задачи регрессии;

2) тексты должны быть на понятном языке (исследование текстов, например, на английском при неуверенном знании языка существенно затрудняет интерпретацию и поиск возможных проблем); и, наконец,
3) тематика должна быть интересна студентам.

В-третьих, рассмотрен инструментарий языка R, позволяющий проводить анализ текста, приведены примеры использования, ссылки на документацию и более подробное описание каждого из методов (раздел Использованные инструменты). Подробный разбор шагов исследования позволяет использовать статью в качестве примера учебного руководства для изучения методов обработки текста и алгоритмов машинного обучения.

В-четвертых, в работе рассмотрен один из вариантов построения процесса изучения методов, основанный, в том числе, на особенностях конкретных данных и особенностях конкретных студентов и организации процесса обучения. В частности, пререквизитами проекта выступали базовые курсы по анализу данных на языке R, то есть во время работы над проектом не нужно было останавливаться подробно на основах языка R, принципах работы в RStudio и базовых алгоритмах машинного обучения. Случаи, когда это условие не соблюдается, потребуют добавления в начало проекта блоков для изучения этих методов. К особенностям конкретных данных можно отнести несбалансированность выборки, что позволило включить раздел по работе с данными соответствующего типа. В зависимости от задачи среди таких особенностей может быть необходимость мультиклассовой, а не бинарной классификации или потребность в нормировании данных.

Также при изменении скорости выполнения проекта и/или ограничений по времени предложенный проект можно расширить. Так, среди направлений развития как исследовательской, так и методологической части можно выделить возможность наращивания обучающей выборки для увеличения точности классификаторов, учет грамматических характеристик слов (например повелительного наклонения для побудительных постов), извлекаемых с помощью MyStem, применение словаря тональностей русского языка, рассмотрение более сложных моделей классификации, например ансамблей.

Список литературы

1. *Абрамов М. В.* Автоматизация анализа социальных сетей для оценивания защищённости от социоинженерных атак // Автоматизация процессов управления. 2018. № 1(51). С. 34–40.
2. *Азаров А. А., Тулупьева Т. В., Суворова А. В., Тулупьев А. Л., Абрамов М. В., Юсупов Р. М.* Социоинженерные атаки. Проблемы анализа. Наука, 2016. 352 с.
3. *Батура Т. В.* Методы автоматической классификации текстов // Программные продукты и системы. 2017. Т. 30. № 1. doi: 10.15827/0236- 235X.030.1.085-099
4. *Бордовская Н. В., Тулупьева Т. В., Тулупьев А. Л., Азаров А. А.* Возможности электронной социальной сети в решении профессиональных задач вузовского преподавателя // Психологическая наука и образование. 2016. Т. 21. № 4. С. 32–39. doi: 10.17759/pse.2016210403
5. *Мухин А. М., Чернышев Г. А.* MiniValgrind: простой детектор утечек памяти //Компьютерные инструменты в образовании. 2017. № 2. С. 5–15.
6. *Осипова Ю. А., Лавров Д. Н.* Применение кластерного анализа методом k-средних для классификации текстов научной направленности //Математические структуры и моделирование. 2017. № 3 (43). С. 108–121. doi: 10.25513/2222-8772.2017.3.108-121
7. *Полячков А. А.* Классификация слабоструктурированного текста малого размера // Журнал научных и прикладных исследований. 2015. № 5. С. 124–125.

