Методы классификации текстовых данных: можно ли потенциал количественного анализа использовать в качественном исследовании?

DOI: 10.19181/inter.2021.13.2.5
Ссылка для цитирования:
Александрова М. Ю. Методы классификации текстовых данных: можно ли потенциал количественного анализа использовать в качественном исследовании? // Интеракция. Интервью. Интерпретация. 2021. Т. 13. № 2. С. 81–96. DOI: https://doi.org/10.19181/inter.2021.13.2.5
For citation:
Aleksandrova M. Yu. (2021) Methods for Classification of Text Data: Can the Potential of Quantitative Analysis Be Applied to Qualitative Research? *Interaction. Interview. Interpretation*. Vol. 13. No. 2. P. 81–96. DOI: https://doi.org/10.19181/inter.2021.13.2.5

Александрова Марина Юрьевна
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
E-mail: myaleksandrova@hse.ru

Интеллектуальный анализ текстовых данных, или текст-майнинг, продемонстрировал активное развитие в последние годы. В статье в качестве примера сравниваются методы классификации, пригодные для решения задач по прогнозированию частичных неответов, и на этом материале автор строит рассуждения о том, как может быть реализован анализ текстовых данных в более широком исследовательском поле. Автор рассматривает ряд метрик, адаптированных для текстового анализа в социальных науках: правильность (accuracy), точность (precision), полноту (recall), F-меру (F1-score), приводит примеры, которые могут помочь исследователю-социологу разобраться, на какую из них стоит обращать внимание в зависимости от поставленной задачи (классифицировать свои текстовые данные с равной точностью или же более полно описать один из интересующих классов). В статье предложена интерпретация результатов, полученных с помощью анализа текстов на материалах Европейского социального исследования (European Social Survey, ESS).

Ключевые слова: текстовые данные; текст-майнинг; текстовый анализ; наивный байесовский классификатор; дерево решений; частичный неответ

В области интеллектуального анализа текстовых данных в последние годы наблюдаются быстрые и значительные изменения. Они происходят благодаря как различным научно-исследовательским лабораториям (развитие подхода в анализе текстовых данных, основанного на векторном представлении слов и разработка модели word2vec на его основе [Mikolov et al., 2017], а также методов машинного обучения и искусственного интеллекта [LeCun et al., 2015]), так и повышенному интересу со стороны индустрии информационных технологий. Например, компания Alphabet разработала языковую модель BERT, которая используется в поисковой системе Google [Devlin et al., 2018], компания OpenAI разработала языковые модели GPT-2 [Radford et al., 2019] и GPT-3 [Brown, 2020], компания Yandex создала Томита-парсера для извлечения структурированной информации из текстовых данных на русском языке[1], Сбербанк подготовил русскоязычную адаптацию языковой модели GPT-3[2], и т.д.

Быстрое развитие области интеллектуального анализа текстовых данных открывает новые возможности для множества других областей научного знания, в частности для социологии. Ученые-социологи тесно взаимодействуют с текстовой информацией, причем как в рамках количественной исследовательской парадигмы (где числовые данные изначально были текстом — в виде анкетных вопросов, инструкций и ответов), так и в рамках качественной парадигмы. Поэтому использование и адаптация методов интеллектуального анализа текстовых данных в социологии представляется перспективным направлением для развития. Методы классификации при работе, например, с транскриптами интервью, можно использовать для определения, описания и сравнения разной тематики, практики описания каких-либо тем, выявления связи между ними и характеристиками информантов. Количественный анализ текстов поспособствует обогащению результатов работы исследователя, поможет обратить его внимание на ранее незамеченные факты, связи или закономерности. Дополнительная польза еще и в том, что появляется еще один, хотя и бездушный, участник процедуры триангуляции для разных этапов качественного исследования. Например, на этапе кодирования транскриптов интервью разными кодировщиками одним из кодировщиков может выступить компьютер, применив латентно-семантический анализ для тематического моделирования — выделения ключевых тем из анализируемых транскриптов с помощью статистических методов. Результат работы компьютера может быть соотнесен с темами, выделенными кодировщиками, послужив базой для дополнительной проверки кодов, полученных кодировщиками, а также предоставляя возможность дополнить их.

В этой статье мы покажем, как может быть реализован анализ текстовых данных и какие могут быть получены результаты с его помощью, на примере

[1] Об учебнике: Tomita Parser // GitHub. 30.05.2019. URL: https://github.com/yandex/tomita-parser (дата обращения: 22.05.2021).

[2] Сбер выложил русскоязычную модель GPT-3 Large с 760 миллионами параметров в открытый доступ // Хабр. 22.10.2020. URL: https://habr.com/ru/company/sberbank/blog/524522/ (дата обращения: 22.05.2021).

