Итоговый отчет о проделанной работе обладателей грантов Государственного Совета Республики Крым молодым ученым Республики Крым имени Н. Я. Данилевского

Исследование способов и автоматизации обработки авторских текстов для определения психотипов

Название научного проекта

Информатика

Номинация

Гаврилова Анна Сергеевна

(Ф.И.О. автора проекта)

Ассистент кафедры математики и информатики Института педагогики, психологии и инклюзивного образования

Ученая степень, ученое звание, должность

Гуманитарно-педагогическая академия (филиал) ФГАОУ ВО «Крымский федеральный университет имени В.И. Вернадского» в г. Ялте

Образовательная организация высшего образовании или научная организация

1. Краткое описание научного проекта, победившего в конкурсе на назначение грантов Государственного Совета Республики Крым молодым ученым Республики Крым имени Н. Я. Данилевского.

Научный проект был посвящен исследованию взаимосвязи психологических и языковых характеристик личности, анализу психотипа (социотипа) человека как параметра его личности по лингвистическому критерию и проверке точности автоматического определения психотипа (социотипа) человека по его речевым характеристикам.

В своей работе «Языковой круг: личность, концепты, дискурс» Карасик В.И. отмечал, что: «лингвистическое изучение языковой личности базируется на тех психологических и социологических особенностях, которые находят выражение в языковой семантике и прагматике и позволяют построить типологию языковых личностей». Такое видение вопроса лишь подтверждает то представление, что посредством письма и речевого аппарата человек демонстрирует свои внутренние черты социальности, этических и моральных норм, видения и представления мира в целом.

Зная, что «язык» человека – это своего рода зеркало, отражающее его внутренний мир, мы применили для определения психологической принадлежности человека методы интеллектуального анализа данных (неструктурированных текстов). При подготовке к проведению данного исследования, необходимо было детально изучить и проанализировать основы психолингвистики и соционики.

Каким бы образом человек не пытался скрыть свои психические черты, он не способен постоянно контролировать выражение своих мыслей в написании, или контролировать их в устном выражении, что как раз позволяет достаточно качественно провести анализ его текстов и сделать соответствующие выводы в психических чертах их автора. Кроме того, существует и прямая связь психологических аспектов личности с ее внешними проявлениями, это касается и языковых особенностей при написании текста. В частности, нас интересовала возможность автоматизации в выявлении взаимосвязи между психотипом человека и его лингвистическими особенностями при написании текста. Например, в «избранных психологических работах» А. Н. Леонтьев представляет наиболее распространенную точку зрения, согласно которой психическая функция восприятия ставится на первый план в характеристике сознания. В своём фундаментальном труде «Психологические типы» К. Г. Юнг детально рассмотрел и заострил внимание на специфике восприятия человеком внешней информации, что, по его мнению, позволяет произвести его классификацию, учитывая личностные особенности психики любого человека. Языковую типологию личности человека достаточно просто построить на основе проводимого лингвистического анализа её языковых характеристик. Что позволяет провести её типирование, основываясь на выделенных данных семантического ядра текста и анализа его языковых особенностей.

В последнее десятилетие областью обработки естественного языка стало NLP (обработка естественного языка) и, в частности, его подраздел «интеллектуальный анализ данных» довольно интенсивно развивается в зарубежных странах. Но, к сожалению, в России эти методы развиты недостаточно массово, что связано с

отсутствием достаточного количества предложений на данный вид деятельности (исключение составляет корпоративный и ИТ-кластеры). Тем не менее, наблюдается постепенный рост активности в этой сфере со стороны различных сфер жизни. Современный мир остро нуждается в совершенствовании и разработке новых алгоритмов и методов обработки больших данных. По результатам исследований количество информации, которое накапливается на физических носителях, увеличивается вдвое ежегодно и данный процесс будет только ускоряться. До 70% из этих данных представляет собой неструктурированный массив текста на различных языках, требующий эффективной обработки и грамотного анализа. Современные же алгоритмы анализа в основе своей работы предполагают увеличение производительности за счёт повышения мощностей аппаратной части.

Как было отмечено выше, имея стремительный рост информации в неструктурированном виде – в форме естественного языка и, учитывая сложности при её обработке, на лицо наличие проблемы анализа подобных массивов данных. Сегодня всё больше и больше сфер деятельности человека выражает нехватку инструментов аналитического разбора текстов. По данным аналитических исследований агентства IDC, к 2025 году предполагается увеличение объёма данных, накопленных человечеством, до 160 ЗБайт. Основой массива этой информации является информация в таких формах, как фотографии, видеозаписи, аудиозаписи, авторские текста на естественных языках.

К задачам современных алгоритмов анализа неструктурированных данных помимо задачи классификации и кластеризации, следует отнести желание добычи из них полезной информации. Целью данной научной работы как раз и являлось совмещение этих двух целей: автоматической классификации авторских текстов в соответствии с заданными критериями на основании извлечённых из него закономерностей и характеристик.

Профилирование автора (классификация авторских текстов) – это задача отнесения автора исследуемого текста к заданной группе, на основании выявления значимых личностных характеристик человека. При решении подобного рода задач и автоматизации данного процесса, в большинстве случает прибегают к использованию методов на основе машинного обучения моделей классификаторов. Методика подобных моделей заключается в получении входного текста, а на выходе – получение некоторых сведений об авторе представленного текста, таких как: пол человека, возрастные характеристики, психологический образ, принадлежность к чему-либо или нечто другое.

Среди задачи профилирования авторов также выделяют такие подзадачи, как: сантимент-анализ и анализ высказываний. Здесь одним из признаков психотипа личности, является его эмоциональная окраска. К примеру, если мы к определённому психотипу автора применим критерий на основе эмоционального показателя то, можно наперёд определить его поведенческие характеристики, основываясь на эмоциональном состоянии на момент проведения исследования. Описанная методика применима в различных направлениях деятельности человека, например, проведение аналитических исследований социальных групп, посетителей онлайн магазинов, настроений избирателей или работников предприятия.

Проведение работ по выявлению семантического портрета пользователя. Анализа текстовой информации для её дальнейшей классификации по заданным правилам требует создания новых качественных алгоритмов и современных подходов. По этой причине, учитывая необходимость человека в быстром поиске и получении информации из неструктурированного вида, крайне актуальной является анализ семантического преобразования подобных текстов. Преобразование представляется как процесс переработки текста, в результате которого будет получен семантически сходный (подобный) с оригинальным документом текстовый массив. Такой массив текста является близким по смыслу к оригиналу. Применяя к практической области, возникает необходимость создания алгоритмов автоматического аннотирования, векторизации и многих иных видов исследования неструктурированных авторских массивов текстовой информации.

На момент написания работы подобные способы классификации авторских текстов с использованием средств автоматизации широко применяются в различных областях сферы деятельности человека.

Определение психотипа (социотипа) человека по написанному им неструктурированному тексту является крайне важной задачей и имеет хорошие перспективы применения в различных областях жизнедеятельности. Одним из ключевых факторов стабильной и прогрессирующей работы современного предприятия является наличие психологически устойчивого, слаженного коллектива работников. Выживание и экономическое состояние современного технологического предприятия напрямую зависит от подобного человеческого фактора. Подбор персонала по анкетам не всегда может дать качественный результат, что обусловлено различными факторами, такими как возбуждение, желание скрыть любую информацию, предать несуществующие черты характера. Кроме того, при массовом анализе с помощью анкет возникает большая нагрузка на аналитика. Гораздо более простым и незаметным способом извлечения информации о респонденте являются его высказывания, публикации, посты в социальных сетях. Существует огромное количество данных о людях в «чистом» виде. Не менее актуально и популярно определение типов личности потребителей при проведении масштабных маркетинговых исследований или рекламных кампаний. Своевременное выявление потребительских желаний и интересов даёт мощные преимущества перед кампаниями-конкурентами, приводит к улучшениям в планировании финансовой деятельности организации, лучшему сервису. Другая область-политическая борьба. Здесь огромную роль играет понимание особенности психики избирателя при проведении точечной агитации.

Как уже отмечалось выше, в Российской Федерации технология классификации и агрегирования данных о личности человека в массовом порядке практически не используется. На русском языке нет грамотных и понятных научных разработок для масс в области интеллектуального анализа данных. В связи с этим считаем, что работа в этом направлении является крайне перспективной и в будущем только станет более актуальным направлением экономического развития. Кроме того, тема исследования актуальна в связи с быстро растущей популярностью психологии в настоящее время. Современный человек начинает искать в себе ответы на многие вечные вопросы, а изучение мышления и его

различных проявлений является одним из важных векторов современных научных интересов.

Актуальность заключается в решении проблемы типирования психотипов на основе больших объёмов данных, которая возникает на предприятиях при необходимости прогнозирования поведения сотрудников, клиентов, обучающихся и иных категорий людей.

2. Заявленные цели и задачи научного проекта, предлагаемые методы, подходы, идеи, рабочие гипотезы, которые предлагались для решения задач научного проекта.

В работе было установлено, что заинтересованные организации в своей деятельности при работе с человеческим ресурсом не применяют автоматизированные подходы, направленные на исследования психологических аспектов личности своих клиентов. В частности, акцентируется внимание на возможности использования методики интеллектуального анализа (Text Mining) неструктурированных авторских текстов в среде RapidMiner с целью определения психотипов их авторов.

Целью научного проекта являлась разработка и внедрение системы автоматизированной обработки авторских текстов для определения психотипов авторов на основе анализа текстовых массивов с использованием методов машинного обучения.

Для реализации поставленной цели планировалось решить следующие поставленные задачи:

– проанализировать существующие классификации психотипов и выбрать наиболее подходящие для текстового анализа;

– исследовать и выбрать эффективные алгоритмы классификации текстов, применимые для определения психотипов;

– разработать и обучить модели машинного обучения для классификации авторских текстов по психотипам;

– применить разработанные модели для типирования текстов из ранее неклассифицированной выборки;

– оценить результаты и доработать модели для повышения точности классификации.

К предлагаемым методам для решения поставленных задач проекта относилось 5 этапов:

1 этап: анализ существующих классификаций психотипов.

