УДК 316.7
DOI: 10.19181/snsp.2024.12.1.3
EDN: LOUOJW

Научная статья

ТЕХНОЛОГИИ TEXT MINING В СОЦИОЛОГИЧЕСКОМ АНАЛИЗЕ (НА ПРИМЕРЕ ИЗУЧЕНИЯ ПРЕДСТАВЛЕНИЙ СТУДЕНТОВ О МИССИИ СОВРЕМЕННОГО ВУЗА)

Антонина Николаевна Пинчук [1]
Светлана Геннадьевна Карепова [2]
Дмитрий Андреевич Тихомиров [3]
[1, 3] РЭУ имени Г. В. Плеханова,
[2] Институт социологии ФНИСЦ РАН,
Москва, Россия,
[1] *antonina.pinchuk27@bk.ru*,
ORCID 0000-0001-7842-7141
[2] *Svetlran@mail.ru*,
ORCID 0000-0002-0472-0924
[3] *dat1983@yandex.ru*,
ORCID 0000-0002-1872-6788

Для цитирования: *Пинчук А. Н., Карепова С. Г., Тихомиров Д. А.* Технологии Text Mining в социологическом анализе (на примере изучения представлений студентов о миссии современного вуза) // Социологическая наука и социальная практика. 2024. Т. 12, № 1. С. 62–79. DOI 10.19181/snsp.2024.12.1.3. EDN LOUOJW.

Аннотация. В статье рассматриваются возможности применения методов Text Mining в практике анализа открытых вопросов анкеты. В работе представлен пример исследования униграмм и биграмм, а также поиска латентных топиков с помощью тематического моделирования. Эмпирическими материалами послужили данные проведённого в 2022 году анкетного опроса, в котором приняло участие 929 студентов одного московского экономического вуза. В открытом вопросе анкеты предлагалось определить миссию университета, что дало возможность представить в плоскости субъективной интерпретации предназначение высшей школы в современных условиях. Частотный анализ униграмм, дополненный качественным исследованием высказываний респондентов, позволил составить словарь студенческого дискурса о миссии вуза. Артикулирование биграмм осуществлялось на основе нескольких статистических метрик, с опорой на которые были проранжированы словосочетания и выделен ключевой набор концептов. Выявлено, что в восприятии студентов первоочередные задачи вуза прежде всего направлены на трансляцию профессиональных знаний и навыков, в широком смысле – подготовку квалифицированных специалистов. Социальные функции университета, ориентированные на удовлетворение потребностей общества и государства, в концепту-

альных интерпретациях опрошенных студентов выражены слабее. На следующем этапе исследования была выдвинута задача анализа латентных топиков с помощью тематического моделирования. Особенностью тематического моделирования является то, что объединённые в один топик слова отражают идентифицированное программой распределение слов, но не в буквальном смысле понятную для человека тему. Учитывая специфику применяемого метода, авторы продемонстрировали результаты поискового анализа в практике обработки открытого вопроса. Как оказалось, ключевые слова, сосредоточенные в ядре основных тем, в основном связаны с обеспечением потребностей самих обучающихся, оставляя на периферии вербализируемых определений понимание значимости вуза как платформы для инноваций, научных разработок, предпринимательских и иных инициатив во благо общества и страны. Результаты представленного исследования могут быть полезны для переосмысления исследовательского инструментария социологов в условиях активного развития цифровых технологий, что требует апробации новых методов, понимания их реальных возможностей и ограничений в решении задач социологического исследования.

Ключевые слова: Text Mining, открытый вопрос, препроцессинг, униграммы, биграммы, тематическое моделирование

Благодарности: исследование выполнено за счёт гранта Российского научного фонда № 24-28-00549 «Культурная маргинальность российских студентов: развитие человеческого потенциала новых поколений как проблема и ресурс развития патриотизма в основных положениях и мерах по реализации государственной молодёжной политики» (руководитель: кандидат социологических наук Д. А. Тихомиров).

Введение

В последнее время особое внимание практикующих социологов-исследователей привлекают технологии интеллектуального анализа текстов (Text Mining), расширяющие и обновляющие классический научно-методологический арсенал новыми методами обработки и анализа социологических данных. Text Mining представляет собой алгоритмический процесс извлечения «не известных ранее знаний из текста, а также выявления основных понятий и взаимосвязей между ними» [1, с. 116]. Интеллектуальный анализ текста как отрасль научного знания развивается на основе междисциплинарных разработок в области информационного поиска, машинного обучения, статистики, вычислительной лингвистики, дата-майнинга [2]. Text Mining – сравнительно молодое научное направление, формированию которого послужило использование технологий искусственного интеллекта для обработки естественного языка. Становление Text Mining как отдельной области исследований отмечается в конце XX в., когда стали разрабатываться характерные для данной отрасли технологии и алгоритмы [3]. Сегодня методы Text Mining получили широкое распространение и используются в различных программных

и информационных системах «и как отдельные приложения, библиотечные модули, и в составе инструментария интеллектуального анализа данных, систем бизнес-аналитики, корпоративного управления и т. д.» [4, с. 478].

Технологии интеллектуального анализа текста могут применяться в социологии для исследования социокультурных явлений и процессов, что неизбежно сопровождается эпистемологическими вопросами и дискуссиями [5]. С одной стороны, к методологической рефлексии учёных подталкивает использование неструктурированных текстовых данных, которые позволяют отразить мнение пользователей сети Интернет без прямых опросов [6; 7]. С другой стороны, интерес вызывает применение методов Text Mining в социальных науках для анализа первичных данных, полученных в результате опросов и интервью [8].