прогнозирования возникновения частичных неответов в зависимости от формулировок анкетных вопросов. В рамках этой задачи мы сравниваем примененные методы и их результаты с помощью метрик качества — специальных коэффициентов, позволяющих оценить качество, с которым обученные данными методами модели могут хорошо угадывать частичный неответ в зависимости от формулировки анкетного вопроса. Нами были использованы такие метрики, как правильность (accuracy), точность (precision), полнота (recall), F-мера (F1-score), а также рассмотрены матрицы ошибок для каждой из обученных моделей. Также мы посмотрим на сами полученные результаты — какие слова в анкетных вопросах, «по мнению» обученных моделей, связаны или не связаны с возникновением частичных неответов. В качестве данных выступили формулировки вопросов Европейского социального исследования (European Social Survey, ESS), использовавшиеся для проведения опросов с первой по девятую волны в Великобритании. Данный выбор был сделан в силу того, что компьютерные методы работы с англоязычными текстами на данный момент более разработаны, чем методы обработки русскоязычных текстов. Для сравнения были отобраны такие методы классификации, как случайный лес, наивный байесовский классификатор, дерево решений и логистическая регрессия. Оцифровка текстовых данных проводилась двумя методами: с помощью «мешка слов» [bag-of-words] (частота встречаемости слов) [Zhang et al., 2010] и метрики важности слов TF-IDF (term frequency — inverse document frequency) [Hirschberg, Manning, 2015]. Все это позволило нам сравнить методы классификации между собой и сформулировать на основе нашего примера некоторые рекомендации для исследователей, которые заинтересованы в применении интеллектуального анализа для своих текстовых данных в качественных исследованиях.

Методы классификации и способы их использования

Методы классификации — это совокупность методов, которые разделяют наблюдения на группы в соответствии с заранее известным критерием [Müller, Guido, 2016: 40]. К методам классификации относятся: деревья решений, наивный байесовский классификатор, случайный лес, бинарная логистическая регрессия и т. д. [Géron, 2019: 34]. Исследователь, планирующий работать с массивом транскриптов интервью, мог бы изучить с помощью методов классификации, например, какие слова или темы свойственны информантам разного пола, материального положения, возрастных групп, а также тональность затрагиваемых тем — любые характеристики информантов могут выступать в качестве зависимой переменной.

Предположим, у нас есть зависимая переменная, которая принимает два значения — назовем их классами. Мы обучили некую модель классификации, которая относит наблюдение к одному из этих классов. Процесс работы такой модели, а также насколько результат этой работы оказывается верным или неверным, можно отразить в удобной форме благодаря матрице ошибок

[Stehman, 1997: 81]. Матрица ошибок — это таблица, которую используют для того, чтобы показать качество работы модели классификации на тестовой подвыборке [Lee, 2019: 166] (пример матрицы ошибок см. ниже).

Таблица 1

Общее представление матрицы ошибок

		Предсказание класса моделью		Всего по строкам
		0	1	
Реальное значение класса	0	TN (True Negative)	FP (False Positive)	TN + FP
	1	FN (False Negative)	TP (True Positive)	TP + FN
Всего по столбцам		TN + FN	TP + FP	TN + TP +FN + FP

В строках матрицы ошибок располагаются реальные значения классов, а в столбцах — предсказанные какой-либо обученной моделью значения классов зависимой переменной [Müller, Guido, 2016: 280] для наблюдений. В случае текстового анализа в качестве наблюдений могут выступать отдельные слова, словосочетания или тексты (например, целые транскрипты интервью).

Рассмотрим значения в ячейках представленной матрицы (см. табл. 1). Истинно положительные и истинно отрицательные значения представляют собой наблюдения, чьи классы были верно предсказаны моделью классификации [Lee, 2019: 166]. Это могут быть слова из интервью, которые были отнесены обученной моделью к той категории, которая соответствует реальной категории этих слов.

Ложно положительные и ложно отрицательные значения представляют собой наблюдения, чьи реальный и предсказанный классы не совпали [Lee, 2019: 166]. Например, это слова из интервью, отнесенные обученной моделью к той категории, которая не соответствовала реальной категории этих слов.

Стоит оговориться, что слова «положительные» и «отрицательные» стоило бы писать в кавычках, так как имеется в виду не буквально положительное или отрицательное значение зависимой переменной у наблюдения, а, скорее, то значение, которое было закодировано исследователем как «1» или как «0» — то есть как принадлежность к одному из двух классов зависимой переменной. Содержательно «положительным» классом может быть что угодно, например, наличие выраженного одобрения экологических ценностей, наличие стремления бойкотировать выборы, наличие пропусков в анкетном вопросе, сформулированном определенным образом, и т.д. «Отрицательным» классом, соответственно, будет противоположный по смыслу класс: отсутствие выраженного одобрения экологических ценностей, отсутствие стремления бойкотировать выборы, отсутствие пропусков в анкетном вопросе, сформулированном определенным образом, и т. д.

Истинно положительные значения (TP) — это верно предсказанные «положительные» значения, то есть у соответствующего наблюдения реальное

значение равно единице, и эта же единица была предсказана по данному наблюдению обученной моделью классификации [Lee, 2019: 167]. Например, истинно положительным значением будет совпадение реального наличия частичного неответа и предсказания о наличии частичного неответа.