Применяемые методы:

– комбинаторный подход для выбора классификаций, которые могут быть адаптированы к автоматическому текстовому анализу;

– психолингвистический анализ для выявления взаимосвязи между психотипом и языковыми характеристиками текста.

2 этап: исследование алгоритмов классификации и выбор подходящих методов.

Применяемые методы:

– сравнительный анализ различных алгоритмов машинного обучения (k-ближайших соседей, машины опорных векторов, случайные леса, нейронные сети);

– кросс-валидация для оценки точности и эффективности каждого алгоритма при работе с текстовыми данными;

– оценка сложности вычислений и адаптация алгоритмов для работы с большими текстовыми массивами.

3 этап: разработка и обучение моделей классификации текстов.

Применяемые методы:

– предварительная обработка текста (очистка, нормализация, лемматизация) с помощью методов обработки естественного языка (NLP);

– векторизация текста (TF-IDF, word embeddings) для преобразования текстов в числовые представления;

– обучение моделей машинного обучения (k-ближайших соседей, SVM, нейронные сети) на основе размеченной выборки текстов с психотипами;

– гиперпараметрическая настройка для оптимизации точности моделей.

4 этап: применение моделей для типирования авторов из неклассифицированной выборки

Применяемые методы:

– масштабируемая обработка текстов с использованием обученных моделей для автоматической классификации больших объёмов текстовых данных;

– анализ ошибок и результаты типирования для выявления возможных улучшений в работе модели;

– множественная классификация (применение нескольких алгоритмов для одной задачи) для повышения точности предсказаний.

5 этап: оценка результатов и доработка моделей

Применяемые методы:

– анализ метрик производительности моделей (точность, полнота F1-мера) для оценки;

– анализ ошибок предсказаний для выявления паттернов ошибок и коррекции моделей;

– ретренинг моделей с использованием новых данных и доработанных алгоритмов.

Гипотезы исследования:

– языковые и лингвистические особенности авторских текстов содержат характерные типы признаки, которые позволят с высокой точностью определять психотипы авторов;

– применение методов машинного обучения, таких как k-ближайших соседей, машины опорных векторов, случайные леса и нейронные сети, позволят автоматически классифицировать тексты по психотипам с точностью, сопоставимой с экспертной оценкой;

– комбинирование нескольких алгоритмов машинного обучения и использование методов векторизации текста, таких как TF-IDF и word embeddings, повысит точность классификации авторов по психотипам по сравнению с использованием одного алгоритма;

– эмоциональная окраска текста и лингвистические особенности (например, длина предложений, частота использования определённых слов) тесно связаны с психотипом автора и могут быть использованы в качестве ключевых признаков для автоматической классификации;

– использование алгоритмов интеллектуального анализа данных (Text Mining) позволят значительно сократить время и усилия, необходимые для типирования авторов текстов, по сравнению с традиционными методами, и может быть внедрено в реальные приложения для подбора персонала, маркетинга и социальной аналитики.

3. Все запланированные научные результаты достигнуты: Да.

4. Сведения о фактически проделанной работе, полученные результаты (дать описание методов проведения исследований, кратко изложить основные результаты, полученные в ходе проведения исследований, выводы и заключение по результатам исследований; привести научную новизну, теоретическую и практическую значимость работы).

Психология личности оперирует множеством классификаций, систематизирующих индивидуальные особенности человеческой психики. Эти типологии существенно различаются между собой по степени обобщения, внутренней непротиворечивости и, что наиболее важно, по своим базовым основаниям. В качестве таких оснований могут рассматриваться различные характеристики: от устойчивых черт характера и направленности внимания (вовне или на внутренний мир) до преобладающих психоэмоциональных процессов и свойств.

Были рассмотрены существующие классификации психики человека. На основании чего был сделан акцент на том, что всевозможные систематизации отличаются друг от друга по количеству обобщений, степени внутренней непротиворечивости, основаниям для систематизации и др. Важно помнить, что при классификации психотипов человеческой психики учитываются различные характеристики, такие как различия в нраве, привлекательность и забота отдельных лиц к себе или же к наружным факторам на уровне психоэмоциальной латентности человека, преобладающие в определённых психологических процессах, функциях, свойствах и иных параметров.

Наиболее фундаментальный труд относительно деятельности нервной системы и психики создал советский учёный И.П. Павлов. Равновесие, сила и подвижность возбуждающих и ингибирующих психологических процессов – это самые основные параметры, которые выделил автор. Важным показателем он считает силу нервной системы, которая определяет работоспособность кортикальных клеток и их выносливость.

И.П. Павлов, в результате своих исследований, связал типы темперамента человека со свойствами нервной системы и выделил четыре основных типа:

1) сангвиник – обладает сильной, уравновешенной и подвижной нервной системой;

2) флегматик – его нервная система также сильная и уравновешенная, но инертная;

3) холерик – характеризуется сильной, но неуравновешенной нервной системой, где возбуждение преобладает над торможением;

4) меланхолик – представляет слабый тип нервной системы.

Ещё одну распространённую типологию предложил швейцарский психиатр К. Г. Юнг (1995), описав два основных характера: экстравертный и интровертный. Базой данного подхода служит субъектно-объектная ориентация. В частности, экстравертам свойственна направленность на объект и внешний мир, факторы которого влияют на них значительно сильнее, чем собственное субъективное отношение к действительности. В отличие от них, интроверты ориентированы преимущественно на внутренний мир; их субъективное восприятие преобладает над объективными внешними обстоятельствами, а собственная психическая реальность оказывается для них важнее воздействий извне.

Заслуживает внимания и классификации психотипов К.Ф. Седовой. В её подходе основное внимание уделяется трём видам речевых стратегий, используемых в управлении конфликтами. Согласно данной концепции, выделены три типа личностей:

1) инвективный тип демонстрирует пониженную семиотичность, то есть незначительную разумность и содержательную организованность речевого поведения: коммуникативные проявления здесь становятся прямым отражением эмоциональных и биологических реакций;

2) куртуазный тип характеризуется повышенной степенью семиотичности речевого поведения, что связано со склонностью говорящего к использованию этикетных форм социального взаимодействия;

3) рационально-эвристический тип в конфликтной ситуации опирается на разум и здравомыслие, выражая отрицательные эмоции опосредованно, часто в форме иронии (Седов, 2000).

Таким образом, по мнению К.Ф. Седовой, тип человека, вовлечённого в конфликт, можно определить по избираемой им речевой тактике.

Классификация С. А. Сухих основана на теории К. Г. Юнга. По мнению автора, личность может быть описана через набор доминирующих противоположных черт: интуиция-сенсорика, логика-этика, рациональность-иррациональность. В данной типологии выделяются четыре психотипа: традиционалисты (сенсорные рационалы), реалисты (сенсорные иррационалы), концептуалисты (интуитивные логики) и идеалисты (интуитивные этики). Однако недостатком этого подхода является то, что он основан на сочетании не одной, а нескольких психических функций, что усложняет однозначную классификацию. Таким образом, для анализа будет целесообразно использовать оригинальную общепризнанную типологию К. Г. Юнга, которая, по нашему мнению, позволяет наиболее полно выразить отличительные черты психотипов (ТИМов) и провести детальную классификацию, отражающую принадлежность к той или иной социальной группе.

В рамках научной работы также были рассмотрены основные характеристики социотипов. Анализ показал, что, несмотря на общую типологическую основу, интерпретация и описание конкретных типов могут существенно различаться у разных авторов.

Для обеспечения целостности исследования далее будет использована классическая модель, основанная на типологии К. Г. Юнга, которая позволяет наиболее системно подойти к описанию шестнадцати соционических типов. Их ключевые особенности раскрываются через доминирующий способ восприятия

мира и взаимодействия с ним. Так, Габен живет в мире собственных ощущений – запахи, звуки, тактильные ощущения, как приятные, так и неприятные, формируют ткань его жизни и так откладываются в его памяти. Он физически чувствует состояния других людей, как свои собственные. Для Есенина жизнь – это поток разнообразных событий и изменений; он отслеживает внутреннюю динамику и видит возможности будущего развития, свободно обращаясь со временем. Жизнь для Жукова – поле боя, где он командует ресурсами и соратниками, чтобы достичь цели, и прекрасно чувствует расстановку сил. Вселенная Штирлица состоит из объектов, которые нужно организовать в полезный процесс; он ценит рациональность и ожидает того же от окружающих. Гамлет живет в мире эмоций, запоминая события через их эмоциональную окраску, и умеет тонко воздействовать на настроения людей. Мир для Робеспьера – это поток информации, который нужно систематизировать; он стремится к логическому порядку и справедливости. Максим Горький видит мир как свод правил и логических систем, легко вписываясь в иерархические структуры. Для Достоевского мир состоит из отношений между людьми; он чутко отслеживает их изменения и стремится к гармонии. Драйзер живет отношениями, строго разделяя людей на своих и чужих на основе их поступков. Мир для Гексли наполнен интересными возможностями и людьми, чью внутреннюю суть он стремится понять. Наполеон считает себя вправе распоряжаться всем, гибко расширяя зону влияния, но чувствуя расстановку сил. Бальзак наблюдает тенденции изменений и умеет прогнозировать развитие ситуаций. Дон Кихот живет в мире идей, оценивая все с точки зрения скрытых возможностей. Джек видит мир как поле для продуктивной деятельности, где важно найти верный момент для действия. Дюма живет в мире житейских радостей и стремится к комфорту. Гюго живет в мире радости и хорошего настроения, заряжая окружающих своим энтузиазмом и стремясь к общей гармонии.

Таким образом, в контексте поставленной задачи автоматического определения социотипа по тексту, ключевой проблемой стала адаптация существующих стилометрических методов к специфике русскоязычного контента и целям психолингвистического профилирования. Недостаточная изученность статистических закономерностей русской речи в приложении к задачам атрибуции, отсутствие тематически и стилистически сбалансированных корпусов текстов, а также необходимость учета связи между личностными характеристиками автора и его речевыми паттернами формируют комплексный вызов для исследования.