Особый фон дискуссиям об использовании методов интеллектуального анализа текстов в социологических исследованиях придаёт обсуждение технических нюансов, требующих от исследователей социально-гуманитарного профиля продвинутых знаний и навыков в области программирования, машинного обучения, математической статистики и т. д. Учитывая многопрофильность и сложность такой социологической практики, следует отметить успешный опыт использования методов Text Mining в социологических исследованиях для обработки и анализа русскоязычных текстов. К примеру, в статье В. И. Дудиной и Д. И. Юдиной решаются задачи тематического моделирования и извлечения мнений из сети Интернет [9], О. Ю. Кольцова и К. А. Маслинский выявляют тематическую структуру блогосферы и апробируют кластерный анализ текстов, латентно-семантический анализ и тематическое моделирование [10], М. А. Кашина и С. Ткач, исследуя тенденцию развития социологии ценностей, осуществляют тематическую кластеризацию текстов [11].

Стоит добавить, что помимо классических примеров классификации и кластеризации текстов, учёные также выделяют автоматическое аннотирование, извлечение ключевых понятий, навигацию по тексту, анализ трендов, поиск ассоциаций [12], что расширяет границы социологического анализа текстовых данных. В то же время освоение новых практик анализа качественных данных подталкивает к переосмыслению традиционных подходов к решению насущных проблем исследовательской рутины социологов. Одной из таких проблем является работа с открытыми анкетными вопросами. Так, в нашем случае в сжатые сроки нужно было обработать и проанализировать результаты опроса студентов, в рамках которого изучалось восприятие современной российской молодёжью общественно значимых функций высшей школы. Студентам было предложено в открытой форме высказать своё мнение о миссии вуза в со-

временных условиях. Прочтение и ручная обработка с последующей кодировкой мнений около тысячи человек является весьма трудозатратой задачей. Чтобы упростить работу с корпусом текстов, было принято решение использовать технологии интеллектуального анализа текста, что в плоскости методологической проблематики выдвинуло новые исследовательские задачи. Во-первых, следовало апробировать современные технологии обработки и анализа естественного языка на корпусе документов, полученном из открытого вопроса анкеты, и продемонстрировать существующие возможности и ограничения. Во-вторых, выдвигалась задача – показать латентные переменные, отражающие общие сюжетные линии для данного корпуса текстов, иными словами, выявить тематическую структуру, что недоступно выполнить при ручном анализе.

В социальных науках технологии интеллектуального анализа текстов пока не стали широко распространённым инструментом исследований, поэтому так важно привлечь внимание к новым исследовательским технологиям для изучения социальных явлений и процессов, показать их преимущества и ограничения.

Эмпирическая база и препроцессинг данных

В качестве эмпирической базы послужили материалы анкетного опроса, проведенного в 2022 году в одном из московских вузов. Всего в исследовании приняли участие 929 студентов. Выборка неслучайная, сформирована методом «снежного кома». В задачи исследования не входило обеспечение репрезентативной выборки, поэтому полученные результаты распространяются только на выборочную совокупность.

В рамках данного исследования был сформулирован открытый вопрос, в котором необходимо было определить миссию современного университета: «Университет имеет свою миссию. На Ваш взгляд, в чём состоит миссия университета в современных условиях?». Вопрос предполагал развернутый ответ, в котором студенты в сжатой форме могли выразить свои идеи об основном предназначении вуза в меняющейся социальной реальности.

Для того чтобы решить исследовательские задачи, на первом шаге была проведена предварительная обработка данных (препроцессинг), в рамках которой текстовый корпус был преобразован в доступную для дальнейшего анализа и моделирования форму [13; 14]. В рамках препроцессинга были выполнены следующие виды работ: разбиение текста на токены (языковые единицы для анализа), приведение слов к единому регистру, удаление нетекстовых знаков, лемматизация (приведение словоформы к нормальной форме), удаление стоп-слов (предлоги, междометия, союзы,

служебные слова, частицы) [15]. В некоторых случаях были удалены слова, которые состоят менее, чем из трёх букв.

Для выполнения необходимых процедур использовалось несколько библиотек, в частности, из библиотеки NLTK был импортирован список стоп-слов. Морфологический анализатор для лемматизации был получен из библиотеки pymorphy2, а удаление ненужных символов проводилось с помощью регулярных выражений, что потребовало применения специального модуля re.

Дополнительные работы с корпусом документов потребовались перед тематическим моделированием, в котором следовало учитывать объём полученных текстов. В рамках анализа ответов респондентов на открытый вопрос последнее требование выступает существенным ограничением, т. к. односложные и немногословные ответы нередко составляют основной массив опросных данных. В нашем случае в результате фильтрации текстов для тематического моделирования были исключены комментарии, объём которых составлял менее 50 слов. Таким образом, в дальнейшем использовался сокращённый корпус текстов, поэтому с содержательной точки зрения нужно учитывать возможную потерю информации. Данные обрабатывались с помощью языка программирования Python.

Результаты интеллектуального анализа текстов

Одной из задач количественного анализа текстов является выявление наиболее часто встречающихся слов. Для социолога, который стремится систематизировать ответы респондентов на открытые вопросы и узнать распространённость тех или иных категорий, частотный анализ также представляет актуальную задачу. Однако специфика обработки прямых и многословных высказываний респондентов проявляется ещё на этапе кодирования. Как известно, обрабатывая и анализируя открытые вопросы анкеты, кодировщики могут быть ориентированы на различение значений и различение смыслов. Последнее небезосновательно вызывает сомнения, т. к. понимание смысла полученных ответов неизбежно будет нести отпечаток субъективной интерпретации исследователя и отражать его концептуальное видение категорий, что ограничивает возможность воспроизведения подобной кодировки другими социологами [16]. В рамках нашей работы было решено придерживаться подхода, согласно которому следует выявить эксплицитно артикулируемые значения, а не смыслы. В данном случае можно выделить наиболее часто встречающиеся слова (униграммы) и словосочетания (n-граммы), выражающие основную идею автора текста в определении миссии вуза. Эта задача оперативно решается с помощью различных алгоритмов подсчёта терминов,