Истинно отрицательные значения (TN) — это верно предсказанные «отрицательные» значения, то есть у соответствующего наблюдения реальное значение равно нулю, и этот же ноль был предсказан по данному наблюдению обученной моделью классификации [Lee, 2019: 167]. Например, истинно отрицательным значением будет совпадение реального отсутствия частичного неответа и предсказания об отсутствии частичного неответа.

Ложноположительные значения (FP) — это неверно предсказанные «положительные» значения, то есть у соответствующего наблюдения реальное значение равно нулю, но обученная модель классификации предсказала единицу для данного наблюдения [Lee, 2019: 167]. Например, ложноположительным значением будет совпадение реального отсутствия частичного неответа и предсказания о наличии частичного неответа по тому же наблюдению. Ложноположительные значения также называют ошибкой первого рода [Kelleher et al., 2020: 538].

Ложноотрицательные значения (FN) — это неверно предсказанные «отрицательные» значения, то есть у соответствующего наблюдения реальное значение равно единице, но обученная модель классификации предсказала нулевое значение для данного наблюдения [Lee, 2019: 167]. Например, ложноотрицательным значением будет совпадение реального наличия частичного неответа и предсказания отсутствия частичного неответа по тому же наблюдению. Ложноотрицательные значения также называют ошибкой второго рода [Kelleher et al., 2020: 538].

Обученная модель классификации тем лучше, чем у большего числа наблюдений были верно предсказаны классы и чем меньше было совершено неверных предсказаний [Witten et al., 2011: 164]. В работе с массивом интервью исследователь может обращать внимание на то, много или мало было получено ложноотрицательных и ложноположительных предсказаний относительно истинно положительных и истинно отрицательных, чтобы оценить качество полученных результатов. Кроме того, такие ошибки может анализировать исследователь и уже самостоятельно корректировать. Например, такие формулировки можно определять как те, по которым нельзя однозначно сказать о какой-то классификации или отнести такие слова к определенному классу.

На основе матрицы ошибок могут быть рассчитаны также метрики качества обученной модели классификации — показатели, которые в числовой форме отражают, насколько хорошо способна предсказывать классы обученная модель, а также позволяют получить больше информации из полученной матрицы ошибок [Marsland, 2015: 23]. Такими метриками являются, например, правильность, точность, полнота, F-мера. Рассмотрим подробнее каждую из этих метрик.

Правильность — это сумма всех верных предсказаний, сделанных обученной моделью классификации, деленная на общее количество всех предсказаний, сделанных моделью [Powers, 2020: 3].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \text{ [Lee, 2019: 168].}$$

Например, это будет доля верно определенных моделью предсказания слов, связанных с наличием и отсутствием частичного неответа у анкетного вопроса от общего числа всех предсказаний. Можно ли сказать, что чем выше это значение, тем лучшей предсказательной способностью обладает модель? Такое умозаключение будет не совсем корректным по следующей причине. Если, скажем, модель верно предсказывает в 99 случаях из 100 наличие признака и лишь в 5 случаях из 100 — его отсутствие, то значение правильности в первом случае будет равно 0,99, а во втором — 0,05. Получается, что правильность чувствительна к несбалансированному распределению наблюдений между классами. Поэтому не следует ограничиваться правильностью при оценке обученной модели [Lee, 2019: 168].

Таблица 2

Пример матрицы ошибок

Значения	0	1	Всего
0	5	95	100
1	1	99	100
Всего	6	194	200

Точность — доля истинно положительных предсказаний по отношению ко всем положительным предсказаниям, сделанных моделью [Lee, 2019: 168]. Например, какова доля наблюдений, верно помеченных моделью как связанных с появлением частичного неответа. Чем выше значение точности, тем ниже уровень ложноположительных предсказаний.

$$\text{Precision} = \frac{TP}{TP + FP} \text{ [Lee, 2019: 168].}$$

Полнота — доля истинно положительных предсказаний, сделанных моделью классификации, по отношению ко всем наблюдениям, относящимся к положительному классу [Lee, 2019: 168]. Например, из всех наблюдений, которые были помечены предсказательной моделью как связанные с появлением частичного неответа, какая доля действительно связана с частичными неответами? Чем выше значение полноты, тем ниже уровень ложноотрицательных предсказаний.

$$\text{Recall} = \frac{TP}{TP + FP} \text{ [Lee, 2019: 168].}$$

F1-мера — гармоническое среднее точности и полноты [Lee, 2019: 170], позволяет учитывать значения обеих метрик, уравновешивая между собой величину ложноотрицательных и ложноположительных предсказаний, что делает данную метрику более устойчивой в случае несбалансированных классов [Müller, Guido, 2016: 284]. Гармоническое среднее используется вместо простого среднего арифметического, так как, в отличие от последнего, гармоническое среднее дает больший вес низким значениям, благодаря чему высокое значение F1 получит только та модель классификации, у которой одновременно были получены высокие значения и точности, и полноты [Géron, 2019: 92] (что является преимуществом в сравнении с метрикой правильности).

$$\text{F1 score} = \frac{2 \times [Precision \times Recall]}{Precision \; x \; Recall} = \frac{TP}{TP + \frac{FN + FP}{2}}$$ [Géron, 2019: 92].