8. *Смирнова О. С., Шишков В. В.* Выбор топологии нейронных сетей и их применение для классификации коротких текстов // International Journal of Open Information Technologies. 2016. Т. 4. № 8. С. 50–54.
9. *Тулупьева Т. В., Суворова А. В., Азаров А. А., Тулупьев А. Л., Бордовская Н. В.* Возможности и опыт применения компьютерных инструментов в анализе цифровых следов студентов-пользователей социальной сети // Компьютерные инструменты в образовании. 2015. № 5. С. 3–13.
10. *Тулупьева Т. В., Тафинцева А. С., Тулупьев А. Л.* Подход к анализу отражения особенностей личности в цифровых следах // Вестн. психотерапии. 2016. № 60 (65). С. 124–137.
11. *Тулупьева Т. В., Тулупьев А. Л., Ющенко Н. А.* Проявление ценностных ориентаций пользователей социальных сетей в контенте персональных страниц (на примере сети «ВКонтакте») // Вестник психотерапии. 2014. № 52. С. 37–50.
12. *Фомин В. В., Фомина И. К., Осочкин А. А.* Классификация текстов на основе частотного и морфологического анализов с применением алгоритмов data-mining // Информатизация образования и науки. 2016. № 3. С. 137–152.
13. *Abdallah A., Maarof M. A., Zainal A.* Fraud detection system: A survey // Journal of Network and Computer Applications. 2016. Vol. 68. P. 90–113. doi: 10.1016/j.jnca.2016.04.007
14. *Barnett J., Lodder J., Pengelley D., Pivkina I., Ranjan D.* Designing student projects for teaching and learning discrete mathematics and computer science via primary historical sources // Recent developments on introducing a historical dimension in mathematics education. 2011. Vol. 78. P. 189–201. doi: 10.5948/UPO9781614443001.018
15. *Blei D. M., Ng A. Y., Jordan M. I.* Latent dirichlet allocation // Journal of machine learning research. 2003. Vol. 3. Jan. P. 993–1022.
16. *Bonilla Y., Rosa J.* # Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States //American Ethnologist. 2015. Vol. 42. № 1. P. 4–17. doi: 10.1111/amet.12112
17. *Boulianne S.* Social media use and participation: A meta-analysis of current research // Information, Communication & Society. 2015. Vol. 18. № 5. P. 524–538. doi: 10.1080/1369118X.2015.1008542
18. *Bulmer M., Haladyn J.K.* Life on an Island: A simulated population to support student projects in statistics // Technology Innovations in Statistics Education. 2011. Vol. 5. № 1.
19. *Centola D., van de Rijt A.* Choosing your network: Social preferences in an online health community // Social science & medicine. 2015. Vol. 125. P. 19–31. doi: 10.1016/j.socscimed.2014.05.019
20. *Chawla N. V., Japkowicz N., Kotcz A.* Special issue on learning from imbalanced data sets // ACM Sigkdd Explorations Newsletter. 2004. Vol. 6. № 1. P. 1–6. doi: 10.1145/1007730.1007733
21. *Fellows I.* wordcloud: Word Clouds. R package version 2.5. 2014. URL: https://CRAN.R-project.org/package=wordcloud
22. *Ferreira M. M., Trudel A. R.* The impact of problem-based learning (PBL) on student attitudes toward science, problem-solving skills, and sense of community in the classroom // Journal of classroom interaction. 2012. Vol. 47. № 1. P. 23–30.
23. *Grun B., Hornik K.* topicmodels: An R Package for Fitting Topic Models // Journal of Statistical Software. 2011. Vol. 40. № 13. P. 1–30. doi: 10.18637/jss.v040.i13
24. *Hallinger P., Bridges E. M.* A systematic review of research on the use of problem-based learning in the preparation and development of school leaders // Educational Administration Quarterly. 2017. Vol. 53. № 2. P. 255–288.
25. *Hone K. S., El Said G. R.* Exploring the factors affecting MOOC retention: A survey study // Computers & Education. 2016. Vol. 98. P. 157–168. doi: 10.1016/J.COMPEDU.2016.03.016
26. *Kong L. N., Qin B., Zhou Y. Q., Mou S. Y., Gao H. M.* The effectiveness of problem-based learning on development of nursing students' critical thinking: A systematic review and meta-analysis // International journal of nursing studies. 2014. Vol. 51. № 3. P. 458–469. doi: 10.1016/j.ijnurstu.2013.06.009
27. *Kuhn M.* caret: Classification and Regression Training. R package version 6.0-77. 2017. URL: https://CRAN.R-project.org/package=caret
28. *Loyens S. M., Jones S. H., Mikkers J., van Gog T.* Problem-based learning as a facilitator of conceptual change // Learning and Instruction. 2015. Vol. 38. P. 34–42.