представленных в языке программирования (к примеру, в Python можно использовать CountVectorizer). Прежде чем приступить к частотному анализу слов, следует обратить внимание на проблему зашумлённости текстов малополезными словами, не имеющими содержательной коннотации. Как ранее было отмечено, в ходе препроцессинга были удалены стоп-слова. Несмотря на предварительное исключение стоп-слов, в тексте могут сохраниться малоинформативные термины, которые не были изначально предусмотрены алгоритмом, к примеру, «свой», «который», «весь», «поэтому», «мочь» и другие. Подобные слова следует вручную включить в список стоп-слов. Так, после исследования частотности слов, к перечню стоп-слов были добавлены часто встречающиеся слова, по существу, не имеющие особой смысловой нагрузки. Безусловно, удаление лишних терминов сопряжено с некоторой потерей информации, когда можно столкнуться с коротким, но эмоционально нагруженным ответом. Так, один из респондентов в ответ на вопрос о миссии вуза кратко написал: «А она есть?». Данное высказывание после исключения ненужных символов, пунктуации и стоп-слов было в итоге полностью очищено программой. Это указывает на риски потери информации при автоматической обработке текстов, что может потребовать дополнительной работы исследователя, который осмысленно подойдёт к анализу тезауруса языка респондентов.

Частотный анализ униграмм. Итак, представим результаты анализа униграмм. Частотный анализ как один из самых простых методов обработки текста на естественном языке позволил выделить топ-10 слов, которые чаще всего использовались опрошенными студентами в определении миссии вуза: «знание» (упоминалось 263 раза), «студент» (227), «навык» (120), «образование» (98), «специалист» (97), «профессиональный» (81), «развитие» (75), «человек» (72), «жизнь» (62), «научить» (51), «работа» (49). Как видно, для открытых вопросов, предполагающих развёрнутый ответ, простой подсчёт слов не даёт исчерпывающую информацию, т. к. остаются непрояснёнными понятия «развитие» и «человек». Какое конкретное развитие имелось в виду: личностное, профессиональное, либо развитие страны или экономики? В каких контекстах использовалось слово «человек»: когда речь шла о становлении человека как личности, особого субъекта образования, либо здесь проявился разговорный стиль и говорили о людях, о человеке вместо слов «студент», «обучающийся»? Становится очевидным, что необходим дополнительный качественный анализ высказываний. Плотное чтение комментариев, в которых фигурировали наиболее распространённые слова, позволило обогатить терминологический тезаурус контекстом описания. Прежде всего стоит отметить, что в понимании миссии вуза преобладают интеллектуальная и

образовательная функции, связанные с формированием необходимых знаний и навыков у студентов. По словам респондентов, вуз должен дать знания, навыки, «фундаментальное образование в соответствии со способностями каждого студента», «дать достойное образование в различных сферах жизни». По мнению многих студентов, вуз отвечает за подготовку кадров, он должен «взрастить специалистов», «подготовить будущих специалистов в своей сфере». Анализ данных показал, что в восприятии опрошенной молодёжи ведущие функциональные характеристики вуза также связаны с развитием, которое артикулируется студентами в различных контекстах: некоторые респонденты писали о развитии личности, мышления, навыков, «профессиональных качеств». Иногда речь шла о «развитии молодёжи и молодёжного капитала», общества, страны, либо «населения в целом». Как было отмечено выше, в текстах также часто встречался термин «человек». Данное понятие использовалось опрошенными юношами и девушками, как правило, для обозначения целевой аудитории вуза (студентов, обучающихся), когда утверждалось, что вуз даёт человеку знания, образование, возможность самореализоваться. В нескольких случаях высказывалась мысль о формировании «развитого, интересного и умного человека», «становлении человека как личности», «социализации человека».

Социальные функции вуза, его служение в интересах государства слабо выражены в представлениях опрошенных студентов. Только 35 человек написали о роли вуза как драйвера государственного развития. В этом аспекте звучали мысли, что миссия вуза состоит в том, чтобы подготовить кадры «во благо общества и страны», «выпустить студентов, которые будут способны улучшить положение страны». Примечательно, что диплом упоминался ещё реже, всего 33 раза, при этом на первый план выдвигались собственные цели пребывания в вузе. Так, студенты подчёркивали, что миссия вуза – «получить диплом, работодателям нужно просто его наличие», «просто получить бумажку (диплом) чисто для повышения зарплаты».

Стоит обратить внимание на то, что в представлениях опрошенной молодёжи крайне слабо актуализирована научно-исследовательская функция вуза: только 11 респондентов артикулировали в высказываниях данную сторону деятельности университета. Причём речь в основном шла о формировании научного потенциала обучающихся. О развитии науки в стране написал только один студент. О предпринимательском и инновационном измерении, характерном для третьей миссии вуза, не упоминалось ни разу.

Хотя частотный анализ униграмм позволяет извлечь определённую статистику из текстовых данных, в то же время более объёмную информацию из неструктурированных текстов можно получить, рассматривая

несколько лексем одновременно. Социологи применяют «анализ биграмм для определения семантического поля концепта на основании слов, которые непосредственно встречаются в одном словосочетании с ним» [17, с. 113]. Это позволит ответить на вопрос о том, какие ключевые концепты, состоящие из словосочетаний, больше всего выражены в студенческой риторике о миссии вуза.