Для его преодоления требовалось решение нескольких взаимосвязанных задач:

1) формирование репрезентативного корпуса текстов. Как отмечалось, в качестве источника данных был выбран форум sociforum.su, где пользователи самостоятельно идентифицируют свой социотип. Это позволяет собрать размеченный датасет, где каждому тексту соответствует метка о социотипе его автора. Для повышения качества данных необходим строгий отбор: следует учитывать только тексты пользователей, чья типовая принадлежность не вызывает сомнений у сообщества, и анализировать сообщения, созданные в естественных условиях коммуникации (например, в дискуссиях, а не в специальных анкетах);

2) выявление и верификация стилометрических признаков для русского языка. Требуется эмпирически проверить, какие именно признаки – будь то синтаксические (длина предложений, сложность конструкций), морфологические

(частотность частей речи) или лексико-статистические (богатство словаря, использование маркерных слов) – являются наиболее информативными для различения социотипов в русскоязычной среде. Это предполагает проведение серии экспериментов по отбору признаков с использованием таких методов, как Information Gain или Chi-square.

3) учет тематической вариативности. Чтобы минимизировать влияние темы на классификацию, необходимо либо сформировать корпус из текстов, близких по тематике (например, только обсуждения межличностных отношений), либо использовать методы, устойчивые к смене темы, например, извлекать признаки, нормализованные относительно индивидуального стиля автора;

4) выбор и адаптация алгоритмов классификации. Учитывая успехи, продемонстрированные в зарубежных исследованиях, целесообразно испытать на собранных данных такие алгоритмы, как Метод Опорных Векторов (SVM), Случайный Лес (Random Forest) и нейронные сети. Особое внимание следует уделить интерпретируемости модели – пониманию того, какие именно лингвистические паттерны алгоритм связывает с тем или иным социотипом, что представляет не только прикладной, но и научный интерес с точки зрения психолингвистики.

Основной задачей семантического анализа неструктурированного текстового массива является формализация его смысловой структуры, а именно — выявление семантических единиц и установление связей между ними. Решение этой проблемы заключается в проведении процедуры семантико-синтаксического и концептуального анализа текстов. Важным инструментом автоматической семантической обработки служат мощные словари понятий, представленные преимущественно фразеологическими оборотами.

При проведении аналитической предобработки текстов необходимо учитывать, что одни и те же объекты и процессы могут иметь разный уровень обобщённости и описываться различными языковыми средствами. Следствием столь разнообразной стилистики построения текстов является необходимость учёта этого фактора при решении задач семантической обработки. Также важно принимать во внимание структурные особенности языков, в частности такие явления, как синонимия, гипонимия (родо-видовые отношения) и разнообразные способы выражения взаимосвязей между высказываниями.

Основной структурной единицей текста традиционно считается предложение. Некоторые лингвисты склонны рассматривать его в качестве основной единицы смысла. Лингвистические исследования показывают, что предложения в тексте не изолированы друг от друга, а находятся в тесной семантической связи. Эта связь основана на ментальных образах тех конкретных или абстрактных объектов (ситуаций, явлений), которые человек активирует при порождении речи. Образы этих объектов обладают определённой структурой и дополнительно организуются человеком при описании на естественном языке, что и придает тексту соответствующую структуру.

С учётом того, что тексты отражают авторские ментальные образы, отношения между предложениями выражаются с помощью разнообразных лексических структур. Связующие элементы текста отсылают к ранее упомянутым понятиям — либо буквально, либо через синонимы и эллиптические конструкции, либо через

обобщённые наименования понятий и местоимения. К таким средствам контекстной связности относятся, например, выражения, отсылающие к предыдущим фрагментам текста: «на основании вышеизложенного», «рассматриваемые», «как описано в главе», «в выражении» и т.д.

На схеме (рис. 1) представлена возможная структура базового анализа текстов на примере системы семантического анализа, в результате работы которой формируется формализованная семантическая структура текста.

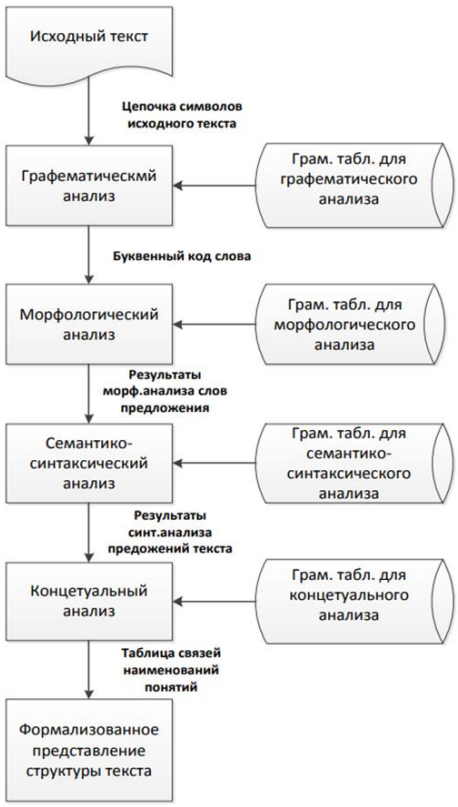


Рис. 1. Структура базового анализа текстов на примере системы семантического анализа

Были рассмотрены подробно основные процедуры семантической обработки и анализа текстов.

Гравиметрический тест специализирован для предварительного анализа слова на основе анализа существующего в нём набора символов. Способы и алгоритмы, реализующие этот тест, описаны в ряде документов. При проведении такого анализа определяется язык слова, уточняются пространства расположения слов, предложений, абзацев, группы фамилий, особых символов, охватывая цифры, даты и адреса электронной почты. Алгоритм при определении данных сведений о формальной структуре слова в соответствующих методах графематического анализа использует грядущий набор грамматических таблиц и словарей (словарь для установки языка текста; таблица элементов для выделения слов и разделителей в тексте; таблица функций для выделения дат в цифровых форматах; таблица элементов для выделения группы имен семейств и др.).

Исходный набор данных дает последовательность символов в исходном тексте.

В итоге выходные данные в качестве конечного итога представляют информацию о местоположении слов, разделителей, дат, установленных групп имен, адресов электронной почты, предложений, заголовков и примечаний.

На рис. 2 показан класс CGraphemAnalytics, который представляет основные методы графемического анализа.

Класс CGraphemAnalysis, реализующий методы графемического анализа слова.

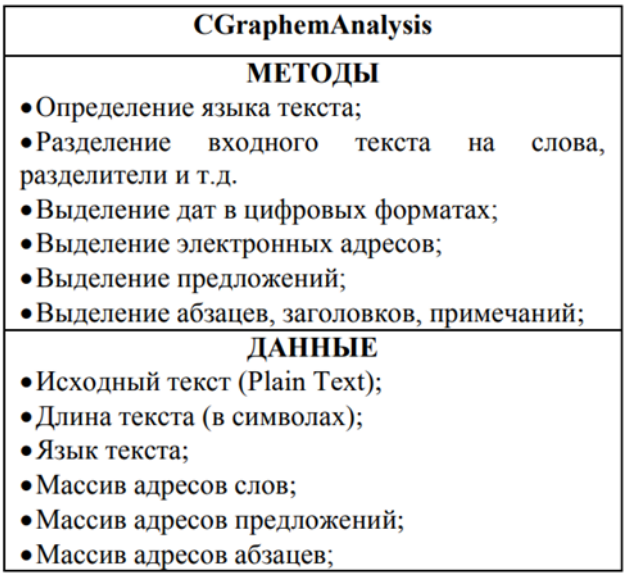


Рис. 2. Класс CGraphemAnalys

Морфологический тест слов неструктурированных текстовых массивов специализирован для определения структуры слов и присвоения им грамматических признаков, необходимых для выполнения последующих процедур автоматически обработки текстовки информации (например, морфологического синтеза слов, синтаксического анализа и синтеза текстов и их концептуального анализа). Целью этого анализа считается определение структуры слова и извлечение данных о грамматической структуре, определяющей слов за пределами контекста. Для автоматического определения указанной информации о слове в соответствующих методах морфологического анализа используются эти заключения, как:

- словарные концевые буквенные комбинации слов русского языка;
- таблица внутренних классов;
- таблица форм подающих слов;
- таблица, устанавливающая присутствие чередований в основах слов;
- поисковая таблица для реализации чередований в базах слов;
- таблица наборов грамматических сведений слов;
- таблица окончаний российских слов;
- словарь исключительных слов (коротких и казенных слов) с присвоенным GI.

Класс CMorphoAnalysis, реализующий методы морфологического анализа текста представлена в рис. 3.

CMorphoAnalysis
МЕТОДЫ • Поиск в словаре слов-исключений; • Поиск в словаре конечных буквосочетаний слов; • Поиск в таблице супплетивных форм слов; • Установления наличия чередований в основах слов; • Определение флективного класса слов; • Назначение слову грамматической информации;
ДАННЫЕ • Буквенный код слова; • Длина слова; • Длина окончания; • Лексикограмматический класс слова; • Флективный класс слова; • Наборы грамматических признаков (род, число, падеж, лицо);

Рис. 3 Класс CMorphoAnalysis

Исходные данные – это набор символов текущего слова.

Вывод – это информация о структуре слова.

Семантически-синтаксический тест предложений в тексте.

Целью семантического и синтаксического анализа текстов считается представление слова в формализованной форме, что достигается путём выделения в нём семантических единиц и установления связей меж ними. При этом конструкция текстов может быть интерпретирована по-разному и описана в различных формализованных языках. При описании синтаксической структуры текстов удобно опираться на некоторую формализованную модель её, к примеру, модель дерева зависимостей. Данная модель предполагает построение каждого предложения в облике дерева, узлами которого считаются слова. Слова соединены краями направленного графа, выражающего отношения прямого господства, и ориентированы от подчиняющего (определяемого) слова к подчиняющему (определяющему). Граница деления слова на набор отношений может быть довольно глубоким. Но, в то же время, глубина дифференцирования впрямую воздействует на будущий процесс описания текстов. В итоге этого анализа определяется синтаксическая конструкция предложения: производится дележ на простые предложения, определяются основные и вторичные члены предложения и уточняются семантические связи меж ними, строится дерево зависимости предложения и для каждого слова определяется однозначная грамматическая информация, соответствующая контексту. Для автоматического определения информации о структуре заданного предложения в соответствующих методах используется грядущий набор грамматических таблиц и словарей:

– таблица описаний правил синтаксического анализа;

– таблица описаний правил установления семантических связей меж словами предложения;

– таблица описаний правил определения однозначной грамматической информации слов с учетом контекста.

