УДК 316.7
DOI: 10.19181/snsp.2025.13.3.9
EDN: BAPPMQ

Научная статья

МЕРА КОСИНУСНОГО СХОДСТВА ДЛЯ ОБРАБОТКИ НЕОКОНЧЕННЫХ ПРЕДЛОЖЕНИЙ (НА ПРИМЕРЕ ИЗУЧЕНИЯ ОБРАЗА ПАТРИОТА)

Антонина Николаевна Пинчук [1]
Дмитрий Андреевич Тихомиров [2]
Егор Васильевич Вахненко [3]
[1, 2, 3] РЭУ имени Г. В. Плеханова,
Москва, Россия,
[1] *antonina.pinchuk27@bk.ru*,
ORCID 0000-0001-7842-7141
[2] *dat1983@yandex.ru*,
ORCID 0000-0002-1872-6788
[3] *egor.vakhnenko@mail.ru*

Для цитирования: *Пинчук А. Н., Тихомиров Д. А., Вахненко Е. В.* Мера косинусного сходства для обработки неоконченных предложений (на примере изучения образа патриота) // Социологическая наука и социальная практика. 2025. Т. 13, № 3. С. 178–196. DOI 10.19181/snsp.2025.13.3.9. EDN BAPPMQ.

Аннотация. В условиях интенсивного развития науки об обработке естественного языка возникает вопрос об интеграции инновационных технологий в рабочие процессы социологов. Социальные учёные нередко сталкиваются с необходимостью обработки текстовых данных, полученных как в рамках собственных исследовательских проектов, так и в сети интернет. Очевидно, что использование в качестве базы данных доступных онлайн-источников выдвигает повышенные требования к техникам и процедурам обработки корпуса документов огромного объёма, нередко превышающего несколько сот тысяч строк. Однако не остаётся за рамками внимания работа с материалами авторских социологических исследований гораздо меньшего объёма, которые часто требуют значительных трудовых и временны х ресурсов, если их обрабатывать вручную. В этом случае возникает проблема согласованности кодирования текстов группой исследователей, где особую роль играет субъективное мнение специалистов при обобщении или группировке данных. В статье показаны возможности и ограничения использования меры косинусного сходства для анализа текстовых данных, полученных методом неоконченных предложений. Эмпирической базой исследования послужили материалы, полученные в ходе изучения образа патриота в одном из московских вузов в марте 2025 г. Всего в исследовании приняло участие 70 студентов. В работе представлена обработка ответов на стимульное предложение, которое респондентам нужно было завершить своими словами: «Патриот всегда…». Результаты расчёта меры косинусного сходства показали, что данная метрика может выступать полезным инструментом в первичном

поиске близких по содержательному контенту утверждений. В случае сомнений и необходимости проверки выводов или решения проблемы согласованности коллективного кодирования использование меры семантической близости может выступить в качестве значимого дополнительного количественного показателя для определения тематической направленности высказывания каждого из респондентов. Так, применяя оценку косинусного сходства, можно сгруппировать тексты, наиболее близкие по семантической нагрузке, тем самым приближая к пониманию общей структуры изучаемого образа и тезауруса участников исследования. В заключении делается вывод о современных требованиях к подготовке специалистов социально-гуманитарного профиля, что порождает новые методологические вопросы и открывает дискуссии об оптимальной интеграции технологических достижений в области обработки естественного языка в аналитические практики социальных учёных и исследователей.

Ключевые слова: метод неоконченных предложений, семантическое сходство, косинусное сходство, языковая модель BERT, образ патриота

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

Современные инструменты интеллектуального анализа текста открывают обширные исследовательские возможности для социологов, позволяя осуществлять автоматическую обработку текстовой информации, извлекать скрытые знания и закономерности в данных. Подобные технологии вызывают особый интерес в рамках обсуждения процедур обработки и анализа качественных данных, получаемых из различных источников. Так, на фоне интенсивного роста объёма текстовых документов в сети интернет повышенное внимание исследователей привлекают нереактивные данные [1]. Ключевой особенностью нереактивных данных является их тенденция к накоплению и агрегации вне зависимости от факта проведения исследования, что в условиях цифровизации ведёт к формированию беспрецедентных размеров информационных массивов, которые также принято называть Большими данными (Big Data), объём которых может превышать 150 Гб в сутки, а содержимое ежесекундно обновляется [1]. Для социолога могут быть интересны данные, воспроизводимые в результате публикаций и действий пользователей в различных социальных сетях, приложениях, сервисах. Это могут быть сообщения и комментарии, фотографии, видео, геолокации и хештеги [2]. Надо сказать, что данные, полученные с веб-ресурсов, которые позволяют зафиксировать поведение людей

[1] *Макаров А., Зуйкова А.* Что такое Big Data и как они устроены // Блог практикума : сайт. 15.12. 2022. URL: https://practicum.yandex.ru/blog/chto-takoe-big-data/ (дата обращения: 01.02.2025).
[2] *Макаров А., Зуйкова А.* Что такое Big Data и как они устроены // Блог практикума : сайт. 15.12. 2022. URL: https://practicum.yandex.ru/blog/chto-takoe-big-data/ (дата обращения: 01.02.2025).

в цифровой среде, называются цифровыми следами (trace data) и становятся особым предметом анализа современных учёных [2]. Тем не менее основным источником данных для многих социологов по-прежнему остаются результаты собственных социологических исследований, то есть реактивные данные, которые предполагают осознанное участие респондентов в исследовательском процессе. Источниками данных в традиционной качественной методологии являются интервью, фокус-группы, социологическое эссе и другие документы, содержащие размышления, высказывания, мнения, представления и ответы людей на задаваемые исследователем вопросы. Полученные такими методами материалы, как правило, подвергаются ручной обработке с последующей кодировкой, выделением категорий и тем, их подсчётом и содержательной интерпретацией в контексте теоретических обобщений. И поскольку классические методы сбора данных остаются распространённой практикой современных социологов, постольку возникает вопрос о применимости новых технологий для анализа и обработки текстовой информации, собранной собственными силами исследователя. Стоит заметить, что в данном случае интересуют не экономия времени и затрачиваемых усилий, хотя и это, безусловно, важно. Прежде всего интересуют эффективность новых алгоритмов для обработки результатов качественных методов и их сравнение с работой экспертов, которая предполагает углублённое чтение и категоризацию данных на основе понимания прочитанного. В этой связи предлагаем рассмотреть результаты методологического эксперимента, где показываются пути интеграции алгоритмов машинного обучения для работы с естественным языком в традиционные техники обработки текстовых данных социологического исследования. В качестве эмпирического материала использованы текстовые данные, полученные посредством метода неоконченных предложений. Обработка результатов метода неоконченных предложений предполагает выделение смысловых компонентов в текстах и сужение пространства интерпретаций до нескольких тематических категорий. Для апробации новых технологий в практике обработки корпуса документов, полученных с помощью метода неоконченных предложений, предлагается рассмотреть специальную метрику семантического сходства – косинусное сходство, – используемую для задач семантического поиска, который «направлен на повышение точности поиска за счёт понимания семантического значения поискового запроса и того, в каком корпусе выполняется поиск» [1].