Ключевые концепты: анализ биграмм. Поиск ключевых словосочетаний осуществлялся с опорой на специальные метрики ассоциации, которые «позволяют вычислять силу связи между элементами словосочетаний и основываются на частотах данных словосочетаний и входящих в них отдельных слов» [18, с. 108]. Перечень таких метрик обширен: простая частота (frequency), коэффициенты сходства Жаккарда (Jaccard), Дайса (the Dice score), коэффициенты поточечной взаимной информации (pointwise mutual information, PMI), логарифмического правдоподобия (LL), хи-квадрат (chi-squared), t-критерий Стьюдента (t-score) и др. [19; 20]. Данные статистические метрики позволяют ранжировать результаты оценивания словосочетаний по степени связи. Однако стоит учитывать определённые ограничения некоторых метрик. В частности, мера количества информации PMI оценивает часто употребляемые слова ниже, чем редкие термины, что не соответствует задаче исследования [21]. В свою очередь t-критерий Стьюдента предполагает нормальное распределение, поэтому для анализа автоматического извлечения биграмм были использованы частотные характеристики, коэффициенты сходства Жаккарда, логарифмического правдоподобия и хи-квадрат. На основе показателей метрик были проанализированы ранги найденных словосочетаний, которые были упорядочены согласно значениям соответствующих мер (см. табл. 1).

Таблица 1

Меры ассоциации для автоматического извлечения биграмм

Частота	Сходство Жаккарда	Логарифмическое правдоподобие	Хи-квадрат
Квалифицированный специалист	Взрослая жизнь	Квалифицированный специалист	Взрослая жизнь
Знание, навык	Квалифицированный специалист	Взрослая жизнь	Квалифицированный специалист
Профессиональный навык	Квалифицированные кадры	Высшее образование	Высшее образование
Взрослая жизнь	Высшее образование	Квалифицированные кадры	Квалифицированные кадры
Профессиональное знание	Будущая профессия	Качественное образование	Будущая профессия

Простая частота показывает, что наиболее распространённые словосочетания выражают потребность в подготовке квалифицированных специалистов, которые обладают профессиональными знаниями и навыками. Метрики ассоциации по-разному ранжируют словосочетания, но в целом выделяют практически идентичный набор биграмм: квалифицированный специалист, квалифицированные кадры, взрослая жизнь, высшее образование, будущая профессия. Коэффициент логарифмического правдоподобия добавляет биграмму «качественное образование». Можно сказать, что в понимании социальной роли вуза прежде всего преобладает прямая задача – дать необходимые знания обучающимся, дать качественное высшее образование, чтобы сформировать квалифицированных специалистов, подготовить студентов к взрослой жизни.

Тематическое моделирование. Следующий этап исследования был направлен на кластеризацию текстов с помощью тематической модели, предполагающей латентное размещение Дирихле (Latent Dirichlet Allocation, LDA). «Цель тематического моделирования – выявление скрытых семантических структур – тем, или топиков (topic), характеризующих содержание исследуемой текстовой коллекции» [22, с. 94]. Для формирования тем модель LDA использует распределения вероятностей таким образом, что отдельные темы предстают как вероятностные распределения слов, у которых есть свои веса для каждой темы [23]. Тематическое моделирование относят к «мягкой» кластеризации, т. к. каждый документ с определёнными вероятностями может быть включён в несколько кластеров-тем [24]. Для реализации тематического моделирования была сформирована матрица термин-документ. Чтобы сократить возможную размерность матрицы и снизить время её обработки, из текстов были удалены слова, которые встречались менее 10 раз. Темы были выявлены с использованием метрики TF-IDF, которая позволяет учитывать не только частоту встречаемости слова, но и важность слова. Критерием отбора количества тематик послужил показатель согласованности терминов (coherence score). В ходе обучения моделей обычно выбирается оптимальная модель из тех, для которых характерна наибольшая по показателю средняя согласованность внутри тем [23]. В нашей работе были построены тематические модели с разным количеством тем в диапазоне от 2 до 20 с шагом 2 и рассчитаны коэффициенты согласованности для каждой из этих моделей. С опорой на расчётные показатели было определено, что оптимальное количество тем для изучаемой коллекции документов составляет 6.

Исследование содержания комментариев респондентов, которые с наибольшей вероятностью отнесены к фиксируемым темам, позволяет выделить общий посыл для связанных в каждой теме слов. Можно сказать, что основные концепции выделенных терминологических класте-

ров сопряжены с функцией вуза дать необходимые знания, образование, подготовить специалистов, сформировать навыки для будущей жизни. В соответствии с общей смысловой направленностью выделенным топикам были даны условные названия. В таблице 2 указаны топики и топ-10 ключевых слов в порядке убывания веса в каждом кластере.

Таблица 2

Результаты тематического моделирования

Номер	Название темы	Ключевые слова
Тема 1	Знания и навыки для будущей работы	Знание, навык, профессиональный, практика, будущее, помощь, работа, деньги, человек, карьера
Тема 2	Квалифицированный специалист	Специалист, квалифицированный, качественный, кадр, образование, общество, условие, современный, страна, развитие
Тема 3	Образование для взрослой жизни	Образование, жизнь, социализация, взрослый, актуальный, навык, высокий, знание, информация, обучение
Тема 4	Обучение и развитие	Обучение, связь, человек, учиться, сфера, саморазвитие, мышление, развитие, интерес, сделать
Тема 5	Дипломированный специалист	Диплом, развитие, хороший, страна, молодёжь, деятельность, специальность, работа, личность, знание
Тема 6	Будущий профессионал	Профессия, будущее, профессионал, обучение, сотрудник, дело, достойный, знание, базовый, работать

Следует признать, что субъективная интерпретация полученных тем оставляет простор для критических размышлений, к чему подталкивают выводы и других авторов, апробировавших данный метод. Как отмечают О. Ю. Кольцова и К. А. Маслинский, «основным достоинством алгоритма оказалась его способность выявлять «яркие» и ясные темы-факторы и определять их вес в коллекции, т. е. определять «повестку дня» текстовой коллекции, а не имитировать то, как люди распределили бы тексты по группам» [10, с. 132]. Западные специалисты также подчёркивают, что топики не отображают реальные социальные явления, а скорее помогают исследователям в более глубоком чтении и качественном анализе текстов [25]. Между тем содержательная интерпретация тематических моделей с разным количеством тем в ходе поиска наиболее оптимального решения позволяет сделать вывод об общих сюжетах в дискурсе студентов о миссии высшей школы. Ключевые слова, сосредоточенные в ядре основных

тем, в основном связаны с обеспечением потребностей самих обучающихся. В этих определениях не выражено понимание значимости вуза как флагмана инноваций, источника научных разработок, платформы для предпринимательских и иных инициатив во благо общества и страны.