На рис. 4 представлен класс CSyntAnalysis, реализующий методы семантико-синтаксического анализа текста.

CSyntAnalysis
МЕТОДЫ • Членение на простые предложения; • Определение главных членов предложения; • Определение второстепенных членов предложения; • Установление однородных второстепенных членов предложения; • Построение дерева зависимостей; • Разрешение многозначности грамматической информации слов;
ДАННЫЕ • Адрес предложения; • Длина предложения; • Массив адресов простых предложений; • Массив адресов словосочетаний; • Массив адресов слов;

Рис. 4 Класс CSyntAnalysis

Входящий материал представляет собой слова предложения и итоги их обработки с помощью процедуры морфологического анализа. Выход – информация о семантически-синтаксической структуре предложения.

Сущность концептуального анализа текстов уменьшена до методологии, задачка которой состоит в том, чтобы определить семантическую структуру текстов, определить их концептуальный (концептуальный) состав и установить связи меж наименованиями понятий. Данную проблему невозможно решить, только проанализировав синтаксическую структуру текстов без вербования семантических особенностей. Сложность этой проблемы связана, до этого всего, с изменчивостью форм представления заглавий понятий в текстах. Наиболее действенный заключением стал метод концептуального анализа текстов с контролем над тезаурусом (справочным словарём), подключающий более 1,8 млн понятий и более 400 тыс. связей меж ними. В итоге этого анализа в тексте открываются наименования понятий, уточняется концептуальная конструкция слова и строится таблица связей меж наименованиями понятий. Для определения базовых отношений в тексте в этом подходе используются грамматические таблицы и словари:

– справочный словарь заглавий понятий;
– словарь семантических отношений меж наименованиями понятий;
– словарь семантических связей слов.

На рис. 5 представлен класс Cconcertananalysis, реализующий методы концептуального анализа слова.

Таблица отношений заглавий понятий – не что иное, как детальнее представление семантической структуры слова. При визуализации этой семантической структуры формируют семантическую карту в облике направленного графа, узловыми веществами которого считаются объекты, события или же темы документов, а дуги – семантические отношения меж ними. Соединения могут быть либо типизированными (определён семантический образ соединения), либо логическими (установлен прецедент их существования). Семантическую карту

можно построить с помощью пакетов утилит Graphviz, SPSS, RapidMiner. Это еще позволяет получать более сложные выходные данные, к примеру, использовать координатную сетку, которая вслед за тем может использоваться для указания областей при отображении на гипертекстовой страничке.

Рис. 5. Класс CConceptAnalysis

Замеры быстродействия выполнения процедур семантического анализа текстов и длительности их выполнения представлены в таблице 1.

Таблица 1

Название процедуры семантического анализа текста	Скорость обработки (слов/сек)	Длительность процесса (в %)	Скорость обработки (слов/сек)	Длительность процесса (в %)
	Текущая версия		Перспективная версия	
Графематический анализ	52496	5%	27342	4%
Морфологический анализ	43746	6%	36455	3%
Семантико-синтаксический анализ	7954	33%	7291	15%
Концептуальный анализ	4687	56%	1402	78%
Система семантического анализа (общее быстродействие)	**4780**	**100%**	**1093**	**100%**

Был рассмотрен тест классификаторов неавтоматического обучения.

Классификатор – это инструмент интеллектуального анализа данных, который воспринимает определённый объём данных, представляющих то, что мы хотим классифицировать, и пробует предвещать, к какому классу эти новые данные должны принадлежать. Рассмотрим наиболее часто используемые, согласно исследованиям, алгоритмы систематизации C4.5.

C4.5 – это алгоритм, используемый для формирования дерева решений, разработанного Р. Квинланом. C4.5 является расширением более раннего алгоритма ID3 Квинлана. Для классификации могут использоваться деревья решений, созданные с помощью C4.5, и по этой причине C4.5 часто называют статистическим классификатором. В 2011 году авторы программного обеспечения машинного обучения Weka описали алгоритма C4.5 как «знаковую программу дерева решений,

которая, вероятно, является рабочей лошадки машинного обучения, наиболее широко используемой на практике на сегодняшний день».

C4.5 строит деревья решений из набора обучающих данных таким же образом, как и ID3, используя концепцию информационной энтропии. Обучающие данные представляют собой набор $S = s_{1,}s_{2,} \ldots, s_n$уже классифицированных выборок. Каждый образец s_n {s_{i}} состоит из p-мерного вектора $(x_{1,i,}x_{2,i,} \ldots, s_{p,i})$, в котором x_i представлены значения атрибутов или признаков образца, а также класс, в который s_i он попадает.

В каждом узле дерева C4.5 выбирает атрибут данных, который наиболее эффективно разбивает его набор выборок на подмножества, обогащенные в том или ином классе. Критерием расщепления является нормированный прирост информации (разница в энтропии). Для принятия решения выбирается атрибут с наибольшим нормированным информационным выигрышем. Затем алгоритм C4.5 рекурсирует в секционированных подлистах.

Этот алгоритм имеет несколько базовых случаев.

Все образцы в списке принадлежат к одному классу. Когда это происходит, он просто создает конечный узел для дерева решений, говорящего, чтобы выбрать этот класс.

Ни одна из функций не обеспечивает получение какой-либо информации. В этом случае C4.5 создает узел решения выше по дереву, используя ожидаемое значение класса.

Обнаружен экземпляр ранее невидимого класса. Опять же, C4.5 создает узел решения выше по дереву, используя ожидаемое значение.

Рассмотрим псевдокод. В псевдокоде общий алгоритм построения деревьев решений представлен следующим образом:

1) проверьте для вышеуказанных базовых случаев;

2) для каждого атрибута *a* найдите нормализованный коэффициент усиления информации от разбиения на *a*;

3) пусть *a_best*-атрибут с наибольшим нормализованным коэффициентам усиления информации;

4) создайте узел принятия решений, который разделяется на *a_best;*

5) повторите на подписках, полученных путем разбиения на *a_best,* и добавьте эти узлы в качестве дочерних узлов узла.

Наибольшим преимуществом использования деревьев решений, является относительная простота понимания принципа их работы и использования. Кроме того, они достаточно быстры при использовании, весьма распространены и их вывод удобочитаемый для человека.

Метод опорных векторов (SVM) строит гиперплоскость для классификации данных на несколько классов. Работа алгоритма SVM схожа с C4.5, отличается в основном тем, сто при выполнении задачи не использует деревья решений.

Прежде чем классификатор приступит к выполнению задачи классификации его необходимо обучить на множестве специально подготовленной выборке документов (примеров), которая состоит из двух частей – примеры с положительной и негативной окраской. Алгоритм использует метод обучения с учителем «обучение с учителем т.е. каждый учебный пример представляет собой пару <учебный вход, правильный ответ>. При работе с подготовленной выборкой

алгоритм постоянно производит коррекцию показателей SVM, подстраиваясь под учебный набор.

Алгоритм обучения находит между элементов учебного набора точки, лежащие на границе 2-ух подмножеств (положительных и отрицательных) и строит разделительную плоскость меж данными тачками. В определениях SVM эти точки именуются опорными векторами.

В рамках научной работы была рассмотрена работа алгоритма на простом примере. Так, пусть имеется некоторое количество красных и синих шаров на поверхности. Если шары не слишком сильно перемешаны, то между ними с лёгкостью можно расположить линию разделения. Когда на стол помещается новый шар, принимая во внимание, с какой стороны палки он в данный момент, можно предвещать его оттенок.

На рис. 6 представлена модель опорных векторов. Данный классификатор описывается следующими соотношениями:

$$h(x) = sign(u(x))$$

$$u(x) = \sum_{i=1}^{m} \lambda_i y_i K(x_i, x) - w_0$$

где w_0 – порог (свободный член); λ_i – коэффициент; x_i, y_i – пара из учебной выборки, (векторы x_i, для которых $\lambda_i > 0$ называются опорными векторами); $K(x_i, x)$ – функция ядра.

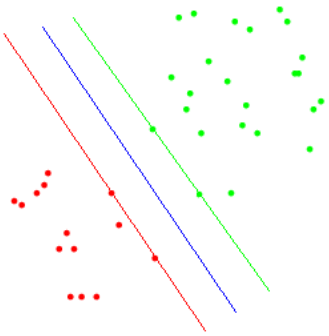


Рис. 6. Модель опорных векторов

Прямая – это линия, круги – точки данных, где синий и алый – два класса.

Если шары перемешаны между собой, прямая гиперплоскость не сработает.

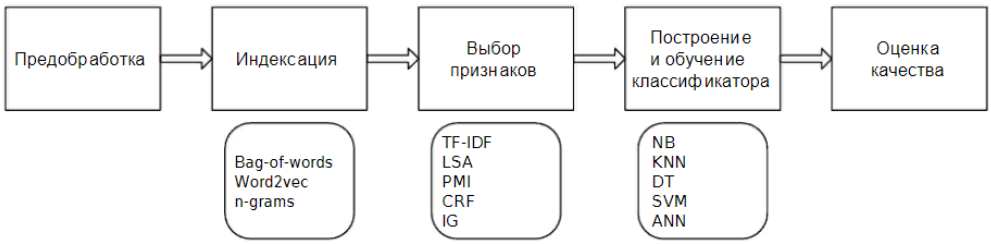
Рассмотрим пример с пациентов клиники. Пусть у больного есть набор данных, у этих данных много параметров пульс, холестирин, артериальное давление и т.д. Каждый из данных является измерением n-мерного пространства.

SVM сравнит их с высшим измерением, а затем найдет гиперпластическую кость для разделения классов.

Являясь по своей сути алгоритмом, предполагающим обучения с учителем, то заранее классифицированная выборка данных используется алгоритмом для проведения обучения SVM заданным классам. SVM сможет производить классификацию новых текстов только после подобного обучения.

