DOI: 10.14515/monitoring.2021.1.1756

М. Ю. Александрова

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ В СОЦИОЛОГИЧЕСКОМ ИСССЛЕДОВАНИИ: ПРЕДСКАЗАНИЕ ЧАСТИЧНОГО НЕОТВЕТА С ИСПОЛЬЗОВАНИЕМ НАИВНОГО БАЙЕСОВСКОГО КЛАССИФИКАТОРА

Правильная ссылка на статью:
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. https://doi.org/10.14515/monitoring.2021.1.1756.
For citation:
Aleksandrova M. Y. (2021) Machine Learning in Social Research: Predicting Item Nonresponse Error Using Naive Bayes Classifier. *Monitoring of Public Opinion: Economic and Social Changes.* No. 1. P. 329–350. https://doi.org/10.14515/monitoring.2021.1.1756. (In Russ.)

DOI: 10.14515/monitoring.2021.1.1756

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ В СОЦИОЛОГИЧЕСКОМ ИССЛЕДОВАНИИ: ПРЕДСКАЗАНИЕ ЧАСТИЧНОГО НЕОТВЕТА С ИСПОЛЬЗОВАНИЕМ НАИВНОГО БАЙЕСОВСКОГО КЛАССИФИКАТОРА

АЛЕКСАНДРОВА Марина Юрьевна — преподаватель, аспирант кафедры методов сбора и анализа социологической информации, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
E-MAIL: myaleksandrova@hse.ru
https://orcid.org/0000-0002-7683-7750

Аннотация. Пропущенные данные в социологических исследованиях могут быть связаны с различными причинами, и в данной статье рассматриваются те из них, что появляются в результате незнания, нежелания или затруднения с поиском ответа на отдельные вопросы анкеты у респондента, — частичные неответы (item nonresponse). Остро стоит вопрос о предсказании частичных неответов, решение которого позволило бы сократить вероятность появления пропусков в собираемых данных.

В статье показано, как возникновение частичного неответа можно прогнозировать с помощью современных методов текст-майнинга и машинного обучения на примере данных Европейского социального исследования (European Social Survey) по Великобритании. Для решения поставленной задачи использовался метод наивного байесовского классификатора (Naive Bayes Classifier) — популярный метод предсказания класса зависимой переменной на основе текстовых данных. С опорой на научную литера-

MACHINE LEARNING IN SOCIAL RESEARCH: PREDICTING ITEM NONRESPONSE ERROR USING NAIVE BAYES CLASSIFIER

Marina Yu. ALEKSANDROVA[1] *— Lecturer, Doctoral Student at the Department of Collection and Analysis of Sociological Information*
E-MAIL: myaleksandrova@hse.ru
https://orcid.org/0000-0002-7683-7750

[1] National Research University Higher School of Economics, Moscow, Russia

Abstract. Various reasons may cause missing data in social research. The article highlights the non-response errors caused by ignorance, the lack of desire, or difficulty searching for answers to specific questionnaire questions. Predicting item nonresponse, which would help reduce missing data, poses particular concerns. Based on the data from the European Social Survey (UK respondents) this article shows how text mining and machine learning can predict item nonresponse. The study employs the Naive Bayes Classifier, a popular method to predict the class of dependent variables based on textual data. It relies on scientific literature to show how this method performs. The author provides a database combining full wordings of questions, answers, and instructions, and the ESS survey results in the UK. The paper shows how separate models for predicting the occurrence of item nonresponse were trained using the Naive Bayes technique based on the word frequency and TF — IDF weights (their calculations are also provided). The authors evaluated each model for the frequency of error occurrence. As a result,

DOI: 10.14515/monitoring.2021.1.1756

туру показываем, как работает этот метод. Мы подготовили базу данных, объединяющую полные формулировки вопросов, ответов, инструкций и результатов опросов исследования European Social Survey по Великобритании. Нами показано, как отдельные модели для предсказания появления частичных неответов были обучены с помощью метода наивного байесовского классификатора на основе частот слов и метрики важности слов TF — IDF, процессу расчета которых мы также приводим подробное описание. Каждая из моделей предсказания частичного неответа оценивалась нами с точки зрения частоты возникновения ошибок при получении прогнозов с их помощью. Мы получили списки слов, наличие в вопросах которых статистически чаще сопровождается или не сопровождается частичными неответами. Наши результаты показали, что респонденты менее охотно отвечают на сенситивные вопросы, а некоторые слова, имеющие отношение к процедуре получения ответа на вопрос, статистически чаще пропускаются респондентами.

Ключевые слова: частичный неответ, отказ от ответа, отсутствие ответа, «затрудняюсь ответить», наивный байесовский классификатор, текст-майнинг, европейское социальное исследование, машинное обучение, качество измерения

lists of words causing or not causing item nonresponse errors were obtained. The results show that respondents are less likely to answer sensitive questions; certain words related to the procedure of getting an answer to a question can also lead to high levels of item nonresponse.

Keywords: item nonresponse, refusal to answer, no answer, “Don't know” option, naive Bayes classifier, text-mining, European Social Survey, ESS, machine learning, measurement quality

Введение. Проблема и ее актуальность

Проблема пропусков в данных часто рассматривается в литературе в силу понятных причин, и ее решение сохраняет актуальность для социологических исследований. В качестве пропусков в данных в социологии чаще всего подразумевается отсутствие ответа респондента в ситуации, когда этот ответ *должен* был бы присутствовать — на отдельные вопросы анкеты (частичный неответ, item

DOI: 10.14515/monitoring.2021.1.1756

nonresponse) или на анкету целиком (полный неответ, unit nonresponse). В данной статье мы рассматриваем только частичные неответы.

Исследование причин возникновения отказа от ответа необходимо для того, чтобы понимать, что именно в разработанной анкете может быть связано с отказами от ответа и какими способами возможно снизить предсказываемую долю таких отказов. Мы поставили перед собой цель построить модель прогнозирования связи между появлением частичных неответов и формулировками вопросов. Модель прогнозирования — это представленная в математической форме информация о влиянии независимых переменных на зависимую в данных, которыми располагает исследователь. Эта информация представляет ценность, так как позволяет строить предположения о поведении зависимой переменной в ситуациях, данные о которых отсутствуют. Модели прогнозирования различаются в зависимости от методов, с помощью которых они могут быть построены (для построения прогнозных моделей используются различные виды регрессий, деревья решений, нейронные сети и многие другие методы), поэтому они могут быть достаточно разнообразны. Благодаря построению модели прогнозирования появления частичных неответов в связи с определенными формулировками вопросов можно понять, какие именно слова в этих вопросах способны повышать вероятность возникновения неответа, а какие, наоборот, могут снижать эту вероятность. Под частичным неответом мы понимаем отказ от ответа, затруднение с ответом и отсутствие ответа на вопросы анкет. Для того чтобы прогнозировать связи между появлением частичных неответов и формулировками вопросов, был выбран активно применяемый в последнее время для работы с текстовыми данными метод — наивный байесовский классификатор.

