САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ (СПбГУ)
Математико-механический факультет
Искусственный интеллект и наука о данных

НАУЧНО-ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА

на тему:

Алгоритмы обучения без учителя для выявления психологических особенностей пользователей по текстовым данным социальных сетей

Выполнил:
студент группы 23.Б16-мм
Салман Эль Фарисси Атмания

Научный руководитель:
О.Р. Гавриленко

Санкт-Петербург
2025

Содержание

1 Введение **1**
1.1 Актуальность исследования . . . 1
1.2 Цель исследования . . . 2
1.3 Задачи исследования . . . 2
1.4 Объект и предмет исследования . . . 4
1.5 Методы исследования . . . 4
1.6 Научная новизна . . . 4
1.7 Практическая значимость . . . 4
1.8 Структура работы . . . 5

2 Обзор литературы **5**
2.1 Прогнозирование личности по цифровым следам . . . 5
2.2 Векторные представления текстов и трансформеры . . . 6
2.3 Методы снижения размерности и визуализации . . . 7
2.4 Современные алгоритмы кластеризации . . . 7
2.5 Методы обучения без учителя в анализе личности . . . 8
2.6 Статистические методы анализа различий между группами . . . 8
2.7 Выводы по обзору литературы . . . 9

3 Данные и методы **9**
3.1 Предварительная обработка данных . . . 9
3.2 Используемые алгоритмы . . . 10
3.3 Используемые метрики качества . . . 12

4 Результаты **13**
4.1 Активность пользователей и структура корпуса . . . 13
4.2 Качество эмбеддингов Sentence–BERT . . . 14
4.3 Выявление латентных признаков и кластеризация . . . 14
4.4 Интерпретация кластеров . . . 16
4.5 Распределение оценок выраженности психологических особенностей пользователей . . . 17
4.6 Предсказание оценок выраженности психологических особенностей пользователей . . . 18

5 Обсуждение результатов **20**

1 Введение

1.1 Актуальность исследования

В современном мире социальные сети стали неотъемлемой частью повседневной жизни миллиардов людей. Исследования показывают, что более 70% взрослого населения развитых стран ежедневно используют хотя бы одну социальную платформу, проводя в них в среднем 2,5 часа в день [1]. Пользователи ежедневно оставляют цифровые следы в виде текстовых сообщений, комментариев, постов и реакций, которые отражают их мысли, эмоции и поведенческие паттерны. Эти данные представляют значительный интерес для исследователей в области психологии личности, маркетинга, социологии и компьютерных наук.

Оценка психологических особенностей личности востребована в широком спектре областей, включая персонализацию контента [2], таргетированную рекламу [3] и подбор персонала [4]. Традиционно для этой цели используются стандартизированные психологические опросники, такие как MBTI или HEXACO [5,6]. Среди них одной из наиболее эмпирически обоснованных и широко признанных в научных исследованиях является пятифакторная модель личности, или «Большая пятерка» (Big Five) [7]. Она описывает личность через пять основных черт: открытость опыту (Openness), добросовестность (Conscientiousness), экстраверсию (Extraversion), доброжелательность (Agreeableness) и нейротизм (Neuroticism). Стандартная процедура оценки по этой модели предполагает заполнение опросников (например, NEO-PI-R), что требует значительных временных затрат со стороны респондентов и ограничивает масштаб исследований.

С развитием методов машинного обучения и обработки естественного языка появилась возможность автоматического прогнозирования личностных черт на основе анализа текстов в социальных сетях [8]. Однако большинство существующих подходов основаны на контролируемом обучении (supervised learning) и требуют размеченных данных с известными оценками личностных черт. Такие методы эффективны для предсказания, но не позволяют обнаружить скрытые паттерны и новые типы коммуникативного поведения, не заложенные в исходной разметке.

Применение методов обучения без учителя открывает альтернативный исследовательский путь. Вместо прямого предсказания оценок большой пятерки, можно сначала выявить текстовые признаки и их эмоциональную окраску путем кластеризации текстов пользователей, а затем исследовать корреляцию этих признаков с личностными чертами.

Современные достижения в области обработки естественного языка, в частности модели трансформеров типа Sentence-BERT [9], позволяют получать высококачественные векторные представления текстов, сохраняющие их семантический смысл. Применение алгоритмов снижения размерности (UMAP [10]) и продвинутых методов кластеризации (HDBSCAN [11]) к этим представлениям дает возможность выявлять устойчивые группы со схожими стилями коммуникации.