Данный метод реализован в следующих платформах: IBM SPSS MODELLER; RapidMiner; Orange; Weka; MATLAB; SAS.

На этапе предварительной подготовки текстового массива необходимо было выполнить следующие преобразования: проведение токенизации слов с помощью встроенных токенайзеров, удаление функционально значимых слов – не имеющих влияния на общую картину классификации: союзы, предлоги, артикли, имена собственные и др. После производятся множественные морфологические преобразования, состоящие в разметке текстов по частям речи и стемминге. Цель данных преобразований является оптимизация векторного пространства свойств



объекта.

В следствии данных преобразований остаются лишь значимые слова – токены, определяющие весомые признаки принадлежности. После необходимо было провекторизировать тексты. Данная процедура представляет из себя преобразование текстовой информации в числовой вид (числовые векторы), что даст возможность проведения дальнейшей машинной обработки.

Рис. 7. Этапы классификации текста

Проанализировав множество алгоритмов и методов проведения классификации текстов, мы остановили свой выбор на четырёх методах векторизации.

При этом необходимо было выполнить следующие шаги:

1) перевод текста в набор числовых векторов, каждый из которых представляет частотную характеристику какого-либо из явлений;

2) второй вариант предполагал создание матрицы слов, которых встречались во всех текстах не менее одного раза, после чего с помощью метрики TF-IDF происходило вычисление их частоты;

3) перевод всех базисных слов из текста в семантические вектора, состоящие из некоторого количества чисел.

Суть первого способа обращения текстов в векторное пространство заключается в представлении каждого из них в виде характеристик относительной частотности явлений таких как: употребление союзов, глаголов и отглагольных наречий, сложных и простых предложений и др., характерных для каждого из социотипов характеристик.

На основе подобных маркеров можно, например, делать выводы о принадлежности автора к группе логиков или этиков.

Путём длительной обработки текстов был составлен набор признаков наиболее характеризующих социотипы, при котором сохраняется баланс между высокой точностью и достаточным временем работы классификатора.

В набор попали, к примеру, такие термины (ниже более детальное описание): знаки препинания, служебные части речи, униграммы, ласкательные суффиксы,

средняя длина слова, средняя длина предложения, правильные глаголы, именные части речи.

Следующий способ обработки текстов носит название «мешок слов» – Bag-of-Words. Это наиболее простой и, в тоже время, наиболее стабильный алгоритм, предполагающий выделение наиболее часто встречающихся и значимых слов. Характерных каждому социотипу. При этом совершенно не учитывается порядок следования элементов (слов).

Другими словами, каждый документ – это вектор в многомерном пространстве, координаты которого соответствуют номерам слов, а значения координат – значениям весов.

Ниже приведены основные этапы подготовки (индексация и взвешенные признаки):

1) создать массив слов, встречающихся в текстовом корпусе обучающей выборки;

2) определить показатель матрицы TF-IDF для каждого слова из обучающего корпуса текстов для каждого социотипа. Метрика TF-IDF выявляет частотность слов в тексте, беря за основу меру «уникальности» каждого из слов во всем корпусе:

– ITF-IDF = ITF * IDF, где ITF – относительная частота слова в тексте, а IDF можно представить следующим образом;

– WIDF = $log\left(\frac{n}{a}\right)$, где n – размер корпуса, а – число текстов, в которых встречается данное слово.

Идея следующей методики базируется на идее представления слов в векторное пространство – «погружение» слов. При этом слова естественного языка посредством определённых моделей транспонируются в числовое векторное представление, причём, различной (произвольной) размерности. Среди различных критериев близости нами был выбран критерий отображения векторов на основе их семантической близости. То есть, задача векторизации слов заключается в том, чтобы векторы наиболее похожих в семантическом плане лексем расположились как можно близка друг к другу в некоем заданном n-мерном векторном пространстве.

Разработчиком данной методики является один из ведущих специалистов компании Google – Томас Миколов. Его идея состоит в том, чтобы располагать тем ближе друг к другу вектора слов, чем более схожи контексты, в которых они употребляются.

Для векторизации слов корпуса текстов по методике Т. Миколова, была обучена модель word2vec. Для её обучения пришлось воспользоваться готовыми векторными репрезентациями, представленными в 300-мерном пространстве для более, чем 150 000 русских слов, расположенными на ресурсе национального корпуса русского языка.

Рис. 8. Графическая интерпретация семантических связей

Альтернативный вариант, рассмотренный в нашем исследовании, уменьшения векторного пространства называется методом латентно-семантического анализа (LSA). Данный алгоритм выполняет сингулярное разложение матриц.

Векторные представления fastText в достаточной степени сходно с технологией, применяющейся в пространстве word2vec, за исключением некоторого отличия, которое состоит том, что fastText учитывает не только векторизованные слова, но и символьные n-граммы. Благодаря чему представляется возможным вычислить векторные представления неизвестных слов. Необходимо акцентировать внимание на том, что данная модель была обучена на корпусе слов русской Википедии.

Таким образом, в работе использовалось всего 4 способа векторного представления текста на русском языке (см. табл. 2).

Таблица 2

Алгоритм	**Характеристика**
n-gram	представление в виде ряда статистических морфологических, синтаксических, экстралингвистических и некоторых других характеристик
TF-IDF	разреженная матрица на основе технологии bag-of-words
word2vec	модель погружений слов
fastText	модель погружений слов

Далее необходимо было провести векторизацию текстов по данным моделям (алгоритмам) и определить эффективности при классификации, применяя так же различные методы классификации.

Перед преобразованием в векторный вид все текста были подвержены предварительной обработке:

1) все символы были приведены к строчному представлению;

2) все слова были очищены от лишних символов и отделены от знаков препинания;

3) все слова были лемманизированы и подвержены грамматическому анализу с помощью блока морфологического анализатора.

В дополнении к этому, в случае 3 и 4 метода векторизации, тексты были очищены от стоп-слов – всех слов короче 3 символов.

Помимо непосредственно исследования способов и применения методик к автоматической классификации авторских текстов, важнейшим моментом данной работы является построение моделей, использующих в своей работе не один, а несколько методов исследования текстовой информации. И, соответственно,

получения наиболее качественной методики определения соципотипов авторства. Учитывая большое разнообразие социотипов, и как следствие, наличие достаточно малого люфта между основными показателями речевых характеристик, интерес заключался в проведении сревнительного анализа в комбинации разных классификаторов и способов векторизации. В данном исследовании классификация осуществлялась способами, которые можно поделить на две группы:

1) случайный лес, метод ближайших соседей, машина опорных векторов;

2) LSTM, GRU, полносвязная многослойная нейронная сеть. В первую группу вошли классификационные методы собственно машинного обучения, во вторую – нейронные сети различной архитектуры.

Перед использованием методов первой группы вся текстовая база (выборка) делится на две части: тестовую и тренировочную. После чего процесс проводится в два этапа: сначала классификатор «обучается» на тренировочной выборке, выделяя ключевые признаки и соотнося их числовые значения 16 социотипам авторов всех текстов заданной группы, а затем тестируется на тестовых текстах, отсутствующих в базе данных. В следствии чего, первый этап работы классификатора может быть назван тренировочным, а второй – тестовым.

Перед началом обучения значения на тестовой выборке, векторизованные леммы, необходимо нормализовать, т.е. привести их весовой коэффициент к единому стандарту. В данном исследовании использовалась минимаксная нормализация. Это значит, что нормализация каждого параметра проводился по формуле $V' = (V_i – V_{min}) / V_{max} – V_{min})$, где i – число всех параметров (в данном случае 13), V_i – значение текущего параметра, V_{max} – максимальное значение данного параметра среди всех текстов, V_{min} – минимальное значение данного параметра среди всех текстов. Результатом каждого преобразования становится «нормализованное», т.е. принадлежащее промежутку [0,1] значение V'.

Следующий шаг – проведение тестирования алгоритмов на второй части текстов, не вошедших в базу тестовых данных. Эта процедура включает вычисление значений различных параметров векторов, их нормализацию и, основанное на полученных данных, попытке предсказания принадлежности автора текста к одному из описанных социотипов. В данном алгоритме валидация работы алгоритма на «новых» текстах производилась по методу кросс-валидации:

1) 320 текстов делятся на 16 частей по 20 текстов;

2) каждая часть поочередно удаляется из базы данных, после чего программа заново обучается на измененной базе и выдвигает гипотезу о каждом из параметров социотипа автора: то есть, проводит бинарную классификацию по шкалам на соответствие одному из 16 социотипов;

3) успешность установления психологического параметра в каждом случае определяется по метрике F1.

Модель выдвигает предположение о принадлежности автора текста заданному классу, сравнивая его с текстами из базы данных. Таким образом, задача классификации сводится к определению степени схожести числового вектора анализируемого текста с векторами текстов разных типов. В данном исследовании эта задача решена 3 методами машинного обучения.

Рассмотрим метод нейронных сетей для классификации текстов.

Концепция нейронных сетей основана на модели обработки информации человеческим мозгом. Искусственная нейронная сеть состоит из нейронов-неких абстрактных ячеек, которые могут принимать информацию в числовом виде, выполнять над ней определенные арифметические операции, нормализовывать и передавать полученные значения дальше. В общем случае нейроны в структуре сети объединяются в слои нескольких типов: нейроны входного слоя получают обучающие данные и передают их в неизменном виде дальше-на вход нейронов скрытых слоев. Именно на уровне скрытых слоев происходят основные вычисления, позволяющие выявить закономерности между данными, с помощью которых можно, например, осуществить их классификацию. В этом случае нейроны выходного слоя дают информацию о принадлежности объекта к определенному классу.

Существуют связи между нейронами разных слоев, которые имеют некоторый вес. Соответственно, значения самих нейронов при обучении умножаются на вес связей, и задача нейронной сети (как, собственно, и любой другой системы машинного обучения) сводится к определению наиболее удачной комбинации весов, позволяющей добиться максимальной точности и минимальной погрешности классификатора. Если значение входных нейронов, т. е. обучающих данных, является постоянным, то веса выступают в качестве коэффициентов, значения которых изменяются на каждой итерации (один цикл обучения). Помимо коэффициентов по аналогии со свободным членом в других системах машинного обучения некоторые ячейки нейронной сети называются нейронами смещения и выполняют аналогичную функцию. Типы нейронных сетей различаются в зависимости от их архитектуры, которая определяется количеством, направлением, характером связей между нейронами разных слоев.

