Проблемы прогнозирования ошибок измерения в опросах с помощью методов текст-майнинга

Александрова Марина Юрьевна

Аспирант, преподаватель

НИУ ВШЭ

Содержание

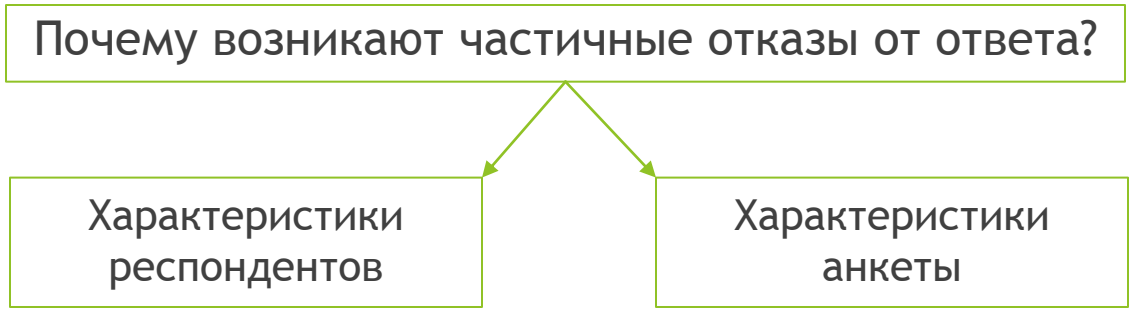
- Качество измерения и возможности прогнозирования частичного отказа от ответа
- Процесс подготовки данных и методология исследования
- Основные результаты и выводы

Качество измерения

- Отсутствие однозначной терминологии.
- Многообразие свойств измерения – необходимость проверки множества свойств измерения.
- Разнообразие видов ошибок измерения и причин их возникновения -> необходимо ввести ограничение, начав с какого-то одного типа ошибки.
- В данном исследовании совершена попытка предсказания частичного отказа от ответа на примере данных European Social Survey с использованием Наивного байесовского классификатора.

Отказ от ответа: возможности для построения предсказательной модели

- Отказ от ответа: полный и частичный (незнание, отсутствие выраженного мнения, нежелание отвечать на вопрос)



- Построение предсказательных моделей затруднено: высокий уровень образования снижает число отказов от ответа (Гровс, 1979) <-> высокий уровень образование увеличивает число отказов от ответа (Шуман и Прессер, 1980) <-> связи между уровнем образования и отказом от ответа нет вообще (Мессмер, Сеймур, 1982)

- Построение предсказательных моделей возможно: сложные предложения (Scherpenzeel, Saris, 1997), сложные слова (Свафорд, Косолапова, 1999), определенные темы (Bradburn, 1978)

Текстовые характеристики вопросов

Подготовка данных



- Данные: European Social Survey (только Великобритания, все волны)
- Частичный неответ в трех форматах: отказ от ответа, затрудняюсь ответить, отсутствие ответа
- (1) или (0) для каждого вопроса

- Скрейпинг формулировок вопросов и ответов
- Данные: формулировки вопросов и ответов, соответствующие переменным

Сводная информация о подготовленных данных

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

Метод для обучения модели

- Метод обучения - мультиномиальный наивный байесовский классификатор на основе частот слов и TF-IDF
- И дальнейшее сравнение моделей, обученных на частотах слов и TF-IDF

Почему?

Потому что позволяет получать качественные результаты и на небольших выборках и устойчив к переобучению

Частота встречаемости слов

словарь всех слов, которые содержатся в корпусе документов

все уникальные слова, встретившиеся в корпусе

все предложения, имеющиеся в корпусе

Далее строится матрица:

	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	1	1	1	…
…	…	…	…	…	…	…	…	…

количество раз, которое каждое слово встречалось в каждом предложении

TF-IDF [term frequency — inverse document frequency]

Статистическая мера важности отдельного слова в тексте, который является частью коллекции документов - корпуса

$$\mathrm{tf-idf}(t,d) = \mathrm{tf}(t,d) \times \mathrm{idf}(t)$$

Присваивает каждому слову в корпусе документов вес, который:

- Тем выше, чем чаще слово t встречается в небольшом количестве документов (это слово помогает лучше понять отличие данного документа от других).
- Тем ниже, чем реже слово t встречается в документе или встречается в большом количество документов (это слово плохо отличает данные документы от остальных)
- Самый низкий вес - у тех слов t, которые встретились во всех документах (самые общие, часто употребляемые слова).