В качестве данных для нашего исследования использовались результаты опроса Европейского социального исследования (European Social Survey, ESS) — известное научное сравнительное межстрановое исследование, проводящееся каждые два года во многих европейских странах. Для анализа мы использовали данные опроса ESS по Великобритании. При подготовке к анализу перед нами встала задача объединения данных, собиравшихся ESS в ходе опросов, — ответов респондентов с самими формулировками вопросов, так как в исходной базе данных ESS содержатся только краткие формулировки, позволяющие получить лишь примерное представление о хранимой информации в соответствующих им переменных, в то время как для прогнозирования связи между появлением частичных неответов и формулировками вопросов нам были нужны точные формулировки вопросов в том виде, в каком они задавались респондентам. Для этого мы воспользовались технологией веб-скрейпинга, с помощью которой автоматически собрали открытые данные с веб-сайтов, содержащие формулировки вопросов и данные с ответами на эти вопросы, и потом объединили их.

Текстовые данные переводились в числовой формат, который необходим для того, чтобы анализ текста стал возможен, с помощью расчета частоты встречаемости слов (сколько раз то или иное слово встречается в одном тексте) и меры важности слов TF — IDF (частота слова — обратная частота документа, «term frequency — inverse document frequency»). Например, слово «она» может часто встречаться в тексте, а «шапочка» — реже, но последняя все равно будет обладать

DOI: 10.14515/monitoring.2021.1.1756

высоким значением меры важности TF — IDF для текста сказки «Красная шапочка» в корпусе текстов детских сказок, потому что лучше описывает содержание этой сказки. Модели предсказания частичных неответов сравнивались с точки зрения их точности — доли ошибок первого и второго рода в получаемых с их помощью прогнозах возникновения неответов. Нами были получены списки слов, наличие которых в вопросах анкеты показывает связь с одним из типов частичного неответа, что позволило сделать выводы о том, какие темы могут статистически чаще встречаться вместе с частичными неответами.

Работа метода наивного байесовского классификатора

Интуитивно ясно, что лучше всего задачу прогнозирования связи между частичными неответами и формулировками вопросов было бы решать с помощью анализа текстов, заложенных в анкете: формулировок вопросов, вариантов ответов и т. д. Под «формулировкой» вопросов мы будем понимать наличие или отсутствие определенных слов в вопросе — для выражения одной и той же мысли можно выбирать разные слова, и именно благодаря выбору слов вопрос может быть понят по-разному, он может вызвать разную реакцию. Это прекрасно иллюстрируется примером того, как по-разному можно сформулировать сенситивный вопрос «Вы убили свою жену?», приведенным со ссылкой на А. Бартона [Barton, 1958] С. Садменом и Н. Брэдберном, авторами книги «Как правильно задавать вопросы» [Садмен, Брэдберн 2002: 65].

Для решения поставленной нами задачи необходимо исследовать связь формулировок вопросов с наличием или отсутствием частичных неответов у этих вопросов. Например, в вопросе А присутствует слово «пожалуйста», и у этого вопроса нет частичных неответов, а в вопросе Б есть слово «возраст», и на этот вопрос некоторые респонденты не дали ответ. Мы можем предположить, что вопросы со словом «пожалуйста» будут статистически реже оставаться без ответа, в то время как вопрос со словом «возраст» — чаще оказываться неотвеченным. Конечно, не все вопросы со словом «пожалуйста» обязательно будут иметь ответ, в то время как вопросы со словом «возраст» — неответ, поэтому мы можем говорить только о какой-то вычисляемой вероятности, показывающей, что вопрос с одним словом может статистически чаще оказываться отвеченным, а другой вопрос с другим словом будет чаще пропускаться респондентом. Проанализировав все слова из вопросов таким образом, мы можем в итоге разделить слова на две группы: те, которые скорее связаны с неответами, и те, которые скорее будут связаны с ответами. Можно предположить, что внутри этих совокупностей также какие-то слова будут с большей или с меньшей вероятностью связаны с неответом (например, это могут оказаться слова, связанные с сенситивными темами [Sakshaug, Yan, Tourangeau, 2010]).

Мы описали примерную логику условной вероятности, которая свойственна статистике Байеса [Айвазян и др., 1989: 70]. Например, мы могли сделать такой вывод: при условии, что в вопросе присутствует слово «пожалуйста», вероятность пропуска этого вопроса респондентом будет низкой. Более того, мы можем получить классификацию. На основе всех проанализированных нами вопросов, по которым имелись данные о наличии частичных неответов, мы получаем два

DOI: 10.14515/monitoring.2021.1.1756

класса: слова, с которыми вопросы анкеты статистически чаще оказываются неотвеченными, и слова, с которыми вопросы анкеты оказываются статистически чаще отвеченными.

Среди байесовских методов классификации наиболее известны наивный байесовский классификатор и знакомая социологам логистическая регрессия. Мы решили использовать в данной работе метод наивного байесовского классификатора, так как он часто используется в задачах классификации текстов. Большинство методов анализа текстовых данных, как упоминается в научной литературе, требуют большого количества данных [Zhang, 2005: 187]. Одним из достоинств наивного байесовского классификатора называют то, что ему не требуются большие обучающие выборки для оценки параметров, важных для классификации [ibid.: 191]. Наивный байесовский классификатор основывается на модели условной вероятности. Условная вероятность предполагает измерение вероятности возникновения какого-то события при условии возникновения другого события. Данный метод исследуется уже достаточно давно — начиная с 1960-х годов, и он начал использоваться в первую очередь для решения задач информационного поиска (information retrieval) [Maron, 1961: 410]. Информационный поиск — это автоматизированный поиск в коллекции текстовых документов тех, которые подходят по интересующей теме [Manning, Raghavan, Schütze, 2008: 114] (эту задачу выполняют, например, современные поисковые системы — Google, Yandex, Yahoo и т. д.). Байесовские модели использовались в известном исследовании Ф. Мостеллера и Д. Уоллеса, посвященном определению авторов американского сборника статей в поддержку утверждения Конституции США под названием «Записки Федералиста» [Mosteller, Wallace, 1964]. Долгое время оставалось неизвестным, кто авторы этих статей, так как все они выходили в газетах под псевдонимом «Публий». Благодаря расчету самых употребляемых слов удалось определить реальных авторов «Записок» [Stine, 2019: 294]. Сегодня наивный байесовский классификатор используется также для решения проблемы распределения текстов по различным категориям (поиск спама, определение научных и публицистических статей, тональности публикаций и т. д.) (см., например, [Rennie, 2003; Vadivukarassi, Puviarasan, Aruna, 2017; Ting, Ip, Tsang, 2011]). Таким образом, наивный байесовский классификатор используется для решения весьма разнообразного круга задач, требующих применения коллекций текстов.

Опыт говорит, что читатель-социолог не всегда готов воспринимать научные статьи, подобные цитируемым нами ниже (описывающие работу интересующего нас алгоритма). Поэтому мы решили уделить внимание «переводу» содержания рассматриваемых статей на «язык» задачи о пропусках в ответах на вопросы социологической анкеты.