29. *Lunardon N., Menardi G., Torelli N.* ROSE: a Package for Binary Imbalanced Learning // R Journal. 2014. Vol. 6(1). P. 82–92.
30. MyStem Технологии Яндекса. URL: https://tech.yandex.ru/mystem/
31. *Park G., Schwartz H. A., Eichstaedt J. C., Kern M. L., Kosinski M., Stillwell D. J., Seligman M. E.* Automatic personality assessment through social media language //Journal of personality and social psychology. 2015. Vol. 108. № 6. P. 934–952. doi: 10.1037/pspp0000020
32. *Prain V., Cox P., Deed C., Dorman J., Edwards D., Farrelly C., Waldrip B.* Personalised learning: Lessons to be learnt // British Educational Research Journal. 2013. Vol. 39. № 4. P. 654–676. doi: 10.1080/18334105.2014.11082020
33. *R Core Team* R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL: https://www.R-project.org/
34. *Richter E., Nehorai A.* Enriching the Undergraduate Program with Research Projects [SP Education] // IEEE Signal Processing Magazine. 2016. Vol. 33. № 6. P. 123–127. doi: 10.1109/MSP.2016.2601652
35. *Savery J. R.* Overview of problem-based learning: Definitions and distinctions // Essential readings in problem-based learning: Exploring and extending the legacy of Howard S. Barrows. 2015. Vol. 9. P. 5–15. doi: 10.7771/1541-5015.1002
36. *Scherman A., Arriagada A., Valenzuela S.* Student and environmental protests in Chile: The role of social media // Politics. 2015. Vol. 35. № 2. P. 151–171. doi: 10.1111/1467-9256.12072
37. *Schmidt H. G., Rotgans J. I., Yew E. H. J.* The process of problem-based learning: what works and why // Medical education. 2011. Vol. 45. № 8. P. 792–806. doi: 10.1111/j.1365-2923.2011.04035.x
38. *Silge J., Robinson D.* tidytext: Text Mining and Analysis Using Tidy Data Principles in R // Journal of Statistical Software. 2016. Vol. 1. № 3. doi: 10.21105/joss.00037
39. *Spyratos S., Vespe M., Natale F., Weber I., Zagheni E., Rango M.* Migration Data using Social Media. JRC Science Hub, 2018. 34 p. doi: 10.2760/964282
40. *Therneau T., Atkinson B., Ripley B.* rpart: Recursive Partitioning and Regression Trees. R package version 4.1-11. 2017. URL: https://CRAN.R-project.org/package=rpart
41. *Wickham H., Henry L.* tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.7.1. 2017. URL: https://CRAN.R-project.org/package=tidyr
42. *Wickham H., Francois R., Henry L., Kirill Muller K.* dplyr: A Grammar of Data Manipulation. R package version 0.7.3. 2017. URL: https://CRAN.R-project.org/package=dplyr
43. *Wickham H.* stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.2.0. 2017. URL: https://CRAN.R-project.org/package=stringr

Поступила в редакцию 07.05.2018, окончательный вариант — 07.06.2018.

Computer tools in education, 2018
№ 3: 49–64
http://ipo.spb.ru/journal
doi:10.32603/2071-2340-3-49-64

RESEARCH PROJECT AS A TOOL FOR TEACHING TEXT ANALYSIS METHODS: PREDICTING THE POST CLASS IN THE SOCIAL NETWORK

Suvorova A. V.[1,2], Smirnova K. R.[3], Budin E. A.[3], Tulupyeva T. V.[1,3,4], Tulupyev A. L.[1,4], Abramov M. V.[1,4]