Рассмотрим понятия точности и полноты на примерах предсказания возникновения частичных неответов в связи с формулировками анкетных вопросов. В качестве положительного класса (1) принимаем наличие частичного неответа, а отсутствие частичного неответа — как отрицательный класс (0), тогда:

- Если коэффициент точности или полноты высокий, это значит, что большее количество формулировок вопросов, связанных с частичными неответами, и было определено моделью предсказания как связанное с частичными неответами.
- Если коэффициент точности низкий, то это означает, что большее количество формулировок вопросов, связанных с отсутствием частичного неответа, было определено предсказательной моделью как связанное с наличием частичного неответа.
- Если коэффициент полноты низкий, то это означает, что большее количество формулировок неответов, связанных с наличием частичного неответа, было определено предсказательной моделью как связанное с отсутствием частичного неответа.

Почему же, если F1 учитывает и точность, и полноту, не следует отказываться от двух последних метрик? Дело в том, что F1 отдает предпочтение моделям, обладающим примерно равными и достаточно высокими значениями точности и полноты. Тем не менее это не всегда то, что необходимо исследователю: в одном случае может быть важнее высокое значение точности, а в другом — высокое значение полноты [Géron, 2019: 93]. Так, если исследователь обучил модель классификации определять слова, которыми пользуются жертвы буллинга, то ему, скорее всего, будет важнее не упустить ни одного человека, который является такой жертвой (высокая полнота), чем по ошибке причислить к жертвам буллинга кого-то, кто таковым не является (низкая точность). Или, например, для другого исследователя, обучившего модель для определения слов, которыми пользуются при описании тех или иных тем в интервью мужчины, будет важно определить не столько все

слова, которыми пользуются именно мужчины (низкая полнота), сколько как можно более точно определить типичные для мужчин слова, при этом минимизировав ошибки и не относя к ним слова, более типичные для женщин (высокая точность).

Подготовка текстовых данных и их предварительный анализ

Собранные нами данные представляли собой полные формулировки заданных респондентам вопросов, объединенных с информацией о наличии и отсутствии трех типов частичного неответа: отказа от ответа, затруднения с ответом и отсутствием ответа. Аналогично массив с транскриптами интервью может быть объединен с информацией, по которой требуется классифицировать имеющиеся тексты.

Тем не менее в таком виде данные все еще требуют дополнительной подготовки для дальнейшего анализа. Текст в процессе обработки должен пройти два этапа: перевод из неструктурированного вида в структурированный вид и далее — перевод структурированных текстовых данных в цифровой формат. Текстовые данные как просто текст являются неструктурированными [Jurafsky, Martin, 2020: 327]. Что в данном случае подразумевается под отсутствием структуры? Безусловно, имеется в виду не та структура, которую мы обычно понимаем, когда говорим о каком-то тексте — его организации, которая помогает лучшему его усвоению читателями благодаря последовательному, понятному изложению мыслей автора, содержательной целостности текста. С точки зрения представления данных текст сам по себе является неструктурированным набором значений, не поддающимся количественному анализу. Чтобы сделать из текста именно текстовые данные, необходимо прибегнуть к дроблению текста на элементы (отдельные предложения, словосочетания или слова) с последующим объединением схожих элементов.

Каким же образом неструктурированный текст переводится в структурированный вид? Сначала проводится сегментация — разбиение текста на отдельные предложения и/или токенизация — разбиение текста на отдельные слова [Bird et al., 2009]. Далее все слова лемматизируются — приводятся к начальной форме, соответствующей той части речи, к которой они относятся. Например, для существительного начальной формой будет единственное число и именительный падеж, для глагола — неопределенная форма, для прилагательного — именительный падеж, единственное число и мужской род (примеры лемматизации см. в табл. 3). Вместо лемматизации может применяться стемминг — выделение основы слова или удаление у него окончания и суффикса [Jurafsky, Martin, 2020: 11]. И лемматизация, и стемминг позволяют разбить любой массив текстов (например, интервью) на отдельные слова, убирая несущественные для анализа различия между ними — в падежах, числе, формах, родах, временах и т. д.

Таблица 3

Примеры лемматизации и стемминга разных частей речи

Слово	Лемматизация	Стемминг
доверия	доверие	довер
проверялась	проверять	провер
печальных	печальный	печаль
trust	trust	trust
checked	check	check
mournful	mournful	mourn

После пролемматизированные слова проверяются с помощью списков стоп-слов, который содержит слова, которые не несут какой-либо смысловой нагрузки. В такие списки, как правило, входят предлоги, междометия, союзы, частицы [Bird et al., 2009: 236]. Список стоп-слов может быть дополнен и самостоятельно — например, исследователь может удалить ненужные ему метки, обозначающие реплики интервьюера и информанта.