В основе наивного байесовского классификатора лежит следующий алгоритм. Допустим, есть некий классифицируемый объект (корпус текстов). Объект или текст может быть отнесен к одному из двух каких-то классов (класс выступает в качестве зависимой переменной) — например, это электронные письма, которые являются или не являются спамом, рецензии на книги — положительные и отрицательные, а также вопросы из анкет социологических исследований, при ответе на которые возникали или не возникали частичные неответы. Для

DOI: 10.14515/monitoring.2021.1.1756

каждого из классов объекта вычисляются функции правдоподобия, на основе которых рассчитываются апостериорные вероятности этих классов [Ting, Ip, Tsang, 2011: 39].

Наивный байесовский классификатор назван наивным, так как в его основе лежит допущение о независимости признаков объектов (признаки объектов — это, иными словами, независимые переменные): предполагается, что наличие одного признака у какого-то класса не влияет на наличие другого признака у этого же класса. Признаки могут зависеть друг от друга или от других признаков, но их вклад в вероятность отнесения объекта к одному из классов остается независимым [Zulfikar et al., 2017: 3]. Например, в случае прогнозирования частичного неответа на основе формулировок вопросов признаками будут выступать слова, из которых состоят анкетные вопросы, а классом — наличие или отсутствие частичного неответа на соответствующий анкетный вопрос с определенными признаками — словами, с помощью которых он был сформулирован. Так, на вопрос А («Укажите, пожалуйста, все марки мороженого, которые вы приобретали в течение последнего месяца хотя бы один раз»), в котором есть слово «мороженое», ответили все респонденты, в то время как на вопрос Б («Скажите, пожалуйста, приходилось ли вам когда-либо пробовать наркотические вещества?»), в котором присутствовало слово «наркотический», респонденты часто отказывались отвечать. При этом наличие признака (слова) «Укажите» в вопросе А или «Скажите» в вопросе Б никак не влияет на то, что в вопросе А присутствовало слово «мороженое», а в вопросе Б — «наркотический». Именно это и является допущением о независимости признаков, в связи с которым рассматриваемый нами метод называется наивным байесовским классификатором.

Исходя из теоремы Байеса, которая лежит в основе наивного байесовского классификатора, апостериорная вероятность того, что наш объект будет отнесен к определенному классу при соответствующем значении признака, рассчитывается следующим образом [Lynch, Bartlett 2019: 55]:

$$P(c_i|d_i) = \frac{P(d_i|c_i)P(c_i)}{P(d_i)},$$

где $P(c_i|d_i)$ — условная вероятность того, что объект i с признаком d принадлежит к классу c;

$P(d_i|c_i)$ — условная вероятность того, что объект i, принадлежащий к классу c, обладает признаком d;

$P(c_i)$ — априорная вероятность класса c;

$P(d_i)$ — априорная вероятность признака d;

i — некий объект, у которого могут быть признаки c и который может быть отнесен к какому-то классу d.

Для нашей задачи прогнозирования появления частичных неответов на основе формулировок вопросов наивный байесовский классификатор будет работать следующим образом. Мы имеем набор данных — формулировок вопросов, в котором содержится один признак — отдельные слова из этих вопросов, и два класса — «частичный неответ есть» («Да») и «частичного неответа нет» («Нет»). Мы предполагаем, что наличие определенных слов в вопросе (признаков), будет вли-

DOI: 10.14515/monitoring.2021.1.1756

ять на то, появится ли частичный неответ в данном вопросе или не появится. Таким образом, мы имеем информацию по каждому вопросу (см. табл. 1)

Таблица 1. ***Набор данных о формулировках вопросов***

Слово из вопроса	Наличие частичного неответа в вопросе с указанным словом
Птица	Да
Доверять	Нет
Доверять	Да
Президент	Нет
Птица	Да
Доверять	Да
Наркотик	Нет
Президент	Да
Наркотик	Нет

Теперь имеющийся набор данных необходимо преобразовать в частотную таблицу, резюмирующую, какое число вопросов вызвало неответы, а какое — не вызвало, для каждого из списка слов (см. табл. 2).

Таблица 2. ***Частотное распределение вопросов по словам и наличию частичного неответа***

Слово из вопроса	Частичный неответ есть	Частичного неответа нет
Птица	2	0
Доверять	2	1
Президент	1	1
Наркотик	0	2
Всего	5	4

Так как данных по всем вопросам в нашем примере немного, то уже на этапе подготовки частотной таблицы можно предположить, что есть слова, которые принадлежат вопросам с бóльшим числом неответов, и слова, которые принадлежат вопросам с меньшим числом неответов.

Теперь могут быть рассчитаны вероятности значений признака (наличие определенного слова в вопросе) при соответствующем классе (частичный неответ есть/нет). Данная вероятность также называется правдоподобием [Lynch, Bartlett, 2019: 55] и рассчитывается следующим образом:

$$P(\text{Есть частичный неответ} \mid \text{Доверять}) =$$

$$= \frac{P(\text{Доверять} \mid \text{Есть частичный неответ}) \times P(\text{Есть частичный неответ})}{P(\text{Доверять})};$$

DOI: 10.14515/monitoring.2021.1.1756

Таким образом, апостериорная вероятность класса «Есть частичный неответ» при значении признака «Доверять», будет равна:

$$P(\text{Доверять} \mid \text{Есть частичный неответ}) = 2/5 = 0{,}4;$$

$$P(\text{Есть частичный неответ}) = 5/9 = 0{,}56;$$

$$P(\text{Доверять}) = 3/9 = 0{,}33.$$

Теперь можно подставить полученные значения для расчета правдоподобия для слова «доверять» быть связанным с частичным неответом в вопросе, в котором это слово встретится:

$$P(\text{Есть частичный неответ} \mid \text{Доверять}) = 0{,}4 \times 0{,}56/0{,}33 = 0{,}68.$$

Следовательно, слово «доверять» в нашем примере окажется связанным с возникновением неответа на вопрос, если оно присутствует в его формулировке.

Таким образом с помощью наивного байесовского классификатора для каждого слова может быть рассчитана вероятность того, что его наличие в тексте будет связывать появление данного слова скорее с одним классом, чем с другим. В качестве подготовительного этапа выступает только преобразование текстовой информации в числовую, которая может происходить различным образом (например, на основе частот встречаемости слов или метрики TF — IDF) [Sharma, Singh, 2016], о чем мы расскажем в следующей части нашей статьи.