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

Матрицы ошибок для построенных моделей

Предсказание отказа от ответа

Counts	0	1
0	267	78
1	74	65

TF-IDF	0	1
0	314	31
1	95	44

Предсказание «Затрудняюсь ответить»

Counts	0	1
0	18	40
1	36	390

TF-IDF	0	1
0	1	57
1	2	424

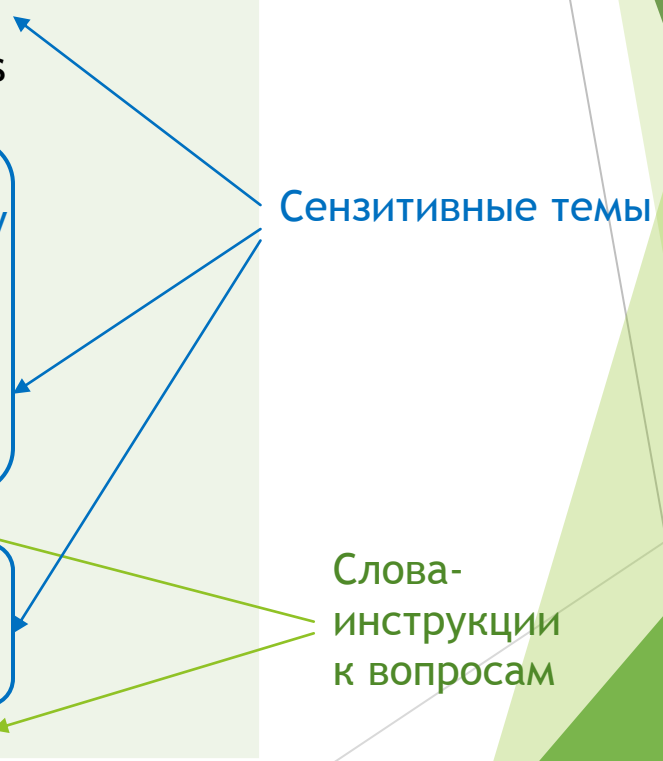
Предсказание отсутствия ответа

Counts	0	1
0	296	57
1	57	74

TF-IDF	0	1
0	324	29
1	86	45

Примеры слов, которые ведут или не ведут к отказу от ответа

Не ведут к отказу от ответа	Ведут к отказу от ответа
accommodation	divorced
accomplishment	examinations
actively	economy
automobile	unemployed
behaviour	responsibility
charitable	politicians
cheerful	religion
childcare	England
church	Ireland
citizenship	Scotland
consumer	overall
friendly	democracy
grandmother	mother
hobby	father
housework	card



Примеры слов, которые ведут или не ведут к затруднению респондентов с ответом

Не ведут к затруднению с ответом	Ведут к затруднению с ответом
aid	woman
apprenticeship	better
church	unemployed
farmers	difficult
hobby	democracy
languages	police
medication	trust
participated	old
relationship	housework
sports	area
students	good
unable	family
vacation	supervising
street	health
transport	card

Сензитивные темы

Слова-инструкции к вопросам

Примеры слов, которые ведут или не ведут к отсутствию ответа

Не ведут к отсутствию ответа	Ведут к отсутствию ответа
able	answer
accept	competent
accommodation	pity
accomplishment	overall
achievements	employed
adult	unemployed
bank	child
buy	retire
calm	money
careful	hours
criticise	husband
discussed	education
discussions	mother
expect	father
expenses	income

Слова-инструкции к вопросам

Сензитивные темы

Выводы и планы по дальнейшим исследованиям

- Подтвердилось предположение, что респонденты менее охотно отвечают на вопросы, связанные сензитивными темами
- Некоторые слова, относящиеся к инструкции к вопросам, вероятно, могут приводить к росту отказа от ответа (использование карточек, необходимость что-то оценить «в целом»).
- Дальнейшая работа с результатами:
  - Увеличить выборку за счет добавления всех англоговорящих респондентов + удаленных в результате скрейпинга некоторых вопросов
  - Сравнить работу наивного байесовского классификатора с другими методами машинного обучения
  - Кластеризация или классификация полученных слов

Спасибо за внимание!

myaleksandrova@hse.ru