Эти шаги позволяют привести текст к структурированному виду, чтобы далее перевести его в цифровой формат. В решении задачи прогнозирования появления частичных неответов на основе формулировок анкетных вопросов мы воспользовались уже ранее описанными [Александрова, 2021] методами оцифровки текстовых данных — «мешком слов» (bag-of-words) (частота встречаемости слов) [Zhang et al., 2010] и мерой важности слов TF-IDF (term frequency — inverse document frequency) [Hirschberg, Manning, 2015]. «Мешок слов» предполагает расчет количества раз, которое каждое слово встречалось в каждом из текстов корпуса [Baayen, 2001]. Корпус текстов — это все текстовые данные исследователя, которые он планирует анализировать. Это могут быть транскрипты интервью, публикации в СМИ, социальных медиа и т. д. Так, расчет частоты встречаемости всех слов, которые есть в коллекции транскриптов интервью, будет представлять собой подсчет числа раз, которое каждое из слов встретилось в каждом из этих транскриптов. Таким образом, большим весом будут обладать те слова, которые использовались информантами чаще, и меньшим весом будут обладать редко упоминавшиеся слова. Соответственно, расчет частоты встречаемости слов может быть полезен для поиска каких-то общих для большинства интервью сюжетов, тем.

Мера важности слов TF-IDF учитывает частоту встречаемости каждого слова в одном тексте с учетом его частоты встречаемости во всех текстах корпуса [Evans, Aceves, 2016: 41]. Таким образом, при работе с транскриптами интервью, при расчете TF-IDF большим весом будут обладать те слова, которые как можно чаще упоминались в как можно меньшем числе транскриптов (то есть это слова, позволяющие увидеть, какими темами или описаниями одно интервью отличается от остальных), меньшим — слова, которые редко использовались в отдельно взятом интервью (то есть те, что мало информативны для описания именно этого интервью) или встречались во многих

имеющихся интервью, а самым низким весом будут обладать слова, присутствующие во всех интервью (т.е. те, что могут быть использованы для описания всех интервью в целом). Соответственно, расчет TF-IDF может быть полезен для поиска специфичных для отдельных интервью сюжетов, тем.

Данные методы перевода текстовых данных в цифровой формат уже использовались нами на формулировках анкетных вопросов исследования ESS в Великобритании в предыдущей работе, где была описана первая попытка обучения модели предсказания частичного неответа с использованием только метода наивного байесовского классификатора [Александрова, 2021]. Модели предсказания частичного неответа в зависимости от формулировок анкетных вопросов обучались как на основе рассчитанных частот встречаемости слов, так и на основе TF-IDF.

Предварительный анализ подготовленных нами данных дал следующие результаты. Распределение частичных неответов представлено в табл. 4. Выборка делилась на обучающую и тестовую в соотношении 70:30.

Таблица 4

Распределение разных типов частичных неответов в собранных данных

	Отказ от ответа	**Затрудняюсь ответить**	**Отсутствие ответа**
Частичного неответа нет	1012	191	1043
Частичный неответ есть	453	1274	422
Обучающая подвыборка	1025	1025	1025
Тестовая подвыборка	440	440	440

На обучающей подвыборке мы строили модели предсказания разных типов частичных неответов в зависимости от формулировок вопросов, а на тестовой подвыборке проверяли их — насколько хорошо они способны «угадывать» слова, содержащиеся в анкетных вопросах, на которые часть респондентов действительно не отвечали. Для обученных моделей рассчитывались коэффициенты метрик качества классификации, позволяющие оценить предсказательную способность данных моделей.

Деление на обучающую и тестовую подвыборки в текстовом анализе помогает исследователю проверить, можно ли доверять результатам, полученным с помощью обученной модели классификации, — действительно ли она видит различия между определенными словами и формулировками интервью или делает это случайным образом, либо эти результаты применимы к проанализированным интервью, но не могут быть использованы для какой-то экстраполяции.

Результаты обучения моделей прогнозирования частичных неответов

Для обучения моделей классификации, предсказывающих наличие или отсутствие частичного неответа, нами использовались методы дерева решений, логистической регрессии, наивного байесовского классификатора и случайного леса. Модели классификации обучались на частотах слов и на коэффициентах TF-IDF. Полученные модели сравнивались с помощью соотнесения метрик качества для каждой из моделей. Наличие частичного неответа кодировалось как положительный класс (1), а его отсутствие — как отрицательный класс (0). В качестве частичных неответов выступали отказ от ответа, затруднение с ответом и отсутствие ответа.