[1]St. Petersburg Institute for Informatics and Automation of the RAS, Saint Petersburg, Russia
[2]National Research University Higher School of Economics, Saint Petersburg, Russia
[3]North-West Institute of Management, branch of RANEPA, Saint Petersburg, Russia
[4]St. Petersburg State University, Saint Petersburg, Russia

Abstract

The article describes a student research project on predicting the class of a post on a social network based on its textual content. The features of the project are discussed as an integral part of the trajectory of teaching data analysis methods, including text analysis methods and tools that are often not included in machine learning courses. The formulation of the problem, the stages of its solution, the sequence of considering new methods as a way for solving students' problems, as well as the used tool of the R environment are described. The possibilities of expanding the task and its modifications depending on the level of training of students are given.

Keywords: *problem-based learning, social networks, machine learning, text analysis, classification, research automation, R language.*

Citation: A. V. Suvorova, K. R. Smirnova, E. A. Budin, T. V. Tulupyeva, A. L. Tulupyev and M. V. Abramov, "Research Project as a Tool for Teaching Text Analysis Methods: Predicting the Post Class in the Social Network," *Computer tools in education*, no. 3, pp. 49–64, 2018 (in Russian).

Acknowledgements: *This work was partially supported by the by RFBR according to the research projects No. 16-31-60063, No. 18-01-00626, No. 18-37-00323 and Governmental contract (SPIIRAS) No. 0073-2018-0001.*

Received 07.05.2018, the final version — 07.06.2018.

Alena V. Suvorova, PhD, Senior Researcher, Theoretical and Interdiciplinary Computer Science Laboratory, SPIIRAS; Associate Professor, HSE University;199178, Russia, St. Petersburg,14-th Line VO, 39, suvalv@gmail.com
Karina R. Smirnova, student, NWIM RANEPA, Smirnova.KR@mail.ru
Evgeniy A. Budin, student, NWIM RANEPA, moyapochta456@gmail.com
Tatiana V. Tulupyeva, PhD, Associate Professor, Senior Researcher, Theoretical and Interdiciplinary Computer Science Laboratory. SPIIRAS; Associate Professor, NWIM RANEPA; Associate Professor, Computer Science Department, SPSU, tvt100a@mail.ru

Alexander L. Tulupyev, PhD, Dc. Sci., Associate Professor, Leading Researcher, Theoretical and Interdiciplinary Computer Science Laboratory, SPIIRAS; Professor, Computer Science Department, SPSU, alt@iias.spb.ru
Maxim V. Abramov, PhD, Researcher, Theoretical and Interdiciplinary Computer Science Laboratory, SPIIRAS; Senior Lecturer, Computer Science Department, SPSU, mva16@list.ru

Суворова Алёна Владимировна,
кандидат физико-математичских наук, старший научный сотрудник, лаборатории теоретических и междисциплинарных проблем информатики, СПИИРАН; доцент НИУ ВШЭ; 199178, Санкт-Петербург, 14-я линия В.О., д. 39,
suvalv@gmail.com

Смирнова Карина Руслановна,
студент, СЗИУ РАНХиГС,
Smirnova.KR@mail.ru

Будин Евгений Александрович,
студент, СЗИУ РАНХиГС,
moyapochta456@gmail.com

Тулупьева Татьяна Валентиновна,
кандидат психологических наук, доцент, старший научный сотрудник, лаборатории теоретических и междисциплинарных проблем информатики, СПИИРАН; доцент СЗИУ РАНХиГС; доцент, кафедра информатики, СПбГУ,
tvt100a@mail.ru

Тулупьев Александр Львович,
доктор физико-математических наук, доцент, главный научный сотрудник, лаб. теоретических и междисциплинарных проблем информатики, СПИИРАН; профессор, кафедра информатики, СПбГУ,
alt@iias.spb.ru

Абрамов Максим Викторович,
кандидат техническх наук, научный сотрудник, лаб. теоретических и междисциплинарных проблем информатики, СПИИРАН; старший преподаватель, кафедра информатики, СПбГУ,
mva16@list.ru