Процесс перевода текстовых данных в числовой формат для прогнозирования частичных неответов

Чтобы перевести наши формулировки вопросов в формат, который позволит применить наивный байесовский классификатор, необходимо пройти подготовительный этап. Вопрос в виде текста нам как людям, владеющим языком, на котором он написан, понятен, однако для компьютерного анализа он не подходит — это язык, который в первозданном виде будет непонятен и для компьютера, и для наивного байесовского классификатора. Поэтому нам нужно каким-то образом перевести формулировки вопросов в цифры. Вопрос в том, как это можно сделать. Для решения этого вопроса существуют различные методы «перевода» текстовой информации в числовую. Чаще всего для этого применяются расчет частот встречаемости слов и расчет метрики TF — IDF. TF — IDF расшифровывается как «term frequency — inverse document frequency», дословно перевести можно так: «частота слова — обратная частота документа» — она позволяет находить часто упоминаемые и специфичные в анализируемом тексте слова. Например, в уже упоминавшемся примере со сказкой «Красная шапочка» слово «она» будет часто встречаться, как и во многих других текстах, из-за чего мера важности этого слова будет низкой. Зато даже если слово «шапочка» встретится в этой сказке меньшее число раз, чем слово «она», мы с трудом найдем текст с такой же частотой упоминания «шапочки». Поэтому данное слово будет обладать высоким значением меры важности TF — IDF.

DOI: 10.14515/monitoring.2021.1.1756

Упомянутые метрики мы использовали для оцифровки формулировок вопросов ESS.

Описание расчета частот встречаемости слов

Частоты слов, встречающихся в формулировках вопросов, которые были собраны из анкет ESS, рассчитываются следующим образом. На основе всего анализируемого корпуса — коллекции текстов (текстами в нашем случае выступают отдельные вопросы) создается словарь слов, содержащихся в нем, — набор слов, которые встречались во всех наших вопросах. Далее строится матрица, где в столбцах располагаются слова, встретившиеся в нашем массиве, а в строках — все вопросы, имеющиеся в массиве. На пересечении строки и столбца размещаются цифры — количество раз, которое каждое слово встречалось в каждом вопросе [Baayen 2002: 38]. Пример можно увидеть ниже (см. табл. 3).

Таблица 3. ***Матрица встречаемости слов в массиве текстовых данных***

	Are	you	a	Citizen	of	the	UK	…
Are you a citizen of the UK?	1	1	1	1	1	1	1	…
in your main job are you…	1	1	0	0	0	0	0	…
do you have any friends who have come to live in the UK from another country?	0	1	0	0	0	0	0	…
…	…	…	…	…	…	…	…	…

Описание расчета меры важности TF — IDF слов

Частота встречаемости слова позволяет определить наиболее и наименее часто встречающиеся слова в изучаемых текстах. Наиболее часто встречающиеся слова могут быть достаточно важными для описания текстов, в которых они встретились. Но в расчете частот встречаемости слов кроется определенный риск, который заключается в том, что самыми часто встречающимися могут оказаться наиболее общеупотребимые, распространенные слова, которые на самом деле не позволят понять, в чем отличие одного текста от всех остальных. Преобразование текста в числовой формат с учетом различия тематик может помочь нам определить, существуют ли темы, связанные с частичными неответами. Для поиска одновременно и часто употребляемых, но при этом достаточно специфичных слов, используется метрика важности слова TF — IDF [Robertson, 2004].

TF — IDF — это статистическая мера важности отдельного слова в тексте, который является частью некой коллекции текстов — корпуса текстов [Evans, Aceves, 2016: 41]. Рассчитывается данная мера по следующей формуле, представляющей собой произведение частоты слова и обратной частоты документа [Hirschberg, Manning, 2015: 263]:

DOI: 10.14515/monitoring.2021.1.1756

$$TF — IDF(t,k) = TF(t,k) \times IDF(t,k),$$

где $TF — IDF(t,k)$ — в обозначении метрики подразумевается не минус, а тире;
t — отдельно взятое слово из какого-то текста из корпуса текстов. Это слово может встречаться и в других текстах данного корпуса тоже, но может и оказаться уникальным для какого-то одного текста;
k — отдельно взятый текст, в котором встретилось данное слово t;
$TF(t,d)$ — частота слова (term frequency) — как часто слово t встречалось в тексте d, которая вычисляется по формуле:

$$TF(t,d) = n_{t,d}/(\Sigma_k n_{k,d}),$$

где $n_{t,d}$ — количество упоминаний слова t в тексте d,
$\Sigma_k n_{k,d}$ — сумма всех слов k, которые есть в тексте d.

Делитель в данной формуле вводится, чтобы избежать смещения из-за того, что анализируемые тексты могут сильно различаться своей длиной: в более длинных текстах количество упоминаний какого-то слова может быть больше просто потому, что сам текст длинный, а не потому, что это слово важно для данного текста.

IDF — обратная частота документа (inverse document frequency), с помощью которой оценивается, насколько редко слово встречается во всем корпусе текстов (если слово присутствует во всех текстах, то данная метрика будет равна нулю). Обратная частота документа вычисляется по формуле:

$$IDF(t,d) = \log \frac{n_d}{|\{d_i \in D \mid t \in D_i\}|},$$

где n_d — количество всех текстов k в корпусе, а выражение $|\{d_i \in D \mid t \in D_i\}|$ означает количество только тех текстов d_i из всего корпуса текстов D, в которых обязательно встретилось слово t как минимум один раз [Manning, Raghavan, Schütze, 2008: 3].

Таким образом, TF — IDF использует функцию частоты встречаемости термина в тексте, деленную на логарифмически масштабированную обратную долю текстов, содержащих конкретное слово — общее количество текстов в корпусе, деленные на те, в которых данное слово упоминается хотя бы один раз [Hirschberg, Manning, 2015: 264].

Рассмотрим на примере расчет по данной формуле. Предположим, что у нас есть корпус, состоящий из 100 000 вопросов. Слово «доверять» присутствует в 1000 вопросах. В одном из вопросов, состоящем из 100 слов, слово «доверять» встретилось 5 раз. Частота слова «доверять» в данном вопросе (например, это некий седьмой вопрос из абстрактной анкеты) будет равна 0,05, если посчитать по приведенной выше формуле:

$$TF(\text{доверять, вопрос 7}) = 5/100 = 0{,}05.$$

Обратная частота текста — соотношение количества всех вопросов и количества вопросов, в которых встретилось слово «доверять», будет равна 2, исходя из уже приведенной формулы для ее расчета:

$$IDF(\text{доверять, вопрос 7}) = \log \frac{100\,000}{1\,000} = 2.$$

DOI: 10.14515/monitoring.2021.1.1756

Вес *TF — IDF* для слова «доверять» в нашем корпусе из 100 000 вопросов составит 0,1 в седьмом вопросе как результат произведения уже посчитанных частоты слова и обратной частоты текста:

$$TF — IDF(t,k) = 0{,}05 \times 2 = 0{,}1.$$

Таким образом, с помощью меры важности *TF — IDF* можно присвоить каждому слову в корпусе текстов вес, который:

— тем выше, чем чаще слово *t* встречается в небольшом количестве текстов (таким образом, данное слово помогает лучше понять отличие данных текстов от остальных);
— тем ниже, чем реже слово *t* встречается в отдельно взятом тексте или встречается в большом количество текстов (из-за чего это слово плохо отличает данные тексты от остальных);
— ниже всего у тех слов *t*, которые встретились во всех текстах (самые общие, часто употребляемые и неспецифичные слова).