В работе использовались семантические сети 3-х типов:

1) многослойная полносвязная сеть прямого распространения (Naïve Base), в которой каждый нейрон соединен со всеми нейронами предыдущего слоя (Хайкин, 2006). Экспериментально были выбраны следующие параметры, необходимые для настройки и построения нейронной сети данного типа:

– 3 скрытых слоя;

– количество нейронов в скрытых слоях (300, 250 и 150);

– функции активации чередуются в скрытых слоях: ReLu (линейная ректификационная единица) по формуле f(x) = max (0, x); сигмоидальная функция активации определяется формулой $\sigma(x) = 1/(1+e^{(-x)})$;

– 10 эпох обучения;

– 5 итераций на каждую эпоху;

– размер пакета (размер так называемого «batch size», т.е. единого обучающего набора пакета данных) 24;

– loss (функция потери, т. е. функция вычисления ошибки классификации) – двоичная перекрестная энтропия. Если модель предсказывает значение p, тогда как истинное значение равно t, то двоичная перекрестная энтропия может быть вычислена по формуле – *t*log(p)-(1-t)log(1-p);*

– оптимизатор (функция оптимизации нейронной сети) – Adam (один из самых популярных встроенных оптимизаторов в keras).

2) рекуррентная нейронная сеть (LSTM), основанная на механизме обратного распространения ошибок – это означает, что информация, полученная на одном из скрытых слоев, в некоторой степени сохраняется и учитывается при расчетах на последующих слоях. Такие сети используются при обработке последовательностей, в которых важен порядок элементов. Примером такой последовательности является текст на естественном языке, который является объектом нашего исследования.

На самом деле, существует несколько типов рекуррентных нейронных сетей, и biLSTM и GRU используются в этом исследовании. Структура сети biLSTM (bidirectional long short – term memory – двунаправленная длительная краткосрочная память) усложняется следующим образом: во-первых, каждый нейрон характеризуется дополнительными фильтрами/клапанами ввода, забывания и вывода.

С помощью этих фильтров сеть учится определять, какая информация из предыдущих слоев хранится и передается дальше, ориентируясь на специальные параметры-постоянное и скрытое состояние каждой клетки, т. е. каждого нейрона. Внутреннее скрытое состояние вычисляется на основе текущего входного сигнала и предыдущего скрытого состояния. Структура хранения информации о скрытых состояниях ячеек делает такую сеть устойчивой к проблеме исчезающего градиента.

В нашем случае сеть (LSTM) также является двунаправленной – она учится учитывать информацию как с предыдущего, так и с последующего уровней, как бы производя обучение в обоих направлениях. Это реализуется с помощью двунаправленной оболочки для слоя lstm из библиотеки keras. В дополнение к двунаправленному рекуррентному слою сеть имеет следующие характеристики:

– длина входной последовательности равна 51 для частотных характеристик (по числу характеристик данных, перечисленных ранее) и 15091 для представления TF-IDF (по числу лексем в словаре);

– размер встраивания (глубина погружения) 128;

– 3 скрытых слоя с числом нейронов 256, 128, 64;

– отсев («сбрасывание» некоторых случайных величин во время тренировки, чтобы предотвратить перетренированность) 0,3;

– 10 эпох обучения;

– 5 итераций на каждую эпоху;

– размер пакета (размер так называемого «batch size», т.е. единого обучающего набора, пакета данных) 24;

– loss (функция потери, т.е. функция вычисления ошибки классификации) – двоичная перекрестная энтропия. Если модель предсказывает значение p, тогда как истинное значение равно t, то двоичная перекрестная энтропия может быть вычислена по формуле: $t*log(p)-(1-t)log(1-p)$.

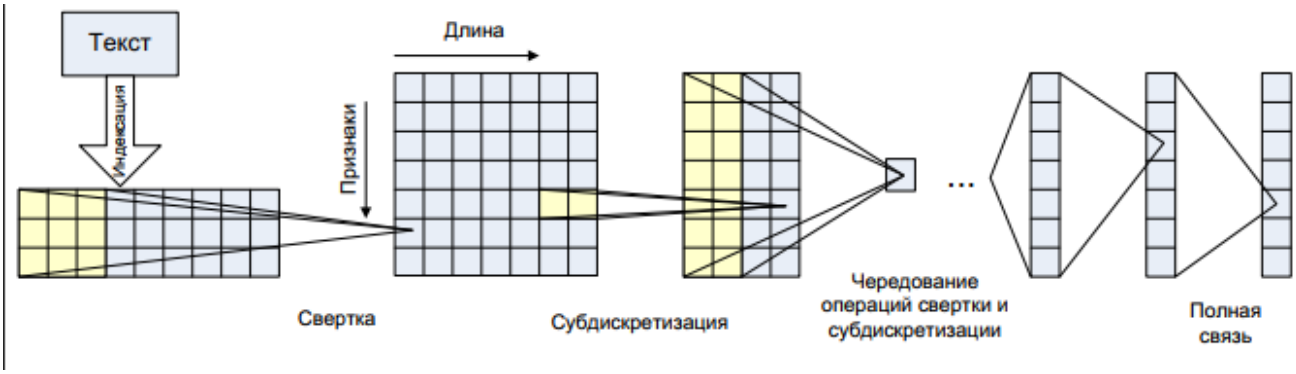


Рис. 9. Обобщённая схема сверхточной нейронной сети (LSTM сети)

3) сеть GRU (gated rucurrent unit – управляемый рекуррентный блок) имеет более простое устройство для реализации механизма запоминания, чем LSTM, и учится быстрее, но также устойчива к проблеме исчезающего градиента. В ячейке GRU есть только два клапана: сброс и обновление. Управляемый блок обновления определяют, какую часть предыдущего сохраненного значения сохранить, а управляемый блок передачи определяют, как объединить предыдущую сетевую память с новым входом. Эта сеть задается параметрами, аналогичными параметрам LSTM, но она не является двунаправленной. Таким образом, при обучении учитывается только прямой порядок элементов.

В результате научной работы были построены модели классификаторов, основанные на использовании методов машинного обучения и иным методах.

Результаты классификации удобно представлены в двух таблицах, каждая из которых относится к одному бинарному психологическому параметру: интуиция-сенсорика или логика-этика. Размерность обеих таблиц составляет 4x5, столбцы представляют методы векторизации, а строки относятся к методам классификации, т. е. методам машинного обучения и типам архитектуры нейронных сетей, которые применялись в практической части исследования.

В таблицах жирным шрифтом выделены те методы классификации и методы векторного представления текстов, при которых достигаются наиболее высокие показатели точности. Последние, соответственно, представлены на пересечении столбцов с методами классификации и строк с методами векторизации.

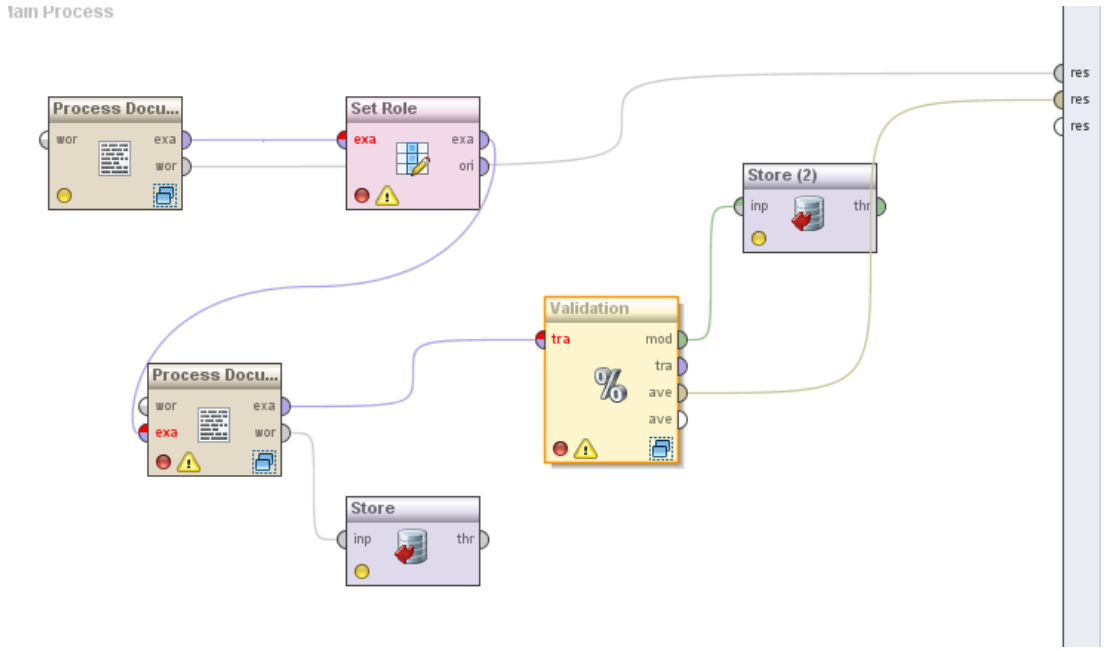


Рис. 10. Алгоритм предобработки и векторизации текстов

Таблица 3

Частотные показатели распределения токенов в соответствии с правилами

Text View | Plot View | Distribution Table | Annotations

Attribute	Parameter	жуков	максим	штирлиц	габен	джек	робеспьер	дюма	дрейзер	наполеон	гамлет	есенин	дон кихот	бальзак	дост
chilli	mean	0	0	0	0	0	0.028	0	0	0	0	0	0	0	0
chilli	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
redic	mean	0	0	0	0.031	0	0	0	0	0.029	0	0	0	0	0
redic	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
simtim	mean	0	0	0	0	0	0.046	0	0	0.024	0.019	0	0	0	0
simtim	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
stud	mean	0.074	0	0	0	0	0	0	0	0	0.074	0	0.022	0	0
stud	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
абзаца	mean	0	0	0.030	0	0	0.028	0	0	0	0	0	0	0	0
абзаца	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
абсолютно	mean	0.025	0.006	0.011	0.006	0	0.010	0	0	0.005	0.013	0.023	0.005	0.004	0.00
абсолютно	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
абсолютное	mean	0	0	0	0	0.030	0	0	0	0	0	0	0	0	0
абсолютное	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
абстрактно	mean	0	0	0	0	0	0	0	0	0	0.019	0.050	0	0.019	0
абстрактно	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
абстрактног	mean	0	0	0.030	0	0	0	0	0	0	0	0	0	0.023	0
абстрактног	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
абстрактны	mean	0	0	0	0	0	0	0	0	0	0	0.025	0.022	0	0
абстрактны	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
абстрактны	mean	0	0	0	0	0	0.028	0	0	0	0	0	0	0.023	0
абстрактны	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
автоматиче	mean	0	0	0	0	0	0	0	0	0	0	0.031	0	0.023	0
автоматиче	standard de	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.00
автор	mean	0	0	0.017	0.036	0	0.033	0	0	0.017	0	0	0	0	0