Заключение

Изучая текст, учёные из социальных наук чаще всего стремятся получить информацию о представлениях, мнениях, воззрениях представителей определённых социальных групп, что позволяет раскрыть содержательную наполненность так называемого тезауруса и составляющих его социальных концептов [26]. Нередко прочтение текстов сопряжено с поиском контекста их написания, социокультурных условий, которые нашли своё отражение в риторике анализируемого языка, в содержащихся в нём метафорах и ссылках на актуальную повестку дня. До недавнего времени технологии решения подобных социологических задач были крайне ограничены и нередко сводились к внимательному прочтению и ручной обработке доступной совокупности документов. С развитием цифровых технологий инструментарий социологического анализа значительно расширился, а вместе с тем появились новые ориентиры исследовательского поиска, направленные на выявление латентных содержательных и структурных характеристик текстов, ранее недоступных при ручной обработке.

В нашей работе был продемонстрирован пример обработки и анализа открытого вопроса анкеты, который начался с выявления униграмм и биграмм и завершился поиском латентных тем. Частотный анализ униграмм и визуализация его результатов с помощью облака слов позволили отразить словарь студенческого дискурса о миссии вуза, однако раскрытие смыслового контента потребовало изучения соответствующих фрагментов в тексте и прочтения конкретных высказываний. Более содержательные результаты отражают биграммы, которые позволяют выявить наиболее распространённые концепты в определении миссии вуза. Можно сказать, что в размышлениях студентов первоочередные задачи вуза направлены на трансляцию профессиональных знаний и навыков, в широком смысле – подготовку квалифицированных специалистов. В этих утверждениях слабо выражены ориентиры на удовлетворение потребностей общества и государства. Многие респонденты писали о формировании профессионалов, что чаще всего было сопряжено с личными траекториями становления и взросления. Довольно редко данная мысль обретала концептуальную идею, в которой подготовка будущих специалистов прежде всего нужна для развития экономики и страны. Таким об-

разом, частотный анализ позволил оперативно отразить статистику текста. Вместе с тем с помощью тематического моделирования были получены латентные топики. Косвенно выявленные темы характеризуют дискурс студентов о миссии вуза, которая, как оказалось, преимущественно сводится к удовлетворению потребностей обучающихся. Стоит признать, что топики, по сути, представляют собой совокупность слов и их содержательная интерпретация становится творческой задачей для аналитика. Иногда она невозможна в силу того, что объединённые в один топик слова отражают идентифицированное программой распределение слов, но не в буквальном смысле понятную для человека тему. Это одна из существенных сложностей, с которыми могут столкнуться практикующие социологи-аналитики. Стоит также помнить о том, что обработка текстов с помощью компьютерных алгоритмов всегда сопряжена с проблемой неопределённости оценивания результатов в тех случаях, когда смысл анализируемых слов выражают метафоры, ирония, либо сарказм.

Представленное исследование скорее является экспериментальным, позволяющим высветить дополнительные возможности цифровой методологии и в то же время обратить внимание на существующие ограничения. Возможно, подобные работы, помимо заявленной цели, также отражают актуальные тенденции развития социологии, которая в эпоху больших данных обновляет исследовательский инструментарий и приобретает новые возможности для изучения социальных явлений и процессов.

СПИСОК ИСТОЧНИКОВ

1. Классификация текстовых документов на основе Text Minig / А. А. Алексеев, А. С. Катасёв, А. Е. Кириллов, А. П. Кирпичников // Вестник технологического университета. 2016. Т. 19, № 18. С. 116–119. EDN WYBSGN.
2. *Hotho A., Nürnberger A., Paaß G.* A Brief Survey of Text Mining // Journal for Language Technology and Computational Linguistics. 2005. Vol. 20, № 1. P. 19–62. DOI 10.21248/jlcl.20.2005.68.
3. *Isaeva E., Aldarova D.* Text-Mining in Terms of Methodology and Development // Proceedings of 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). (Moscow, 26–29 January 2021). Moscow : IEEE, 2021. P. 413–416. DOI 10.1109/ElConRus51938.2021.9396437. EDN SECGLN.
4. *Осочкин А. А., Фомин В. В., Флегонтов А. В.* Метод частотно-морфологической классификации текстов // Программные продукты и системы. 2017. Т. 30, № 3. С. 478–486. DOI 10.15827/0236-235X.030.3.478-486. EDN ZDUXZD.
5. *Macanovic A.* Text mining for social science – The state and the future of computational text analysis in sociology // Social Science Research. 2022. Vol. 108. P. 1–16. DOI 10.1016/j.ssresearch.2022.102784. EDN SXELZJ.
6. *Evans J. A., Aceves P.* Machine Translation: Mining Text for Social Theory // Annual Review of Sociology. 2016. Vol. 42. P. 21–50. DOI 10.1146/annurev-soc-081715-074206.