Мы рассмотрели два способа перевода текстовой информации в числовую. После того как этот переход был совершен, можно приступать к использованию наивного байесовского классификатора на полученных данных для прогнозирования связи между появлением частичных неответов и формулировками вопросов (наличия определенных слов в содержании вопроса).

Особенности подготовки данных для дальнейшего обучения модели предсказания частичных неответов

В нашем исследовании использовались данные Европейского социального исследования (European Social Survey, ESS), собранные в Великобритании на английском языке и включающие не только ответы и неответы респондентов, но и полные формулировки вопросов анкет в том виде, в котором они задавались респондентам. В этих данных содержится информация о трех вариантах частичного неответа: отказ от ответа, «нет ответа» и «затрудняюсь ответить».

Для подготовки к анализу мы объединили данные, собранные во время опроса: ответы респондентов с самими формулировками вопросов, так как в исходной базе данных ESS содержатся только краткие формулировки, позволяющие лишь получить представление о хранимой информации в соответствующих им переменных, в то время как для прогнозирования связи между появлением частичных неответов и формулировками вопросов нам нужны именно формулировки вопросов в том виде, в котором они задавались респонденту. Такие исходные формулировки вопросов можно найти на сайте ESS, однако содержащие их анкеты хранятся в неудобном для автоматической обработки формате pdf, а ручная обработка файлов происходила бы значительно медленнее и не смогла бы застраховать от ошибок при вводе данных. Поэтому формулировки вопросов разных волн ESS мы автоматически собрали с сайта инструмента SQP (Survey Quality Predictor)[1]. Данный инструмент был разработан для прогнозирования качества измерения анкетных вопросов при участии ESRA — организации, занимающейся проведением

[1] SQP 2.1. URL: http://sqp.upf.edu (дата обращения: 12.01.2021).

DOI: 10.14515/monitoring.2021.1.1756

исследования ESS. В настоящий момент SQP используется в ESS при разработке социологических анкет и в процессе ее перевода на другие языки.

Информация по вопросам хранится в SQP в структурированном виде (см. рис. 1): вопросы могут быть отфильтрованы по источнику — исследованию, в котором была использована та или иная формулировка, по языку, на котором сформулированы вопросы, и по стране, в которой проводилось само социологическое исследование. Основную часть окна занимает таблица, содержащая информацию по каждому из вопросов интересующего исследования, при нажатии на который появляется подробная информация о формулировке соответствующего вопроса и вариантов ответа на него.

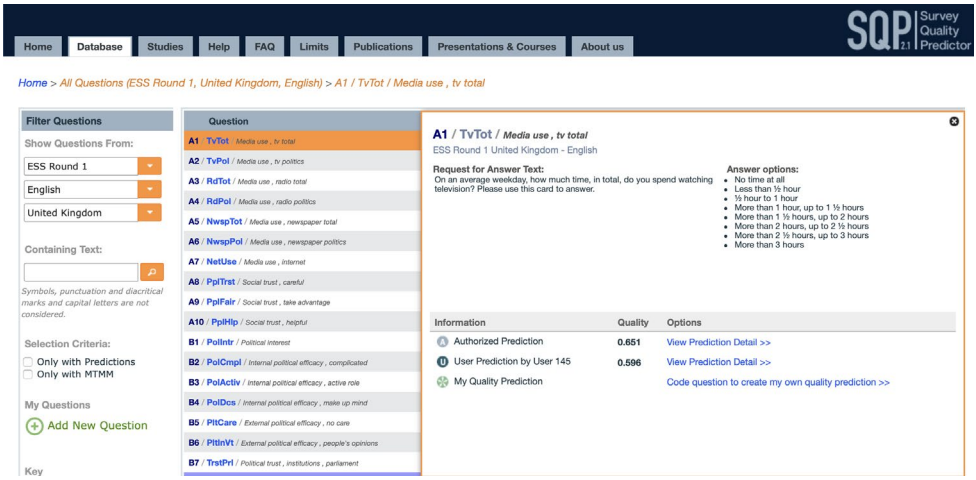

Рис. 1. Вид хранящихся формулировок вопросов в SQP на официальном сайте

После того как формулировки вопросов из исследования ESS были собраны, они объединялись с собранными ответами респондентов. Для данного исследования в ответах респондентов наиболее важны наличие или отсутствие частичного неответа: выбор респондента между вариантами отказом от ответа (refusal), отсутствием ответа (no answer) и затруднением с ответом (don't know).

Тем не менее на этапе объединения данных с ответами респондентов и соответствующих им формулировок из анкеты было выявлено, что на сайте SQP представлены не все вопросы волн исследования ESS. В результате была получена база с 1456 наблюдениями-вопросами.

Сводная информация по подготовленным данным представлена в таблице 4.

Таблица 4. ***Сводная информация о данных***

	Частичный неответ: есть	**Частичный неответ: нет**	**Обучающая подвыборка**	**Тестовая подвыборка**
Затрудняюсь ответить	1274	182	981	484
Отказ от ответа	453	1003	981	484
Отсутствие ответа	422	1034	981	484

DOI: 10.14515/monitoring.2021.1.1756

Обучающая и тестовая подвыборки нам нужны для того, чтобы сначала позволить наивному байесовскому классификатору «проанализировать» наши данные и найти закономерности в формулировках вопросов, которые могут влиять на появление или непоявление частичных неответов, а затем проверить полученные результаты. Приведем пример. Пусть в нашей выборке вопросов есть десять вопросов, в которых встретилось слово «доверять». Между обучающей и тестовой подвыборками эти вопросы распределились случайным образом: ничто не влияло на то, что в обучающую подвыборку попали семь вопросов о доверии, а в тестовую — оставшиеся три вопроса о доверии (влияло только соотношение величин подвыборок, которое мы определили в данном случае — 70: 30, то есть 70 % — обучающая и 30 % — тестовая подвыборка). В результате мы будем получать информацию о том, связано или не связано статистически слово «доверять» с частичным неответом на основе тех семи вопросов, которые попали в обучающую подвыборку. Получив эту информацию, мы сможем проверить ее на трех вопросах — будет ли предсказание о частичных неответах на основе семи вопросов подходить для трех вопросов, которые оказались в тестовой подвыборке.

Обучающая подвыборка представляет собой данные, на основе которых производится поиск связей между независимыми и зависимыми переменными. Тестовая подвыборка — это совокупность данных, на которых проверяются связи, выявленные в обучающей подвыборке [James et al., 2013: 73]. Если тестовая подвыборка подтверждает связи между переменными, которые показала обучающая подвыборка, то это позволяет, во-первых, сравнивать между собой построенные модели, а во-вторых, таким образом осуществляется проверка устойчивости полученных на обучающей подвыборке результатов. Процесс обучения модели проходит следующим образом: на обучающей подвыборке производится поиск связей между переменными. Далее на основе этих связей формируются предположения о поведении зависимой переменной — предсказания. Полученные предсказания проверяются на основе тестовой подвыборки — подтверждают ли данные в ней предсказания о поведении переменных или нет [Kuhn, Johnson, 2013].