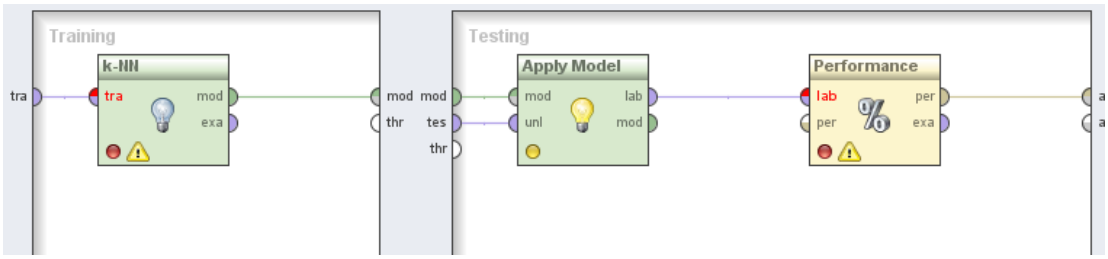


Рис. 11. Модели классификации на основе KNN

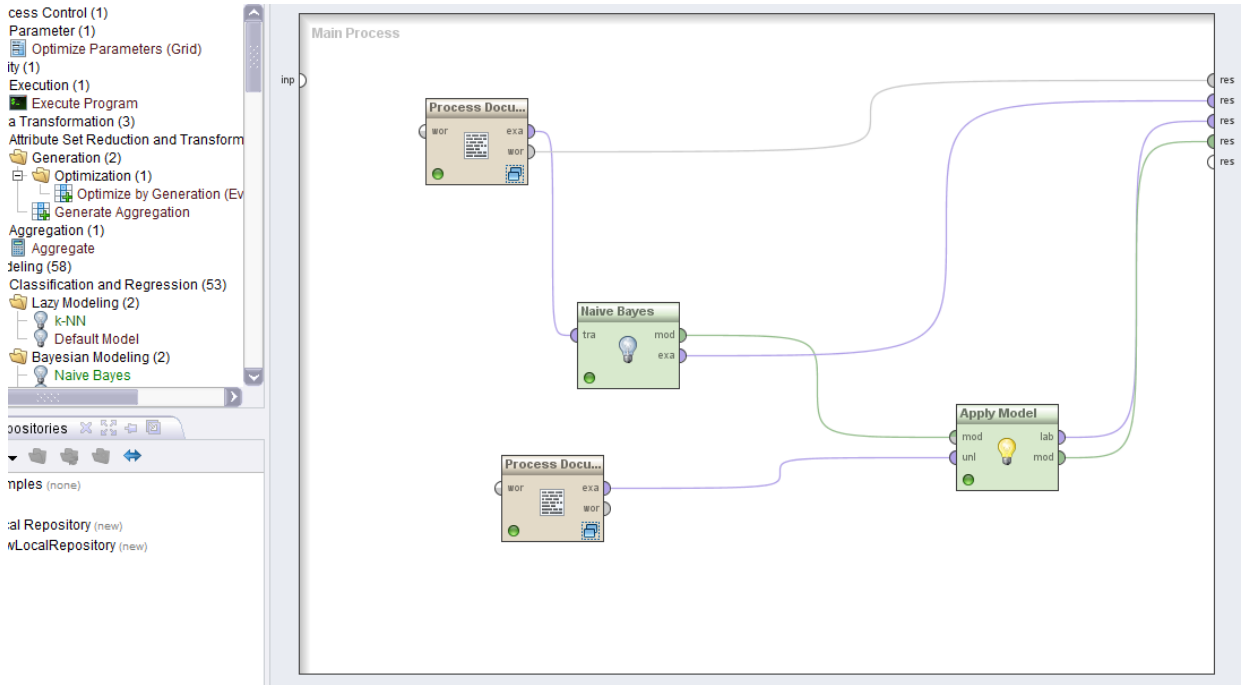


Рис. 12. Модели классификации на основе Naive Base

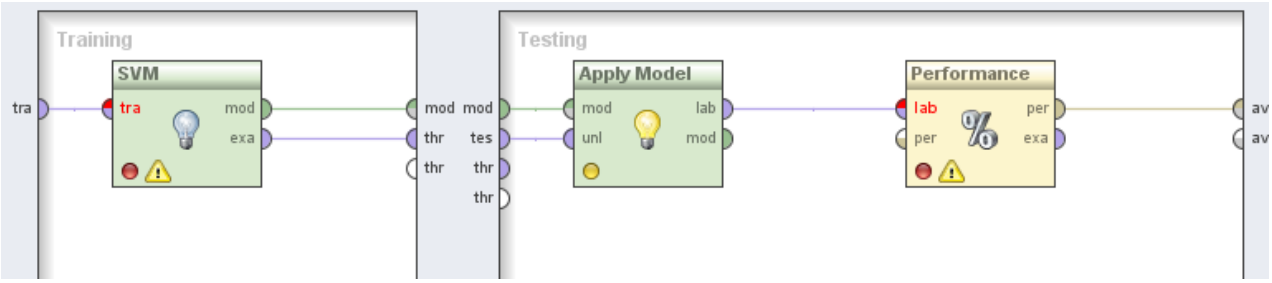


Рис. 13. Модели классификации на основе LSTM

Таблица 4

Результаты классификации по шкале логики-этики на основе методов машинного обучения

	TF-IDF	**Частотные черты**	**Word2Vec**	**FastText**
Naive Base	0,53	0,46		
biLSTM	0,43	0,66	0,52	0,54
GRU	0,4	0,61	0,41	0,38

Как видно из таблицы, наилучший результат в 66% достигается при использовании сети LSTM на материале текстов, преобразованных в числовые векторы, путем выделения 50 грамматических, синтаксических и экстралингвистических характеристик и подсчета их частотных индексов. Следующий результат 61% достигается при использовании сети GRU с тем же методом векторизации.

Вспомним полный перечень используемых признаков: ряды местоимений; сравнительные и переходные степени прилагательных и наречий; тип, наклонение и переходность глаголов; причастные залоги; композиционные и подчинительные союзы; знаки препинания; символические униграммы; улыбки; средняя длина слов

и предложений; соотношение глагольной и именной частей речи; соотношение служебной и самостоятельной частей речи.

Таким образом, перечисленные признаки (или, по крайней мере, некоторые из них) обозначают такой психологический признак, как логическая или этическая направленность сознания.

Таблица 5

Результаты классификации по шкале интуиции-сенсорики на основе методов машинного обучения

	TF-IDF	**Частотные черты**	**Word2Vec**	**FastText**
Naive Base	0,69	0,49		
biLSTM	0,57	0,45	0,7	0,61
GRU	0,54	0,42	0,66	0,62

Небольшая классификация – 72% – достигается при классификации по алгоритму случайного леса с векторизацией текста методом TF-IDF. Следующий результат в 70% показывает рекуррентная нейронная сеть LSTM с векторизацией по модели Word2Vec. Таким образом, печатать на шкале интуиции-сенсорики можно с большей точностью.

Известно, что метод Word2Vec позволяет сопоставить слово с некоторым контекстным вектором в многомерном семантическом пространстве. Таким образом, на выходе получаем матрицу распределения семантических полей текста. Можно сделать вывод, что это распределение маркирует признак интуиции-сенсорики.

Наконец, высокая точность метрики TF-IDF, ранжирующей слова по критерию информационной значимости, т. е. их редкости в корпусе и частотности в конкретном тексте, свидетельствует о корреляции этой психической функции с лексикой автора, соотношении редких и уникальных для корпуса в целом слов, используемых им в тексте. Конечно, в этом случае необходимо учитывать размеры собранного корпуса и его возможную непредставительность по указанному критерию.

При сравнении моделей были выявлены следующие достоинства и недостатки:

Таблица 6

Достоинства и недостатки наивного байесовского классификатора

Достоинства	Недостатки
Несмотря на свою простоту, данный классификатор часто работает не хуже, а то и лучше более сложных алгоритмов	Классификатор считает, что все атрибуты объектов являются независимыми между собой, что на практике труднодостижимо

Таблица 7

Достоинства и недостатки нейронной сети

Достоинства	Недостатки
Нейронная сеть может найти закономерности в достаточно сложно структурированных данных, при этом показать очень неплохие результаты.	Алгоритм не так прост, как тот же баесовский классификатор, нужно быть аккуратным.

Устойчивость при большом числе шумовых входных сигналов. Нет необходимости как-то предварительно просеивать данные, выделять среди них эталонные, нейросеть сама в состоянии определить малопригодность данных для стоящей перед ней задачи.	Трудное представление процесса формирования результатов, в отличии от того же дерева решений.
Благодаря распараллеливанию вычислений есть возможность реализовать достаточно эффективный алгоритм.	Отсутствие строгой теории. На сегодняшний день создано большое количество модификаций стандартной нейросети.

В рамках выполнения научного проекта «Исследование способов и автоматизации обработки авторских текстов для определения психотипов» были успешно решены поставленные задачи и достигнута основная цель – разработана и апробирована система автоматизированной обработки авторских текстов для определения психотипов (социотипов) их авторов с использованием методов машинного обучения и интеллектуального анализа данных.