7. Does counting emotion words on online social networks provide a window into people's subjective experience of emotion? A case study on Facebook / E. Kross, P. Verduyn, M. Boyer [et al.] // Emotion. 2019. Vol. 19, № 1. P. 97–107. DOI 10.1037/emo0000416.
8. *Karlgren J., Li R., Meyersson Milgrom E. M.* Text mining for processing interview data in computational social science // arXiv : [сайт]. 28 Nov 2020. URL: https://arxiv.org/abs/2011.14037 (дата обращения: 26.10.2023). DOI 10.48550/arXiv.2011.14037.
9. *Дудина В. И., Юдина Д. И.* Извлекая мнения из сети Интернет: могут ли методы анализа текстов заменить опросы общественного мнения? // Мониторинг общественного мнения: Экономические и социальные перемены. 2017. № 5 (141). С. 63–78. DOI 10.14515/monitoring.2017.5.05. EDN VTHJMT.
10. *Кольцова О. Ю., Маслинский К. А.* Выявление тематической структуры российской блогосферы: автоматические методы анализа текстов // Социология: 4М. 2013. № 36. С. 113–139. EDN RCFOWJ.
11. *Кашина М. А., Ткач С.* Социология ценностей: опыт построения таксономии путём использования технологии анализа естественного языка // Цифровая социология. 2023. Т. 6, № 1. С. 48–58. DOI 10.26425/2658-347X-2023-6-1-48-58. EDN YROQXD.
12. Оценка соответствия приоритетов стратегического развития регионов их отраслевой специализации на основе Text Mining / Е. В. Козоногова, Ю. В. Дубровская, М. Р. Русинова, П. В. Иванов // Вопросы государственного и муниципального управления. 2022. № 2. С. 106–133. DOI 10.17323/1999-5431-2022-0-2-106-133. EDN JRFOUQ.
13. *Kotsiantis S. B., Kanellopoulos D., Pintelas P. E.* Data Preprocessing for Supervised Leaning // International Journal of Computer and Information Engineering. 2007. Vol. 1, № 12. P. 4091–4096.
14. *Bird S., Klein E., Loper E.* Natural language processing with Python. Sebastopol : O'Reilly Media, 2009. 479 p. ISBN 978-0-596-51649-9.
15. *Воронцов К. В.* Вероятностное тематическое моделирование. 2013. 28 с. URL: https://mathprofi.com/uploads/files/3314_f_41_veroyatnostnoe-tematicheskoe-modelirovanie.-k.v.voroncov-2013g.pdf?key=19789ad13cac2399925acb68b1e18d8e/ (дата обращения: 26.10.2023).
16. *Оберемко О. А.* К типологии открытых вопросов // Мониторинг общественного мнения: Экономические и социальные перемены. 2018. № 4 (146). С. 97–108. DOI 10.14515/monitoring.2018.4.06. EDN UZQQIE.
17. *Ненько А. Е., Недосека Е. В., Галактионова А. А.* Возможности семантического анализа ключевых биграмм для исследования дискурса соседского онлайн сообщества // International Journal of Open Information Technologies. 2021. Т. 9, № 12. С. 111–118. DOI 10.25559/INJOIT.2307-8162.09.202112.111-118. EDN QTJRPZ.
18. *Хохлова М. В.* Статистический подход применительно к исследованию сочетаемости: от мер ассоциации к машинному обучению // Структурная и прикладная лингвистика: межвуз. сб. / Отв. ред. И. С. Николаев. СПб : Изд-во С.-Петерб. ун-та, 2019. Вып. 13. С. 106–122. EDN GKFUJY.
19. *Хохлова М. В.* К вопросу о сходстве мер ассоциации применительно к задаче автоматического извлечения глагольных коллокаций // Компьютерная лингвистика и вычислительные онтологии. 2019. № 3. С. 9–18. DOI 10.17586/2541-9781-2019-3-9-18. EDN LCONAI.

20. *Kormacheva D., Pivovarova L., Kopotev M.* Evaluation of collocation extraction methods for the Russian language // Quantitative approaches to the Russian language. New York : Routledge, 2018. P. 137–157. DOI 10.4324/9781315105048-7.
21. *Рассел М., Классен М.* Data Mining. 3-е изд. СПб. : Питер, 2020. 464 с. ISBN 978-5-4461-1246-3.
22. *Кирина М. А.* Сравнение тематических моделей на основе LDA, STM и NMF для качественного анализа русской художественной прозы малой формы // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2022. Т. 20, № 2. С. 93–109. DOI 10.25205/1818-7935-2022-202-93-109. EDN MWZRKH.
23. Тематическое моделирование в контексте медицинских текстов / С. А. Землянский, С. В. Аксёнов, И. А. Лызин, О. Г. Берестнева // Доклады ТУСУР. 2021. Т. 24, № 4. С. 58–64. DOI 10.21293/1818-0442-2021-24-4-58-64. EDN PWQTGR.
24. *Воронцов К. В., Потапенко А. А.* Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. 2012. Т. 4, № 4. С. 693–706. EDN PWNZXV.
25. *Pääkkönen J., Ylikoski P.* Humanistic interpretation and machine learning // Synthese. 2021. Vol. 199, № 1. P. 1461–1497. DOI 10.1007/s11229-020-02806-w. EDN CDPQZP.
26. *Луков В. А.* Тезаурусная социология : в 4 т. М. : Изд-во Моск. гуманит. ун-та, 2018. Т. 1. 608 с. ISBN 978-5-907017-45-0.

Сведения об авторах

А. Н. Пинчук
кандидат социологических наук, доцент,
старший научный сотрудник
AuthorID РИНЦ: 898003

С. Г. Карепова
кандидат социологических наук,
ведущий научный сотрудник
AuthorID РИНЦ: 717557

Д. А. Тихомиров
кандидат социологических наук, доцент,
ведущий научный сотрудник
AuthorID РИНЦ: 560487

Вклад авторов в подготовку публикации:

А. Н. Пинчук – 60% (подготовка общетеоретической и методологической основы исследования, участие в написании всех разделов статьи, обработка статистических данных). С. Г. Карепова – 20% (участие в написании всех разделов статьи, оформление публикации в соответствии с требованиями журнала). Д. А. Тихомиров – 20% (организация сбора и обработки социологических данных в ходе исследования, осуществление критического анализа и доработка текста статьи).

У авторов нет конфликта интересов для декларации.

Статья поступила в редакцию 12.11.2023; одобрена после рецензирования 12.01.2024; принята к публикации 21.01.2024.