Цель статьи – показать возможности и ограничения использования меры косинусного сходства для обработки текстовых данных, полученных методом неоконченных предложений и выявить оптимальные пути применения новых методов обработки данных в работе современных социальных учёных и исследователей на прикладном уровне.

Для достижения поставленной цели использованы данные, полученные в ходе изучения образа патриота в восприятии московской молодёжи. Для

[1] *Тищенко Д.* Семантический поиск (homemade) // Хабр : сайт. 07.08.2024. URL: https://habr.com/ru/articles/834356/ (дата обращения: 01.02.2025).

анализа образа патриота метод неоконченных предложений представляет удачное решение, если необходимо выявить особенности повседневного понимания данного явления вне контекста официального дискурса, где этот образ преимущественно имеет положительные черты. По существу, речь идёт о сложных явлениях, многообразие и противоречивость в определении которых лучше всего выявляется с помощью «мягких» неформализованных методов, среди которых особое место занимают проективные методики. Известно, что проективные методы позволяют минимизировать эффект интервьюера и получить развёрнутые рассуждения и комментарии участников исследования, не ограничиваясь заранее заданным перечнем ответов и отчасти решая проблему социально одобряемых откликов [3]. Так как метод неоконченных предложений относится к проективным методикам, то он позволяет отразить вербальную реакцию людей на стимульные предложения и выделить личностные смыслы и критерии, используемые респондентами для описания значений определяемых понятий и жизненного опыта [4]. Как отмечает Г. Г. Татарова, «на этапе сбора эмпирических данных вербальное поведение респондента не блокируется жёстко заданной схемой, он находится в системе своих личностных конструктов, отвечая на вопросы» [5, с. 155].

Следует заметить, что основной трудностью проективных методик является низкий уровень стандартизации и зависимость интерпретации полученных данных от личности исследователя [3]. Но возможно ли заменить обработку текстовых данных цифровыми вычислительными моделями или следует дополнять работу исследователя новыми методами анализа данных? В поиске ответа на этот вопрос рассмотрим традиционный способ обработки метода неоконченных предложений и семантические меры сходства, которые можно использовать в практике социальных учёных.

Обработка результатов метода неоконченных предложений

Метод неоконченных предложений является источником слабо структурированных данных и относится к проективным методикам, изначально разрабатываемым в психологии [5]. Однако если в психологии метод неоконченных предложений направлен на анализ латентных внутренних переживаний посредством косвенных воздействий, то в социологии предполагается изучение социальных явлений именно в том контексте, который осознаётся и подразумевается респондентами, когда они высказываются в ответ на стимульные фразы, и «предлагаемые ими окончания фраз создают определённое смысловое пространство, спектр возможных ответов и их обоснований» [6, с. 789].

Работу по созданию методики неоконченных предложений в социологическом исследовании начал В. Б. Ольшанский совместно с исследовательским коллективом, который ещё в начале 80-х гг. прошлого века одним из первых использовал неоконченные предложения с целью получения спонтанных

реакций, подобных репликам в ежедневных разговорах, для изучения жизненных и повседневных проблем [7; 8]. Сам автор делится воспоминаниями о непростой процедуре кодирования полученных ответов, которые сначала вручную распределялись по «признакам», перемешивались, а затем объединялись в «рубрики». Так, для каждого стимульного предложения создавался кодификатор, с помощью которого высказываниям присваивались четыре цифры: первые две из них означали номер предложения, третья цифра отражала номер категории, а четвертая – номер рубрики. Примечательно, что процедура обработки текстов требовала неоднократного возвращения к данным. Как отмечает учёный: «Кодификаторы перерабатывались три раза: уточнялись названия классов и первичных смысловых групп» [8, с. 89]. С. Г. Климова в своё время принимала участие в работе группы, которой руководил В. Б. Ольшанский, и по той же методике осуществила повторное исследование в 1993–1994 гг. Описывая процедуру обработки данных, она также указывает на сложности кодирования и отмечает, что «иногда возникали споры, поскольку многозначность ответов не позволяла принять определённое решение» [7, с. 55].

Институционализации метода неоконченных предложений как прикладного инструментария социологического исследования особо способствовали работы Г. Г. Татаровой и А. В. Бурлова [5; 9; 10; 11], посвящённые вопросам стратегии применения данного метода и логике анализа полученных данных.

Стоит отметить, что в социологическом дискурсе интерес к методу неоконченных предложений продолжает поддерживаться, и за последние десять лет в свет был выпущен ряд работ с результатами социологических исследований с применением данного метода. Хотя выборка подобных публикаций немногочисленна, в ней можно отметить как устоявшиеся исследовательские практики, так и авторские модификации, которые развивают прикладной потенциал метода неоконченных предложений [6; 12; 13; 14; 15; 16; 17].

Неоднократно авторы, использующие в своей работе метод неоконченных предложений, отмечали трудности обработки полученного массива данных [6; 18], на что указывали и первопроходцы в данной области. Безусловно, ответы, написанные самими участниками исследования в произвольной и приемлемой для них формулировке, даже в небольших объёмах (до 100 ответов) представляют особую сложность для обобщений в рамках предварительной обработки и подготовки для предстоящего анализа. Здесь следует учитывать, что формулировка незавершённых предложений может быть разной: от одного или нескольких слов, когда предполагается охват обширного смыслового поля и респонденты могут двигаться в разных направлениях в своих суждениях, до развёрнутых предложений, окончание которых будет кратким и узконаправленным, состоящим из нескольких слов [18].