Таблица 5

Матрица ошибок предсказания отказа от ответа, затруднения с ответом, отсутствия ответа для моделей, обученных с помощью дерева решений, наивного байесовского классификатора, случайного леса и логистической регрессии

	Отказ от ответа			Затруднение с ответом			Отсутствие ответа		
Дерево решений		0	1		0	1		0	1
	0	326	89	0	48	31	0	365	64
	1	97	74	1	81	426	1	73	84
Наивный Байесовский классификатор		0	1		0	1		0	1
	0	393	22	0	26	53	0	324	105
	1	131	40	1	36	471	1	49	108
Случайный лес		0	1		0	1		0	1
	0	315	100	0	48	31	0	324	105
	1	89	82	1	88	419	1	60	97
Логистическая регрессия		0	1		0	1		0	1
	0	281	134	0	48	31	0	302	127
	1	75	96	1	81	426	1	49	108

Матрицы ошибок для обученных моделей классификации показаны в табл. 5. Наличие отказа от ответа лучше всего предсказывается моделью логистической регрессии (96 истинно положительных предсказаний), хуже всего — с помощью наивного байесовского классификатора (40 истинно положительных предсказаний). Соответственно, отсутствие отказа от ответа в зависимости от формулировки вопроса предсказывается лучше всего моделью

наивного байесовского классификатора (393 истинно отрицательных предсказаний) и хуже всего — логистической регрессией (281 истинно отрицательное предсказание). Наличие затруднения с ответом лучше всего предсказывается моделью наивного байесовского классификатора (471 истинно положительное предсказание). Отсутствие затруднений с ответом предсказывается одинаково почти всеми моделями, кроме модели наивного байесовского классификатора (26 истинно отрицательных предсказаний). Отсутствие ответа лучше всего предсказывается с помощью модели случайного леса и логистической регрессией (по 108 истинно положительных предсказаний), хуже всего — моделью дерева решений (84 истинно положительных предсказаний). Наличие ответа лучше предсказывалось с помощью дерева решений (365 истинно отрицательных предсказаний), и хуже себя показала модель логистической регрессии (302 истинно отрицательных предсказаний). Таким образом, изучение матриц ошибок позволяет предположить, что лучше показывали себя модели наивного байесовского классификатора и бинарной логистической регрессии. Но стоит учитывать, что зачастую лучшее предсказание одного класса приводит к ухудшению способности угадывать той же моделью другой класс. Поэтому, если необходимо одинаковое качество предсказаний и для наличия, и для отсутствия частичного неответа, то следует обратить внимание на модели дерева решений и случайного леса. Кроме того, можно заметить, что при дисбалансе классов модель наивного байесовского классификатора лучше работает с предсказанием более наполненного класса.

Ниже представлены результаты расчета метрик качества классификации (правильность, точность, полнота и F1) для обученных моделей дерева решений, наивного байесовского классификатора, бинарной логистической регрессии и случайного леса для предсказания возникновения разных типов частичного неответа, связанных с формулировками анкетных вопросов (см. табл. 6).

Таблица 6

Метрики качества предсказания отказа от ответа

Модель	Правильность	Точность	Полнота	F1
Отказ от ответа				
Дерево решений	0.683	0.678	0.683	0.680
Наивный байесовский классификатор	0.719	0.719	0.739	0.693
Случайный лес	0.677	0.684	0.677	0.680
Логистическая регрессия	0.643	0.681	0. 643	0.656
Затруднение с ответом				
Дерево решений	0.809	0.857	0.809	0.827
Наивный байесовский классификатор	0.848	0.834	0.848	0.840
Случайный лес	0.797	0.853	0.797	0.818
Логистическая регрессия	0.809	0.857	0.809	0.827

Окончание табл. 6

Модель	Правильность	Точность	Полнота	F1
Отсутствие ответа				
Дерево решений	0.766	0.762	0.766	0.764
Наивный байесовский классификатор	0.737	0.772	0.737	0.748
Случайный лес	0.718	0.746	0.718	0.728
Логистическая регрессия	0.700	0.753	0.700	0.715

Из табл. 6 можно увидеть, что самым высоким качеством обладают модели, предсказывающие затруднение с ответом, хуже всего — отказ от ответа. Модели, обученные с помощью методов наивного байесовского классификатора, в целом показывают лучшее качество, чем модели, обученные методом случайного леса. Стоит отметить, что, хотя мы и пишем о различии моделей прогнозирования разных типов частичных неответов, обученных разными методами машинного обучения, тем не менее разница в качестве этих моделей не так уж и велика — самая большая разница в метриках качества составляет не более 0,06.

Таким образом, построив несколько моделей классификации, мы имеем возможность посмотреть, как хорошо они могут «угадывать» интересующие нас классы, а также сравнивать их между собой и, в зависимости от стоящих перед исследователем задач, выбирать наиболее подходящую модель.

Заключение

На примере решения задачи прогнозирования частичных неответов мы продемонстрировали, как может быть реализован анализ текстовых данных и как может происходить оценка результатов обучения моделей классификации на текстовых данных. Наше описание матриц ошибок и основанных на них метрик правильности, точности, полноты и F1-меры было адаптировано для лучшего понимания их пользы для текстового анализа в социальных науках, что позволило сформулировать рекомендации по решению задач классификации в текстовом анализе, а также увидеть, какие результаты могут быть получены на практике.