Original article

DOI: 10.19181/snsp.2024.12.1.3

TEXT MINING TECHNOLOGIES IN SOCIOLOGICAL ANALYSIS (USING THE EXAMPLE OF STUDYING STUDENTS' IDEAS ABOUT THE MISSION OF A MODERN UNIVERSITY)

Antonina Nikolaeva Pinchuk [1]
Svetlana Gennadievna Karepova [2]
Dmitry Andreevich Tikhomirov [3]
[1, 3] Plekhanov Russian University of Economics,
[2] Institute of Sociology of FCTAS RAS,
Moscow, Russia,
[1] *antonina.pinchuk27@bk.ru*,
ORCID 0000-0001-7842-7141

[2] *Svetlran@mail.ru*,
ORCID 0000-0002-0472-0924

[3] *dat1983@yandex.ru*,
ORCID 0000-0002-1872-6788

For citation: Pinchuk A. N., Karepova S. G., Tikhomirov D. A. Text Mining technologies in sociological analysis (using the example of studying students`ideas about the mission of a modern university. *Sociologicheskaja nauka i social'naja praktika*. 2024;12(1):62–79. (In Russ.). DOI 10.19181/snsp.2024.12.1.3.

Abstract. There are discussed in the article the possibilities of using Text Mining methods in the practice of analyzing the information received on the base of open questionnaire questions. The paper presents an example of unigrams and bigrams analysis, as well as the search for latent topic using thematic modeling. Empirical materials present the data of survey conducted in 2022, in which 929 students of one Moscow economics university took part. In the open question of the questionnaire, it was proposed to define the mission of the university. Information made it possible to get the subjective interpretation of the main significancy of higher education in modern conditions. The frequency analysis of unigrams, supplemented by a qualitative analysis of respondents' statements, allowed reflecting the vocabulary of student discourse about the mission of the university. The articulation of bigrams was carried out on the basis of several statistical metrics, which made it possible to rank phrases and highlight a key set of concepts. The procedure revealed that in the perception of students, the priorities of the university are aimed at the transferring of professional knowledge and skills, in a broad sense – the training of qualified specialists. The social functions of the university, focused on meeting the needs of society and the state, are less pronounced in the conceptual interpretations of the interviewed students. At the next stage of the study the task of articulation and research of latent topics was put forward. The specific feature of thematic modeling is that the words combined into one topic reflect the distribution of words identified by the program, but not a topic that is literally understandable to a person. Taking into account the specifics of

the method used, the authors demonstrated the results of search analysis in the practice of processing an open question. As it turned out, the keywords concentrated in the core of the main topics are mainly related to meeting the needs of the students themselves, leaving on the periphery of the verbalized definitions any understanding of the importance of the university as a platform for innovation, scientific research, entrepreneurial and other initiatives for the benefit of society and the country. The results of the presented research can be useful in rethinking the research tools of sociologists in the context of the active development of digital technologies, which requires testing new methods, understanding their real capabilities and limitations in solving the tasks of sociological research.

Keywords: Text Mining, open question, preprocessing, unigrams, bigrams, thematic modeling

Acknowledgments: the research was carried out with the grant No. 24-28-00549 support of the Russian Science Foundation “Cultural marginality of Russian students: human potential of new generations as a problem and resource for developing patriotism in the main provisions and measures for implementing the state youth policy” (principal investigator candidate of sociology D. A. Tikhomirov).

REFERENCES

1. Alekseev A. A., Katasev A. S., Kirillov A. E., Kirpičnikov A. P. Classification of text documents based on Text Minig. *Vestnik tehnologičeskogo universiteta=Bulletin of the Technological University*. 2016;19(18):116–119. (In Russ.).
2. Hotho A., Nürnberger A., Paaß G. A Brief survey of Text Mining. *Journal for Language Technology and Computational Linguistics*. 2005;20(1):19–62. DOI 10.21248/jlcl.20.2005.68.
3. Isaeva E., Aldarova D. Text-Mining in terms of methodology and development. In: Proceedings of 2021 IEEE conference of Russian young researchers in electrical and electronic engineering (ElConRus). (Moscow, 26–29 January 2021). Moscow: IEEE; 2021. P. 413–416. DOI 10.1109/ElConRus51938.2021.9396437.
4. Osochkin A. A., Fomin V. V., Flegontov A. V. Method of frequency-morphological classification of texts. *Software products and systems=Programmny’e produkty’ i sistemy’*. 2017;30(3):478–486. (In Russ.). DOI 10.15827/0236-235X.030.3.478-486.
5. Macanovic A. Text mining for social science – The state and the future of computational text analysis in sociology. *Social Science Research*. 2022;(108):1–16. DOI 10.1016/j.ssresearch.2022.102784.
6. Evans J. A., Aceves P. Machine translation: Mining text for social theory. *Annual Review of Sociology*. 2016;(42):21–50. DOI 10.1146/annurev-soc-081715-074206.
7. Kross E., Verduyn P., Boyer M. [et al]. Does counting emotion words on online social networks provide a window into people’s subjective experience of emotion? A case study on Facebook. *Emotion*. 2019;19(1):97–107. DOI 10.1037/emo0000416.
8. Karlgren J., Li R., Meyersson Milgrom E. M. Text mining for processing interview data in computational social science. arXiv. 28 Nov 2020. Available at: https://arxiv.org/abs/2011.14037 (accessed: 26.10.2023). DOI 10.48550/arXiv.2011.14037.
9. Dudina V. I., Iudina D. I. Mining opinions on the Internet: can the text analysis methods replace public opinion polls? *Monitoring obshchestvennogo mneniya:*

ekonomicheskie i social'nye peremeny=Monitoring of public opinion: Economic and social changes. 2017;5(141):63–78. (In Russ.). DOI 10.14515/monitoring.2017.5.05.