Чаще всего процедура обработки данных предполагает контент-анализ и подсчёт частоты используемых слов. Например, З. В. Сикевич в рамках исследования этнической идентичности с помощью метода неоконченных предложений, осуществляла группировку утверждений респондентов в модальные

конструкты и представила те из них, которые встречались в высказываниях более чем у 5% респондентов, ответивших на предложение [18]. Есть и другие авторские подходы к агрегированию полученных ответов респондентов. В частности, Г. Г. Татарова и А. В. Бурлов сначала разделили всю совокупность неоконченных предложений на смысловые блоки, затем в каждом из них обобщили полученные данные посредством выделения элементарных обоснований (смысловой основы) высказывания респондента с их последующим объединением в элементы, а на более высоком уровне типологизации – в компоненты [9]. По мнению авторов, на этапе интеграции первичных обоснований в элементы обязательно следует привлекать самих участников исследования: «Это самая трудоёмкая часть исследования, включающая долгие и громкие дебаты некоторой группы экспертов из числа респондентов» [9, с. 14].

При отсутствии должных ресурсов исследователи самостоятельно читают полученные ответы, привлекая по возможности других исследователей, для кластеризации разнообразия мнений респондентов. В любом случае вопрос о роли субъективного мнения людей, которые распределяют ответы респондентов по понятным им смысловым категориям, остаётся дискуссионным и слабой стороной неформализованных методов. Трудности, связанные с групповым кодированием, по существу, отражают одну из важнейших проблем качественной методологии, где роль субъективного мнения экспертов становится ключевой при интерпретации переживаний, мнений, ожиданий и опасений респондентов. Неудивительно, что процедура обработки текстовых данных, полученных посредством метода неоконченных предложений, требует существенных трудовых и временны́х затрат исследователя. В этой связи видится актуальным апробировать новые технологии в оказании помощи в обработке корпуса документов на естественном языке. С этой целью рассмотрим меры семантического сходства.

Оценка семантической близости: косинусное сходство. Анализ семантического сходства между парами документов на естественном языке представляет одну из важнейших задач для автоматической обработки текстов. Поиск семантического сходства направлен на выявление смысловой связности двух текстов [19], то есть происходит поиск схожести пар объектов по содержанию [20]. «Причём, сходство должно быть выражено конкретным значением... В целом, концепция релевантности информации основана на её количественной оценке» [21, с. 20]. Для оценки семантической близости используют специальные меры, которые в числовом выражении позволяют определить меру сходства.

На данный момент можно выделить ряд мер, используемых учёными для оценки релевантности документов: сходство Жаккара, алгоритм шинглов, расстояние Левенштейна и другие [1]. Одной из самых распространённых мер является косинусное сходство. «Косинусное сходство представляет собой меру сходства, которая используется для векторных моделей. Векторная модель

[1] Семантический поиск: от простого сходства Жаккара к сложному SBERT // Хабр : сайт. 06.07.2021. URL: https://habr.com/ru/companies/skillfa000ctory/articles/566414/ (дата обращения: 01.02.2025).

позволяет представить текст документа и запроса в виде векторов одного пространства, а степень схожести этих векторов выражается через косинус угла между ними» [22, с. 7]. Здесь ставится задача векторизовать имеющиеся текстовые данные, где слова (в информационном поиске их называют термами) получают определённый вес. Представление документов в виде векторов позволяет определять расстояние между точками, фиксирующее их расположение по отношению друг к другу в пространстве [22]. Чем ближе расположены два вектора, тем меньше угол между ними и тем больше косинус угла между ними, что указывает на степень релевантности текстов. Так как используется скалярное произведение единичных векторов, то два одинаковых вектора будут иметь угол 0 градусов и мера косинусного сходства будет равна 1,0, тогда как два ортогональных вектора будут находиться под углом 90 градусов и давать метрику сходства 0.0 [23].

Остаётся открытым вопрос: каким образом текст можно перевести в вектор? В решении подобной задачи особую эффективность демонстрирует архитектура трансформера BERT (Bidirectional Encoder Representations from Transformers), основанная на мощной структуре нейронной сети [1]. BERT способен улавливать контекстуальные связи между словами, что используется в широком спектре задач обработки естественного языка: анализ тональности текста, автоматические ответы на вопросы, обобщение текста, тематическое моделирование, обнаружение спама, генерация естественного языка, машинный перевод, поиск информации и многое другое [24]. Как отмечают специалисты, предварительно обученный BERT способен понимать смысл и контекст документов на естественном языке в силу своей особой архитектуры [2]. BERT «кодирует огромное количество информации в набор плотных векторов» [3]. Сначала токенизатор разбивает текст на токены (фрагменты предложения от буквы до целого слова), которые поступают на вход модели, затем для токенов из специальной таблицы берутся эмбеддинги, которые получают на выходе. Причём эмбеддинги обновляются с целью распознавания контекста (соседних токенов) [4]. То есть BERT генерирует контекстуальные эмбеддинги, которые отражают значения слов на основе их использования в определённых контекстах, что значительно улучшает обработку естественного языка за счёт сохранения семантической информации [25]. «То есть для определённого количества входных токенов мы получаем соответствующее количество их векторных представлений» [5]. Так, в результате векторизации текстовых документов можно использовать

[1] Israa Hamdine Fine-tuning BERT for Semantic Textual Similarity with Transformers in Python // The Python Code : сайт. Updated: Jun. 2023. URL: https://thepythoncode.com/article/finetune-bert-for-semantic-textual-similarity-in-python (дата обращения: 01.02.2025).

[2] *Тищенко Д.* Указ. соч.

[3] Семантический поиск: от простого сходства Жаккара к сложному SBERT. Указ. соч.

[4] *Дале Д.* Маленький и быстрый BERT для русского языка // Хабр : сайт. 10.06.2021. URL: https://habr.com/ru/articles/562064/ (дата обращения: 01.02.2025).

[5] *Тищенко Д.* Указ. соч.

полученные числовые последовательности для подсчёта меры косинусного сходства. Здесь же следует сказать о практике применения BERT для поиска семантической близости на основе косинусного сходства [26; 27].

Методология и методы исследования

В рамках социологического исследования образа патриота в марте 2025 г. был реализован авторский исследовательский проект с применением метода неоконченных предложений для сбора данных. Всего в исследовании приняло участие 70 студентов социально-гуманитарного профиля подготовки, обучающихся в одном из московских вузов. Выборка невероятностная, формировалась методом «снежного кома», поэтому результаты исследования не могут быть распространены на генеральную совокупность. В исследовании приняли участие 69% девушек и 31% юношей. Возраст респондентов варьировался от 19 до 22 лет с медианным показателем 20 лет.