Теоретическая значимость исследования заключается в изучении взаимосвязи между скрытыми зависимостями в текстовых данных пользователей и оценками их психологических особенностей. Практическая значимость состоит в разработке алгоритма для автоматического определения этих особенностей по текстовым данным, который может быть применен для персонализации контента [12], таргетированной рекламы [3], подбора персонала [4] и разработки рекомендательных систем [2].

1.2 Цель исследования

Применить методы обучения без учителя для обнаружения латентных стилей текстовый контент в социальных сетях и исследовать их взаимосвязь с личностными чертами теста Big Five.

1.3 Задачи исследования

Для достижения поставленной цели необходимо решить следующие задачи:

1. **Обзор литературы и поиск аналогов**
    - Провести обзор научных статей по применению методов обучения без учителя для анализа психологических характеристик пользователей;
    - Исследовать существующие подходы к прогнозированию личностных черт на основе текстовых данных;

- Выявить преимущества и недостатки существующих методов.

2. **Подготовка и обработка данных**

   - Проанализировать предоставленный датасет (оценки Big Five и VK ID), отфильтровать дубликаты и проверить доступность профилей через API, исключив закрытые страницы;
   - Осуществить сбор текстов постов пользователей (корпуса) через VK API;
   - Провести предобработку текстов: очистка от шума (ссылок, спецсимволов), токенизация и удаление стоп-слов;
   - Выполнить исследовательский анализ данных (EDA) для оценки репрезентативности выборки и характеристик текстового корпуса.

3. **Получение векторных представлений текстов**

   - На основе корпуса текстов пользователей (постов, комментариев) сгенерировать современные текстовые эмбеддинги с использованием модели Sentence-BERT;
   - Оценить качество полученных эмбеддингов через анализ семантической близости текстов.

4. **Визуализация и снижение размерности**

   - Применить методы понижения размерности (UMAP, t-SNE) для визуализации пространства текстовых эмбеддингов в двумерном представлении;
   - Провести сравнительный анализ результатов различных методов снижения размерности;
   - Создать визуализации, позволяющие предварительно оценить наличие групп со схожими стилями.

5. **Кластеризация и выделение латентных стилей коммуникации**

   - Использовать алгоритмы кластеризации (HDBSCAN, Gaussian Mixture Models) для выделения устойчивых групп пользователей со схожими стилями письменной коммуникации;
   - Определить оптимальное количество кластеров и оценить качество кластеризации с помощью внутренних метрик (Silhouette Score, Davies-Bouldin Index);
   - Провести сравнение результатов различных алгоритмов кластеризации.

6. **Интерпретация выявленных стилей коммуникации**

   - Дать содержательную интерпретацию полученным кластерам через анализ наиболее характерных текстов;
   - Выявить ключевые слова и лингвистические особенности каждого кластера с использованием TF-IDF анализа;
   - По возможности применить лингвистический анализ для выявления эмоциональных, когнитивных и стилистических особенностей каждой группы.

7. **Анализ связи стилей коммуникации с психологическими особенностями пользователей**

- проанализировать различия в оценках выраженности психологических особенностей пользователей между выделенными стилями коммуникации на основе агрегированных признаков;
- по полученной информации сформировать новые признаки пользователей социальной сети, отражающие особенности их письменной коммуникации и принадлежность к стилевым группам;
- использовать сформированные признаки для предсказания оценок выраженности психологических особенностей пользователей социальной сети.

1.4 Объект и предмет исследования

Объект исследования: процесс выявления психологических особенностей пользователей на основе их цифровых следов в социальной сети ВКонтакте.

Предмет исследования: использование алгоритмов обучения без учителя для анализа текстовых данных пользователей и предсказания выраженности личностных черт по модели большой пятерки.

1.5 Методы исследования

В работе используются следующие методы:

- предварительная обработка текстов, включающая очистку, токенизацию и лемматизацию;
- получение векторных представлений текстов с использованием модели Sentence-BERT;
- снижение размерности векторных представлений методом анализа главных компонент (Principal Component Analysis, PCA);
- кластеризация пользователей на основе эмбеддингов с применением алгоритмов KMeans и Gaussian Mixture Models (GMM);
- извлечение и использование стилевых и латентных признаков для предсказания оценок выраженности психологических особенностей пользователей.