10. Koltsova O. Y., Maslinsky K. A. Identifying the thematic structure of the Russian blogosphere: automatic text analysis methods. *Sociologiya: 4M=Sociology: 4M*. 2013;(36):113–139. (In Russ.).
11. Kashina M. A., Tkach S. Sociology of values: experience of building a taxonomy by using natural language analysis technology. *Cifrovaya sociologiya=Digital Sociology*. 2023;6(1):48–58. (In Russ.). DOI 10.26425/2658-347X-2023-6-1-48-58.
12. Kozonogova E. V., Dubrovskaya Yu. V., Rusinova M. R., Ivanov P. V. Assessment of compliance of strategic development priorities of regions with their industry specialization based on Text Mining. *Voprosy gosudarstvennogo i municipal'nogo upravleniya=Public administration issues*. 2022;(2):106–133. (In Russ.). DOI 10.17323/1999-5431-2022-0-2-106-133.
13. Kotsiantis S. B., Kanellopoulos D., Pintelas P. E. Data preprocessing for supervised leaning. *International Journal of Computer and Information Engineering*. 2007;1(12):4091–4096.
14. Bird S., Klein E., Loper E. Natural language processing with Python. Sebastopol: O'Reilly Media; 2009. 479 p. ISBN 978-0-596-51649-9.
15. Vorontsov K. V. Probabilistic Topic modeling. 2013. 28 p. Available at: https://mathprofi.com/uploads/files/3314_f_41_veroyatnostnoe-tematicheskoe-modelirovanie.-k.v.voroncov-2013g.pdf?key=19789ad13cac2399925acb68b1e-18d8e/ (accessed: 26.10.2023). (In Russ.).
16. Oberemko O. A. On typology of open-ended questions. *Monitoring obshchestvennogo mneniya: ekonomicheskie i social'nye peremeny=Monitoring of public opinion: Economic and social changes*. 2018;(4):97–108. (In Russ). DOI 10.14515/monitoring.2018.4.06.
17. Nenko A., Nedoseka E., Galaktionova A. Possibilities of the key bigrams semantic analysis for studying the discourse of an online neighbor community. *International Journal of open information technologies*. 2021;9(12):111–118. (In Russ.). DOI 10.25559/INJOIT.2307-8162.09.202112.111-118.
18. Khokhlova M. V. Statistical approach to collocation extraction: from association measures to machine learning. In: Nikolaev I. S. ed. Structural and applied linguistics: interuniversity collection of articles. Issue 13 [Strukturnaya i prikladnaya lingvistika: mezhvuz. Sb.] Saint-Petersburg: Izd-vo S.-Peterb. un-ta; 2019. P. 106–122. (In Russ.).
19. Khokhlova M. V. On the question of the similarity of association measures in relation to the problem of automatic extraction of verb collocations. *Komp'yuternaya lingvistika i vychislitel'nye ontologii=Computer linguistics and computing ontologies*. 2019;(3):9–18. (In Russ.). DOI 10.17586/2541-9781-2019-3-9-18.
20. Kormacheva D., Pivovarova L. Kopotev M. Evaluation of collocation extraction methods for the Russian language. In: Quantitative approaches to the Russian language. New York: Routledge; 2018. P. 137–157. DOI 10.4324/9781315105048-7.
21. Russell M. A., Klassen M. Mining the Social Web: Data Mining. Saint-Petersburg: Piter; 2020. 464 p. (In Russ.). ISBN 978-5-4461-1246-3.
22. Kirina M. A. A Comparison of topic models based on LDA, STM and NMF for qualitative studies of Russian short prose. *Vestnik NGU. Seriya: Lingvistika i mezhkul'turnaya kommunikaciya=Vestnik NSU. Series: Linguistics and intercultural communication*. 2022;20(2):93–109. (In Russ.). DOI 10.25205/1818-7935-2022-202-93-109.

23. Zemlyansky S. A., Axyonov S. V., Lyzin I. A., Berestneva O. G. Topic modeling in the context of medical texts. *Doklady TUSUR=Proceedings of TUSUR University*. 2021;24(4):58–64. (In Russ.). DOI 10.21293/1818-0442-2021-24-4-58-64.
24. Vorontsov K. V., Potapenko A. A. Regularization, robustness and sparsity of probabilistic topic models. *Komp'yuternye issledovaniya i modelirovanie=Computer research and modeling*. 2012;4(4):693–706. (In Russ.).
25. Pääkkönen J., Ylikoski P. Humanistic interpretation and machine learning. *Synthese*. 2021;199(1):1461–1497. DOI 10.1007/s11229-020-02806-w.
26. Lukov Val. A. Thesaurus Sociology: in 4 volumes [Tezaurusnaya sociologiya: v 4 t.]. Moscow: Izd-vo Mosk. gumanit. un-ta; 2018. Vol. 1. 608 p. (In Russ.). ISBN 978-5-907017-45-0.

Information about the Authors

A. N. Pinchuk
Candidate of Sociology,
Associate Professor,
Senior Researcher
Researcher ID: J-8648-2018
Scopus AuthorID: 57207845663

S. G. Karepova
Candidate of Sociology,
Leading Research Associate
Researcher ID: J-8658-2018
Scopus AuthorID: 56685800600

D. A. Tikhomirov
Candidate of Sociology,
Associate Professor,
Leading Research Associate
Researcher ID: AAS-4884-2021
Scopus AuthorID: 57210471226

Contribution of the authors:

A. N. Pinchuk – 60% (preparation of the general theoretical and methodological basis of the study, participation in writing all sections of the article, processing statistical data). S. G. Karepova – 20% (participation in writing all sections of the article, designing the publication in accordance with the requirements of the journal). D. A. Tikhomirov – 20% (organization of collection and processing of sociological data during the study, critical analyzing and final editing).

The article was submitted 12.11.2023; approved after reviewing 12.012024; accepted for publication 22.01.2024.