Инструментарий исследования содержал 10 незавершённых предложений, направленных на выявление структуры образа патриота: *1) «Для меня патриот – это...»; 2) «Патриот всегда –...»; 3) «Патриот никогда...»; 4) «Патриот должен...»; 5) «Патриот не должен...»; 6) «Патриотами становятся, потому что...»; 7) «Чаще всего патриотами являются...»; 8) «Патриот обладает такими чертами характера, как...»; 9) «Патриот в общении с другими людьми...»; 10) «Патриотом я могу назвать...»*. Эти предложения составляли разные смысловые блоки, воссоздающие многомерную структуру образа. Формулировка стимульных предложений происходила с опорой на более ранние работы социологов, в которых показан успешный опыт раскрытия образа исследуемого феномена, в частности, образа «культурного человека» [9], «террориста» [28], «коррупционера» [29], «героя» и «антигероя» [30]. Так как в данной статье основная задача – разобрать новые методы обработки данных, то ограничимся анализом предложения *«Патриот всегда...»,* с помощью которого предполагалось получить утверждения о поведении патриотов, что отчасти подтвердилось при анализе полученных ответов.

Полученные данные обрабатывались в среде разработки Google Colab с использованием языка программирования Python. В ходе обработки и анализа данных был осуществлён частотный анализ слов, затем рассчитано косинусное сходство высказывания каждого респондента с высказываниями других участников исследования. Отдельно была осуществлена ручная обработка полученных данных с прочтением текстов и их группировкой. Для этого сначала выделялись смысловые единицы в каждом высказывании, позволяющие идентифицировать мысли респондента с определённой смысловой категорией. После этого происходил расчёт доли выделенных смысловых категорий в отношении всех категорий.

Результаты исследования

Как было указано выше, на первом этапе были проведены аналитические работы по подсчёту частоты слов и дифференциации предложений на основе косинусного сходства. В ходе первичной обработки данных для статистического анализа символы были приведены к нижнему регистру, а предложения разделены на отдельные слова (токены). После этого была осуществлена лемматизация, то есть слова были приведены к начальной форме. Также предложения были очищены от стоп-слов, которые создают шум, но не несут смысловой нагрузки. Это междометия, союзы, предлоги и иные частицы [31].

Частотный анализ позволил выделить топ-10 слов, которые респонденты чаще всего писали для продолжения стимульного предложения *«Патриот всегда...»*. Это такие слова, как: «страна» (18%), «готов» (5%), «верный» (3%), «защищать» (3%), «хороший» (2%), «уважать» (2%), «благо» (2%), «стремиться» (2%), «история» (2%), «интерес» (2%).

Далее осуществлялся анализ семантического сходства предложений с помощью метрики косинусного сходства, которая показала себя как эффективный способ формирования тематических подвыборок. К примеру, первый респондент завершил предложение, ответив, что патриот всегда «на стороне своей страны при других, пусть даже и знает, что где-то её правительство может быть не право». В корпусе документов были выделены схожие по смыслу ответы, которые получили соответствующий показатель метрики семантической близости и отфильтрованы от большего к меньшему. В таблице 1 приведены семантически схожие с приведённым выше предложения на основе меры косинусного сходства, показатель которого больше 0.5.

Другим примером может послужить развёрнутый ответ, где респондент указывал на роль патриота в развитии страны: патриот всегда «стремится к развитию и процветанию своей Родины». Программа определила, что с данной фразой наиболее схожи следующие утверждения респондентов: «старается внести вклад в развитие и улучшение жизни страны», «старается ради улучшения страны», «заботится о благополучии своей нации». Можно сказать, что в полученную группу объединялись утверждения, которые содержали посыл к процветанию и развитию государства, что отражено в таблице 2.

Следует отметить, что трое человек написали идентичные ответы: патриот всегда «за свою страну». Применяемая модель позволила сразу выделить одинаковые ответы, а также показала смысловую близость других высказываний, позволяя охватить диапазон похожих позиций респондентов (см. табл. 3).

Таким образом, для каждого предложения в корпусе документов можно было вычленить наиболее схожие утверждения и перечитать в полученной подгруппе элементарные обоснования. Это значительно облегчало поиск схожих текстов.

На втором этапе был осуществлён анализ полученных ответов респондентов традиционным методом с прочтением текстов и распределением высказываний

Таблица 1

Значение меры косинусного сходства предложений с высказыванием о поведении патриота

Номер респондента	«Патриот всегда…»	Косинусное сходство
64	«на стороне своей страны при других, пусть даже и знает, что где-то её правительство может быть не право»	1.000000
53	«осознаёт интересы своего народа, которые могут и не совпадать с государственными»	0.806262
51	«защищает свою страну, даже если она не права»	0.795590
50	«действует в интересах страны, но не обязательно в интересах власти»	0.770452
18	«видит достоинства своей страны, говорит о стране уважительно, трудится на благо Отечества»	0.762913
4	«поддерживает свою страну, уважает её историю и стремится сделать её лучше»	0.756263
61	«уважает территорию своей страны, думает о её благе»	0.730184
49	«ищет позитив, а не негатив в процессах развития своей страны»	0.724340
8	«должен понимать, за что он любит свою страну, оценивать картину происходящего трезвым взглядом, а не просто следовать тому, что говорят сверху»	0.709943
17	«готов заступиться за свою страну как в споре, так и в войне»	0.707997
2	«будет за свою Родину, даже когда она не в лучшем состоянии»	0.688778

Таблица 2

Значение меры косинусного сходства предложений с высказыванием о роли патриота в развитии страны

Номер респондента	«Патриот всегда…»	Косинусное сходство
5	«стремится к развитию и процветанию своей Родины»	1.000000
19	«старается внести вклад в развитие и улучшение жизни страны»	0.872370
48	«старается ради улучшения страны»	0.757039
24	«заботится о благополучии своей нации»	0.752189
4	«поддерживает свою страну, уважает её историю и стремится сделать её лучше»	0.748969
39	«должен восхищаться своей страной и делать всё в её благо»	0.742331
49	«ищет позитив, а не негатив в процессах развития своей страны»	0.739976
3	«готов защищать Родину и отстаивать её интересы»	0.731378
55	«стремится к лучшему для своей страны»	0.728330

Таблица 3

Значение меры косинусного сходства предложений с высказыванием о поддержке патриотом страны

Номер респондента	«Патриот всегда…»	Косинусное сходство
29	«за свою страну»	1.000000
47	«за свою страну»	1.000000
40	«за свою страну»	1.000000
16	«выбирает свою страну»	0.918794
52	«остаётся верным своей стране»	0.839288
42	«благодарен своей стране»	0.838592
33	«выступает за свою Родину»	0.837869
7	«верен Родине»	0.828302
56	«уважает Родину»	0.811401
65	«будет на стороне Родины»	0.792960

по смысловым основаниям. Агрегирование схожих по смыслу ответов респондентов в группы происходило на основе экспертной оценки. Надо отметить, что прочтение вариантов ответа респондентов потребовало возвращение к материалу несколько раз, чтобы убедиться в классификации на укрупнённые блоки собранных высказываний. В итоге было выделено 8 элементов для продолжения предложения «Патриот всегда...», которые отражены в таблице 4. В качестве примера приведены некоторые элементарные обоснования, чтобы прояснить контент полученных элементов.