Данные между обучающей и тестовой выборкой распределяются случайным образом, то есть мы не можем повлиять на то, в какую подвыборку попадают наблюдения из всей выборки. Это необходимо нам для того, чтобы тестовая и обучающая подвыборки не имели каких-то латентных особенностей, которые влияли бы на получаемые нами результаты. Мы можем проконтролировать, какую долю составят эти две подвыборки. У нас это соотношение получилось следующее: 70 % — обучающая подвыборка и 30 % — тестовая.

Результаты обучения моделей прогнозирования частичных неответов с помощью наивного байесовского классификатора

Для всех моделей рассчитывались матрицы ошибок (confusion matrix) и коэффициент точности предсказаний (accuracy score), которые являются распространенными способами оценки качества полученных прогнозных моделей. Матрица ошибок — двумерная матрица, показывающая распределение правильных и ошибочных предсказаний, сделанных с помощью обученной модели. Данная матрица показывает количество правильных предсказаний наличия признака (закоди-

DOI: 10.14515/monitoring.2021.1.1756

рованный как «1») и отсутствия признака (закодированный как «0»), сделанных моделью с использованием отобранных независимых переменных. Коэффициент точности предсказаний резюмирует и обобщает матрицу ошибок, он представляет собой долю тестовой выборки, предсказания для которой оказались верными. Следовательно, чем выше его значение, тем более точные предсказания могут быть получены с помощью построенной модели.

Таблица 5 показывает коэффициенты точности предсказаний отказов от ответа, затруднений с ответом и отсутствия ответов для моделей, обученных на частотах слов и на TF — IDF. Коэффициенты точности показывают, что для предсказания отказов от ответа и затруднений с ответом точнее оказались модели, основанные на метрике TF — IDF, в то время как отсутствие ответа несколько лучше прогнозируется моделью, построенной на частотах слов.

Таблица 5. ***Качество предсказания частичного неответа***

	Отказ от ответа	**«Затрудняюсь ответить»**	**Отсутствие ответа**
Counts	0,686	0,843	0,764
TF — IDF	0,740	0,878	0,762

Таблицы 6, 7, 8, демонстрируют матрицы ошибок предсказания отказов от ответа, затруднений с ответом и отсутствия ответа для моделей, обученных на частотах слов и на коэффициентах TF — IDF. В целом можно заметить, что хотя доля ошибок первого и второго рода суммарно становится меньше для моделей, рассчитанных на основе TF — IDF, тем не менее улучшение это происходит неравномерно, более того — с увеличением по какому-то одному из типов ошибок. Например, хотя доля ложноположительных предсказаний отказа от ответа уменьшилась (см. табл. 6) с 78 до 31, доля ложноотрицательных предсказаний увеличилась — с 74 до 95. Данное увеличение произошло за счет снижения доли верных предсказаний наличия отказа от ответа на вопрос (с 65 верных предсказаний сократилось до 44).

Таблица 6. ***Матрица ошибок предсказания отказа от ответа***

Counts	**Отсутствие неответа**	**Наличие неответа**
Отсутствие неответа	267	78
Наличие неответа	74	65

TF — IDF	**Отсутствие неответа**	**Наличие неответа**
Отсутствие неответа	314	31
Наличие неответа	95	44

Доля ложноположительных предсказаний отсутствия ответа также уменьшилась (с 57 до 29), доля ложноотрицательных предсказаний увеличилась — с 57 до 86 за счет снижения доли верных предсказаний наличия отказа от ответа на вопрос (с 74 верных предсказаний сократилось до 45).

DOI: 10.14515/monitoring.2021.1.1756

Таблица 7. ***Матрица ошибок предсказания отсутствия ответа***

Counts	Отсутствие неответа	Наличие неответа
Отсутствие неответа	296	57
Наличие неответа	57	74

TF — IDF	Отсутствие неответа	Наличие неответа
Отсутствие неответа	324	29
Наличие неответа	86	45

Предсказание затруднений ответов у респондентов показывает несколько иную картину. Количество верных предсказаний наличия затруднений увеличилось, в то время как отсутствие затруднений с ответом в модели, построенной на TF — IDF, стало предсказываться значительно хуже — только 1 случай был верно угадан моделью (см. табл. 8).

Таблица 8. ***Матрица ошибок предсказания «Затрудняюсь ответить»***

Counts	Отсутствие неответа	Наличие неответа
Отсутствие неответа	18	40
Наличие неответа	36	390

TF — IDF	Отсутствие неответа	Наличие неответа
Отсутствие неответа	1	57
Наличие неответа	2	424

Исходя из примеров слов, представленных в таблице 9, можно заметить, что с более частым появлением частичного неответа демонстрируют связь слова — маркеры сенситивных тематик, так как респондентам может быть некомфортно, неловко отвечать на какие-то вопросы анкеты: вопросы, затрагивающие темы развода (divorced), работы и безработицы (unemployed — безработный, money — деньги, income — доход), этнополитические вопросы (England, Ireland, Scotland). Это подтверждается результатами исследований других авторов, которые обращают в целом внимание на то, что сенситивные вопросы могут быть связаны с нежеланием отвечать или желанием отвечать неискренне на такие вопросы (см., например, [Chou, Imai, Rosenfeld, 2020; O'Brien et al., 2006; Sakshaug, Yan, Tourangeau, 2010; Ипатова, Рогозин, 2019 и т. д.]). Более того, отдельные исследования подтверждают получаемые нами результаты. Так, К. Митчелл подтверждает, что именно пропуски, связанные с частичными неответами, являются основной причиной смещений в данных [Mitchell, 2010: 899]. Вопросы, связанные с доходом, нередко ассоциируются со значительным объемом пропущенных данных [Yan, Curtin, Jans, 2010: 152]. Можно предположить, что степень сенситивности этнополитической тематики для британцев тоже высока в связи с историей конфликтов между Англией и Ирландией (знаменитый Ольстерский конфликт, война за независимость Ирландии, теракты Ирландской республиканской армии), а также популярности идей о выходе Шотландии из состава Великобритании (на референдуме о независимости Шотландии 2014 г. мнения

DOI: 10.14515/monitoring.2021.1.1756

поделились почти поровну — только 55,3 % опрошенных пожелали остаться в составе Великобритании [2]).

Кроме того, с более частым появлением частичных неответов показывают связь слова, поясняющие, как ответить на поставленный вопрос: оценить что-либо «в целом» (overall), использовать вспомогательную карточку (card) — визуальные материалы, помогающие понять и дать ответ на анкетный вопрос (например, карточка с перечислением возможных вариантов ответа на вопрос, какие-то изображения, иллюстрации, с опорой на которые респонденту необходимо ответить).