Проведенное исследование подтвердило ключевую рабочую гипотезу о существовании устойчивой связи между лингвистическими особенностями письменной речи и психологическими характеристиками личности. В ходе работы был выполнен комплексный анализ существующих психологических и соционических классификаций, в результате которого для дальнейшего автоматического анализа была выбрана типология К.Г. Юнга, адаптированная в соционике к 16 типам, как наиболее релевантная для задач текстовой атрибуции.

Центральным практическим результатом проекта стало создание и сравнительное тестирование ряда моделей машинного обучения для классификации текстов. В процессе исследования были реализованы и оценены различные подходы:

1) векторизация текста: апробированы четыре метода преобразования текстовой информации в числовые векторы: на основе частотных грамматико-синтаксических признаков, TF-IDF, word2vec и fastText.

2) классификация: применены как традиционные алгоритмы машинного обучения (Наивный Байес, Метод опорных векторов, Случайный лес), так и современные архитектуры нейронных сетей (многослойный перцептрон, двунаправленные LSTM и GRU).

Наиболее значимые результаты были получены при классификации по бинарным шкалам «логика-этика» и «интуиция-сенсорика», являющимся базовыми для юнгианской типологии:

– для шкалы «логика-этика» наивысшая точность классификации (66%) достигнута при использовании нейронной сети типа biLSTM на текстах, векторизованных методом частотных признаков. Это свидетельствует о том, что именно грамматические, синтаксические и экстралингвистические характеристики (типы глаголов, использование союзов, структура предложений) наиболее ярко маркируют данную психологическую дихотомию;

– для шкалы «интуиция-сенсорика» лучший результат (72%) показал алгоритм Случайного леса с векторизацией TF-IDF, а также высокую эффективность (70%) продемонстрировала сеть biLSTM с векторизацией word2vec. Это указывает на то, что данная психическая функция тесно связана с лексическим

составом текста, использованием редких или уникальных слов и их семантическими полями.

Проведенный сравнительный анализ выявил сильные и слабые стороны различных подходов. Нейронные сети, в частности LSTM, показали себя как мощный инструмент для выявления сложных паттернов в последовательностях текстовых данных, в то время как классические методы, такие как Наивный Байес, несмотря на свою простоту, в некоторых конфигурациях демонстрировали конкурентоспособные результаты.

В ходе проекта был собран и размечен репрезентативный корпус текстов с форума socionics.su, что является ценным ресурсом для дальнейших исследований в области русскоязычной психолингвистики и стилометрии. Также была разработана и описана комплексная методика предобработки текстов, включающая графематический, морфологический, синтаксический и концептуальный анализ.

Научная и практическая значимость работы заключается в следующем:

– теоретическая: работа вносит вклад в развитие психолингвистики и computational linguistics, эмпирически подтверждая связь между языком и личностью и выявляя конкретные лингвистические маркеры психотипов для русского языка;

– практическая: разработанные модели и методики имеют высокий потенциал для применения в HR-менеджменте (автоматизированный подбор персонала), маркетинге (анализ потребительских предпочтений и целевых аудиторий), образовании (адаптация учебных программ под психологические особенности учащихся), политическом консалтинге и социальной аналитике;

– технологическая: исследование восполняет дефицит в области интеллектуального анализа русскоязычных текстов и создает основу для разработки коммерческих и научных программных продуктов для автоматического психологического профилирования.

Все планируемые работы по проекту выполнены в полном объеме. Заявленные цели достигнуты, научные результаты, включая публикации в сборниках международной конференции и подготовку статей в журнал ВАК, получены.

Таким образом, выполненный проект доказал принципиальную возможность и продемонстрировал эффективность автоматического определения психотипов по авторским текстам, заложив прочный фундамент для дальнейших исследований и практических разработок в этой перспективной междисциплинарной области.

5. Все планируемые работы выполнены полностью: Да.

6. Перечень публикаций научных статей в специализированных изданиях, программ и тезисов конференций по результатам выполненного проекта.

6.1. Гаврилова, А. С. Исследование методов аналитической обработки авторских текстов для определения психотипов обучающихся с использованием дистанционных образовательных технологий / А. С. Гаврилова, Е. П. Линник, И. И. Линник // Современное педагогическое образование. – 2025. – № 9. – С. 294-299. Режим доступа: https://www.elibrary.ru/item.asp?id=83025324 (статья ВАК).

6.2. Практическая реализация метода автоматизации при определении психотипа авторства неструктурированных текстов. Гаврилова А.С., Линник Е.П., Линник И.И. Режим доступа: Экономика строительства, 2025, N 9 (статья ВАК, в

печати).

6.3. Гаврилова, А. С. Исследование методов аналитической обработки авторских текстов: алгоритмы и методы анализа данных / А. С. Гаврилова, Е. П. Линник // Дистанционные образовательные технологии : СБОРНИК ТРУДОВ X МЕЖДУНАРОДНОЙ ЮБИЛЕЙНОЙ НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ, Ялта, 16–18 сентября 2025 года. – Симферополь: Общество с ограниченной ответственностью «Издательство Типография «Ариал», 2025. – С. 292-294. – EDN KHBKFI. Режим доступа: https://www.elibrary.ru/item.asp?id=83154838&pff=1

6.4. Гаврилова, А. С. Сравнительный анализ классификаторов для аналитической обработки авторских текстов / А. С. Гаврилова, И. И. Линник // Дистанционные образовательные технологии : СБОРНИК ТРУДОВ X МЕЖДУНАРОДНОЙ ЮБИЛЕЙНОЙ НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ, Ялта, 16–18 сентября 2025 года. – Симферополь: Общество с ограниченной ответственностью «Издательство Типография «Ариал», 2025. – С. 295-297. – EDN RVBTIP. Режим доступа: https://www.elibrary.ru/item.asp?id=83154839&pff=1

6.5. Гаврилова А.С. Проявление психотипа в цифровой среде: особенности самопрезентации и коммуникации представителей различных соционических типов в социальных медиа – статья РИНЦ подана на XXIV Международную научно-практическая конференция «Менеджмент XXI века: интеграция в экономике, образовании и науке»? Потому что я ещё знать не знаю сборника? (в печати).

7. В отчетном периоде возникли исключительные права на результаты интеллектуальной деятельности, созданные при выполнении проекта: Нет.

8. Информация о представлении достигнутых научных результатов на научных мероприятиях (конференциях, семинарах и пр.) (в том числе форма представления - приглашенный доклад, устное выступление, стендовый доклад).

8.1. X Международная юбилейная научно-практическая конференция «Дистанционные образовательные технологии», г. Симферополь 16-18 сентября 2025 г. – устный доклад на конференции с темой «Сравнительный анализ классификаторов для аналитической обработки авторских текстов».

8.2. XXIV Международная научно-практическая конференция «Менеджмент XXI века: интеграция в экономике, образовании и науке», г. Санкт-Петербург, 20-21 ноября 2025 г. – устный доклад на конференции с темой «Проявление психотипа в цифровой среде: особенности самопрезентации и коммуникации представителей различных соционических типов в социальных медиа».

9. Информация (при наличии) о публикациях в СМИ, посвященных результатам выполнения проекта - не имеется.

10. Привлекались ли к реализации проекта ученые, добровольцы (волонтеры) и иные специалисты? - нет.

11. Информация о внедрении результатов научного проекта в практическую деятельность.

Результаты исследования легли в основу научной диссертации на тему

«Формирование информационно-коммуникационной компетенции будущих психологов средствами digital-маркетинга».

12. Расходование средств обладателей грантов Государственного Совета Республики Крым молодым ученым Республики Крым имени Н. Я. Данилевского:

№ п.п.	Направления расходования средств гранта	Сумма расходов (тыс. руб.)
1.	**Налог на доходы физических лиц в размере 13%**	**39 тыс. руб. (тридцать девять тысяч рублей)**
2.	**Гаврилова А.С., Линник Е.П. Исследование методов аналитической обработки авторских текстов: алгоритмы и методы анализа данных (статья РИНЦ, в печати)**	**2 тыс. руб. (две тысячи рублей)**
3.	**Гаврилова А.С., Линник И.И. Сравнительный анализ классификаторов для аналитической обработки авторских текстов (статья РИНЦ, в печати)**	**2 тыс. руб. (две тысячи рублей)**
4.	Гаврилова, А. С. Исследование методов аналитической обработки авторских текстов для определения психотипов обучающихся с использованием дистанционных образовательных технологий / А. С. Гаврилова, Е. П. Линник, И. И. Линник // Современное педагогическое образование. – 2025. – № 9. – С. 294-299. Режим доступа: https://www.elibrary.ru/item.asp?id=83025324 (статья ВАК)	17 тыс. 500 руб. (семнадцать тысяч пятьсот рублей)
5.	Гаврилова А.С. Проявление психотипа в цифровой среде: особенности самопрезентации и коммуникации представителей различных соционических типов в социальных медиа (статья РИНЦ, в печати)	2 тыс руб. (две тысячи рублей)
6.	Гаврилова А.С., Линник Е.П., Линник И.И. Практическая реализация метода автоматизации при определении психотипа авторства неструктурированных текстов (статья ВАК, в печати)	17 тыс. 500 руб. (семнадцать тысяч пятьсот рублей)
7.	5 проверок в системе Антиплагиат.ВУЗ.	3 тыс. 450 руб. (три тысячи четыреста пятьдесят рублей)
8.	Проезд Симферополь-Санкт-Петербург (ЖД билет, № билета: 7805 4677 8834 16)	14 тыс. 162 руб. 60 коп. (четырнадцать тысяч сто шестьдесят два рубля шестьдесят копеек)
9.	Проезд Санкт-Петербург-Симферополь (ЖД билет, № билета: 7830 4677 8855 31)	13 тыс. 75 руб. 60 коп. (тринадцать тысяч семьдесят пять рублей шестьдесят копеек)
10.	Проживание в г. Санкт-Петербург (трое суток)	10 тыс. 500 руб. (десять тысяч пятьсот рублей)

_____ _____ А.С. Гаврилова

(Дата) (Подпись) (Расшифровка подписи