Таблица 4

Группировка данных на основе элементарных обоснований для предложения «Патриот всегда…»

Элементы	Элементарные обоснования	%
Защита	«встанет на защиту», «защищает своё Отечество»,«защищает свою страну»	17
Верность	«верен Родине», «остаётся верным своей стране»,«предан своей Родине»	16
Уважение	«уважает историю и культуру своей страны»,«уважает Родину»	14
Развитие страны	«стремится к развитию и процветанию своей Родины»,«старается внести вклад в развитие и улучшение жизни страны»	13
Личностные качества	«соблюдает нормы чести», «честен», «ответственный», «хороший», «прав», «верит в правильность своих поступков»	11
Поддержка	«будет за свою Родину», «выступает за свою Родину», «за свою страну»	8
Готовность к действию	«готов к труду ради Родины», «готов жертвовать ради Родины», «делает, а не говорит»	7
Приоритизация интересов общества	«действует в интересах страны, но не обязательно в интересах власти», «осознаёт интересы своего народа, которые могут и не совпадать с государственными»	5
Другое	«всегда патриот», «нужен», «согласен», «везде патриот»	9
Всего	-	100

Основываясь на результатах исследования, можно сделать вывод, что мера косинусного сходства выступает полезным инструментом в первичном поиске близких по содержательному контенту утверждений. В случае сомнений и необходимости проверки выводов меры семантического сходства могут послужить значимым дополнительным показателем. Так, используя оценку косинусного сходства, можно составить подвыборку текстов, которые по семантической нагрузке имеют наибольшую близость. Чтение фраз в сформированном кластере текста формирует представление о содержательном посыле близких по косинусному сходству утверждений и приближает к пониманию тезауруса изучаемой группы. Тем самым проблема однозначного определения тематической направленности предложения отчасти решается за счёт привлечения дополнительных оценок для прояснения ответа каждого респондента. С опорой на классификацию ответов участников исследования в соответствии с семантической близостью можно сэкономить временны́е ресурсы экспертов и добиться большей согласованности ответов, если работают несколько человек. Это тем более актуально, если исследователь ограничен в собственных ресурсах и работает самостоятельно. В этом случае использование меры косинусного сходства становится значимым подспорьем в принятии решения о смысловой близости полученных утверждений и высказываний.

Заключение

Метод неоконченных предложений остаётся в исследовательском арсенале современных социологов как основным, так и дополнительным способом получения эмпирических данных в случае необходимости анализа многослойных и неоднозначных в интерпретации явлений. Однако проблема во многом заключается в обработке данных, требующей существенных затрат человеческих и временны́х ресурсов. В этом случае цифровые методы могут послужить эффективным инструментом для оценки семантической близости высказываний респондентов, полученных в ответ на стимульные предложения, позволяя тем самым объединять ответы в схожие по смыслу группы. В данной статье был представлен пример использования косинусного сходства, рассчитанного на основе модели BERT. Методологический эксперимент показал, что используемые модели вполне успешно справились с поставленными задачами и выделяли с опорой на количественную оценку из имеющего массива предложения, которые наиболее близки по смыслу. Безусловно, использование количественных показателей позволяет экспертам более обоснованно взвешивать свои решения, что обогащает исследовательские практики и значительно упрощает процедуру кластеризации высказываний респондентов, особенно в случае сомнений экспертов и возникающих дискуссий. Но вместе с тем подобные методы обработки естественного языка, будучи полезными исследователю, не заменяют вдумчивое прочтение, требующее понимания высказываний респондентов с учётом контекста и метафор. Стоит добавить, что информационный поиск

представляет своего рода актуальное направление для социологов. Так, поиск схожих по смыслу высказываний в текстовом массиве представляет одну из задач в рамках обработки социологических данных, полученных не только методом неоконченных предложений, но и другими способами классического инструментария: интервьюирование, фокус-групповые дискуссии, социологические эссе и др. Это значительно расширяет возможности прикладного применения новых методов обработки естественного языка в социологической практике анализа данных. Вместе с тем, затрагивая вопросы работы с качественными данными в более широком плане, следует особо пристальное внимание уделять адаптации новых технологий к прикладным задачам социологического исследования и его методологическим принципам. С одной стороны, промедление в обновлении исследовательских возможностей может привести к отставанию всей области социологической науки от запросов современности, с другой, освоение цифровых методов должно происходить плавно, оптимально сочетая концептуальные основы методологии социологического исследования и инновационные способы обработки информации, что выдвигает новые вопросы для методологических дискуссий.

СПИСОК ИСТОЧНИКОВ

1. *Бызов А. А.* Интеллектуальный анализ текстов в социальных науках // Социология: методология, методы, математическое моделирование (Социология: 4М). 2019. № 49. С. 131–160. EDN GCIIVL.
2. *Hampton K. N.* Studying the Digital: Directions and Challenges for Digital Methods // Annual Review of Sociology. 2017. № 43 (1). P. 167–188. DOI 10.1146/annurev-soc-060116-053505.
3. *Пузанова Ж. В.* «Одиночество» как предмет эмпирического анализа // Социология: методология, методы, математическое моделирование (Социология: 4М). 2009. № 29. С. 132–154. EDN KNOYNZ.
4. *Зубова О. Г.* Проективные методики в социологических исследованиях: теория и практика // Вестник Московского университета. Серия 18. Социология и политология. 2023. № 29 (1). С. 194–218. DOI 10.24290/1029-3736-2023-29-1-194-218. EDN RUIPJM.
5. *Татарова Г. Г.* Основы типологического анализа в социологических исследованиях. М. : Высшее Образование и Наука, 2007. 236 с. ISBN 5-94084-047-7. EDN QOGTDB.
6 *Троцук И. В., Субботина М. В.* «Ядро» и «периферия» понятий «счастье» и «справедливость»: метод неоконченных предложений как инструмент валидизации // Вестник РУДН. Серия: Социология. 2022. Т. 22, № 4. С. 782–801. DOI 10.22363/2313-2272-2022-22-4-782-801. EDN TAPIWN.
7. *Климова С. Г.* Опыт использования методики неоконченных предложений в социологическом исследовании // Социология: методология, методы, математические модели (Социология: 4М). 1995. № 5-6. С. 49–64. EDN PFTWHV.
8. *Ольшанский В. Б.* Становление метода неоконченных предложений в Советском Союзе 70-х гг. // Социология: методология, методы, математические модели (Социология: 4М). 1997. № 9. С. 82–97. EDN PFTWRB.
9. *Татарова Г. Г., Бурлов А. В.* Метод неоконченных предложений в изучении образа («культурный человек») // Социология: методология, методы, математическое моделирование (Социология: 4М). 1997. № 9. С. 5–31. EDN PFTWPN.