Мы описали особенности подготовки текстовых данных к дальнейшему анализу — с опорой на научную литературу совершили «перевод» описания основных этапов этой подготовки: преобразование неструктурированных текстов в структурированный вид (сегментация, токенизация, лемматизация и стемминг, удаление стоп-слов) и последующая оцифровка уже структурированных текстовых данных (методы «мешка слов» и TF-IDF) на «язык» текстового анализа социологического исследования, а также приводим рассуждения о том, в каком случае какие процедуры могут быть полезны для социолога-исследователя.

Опираясь на проведенное исследование, мы можем сделать вывод, что если исследователю важна общая доля всех верных предсказаний в его текстовом анализе, то лучше использовать метрику F1. Если же планируется,

например, как можно более полно описать какой-то один класс, то нужно искать модель с самым высоким значением полноты. А если исследователю не так важна полнота описания определенного класса, но есть задача сделать это описание более точным (то есть как можно реже вносить в описание интересуемого класса не относящиеся к нему элементы), то выбор следует остановить на модели с самым высоким значением точности.

Литература

Александрова М. Ю. Методы машинного обучения в социологическом исследовании: предсказание частичного неответа с использованием наивного байесовского классификатора // Мониторинг общественного мнения: экономические и социальные перемены. 2021. № 1. С. 329–350. DOI: https://doi.org/10.14515/monitoring.2021.1.1756

Baayen R. H. Word Frequency Distributions. Dordrecht: Springer, 2001. DOI: https://doi.org/10.1007/978-94-010-0844-0

Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Sebastopol: O’Reilly Media, 2009.

Brown T. B. et al. Language Models Are Few-Shot Learner. 2020. URL: https://arxiv.org/pdf/2005.14165.pdf (дата обращения: 22.05.2021).

Devlin J. et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. 2018. URL: https://arxiv.org/pdf/1810.04805.pdf (дата обращения: 22.05.2021).

Evans J. A., Aceves P. Machine Translation: Mining Text for Social Theory // Annual Review of Sociology. 2016. № 42. P. 21–50. DOI: https://doi.org/10.1146/annurev-soc-081715-074206

Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Sebastopol: O’Reilly Media, 2019.

Hirschberg J., Manning C. D. Advances in Natural Language Processing // Science. 2015. Vol. 349. № 6245. P. 261–266. DOI: https://doi.org/10.1126/science.aaa8685

Jurafsky D., Martin J. H. Speech and Language Processing (3rd ed. draft). 2020. URL: https://web.stanford.edu/~jurafsky/slp3/ (дата обращения: 20.05.2021).

Kelleher J. D., Mac Namee B., D’arcy A. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. Cambridge: MIT Press, 2020.

LeCun Y., Bengio Y., Hinton G. Deep Learning // Nature. 2015. Vol. 521. № 7553. P. 436–444. DOI: https://doi.org/10.1038/nature14539

Lee W. M. Python Machine Learning. Indianapolis: John Wiley & Sons, 2019. DOI: https://doi.org/10.1002/9781119557500

Marsland S. Machine Learning: An Algorithmic Perspective. Boca Raton: CRC Press, 2015. DOI: https://doi.org/10.1201/b17476

Mikolov T. et al. Advances in Pre-Training Distributed Word Representations. 2017. URL: https://arxiv.org/pdf/1712.09405.pdf (дата обращения: 22.05.2021).

Müller A. C., Guido S. Introduction to Machine Learning with Python: A Guide for Data Scientists. Sebastopol: O’Reilly Media, 2016.

Powers D. M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. 2020. URL: https://arxiv.org/pdf/2010.16061.pdf (дата обращения: 22.05.2021).

Radford A. et al. Language Models Are Unsupervised Multitask Learners // OpenAI blog. 2019. Vol. 1. № 8. URL: http://www.persagen.com/files/misc/radford2019language.pdf (дата обращения: 22.05.2021).

Stehman S. V. Selecting and Interpreting Measures of Thematic Classification Accuracy // Remote Sensing of Environment. 1997. Vol. 62. № 1. P. 77–89. DOI: https://doi.org/10.1016/s0034-4257(97)00083-7

Witten I., Frank E., Hall M. Data Mining: Practical Machine Learning Tools and Techniques. Burlington: Morgan Kaufmann, 2011. DOI: https://doi.org/10.1016/C2009-0-19715-5

Zhang Y., Jin R., Zhou Z. H. Understanding Bag-of-Words Model: A Statistical Framework // International Journal of Machine Learning and Cybernetics. 2010. № 1. P. 43–52. DOI: https://doi.org/10.1007/s13042-010-0001-0

Сведения об авторе:

Александрова Марина Юрьевна — стажер-исследователь Международной лаборатории исследований социальной интеграции, преподаватель кафедры методов сбора и анализа социологической информации, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. **E-mail**: myaleksandrova@hse.ru; **РИНЦ Author ID:** 1112564; **ORCID ID**: 0000-0002-7683-7750; **ResearcherID**: T-9377-2017.

Статья поступила в редакцию: 04.05.2021
Принята к публикации: 10.06.2021

Methods for Classification of Text Data: Can the Potential of Quantitative Analysis Be Applied to Qualitative Research?