Таблица 9. ***Примеры слов, которые связаны с частичными неответами***

Тип частичного неответа	Слова в вопросах, на которые статистически чаще оставляют без ответа	Слова в вопросах, на которые отвечают статистически чаще
Отказ от ответа	accommodation accomplishment actively automobile behaviour charitable cheerful childcare church citizenship consumer friendly grandmother hobby housework	divorced examinations economy unemployed responsibility politicians religion England Ireland Scotland overall democracy mother father card
Затруднение с ответом	aid apprenticeship church farmers hobby languages medication participated relationship sports students unable vacation street transport	woman better unemployed difficult democracy police trust old housework area good family supervising health card

[2] Референдум: Шотландия решила остаться в составе Британии // BBC News. Русская служба. 19.09.2014. URL: https://www.bbc.com/russian/uk/2014/09/140919_scotland_wrap_up.shtml (дата обращения: 21.01.2021).

DOI: 10.14515/monitoring.2021.1.1756

Тип частичного неответа	Слова в вопросах, на которые статистически чаще оставляют без ответа	Слова в вопросах, на которые отвечают статистически чаще
Отсутствие ответа	able accept accommodation accomplishment achievements adult bank buy calm careful criticise discussed discussions expect expenses	answer competent pity overall employed unemployed child retire money hours husband education mother father income

Можно также заметить, что примерно одни и те же слова в вопросах показывают связь с частичными неответами — предсказывается ли отказ от ответа, затруднение с ответом или отсутствие ответа, а также модели, обученные на основе частоты встречаемости слов и TF — IDF дают примерно одинаковые результаты и схожим образом разделяют слова, которые демонстрируют связь с частичным неответом (вопросы с подобными словами статистически чаще оказываются неотвеченными или, наоборот, чаще даются ответы на вопросы с определенными словами).

Выводы и рекомендации

Результаты анализа данных показывают, что респонденты менее охотно отвечают на вопросы, связанные сенситивными темами, что подтверждает выводы других исследований (см., например, [Chou, Imai, Rosenfeld 2020; O'Brien et al., 2006; Sakshaug, Yan, Tourangeau, 2010; Ипатова, Рогозин, 2019] и т. д.). Это может показаться очевидным результатом, но его польза состоит в том, что его получение подтверждает качество полученных предсказательных моделей: они оказались способны делать то, что делает социолог при разработке анкеты, а именно видеть темы, которые в перспективе могут вызывать у респондентов нежелание отвечать на связанные с ними вопросы.

Более того, обучение моделей предсказания частичных неответов показывает, что некоторые слова, имеющие отношение к самой процедуре ответа на вопрос (использование карточек, необходимость что-то оценить «в целом»), могут быть статистически связаны с более частым возникновением частичных неответов. Данные слова могут маркировать сложные вопросы — требующие использования вспомогательных средств для помощи респонденту в поиске ответа на вопрос (использование карточек) или те, которые требуют размышлений, рефлексии, обращения к памяти респондента.

DOI: 10.14515/monitoring.2021.1.1756

Тем не менее у данного исследования существуют и некоторые ограничения. Во-первых, построенные модели не учитывают помимо формулировок вопросов иные параметры, которые, безусловно, тоже влияют на возникновение частичных неответов. Это и способ сбора данных [Hansen, Hurwitz, 1946], и наличие интервьюера [Groves, 1989], и характеристики самих респондентов — пол [Francis, Busch, 1975], возраст и уровень образования [Groves, 1979; Schuman, Presser, 1980; Herzog, Dielman, 1985], наличие или отсутствие работы [Bell, 1984] и даже национальность, расовая принадлежность [Sicinski, 1970]. Один и тот же вопрос может быть задан разными интервьюерами совершенно разным респондентам: мужчине или женщине, молодому или пожилому, с высшим образованием или без него, британцу или испанцу. И этот вопрос может показывать разную долю частичных неответов при наличии этих характеристик. При этом данный вопрос не меняется, в анкете он был сформулирован совершенно одинаково — даже в межстрановых опросах существуют процедуры, направленные на достижение эквивалентности переводов; и все равно в условной Великобритании неответы будут, например, практически отсутствовать для 80 % анкет, в то время как в условной Испании частичных неответов не будет только в 15 %—20 % анкет. Наше исследование не исключает данные исследовательские направления и не противоречит им, но несет в себе цель попытаться посмотреть на проблему работы с частичными неответами с еще одной стороны. Мы также не исключаем возможность и ценность проведения сравнительного исследования, которое учитывало бы влияние характеристик респондентов, имеющейся информации о проведении опроса и формулировок самих вопросов и соотносило бы их влияние между собой.

Во-вторых, анализ слов из анкетных вопросов по отдельности создает риск потери контекста вопроса. Так, слово «возраст» может встретиться в разных по степени своей сенситивности вопросах: вопрос «В каком возрасте вы пошли в школу?» и вопрос «В каком возрасте у вас был первый секс?» будут иметь разный уровень отвечаемости. Проведенный нами анализ позволяет получить усредненное значение отвечаемости между этими вопросами. Как это ограничение может быть снято? Методы автоматической обработки текстовых данных позволяют разбивать текст не только на отдельные слова (униграммы), но и на словосочетания, состоящие из двух (биграммы) и трех слов (триграммы) [Bird, Klein, Loper, 2009: 141]. Анализ словосочетаний, а не отдельных слов, может помочь уйти от проблемы «средней температуры» отвечаемости при анализе только отдельных слов («униграмм»).

Для повышения качества результатов также предполагается выполнение следующей работы. Во-первых, увеличение выборки за счет добавления вопросов, удаленных в процессе веб-скрейпинга (автоматизированный сбор данных со страниц веб-сайтов [Mitchell, 2018]) анкетных вопросов данного исследования, позволит разнообразить данные, увеличив разнообразие формулировок вопросов и тем, которые они затрагивают. Во-вторых, увеличение выборки за счет добавления всех респондентов, проходивших опросы ESS на английском языке, позволит нам проверить наличие влияния межнациональных особенностей на частичные неответы, а также проверить, насколько формулировка вопроса сильнее или слабее может влиять на возникновение частичных неответов, чем межнациональные различия.

DOI: 10.14515/monitoring.2021.1.1756

Более того, возможно провести сравнение работы предсказательных моделей — наивного байесовского классификатора с другими методами машинного обучения для определения лучшего метода для прогнозирования частичного неответа, взяв в качестве основания сравнения точность предсказания. Возможны дальнейшие кластеризация и классификация полученных слов, которые могут показывать связь с частичным неответом, с помощью лучше всего показавшей себя предсказательной модели для формирования рекомендаций по составлению вопросов в анкетах социологических исследованиях.

Список литературы (References)

Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. М.: Финансы и статистика, 1989.
Aivazyan S. A., Bukhshtaber V. M., Enyukov I. S., Meshalkin L. D. (1989) Applied Statistics: Classification and Dimensionality Reduction. Moscow: Finance and Statistics. (In Russ.)