10. *Татарова Г. Г., Бурлов А. В.* Логическая организация анализа данных, полученных методом неоконченных предложений // Социологические исследования. 1999. № 8. С. 123–133. EDN SNBITP.
11. *Бурлов А. В.* Метод неоконченных предложений в социологии: стратегии использования и логика анализа данных : дис. ...канд. соцол. наук : 22.00.01 / Бурлов Антон Вячеславович. М. : ИС РАН, 2001. 179 с. EDN QDMELN.
12. *Тихомиров Д. А., Новицкая К. В.* Представления молодёжи Москвы о гендерных ролях и характеристиках современной женщины // Горизонты гуманитарного знания. 2018. № 3. С. 90–102. DOI 10.17805/ggz.2018.3.6. EDN VMKDDA.
13. *Сикевич З. В., Фёдорова А. А.* «Мы – русские» (ассоциативные этнические образы молодых петербуржцев) // Социологическая наука и социальная практика. 2019. Т. 7, № 3 (27). С. 40–56. DOI 10.19181/snsp.2019.7.3.6688. EDN CPKOVO.
14. *Субботина М. В.* Применение метода неоконченных предложений в изучении понятий со сложными коннотациями: концептуализация героизма и справедливости // Общество: социология, психология, педагогика. 2021. № 5 (85). С. 88–96. DOI 10.24158/spp.2021.5.15. EDN EXIGEF.
15. *Бубнов А. Ю., Савельева М. А.* Память о Великой Отечественной войне: сравнительный анализ взглядов российской и белорусской молодёжи // Наука. Общество. Оборона. 2021. Т. 9, № 2 (27). С. 13. DOI 10.24412/2311-1763-2021-2-13-13. EDN VCTHOA.
16. *Савенкова А. С., Субботина М. В.* Возможности метода неоконченных предложений в изучении «культуры отмены» // Вестник РУДН. Серия: Социология. 2024. Т. 24, № 3. С. 660–683. DOI 10.22363/2313-2272-2024-24-3-660-683. EDN DXLFCJ.
17. *Татарова Г. Г., Чиркова А. В.* Здоровьесберегающее поведение молодёжи: формирование типообразующих признаков методом неоконченных предложений // Социологическая наука и социальная практика. 2024. Т. 12, № 1. С. 25–61. DOI 10.19181/snsp.2024.12.1.2. EDN GWRDZA.
18. *Сикевич З. В.* Опыт применения процедуры неоконченных предложений в социологическом исследовании // Вестник Санкт-Петербургского университета. Социология. 2019. Т. 12, № 4. С. 317–328. DOI 10.21638/spbu12.2019.402. EDN XKAFTS.
19. *Андриевская Н. К.* Гибридная интеллектуальная мера оценки семантической близости // Проблемы искусственного интеллекта. 2021. № 1 (20). С. 4–17. EDN ZDZKGK.
20. Меры семантической близости в онтологии / К. В. Крюков, Л. А. Панкова, В. А. Пронина [и др.] // Проблемы управления. 2010. № 5. С. 2–14. EDN MUVNSP.
21. *Бермудес С. Х. Г.* Метод измерения семантического сходства текстовых документов // Известия ЮФУ. Технические науки. 2017. № 3 (188). С. 17–29. DOI 10.23683/2311-3103-2017-3-17-29. EDN ZDHXJR.
22. *Белова К. М., Судаков В. А.* Исследование эффективности методов оценки релевантности текстов // Препринты ИПМ им. М. В. Келдыша. 2020. № 68. С. 1–16. DOI 10.20948/prepr-2020-68. EDN CYCEWZ.
23. *Рассел М., Классен М.* Data Mining. Извлечение информации из Facebook, Twitter, LinkedIn, Instagram, GitHub. СПб. : Питер, 2020. 464 с. ISBN 978-5-4461-1246-3.
24. *Sarika K., Vijay Kumar A., Vijay R.* Beyond Text: Exploring Multimodal BERT Models // Journal of Computer Science Applications and Information Technology. 2025. № 10 (1). P. 1–6. DOI 10.15226/2474-9257/10/1/00164.
25. BERT applications in natural language processing: a review / N. M. Gardazi, A. Daud, M. K. Malik [et al.] // Artif Intell Rev. 2025. Vol. 58. № 166. DOI 10.1007/s10462-025-11162-5.

26. Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT / F. W. Mutinda, Sh. Yada, Sh. Wakamiya, E. Aramaki // Methods of Information in Medicine. 2021. T. 60, № S01. P. e56–64. DOI 10.1055/s-0041-1731390. EDN QQSZZL.
27. *Syaifudin M. F., Adiatmaja G., Hidayaturrohman B.* Calculation of Similarity between MUI Fatwas: A Comparison of Text Extraction Features and String Matching Algorithms // Halal Research Journal (HRJ). 2025. Vol. 5, № 1. P. 1–13. DOI 10.12962/j22759970.v5i1.1226. EDN SWVYVB.
28. *Пузанова Ж. В., Тертышникова А. Г.* Метод неоконченных предложений в исследовании социальных представлений (на примере образа террориста) // Теория и практика общественного развития. 2015. № 4. С. 12–15. EDN TKAMQH.
29. *Пинчук А. Н., Тихомиров Д. А.* Образ коррупционера в восприятии российской молодёжи: применение метода неоконченных предложений // Вестник Института социологии. 2019. Т. 10, № 2. С. 12–27. DOI 10.19181/vis.2019.29.2.573. EDN UFIZXB.
30. *Желизняк М. Н.* Опыт использования метода неоконченных предложений в изучении образов «героя» и «антигероя» нашего времени // Мониторинг общественного мнения: экономические и социальные перемены. 2024. № 1 (179). С. 257–275. DOI 10.14515/monitoring.2024.1.2460. EDN TKBIIJ.
31. *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.