DOI: 10.19181/inter.2021.13.2.5

Marina Yu. Aleksandrova *HSE University, Moscow, Russia*
E-mail: myaleksandrova@hse.ru

Text mining has developed rapidly in recent years. In this article we compare classification methods that are suitable for solving problems of predicting item nonresponse. The author builds reasoning about how the analysis of textual data can be implemented in a wider research field based on this material. The author considers a number of metrics adapted for textual analysis in the social sciences: accuracy, precision, recall, F1-score, and gives examples that can help a sociologist figure out which of them is worth paying attention depending on the task at hand (classify text data with equal accuracy, or more fully describe one of the classes of interest). The article proposes an analysis of results obtained by analyzing texts based on the materials of the European Social Survey (ESS).

Keywords: text data; text mining; text analysis; Naive Bayes classifier; binary logistic regression; decision tree; random decision forest; item nonresponse

References

Aleksandrova M. Yu. (2021) Metody mashinnogo obucheniya v sociologicheskom issledovanii: predskazanie chastichnogo neotveta s ispol'zovaniem naivnogo bajesovskogo klassifikatora [Machine Learning in Social Research: Predicting Item Nonresponse Error Using Naive Bayes Classifier]. *Monitoring obshchestvennogo mneniya: ekonomicheskie i social'nye peremeny* [Monitoring of Public Opinion: Economic and Social Changes]. No. 1. P. 329–350. (In Russ.) DOI: https://doi.org/10.14515/monitoring.2021.1.1756

Baayen R. H. (2001) *Word Frequency Distributions*. Dordrecht: Springer. DOI: https://doi.org/10.1007/978-94-010-0844-0

Bird S., Klein E., Loper E. (2009) *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol: O'Reilly Media.

Brown T. B. et al. (2020) *Language Models Are Few-Shot Learner*. URL: https://arxiv.org/pdf/2005.14165.pdf (accessed 22 May 2021).

Devlin J. et al. (2018) *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.* URL: https://arxiv.org/pdf/1810.04805.pdf (accessed 22 May 2021).

Evans J. A., Aceves P. (2016) Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology*. No. 42. P. 21–50. DOI: https://doi.org/10.1146/annurev-soc-081715-074206

Géron A. (2019) *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol: O'Reilly Media.

Hirschberg J., Manning C. D. (2015) Advances in Natural Language Processing. *Science*. Vol. 349. No. 6245. P. 261–266. DOI: https://doi.org/10.1126/science.aaa8685

Jurafsky D., Martin J. H. (2020) *Speech and Language Processing (3rd ed. draft)*. URL: https://web.stanford.edu/~jurafsky/slp3/ (accessed 20 May 2021).

Kelleher J. D., Mac Namee B., D'arcy A. (2020) *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.* Cambridge: MIT Press.

LeCun Y., Bengio Y., Hinton G. (2015) Deep Learning. *Nature*. Vol. 521. No. 7553. P. 436–444. DOI: https://doi.org/10.1038/nature14539

Lee W. M. (2019) *Python Machine Learning*. Indianapolis: John Wiley & Sons. DOI: https://doi.org/10.1002/9781119557500

Marsland S. (2015) *Machine Learning: An Algorithmic Perspective*. Boca Raton: CRC Press. DOI: https://doi.org/10.1201/b17476

Mikolov T. et al. (2017) *Advances in Pre-Training Distributed Word Representations*. URL: https://arxiv.org/pdf/1712.09405.pdf (accessed 22 May 2021).

Müller A. C., Guido S. (2016) *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Sebastopol: O'Reilly Media.

Powers D. M.W. (2020) *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation.* URL: https://arxiv.org/pdf/2010.16061.pdf (accessed 22 May 2021).

Radford A. et al. (2019) Language Models Are Unsupervised Multitask Learners. *OpenAI blog.* Vol. 1. No. 8. URL: http://www.persagen.com/files/misc/radford2019language.pdf (accessed 22 May 2021).

Stehman S. *V.* (1997) Selecting and Interpreting Measures of Thematic Classification Accuracy. *Remote Sensing of Environment*. Vol. 62. No. 1. P. 77–89. DOI: https://doi.org/10.1016/s0034-4257(97)00083-7

Witten I., Frank E., Hall M. (2011) *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington: Morgan Kaufmann. DOI: https://doi.org/10.1016/C2009-0-19715-5

Zhang Y., Jin R., Zhou Z. H. (2010) Understanding Bag-of-Words Model: A Statistical Framework. *International Journal of Machine Learning and Cybernetics*. No. 1. P. 43–52. DOI: https://doi.org/10.1007/s13042-010-0001-0

Author bio:

Marina Yu. Aleksandrova — Trainee Researcher, International Laboratory for Social Integration Research; Lecturer, Department of Sociological Research Methods, HSE University, Moscow, Russia. **E-mail**: myaleksandrova@hse.ru; **RSCI Author ID:** 1112564; **ORCID ID**: 0000-0002-7683-7750; **ResearcherID**: T-9377-2017.

Received: 04.05.2021
Accepted: 10.06.2021