Ипатова А. А., Рогозин Д. М. Способы преодоления коммуникативных затруднений в стандартизированном телефонном интервью // Вестник Российского университета дружбы народов. Серия: Социология. 2019. Т. 19. № 1. С. 141—166. https://www.doi.org/10.22363/2313-2272-2019-19-1-144-166.
Ipatova A. A., Rogozin D. M. (2019) Techniques for Communication Repair in the Standardized Telephone Interview. *RUDN Journal of Sociology*. Vol. 19. No. 1. P. 141—166. (In Russ.)

Садмен С., Брэдберн Н. Как правильно задавать вопросы: введение в проектирование массовых обследований. М.: Институт Фонда «Общественное мнение», 2002.
Sudman S., Bradburn N. (2002) Asking Questions: A Practical Guide to Questionnaire Design. Moscow: Institute of the Public Opinion Foundation. (In Russ.)

Baayen R. H. (2002) Word Frequency Distributions. Dordrecht: Springer. https://www.doi.org/10.1007/978-94-010-0844-0.

Barton A. J. (1958) Asking the Embarrassing Question. *Public Opinion Quarterly*. Vol. 22. No. 1. P. 67—68. https://www.doi.org/10.1086/266761.

Bell R. (1984) Item Nonresponse in Telephone Surveys: An Analysis of Who Fails to Report Income. *Social Science Quarterly*. Vol. 65. No. 1. P. 207—215.

Bird S., Klein E., Loper E. (2009) Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit. Beijing Köln O'Reilly June.

Chou W., Imai K., Rosenfeld B. (2020) Sensitive Survey Questions With Auxiliary Information. Sociological Methods & Research. Vol. 49. No. 2. P. 418—454. https://www.doi.org/10.1177/0049124117729711.

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

DOI: 10.14515/monitoring.2021.1.1756

Francis J. D., Busch L. (1975) What We Now Know About “I Don’t Knows”. *Public Opinion Quarterly*. Vol. 39. No. 2. P. 207—218.

Groves R. M. (1979) Actors and Questions in Telephone and Personal Interview Surveys. *Public Opinion Quarterly*. Vol. 43. No. 2. P. 190—205.

Groves R. M. (1989) Survey Costs and Survey Errors. New York, NY: Wiley.

Hansen M. H., Hurwitz W. N. (1946) The Problem of Nonresponse in Sample Surveys. *Journal of the American Statistical Association*. No. 41. P. 517—529.

Herzog A., Dielman L. (1985) Age Differences in Response Accuracy for Factual Survey Questions. *Journal of Gerontology*. Vol. 40. No. 3. P. 350—357.

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

James G., Witten D., Hastie T., Tibshirani R. (2013) An Introduction to Statistical Learning: With Applications In R. New York, NY: Spinger. https://www.doi.org/10.1007/978-1-4614-7138-7.

Kuhn M., Johnson K. (2013) Applied Predictive Modeling. New York, NY: Springer. https://www.doi.org/10.1007/978-1-4614-6849-3.

Lynch S. M., Bartlett B. (2019) Bayesian Statistics in Sociology: Past, Present, and Future. *Annual Review of Sociology*. No. 45. P. 47—68. https://www.doi.org/annurev-soc-073018-022457.

Manning C. D., Raghavan P., Schütze H. (2008) Introduction to Information Retrieval. New York, NY: Cambridge University Press.

Maron M. E. (1961) Automatic Indexing: An Experimental Inquiry. *Journal of the ACM*. Vol. 8. No. 3. P. 404—417.

Mitchell C. (2010) Are Divorce Studies Trustworthy? The Effects of Survey Nonresponse and Response Errors. *Journal of Marriage and Family*. Vol. 72. No. 4. P. 893—905. https://doi.org/10.1111/j.1741-3737.2010.00737.x.

Mitchell R. (2018) Web Scraping With Python: Collecting More Data From the Modern Web. Sebastopol, CA: O’Reilly Media.

Mosteller F., Wallace D. L. (1963) Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers. *Journal of the American Statistical Association*. Vol. 58. No. 302. P. 275—309. https://www.doi.org/10.2307/2283270.

O’Brien E. M., Black M. C., Carley-Baxter L. R., Simon Th. R. (2006) Sensitive Topics, Survey Nonresponse, and Considerations for Interviewer Training. *American Journal of Preventive Medicine*. Vol. 31. No. 5. P. 419—426. https://www.doi.org/10.1016/j.amepre.2006.07.010.

DOI: 10.14515/monitoring.2021.1.1756

Rennie J. D., Shih L., Teevan J., Karger D. R. (2003) Tackling the Poor Assumptions of Naive Bayes Text Classifiers. Proceedings of the 20th International Conference on Machine Learning (ICML-03). P. 616—623.

Robertson S. (2004) Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. *Journal of Documentation*. https://www.doi.org/10.1108/00220410410560582.

Sakshaug J. W., Yan T., Tourangeau R. (2010) Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-Mode Survey of Sensitive and Non-Sensitive Items. *Public Opinion Quarterly*. Vol. 74. No. 5. P. 907—933. https://www.doi.org/10.1093/poq/nfq057.

Sharma N., Singh M. (2016) Modifying Naive Bayes Classifier for Multinomial Text Classification. *2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*. IEEE. P. 1—7.

Schuman H., Presser S. (1980) Public Opinion and Public Ignorance: The Fine Line Between Attitudes and Nonattitudes. *American Journal of Sociology*. Vol. 85. No. 5. P. 1214—1225.

Sicinski A. (1970) "Don't Know" Answers in Cross-National Surveys. *Public Opinion Quarterly*. Vol. 34. No. 1. P. 126—129.

Stine R. A. (2019) Sentiment Analysis. *Annual Review of Statistics and Its Application*. No. 6. P. 287—308. https://www.doi.org/10.1146/annurev-statistics-030718-105242.

Ting S. L., Ip W. H., Tsang A. H. C. (2011) Is Naive Bayes a Good Classifier for Document Classification. *International Journal of Software Engineering and Its Applications*. Vol. 5. No. 3. P. 37—46.

Vadivukarassi M., Puviarasan N., Aruna P. (2017) Sentimental Analysis of Tweets Using Naive Bayes Algorithm. *World Applied Sciences Journal*. Vol. 35. No. 1. P. 54—59. https://www.doi.org/10.5829/idosi.wasj.2017.54.59.

Yan T., Curtin R., Jans M. (2010) Trends in Income Nonresponse Over Two Decades. *Journal of Official Statistics*. Vol. 26. No. 1. P. 145—164.

Zhang H. (2005) Exploring Conditions for the Optimality of Naive Bayes. *International Journal of Pattern Recognition and Artificial Intelligence*. Vol. 19. No. 2. P. 183—198. https://www.doi.org/10.1142/S0218001405003983.

Zulfikar W. B., Irfan M., Alam C. N., Indra M. (2017) The Comparation of Text Mining With Naive Bayes Classifier, Nearest Neighbor, and Decision Tree to Detect Indonesian Swear Words on Twitter. *2017 5th International Conference on Cyber and IT Service Management (CITSM)*. IEEE. P. 1—5. https://www.doi.org/10.1109/CITSM.2017.8089231.