Сведения об авторах

А. Н. Пинчук
кандидат социологических наук,
доцент
SPIN-код: 7853-0878

Д. А. Тихомиров
кандидат социологических наук,
доцент
SPIN-код: 3369-3077

Е. В. Вахненко
студент
SPIN-код: 2707-9952

Вклад авторов в подготовку публикации:
А. Н. Пинчук – 70% (подготовка общетеоретической и методологической основы исследования, участие в написании всех разделов статьи, расчёт косинусного сходства).
Д. А. Тихомиров – 20% (организация сбора и обработки социологических данных в ходе исследования, осуществление критического анализа и доработка текста статьи).
Е. В. Вахненко – 10% (участие в сборе данных, предварительная обработка текстовых данных).
У авторов нет конфликта интересов для декларации

Статья поступила в редакцию 01.05.2025; одобрена после рецензирования 21.05.2025; принята к публикации 25.07.2025.

Original article
DOI: 10.19181/snsp.2025.13.3.9

COSINE SIMILARITY MEASURE TO PROCESS THE UNFINISHED SENTENCES (USING THE EXAMPLE OF STUDYING THE IMAGE OF A PATRIOT)

Antonina Nikolaeva Pinchuk [1]
Dmitry Andreevich Tikhomirov [2]
Egor Vasilyevich Vakhnenko [3]
[1, 2, 3] Plekhanov Russian University of Economics,
Moscow, Russia,
[1] *antonina.pinchuk27@bk.ru*,
ORCID 0000-0001-7842-7141
[2] *dat1983@yandex.ru*,
ORCID 0000-0002-1872-6788
[3] *egor.vakhnenko@mail.ru*

For citation: Pinchuk A. N., Tikhomirov D. A., Vakhnenko E. V. Cosine similarity measure to process the unfinished sentences (using the example of studying the image of a patriot). *Sociologicheskaja nauka i social'naja praktika.* 2025;13(3):178–196. (In Russ.). DOI 10.19181/snsp.2025.13.3.9.

Abstract. In the context of the intensive development of natural language processing methods, the question arises about the integration of innovative technologies into the work processes of sociologists. Social scientists often face the need to process text data obtained both as part of their own research projects and on the Internet. Obviously, using available online sources as a database places increased demands on the techniques and procedures for processing a huge corpus of documents, often exceeding several hundred thousand lines. However, it is not beyond the scope of attention to work with the materials of author's sociological research of a much smaller volume, which often require significant labor and time resources if they are processed manually. In this case, the consistency of collective coding and the role of the subjective opinion of experts in the generalization or grouping of data raises questions. The purpose of the article is to show the possibilities and limitations of using the cosine similarity measure to process the results of the unfinished sentences method. The empirical basis of the study was the materials obtained during the study of the image of a patriot in one of the Moscow universities in March 2025. A total of 70 students participated in the study. The article processed responses to a stimulus sentence, which the respondents had to complete in their own words: "A patriot always..." The results of calculating the cosine similarity measure have shown that this metric can be a useful tool in the initial search for statements that are similar in content. In case of doubt and the need to verify their conclusions or solve the problem of consistency of collective coding, the use of a measure of semantic proximity can act as a significant additional quantitative indicator to determine the thematic focus of each respondent's utterance. Thus, using the cosine similarity assessment, it is possible to group the texts that are closest in semantic load, thereby bringing closer to understanding the general structure of the studied image and the thesaurus of the study participants. In conclusion, a conclusion is drawn about the modern requirements for the training of specialists in the social and humanitarian fields,

which raises new methodological questions and opens up discussions about the optimal integration of technological advances in natural language processing into the analytical practices of social scientists and researchers.

Keywords: unfinished sentence method, semantic similarity, cosine similarity, BERT language model, patriot image

Acknowledgments: the research was carried out with the support of the Russian Science Foundation grant No. 24-28-00549 “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. Byzov A. Text mining in social sciences. *Sociology: 4M=Cociologiya: 4M*. 2019;(49):131–160. (In Russ.).
2. Hampton K. N. Studying the digital: directions and challenges for digital methods. *Annual Review of Sociology*. 2017;43(1):167–188. DOI 10.1146/annurev-soc-060116-053505.
3. Puzanova Zh. V. Loneliness as a subject of empirical analysis. *Sociology: 4M=Cociologiya: 4M*. 2009;(29):132–154. (In Russ.).
4. Zubova O. G. Projective techniques in sociological research: theory and practice. *Moscow State University Bulletin. Series 18. Sociology and Political Science=Vestnik Moskovskogo gosudarstvennogo universiteta. Seriya 18. Sociologiya i politologiya*. 2023;29(1):194–218. (In Russ.). DOI 10.24290/1029-3736-2023-29-1-194-218.
5. Tatarova G. G. Fundamentals of typological analysis in sociological research. [Osnovy tipologicheskogo analiza v sotsiologicheskikh issledovaniyakh]. Moscow: Vy’sshee Obrazovanie i Nauka; 2007. 236 p. (In Russ.). ISBN 5-94084-047-7.
6. Trotsuk I. V., Subbotina M. V. Core and periphery of the concepts happiness and justice: Unfinished sentences technique as a means of validation. *RUDN Journal of Sociology=Vestnik RUDN. Seriya: Sociologiya*. 2022;22(4):782–801. (In Russ.). DOI 10.22363/2313-2272-2022-22-4-782-801.
7. Klimova S. G. Experience of using the sentence completion technique in sociological research [Opy’t ispol’zovaniya metodiki neokonchenny’x predlozhenij v sociologicheskom issledovanii]. *Sociology: 4M=Cociologiya: 4M*. 1995;(5-6):49–64. (In Russ.).
8. Olshansky V. B. The formation of the sentence completion method in the Soviet Union of the 70s. [Stanovlenie metoda neokonchenny’x predlozhenij v Sovetskom Soyuze 70-x gg.]. *Sociology: 4M=Cociologiya: 4M*. 1997;(9):82–97. (In Russ.).
9. Tatarova G. G., Burlov A. V. The method of unfinished sentences in the study of an image (cultured man) [Metod neokonchenny’x predlozhenij v izuchenii obraza («kul’turny’j chelovek»)]. *Sociology: 4M=Cociologiya: 4M*. 1997;(9):5–31. (In Russ.).
10. Tatarova G. G., Burlov A. V. Logical organization of data analysis obtained by the sentence completion method [Logicheskaya organizaciya analiza danny’x, poluchenny’x metodom neokonchenny’x predlozhenij]. *Sociological Studies=Sotsiologicheskie Issledovaniya*. 1999;(8):123–133. (In Russ.).
11. Burlov A. V. The sentence completion method in sociology: strategy of use and logic of data analysis. [Metod neokonchennykh predlozheniy v sotsiologii: strategii ispol’zovaniya i logika analiza dannykh]. Dissertation for the degree of candidate of sociological sciences. Moscow: Institut sociologii RAN; 2001. 179 p. (In Russ.).

12. Tikhomirov D. A., Novitskaya K. V. Moscow youth's representations of gender roles and characteristics of a modern woman. *Horizons of humanitarian knowledge Gorizonty' gumanitarnogo znaniya*. 2018;(3):90–102. (In Russ.). DOI 10.17805/ggz.2018.3.6.
13. Sikevich Z. V., Fedorova A. A. We are Russians (associative ethnic images of young St. Petersburg residents). *Sociological Science and Social Practice=Sociologicheskaja nauka i social'naja praktika*. 2019;(3):40–56. (In Russ.). DOI 10.19181/snsp.2019.7.3.6688.
14. Subbotina M. V. Application of the method of incomplete sentences in the study of concepts with complex connotations: conceptualization of the concepts of «heroism» and «justice». *Society: sociology, psychology, pedagogics=Obshhestvo: sociologiya, psixologiya, pedagogika*. 2021;(5):88–96. (In Russ.). DOI 10.24158/spp.2021.5.15.
15. Bubnov A. Yu., Savelieva M. A. Memory of the Great Patriotic War (World War II): comparative analysis of the views of Russian and Belarusian youth. *Science. Society. Defense=Nauka. Obŝestvo. Oborona*. 2021;9(2):13. (In Russ.). DOI 10.24412/2311-1763-2021-2-13-13.
16. Savenkova A. S., Subbotina M. V. Possibilities of the unfinished sentences technique in the study of cancel culture. *RUDN Journal of Sociology=Vestnik RUDN. Seriya: Sociologiya*. 2024;24(3):660–683. (In Russ.). DOI 10.22363/2313-2272-2024-24-3-660-683.
17. Tatarova G. G., Chirkova A. V. Health behavior of young people: formation of typological attributes using the sentence completion method. *Sociological Science and Social Practice=Sociologicheskaja nauka i social'naja praktika*. 2024;12(1):25–61. (In Russ.). DOI 10.19181/snsp.2024.12.1.2.
18. Sikevich Z. V. The experience of applying the procedure of "unfinished sentences" to sociological research. *Bulletin of Saint Petersburg University. Sociology=Vestnik Sankt-Peterburgskogo universiteta. Sociologiya*. 2019;12(4):317–328. (In Russ.). DOI 10.21638/spbu12.2019.402.
19. Andrievskaya N. K. Hybrid intelligent measure of semantic similarity evaluation. *Problems of Artificial Intelligence=Problemy' iskusstvennogo intellekta*. 2021;1(20):4–17. (In Russ.).
20. Kryukov K. V., Pankova L. A., Pronina V. A. [et al.] Measures of semantic proximity in ontology. *Control Sciences=Problemy Upravleniya*. 2010;(5):2–14. (In Russ.).
21. Bermudez S. J. G. Method for measuring the semantic-similarity of textual documents. *Scientific, technical and practical journal=Izvestiya SFedU. Engineering Sciences*. 2017;(3):17–29. (In Russ.). DOI 10.23683/2311-3103-2017-3-17-29.
22. Belova K. M., Sudakov V. A. Effectiveness of methods for assessing the text relevance. *Preprints of the IPM named after M. V. Keldysh=Preprinty' IPM im. M. V. Keldy'sha* 2020;(68):1–16. (In Russ.). DOI 10.20948/prepr-2020-68.
23. *Russell M. A., Klassen M.* Mining the social web: data mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and More. Saint-Petersburg: Piter; 2020. 464 p. (In Russ.). ISBN 978-5-4461-1246-3.
24. Sarika K., Vijay Kumar A., Vijay R. Beyond text: exploring multimodal BERT models. *Journal of Computer Science Applications and Information Technology*. 2025;(10):1–6. DOI 10.15226/2474-9257/10/1/00164.
25. Gardazi N. M., Daud A., Malik M. K. [et al.] BERT applications in natural language processing: a review. *Artif Intell Rev*. 2025;58(166). DOI 10.1007/s10462-025-11162-5.
26. Mutinda F. W., Yada S, Wakamiya S, Aramaki E. Semantic textual similarity in Japanese clinical domain texts using BERT. *Methods of Information in Medicine*. 2021;60(S01):e56–64. DOI 10.1055/s-0041-1731390.
27. Syaifudin M. F., Adiatmaja G., Hidayaturrohman B. Calculation of similarity between MUI fatwas: a comparison of text extraction features and string matching algorithms. *Halal Research Journal (HRJ)*. 2025;5(1):1–13. DOI 10.12962/j22759970.v5i1.1226.

28. Puzanova Zh. V., Tertyshnikova A. G. The method of incomplete sentences in the study of social representations (the case of the terrorists' image). *Theory and Practice of Social Development=Teoriya i praktika obshhestvennogo razvitiya*. 2015;(4):12–15. (In Russ.).
29. Pinchuk A. N., Tikhomirov D. A. The image of a corrupt official as perceived by Russia's youth: using the unfinished sentences method. *Bulletin of the Institute of Sociology=Vestnik instituta sotziologii*. 2019;10(2):12–27. (In Russ.). DOI 10.19181/vis.2019.29.2.573.
30. Zheliznyk M. N. Using the method of unfinished sentences in studying the images of the "hero" and "anti-hero" of our time. *Monitoring of Public Opinion: Economic and Social Changes=Monitoring obshhestvennogo mneniya: e'konomicheskie i social'ny'e peremeny.* 2024;(1):257–275. (In Russ.). DOI 10.14515/monitoring.2024.1.2460.
31. 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. *Sociological Science and Social Practice=Sociologicheskaja nauka i social'naja praktika*. 2024;12(1):62–79. (In Russ.). DOI 10.19181/snsp.2024.12.1.3.

Information about the Authors

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

D. A. Tikhomirov
Candidate of Sociology,
Associate Professor,
ResearcherID: AAS-4884-2021
Scopus AuthorID: 57210471226

E. V. Vakhnenko
Student

Contribution of the authors:
A. N. Pinchuk – 70% (preparation of the general theoretical and methodological basis of the study, participation in writing all sections of the article, calculation of cosine similarity).
D. A. Tikhomirov – 20% (organization of collection and processing of sociological data during the study, critical analyzing and final editing).
E. V. Vakhnenko – 10% (participation in data collection, preliminary data processing).
The authors declare no conflicts of interests.

The article was submitted 01.05.2025; approved after reviewing 21.05.2025; accepted for publication 25.07.2025.