1.6 Научная новизна

Научная новизна работы заключается в выявлении скрытых зависимостей в текстовых данных пользователей социальной сети ВКонтакте с помощью методов обучения без учителя, которые затем используются для предсказания оценок выраженности психологических особенностей по модели большой пятерки. В отличие от существующих исследований, основанных на прямом анализе текста или контролируемом обучении, предлагаемый подход фокусируется на обнаружении латентных структур и паттернов в текстовых данных без предварительной разметки, что позволяет выявить неочевидные корреляции между характеристиками текста и личностными чертами пользователей.

1.7 Практическая значимость

Результаты исследования могут быть применены для:

- Персонализации контента и рекомендательных системах;

- Таргетированной рекламы и маркетинге;
- Анализа аудитории социальных медиа;
- HR-аналитики при подборе персонала;
- Разработки чат-ботов и виртуальных ассистентов с учетом стилей общения пользователей.

1.8 Структура работы

Работа состоит из введения, четырех глав, заключения, списка литературы и приложений.

Глава 1. «Теоретические основы анализа личностных черт по текстовым данным» содержит обзор литературы по теме исследования. Рассмотрены существующие подходы к прогнозированию личностных черт, методы обучения без учителя и современные достижения в области обработки естественного языка.

Глава 2. «Методология исследования и обработка данных» описывает используемые данные из социальной сети ВКонтакте и методы исследования. Детально представлены этапы предобработки текстов, архитектура Sentence-BERT для получения векторных представлений, алгоритмы кластеризации и статистические критерии.

Глава 3. «Экспериментальное выявление скрытых зависимостей в текстовых данных» представляет результаты экспериментов: анализ кластеризуемости данных, визуализацию текстового пространства, выделенные кластеры и их текстовые характеристики.

Глава 4. «Анализ взаимосвязи текстовых паттернов с психологическими особенностями пользователей» посвящена анализу результатов, полученных на предыдущих этапах работы. В данной главе рассматриваются выявленные стили письменной коммуникации пользователей и их связь с оценками выраженности психологических особенностей. Отдельное внимание уделяется интерпретации полученных кластеров и анализу вклада текстовых и латентных признаков в задачу предсказания.

В разделе **«Обсуждение результатов»** проводится качественный анализ полученных результатов, сопоставление их с выводами предыдущих исследований, а также обсуждается практическая применимость предложенного подхода. Рассматриваются ограничения используемых методов и возможные направления их дальнейшего развития.

В **заключении** подводятся итоги выполненной работы, формулируются основные выводы и обозначаются направления дальнейших исследований.

2 Обзор литературы

2.1 Прогнозирование личности по цифровым следам

С развитием социальных сетей и цифровых технологий появилась возможность анализировать личность на основе цифровых следов пользователей. Исследования в этой области показали, что анализ предсказывать не может. Анализ цифровых следов пользователей позволяет предсказывать личностные черты с высокой точностью

В работе Youyou et al. (2015) [8] была рассмотрена задача регрессии для предсказания личностных черт на основе цифровых следов. Авторы продемонстрировали, что компьютерные алгоритмы, анализирующие лайки пользователей в Facebook, могут превосходить

оценки, данные друзьями и коллегами. В качестве метода использовалась линейная регрессия, обученная на данных 86 220 добровольцев, предоставивших свои лайки и результаты теста «Большая пятерка». Качество моделей оценивалось с помощью корреляции Пирсона, и результаты показали, что предсказания алгоритма достигают корреляции 0.56 с реальными оценками, что сопоставимо даже с точностью оценок супругов ($r = 0.58$).

Аналогичное исследование провели Park et al. (2015) [13], которые решали задачу регрессии для предсказания пяти личностных черт на основе анализа текстов в Twitter. В работе использовались методы машинного обучения с разнообразными лингвистическими признаками, включая словарь LIWC, n-граммы слов и тематическое моделирование (LDA). На датасете, состоящем из твитов 66 732 пользователей и их результатов теста «Большая пятерка», модели достигли корреляции Пирсона в диапазоне от 0.35 до 0.48 с реальными оценками. Это подтвердило, что стиль письма в социальных сетях существенно коррелирует с личностными чертами.

В более поздней работе Stachl et al. (2020) [14] изучалась возможность предсказания личностных черт на основе поведенческих данных, собираемых со смартфонов. В рамках исследования решалась задача регрессии с использованием метода градиентного бустинга (XGBoost). Датасет включал данные 624 участников, чья активность (использование приложений, коммуникации, GPS) отслеживалась в течение 30 дней. Результаты показали, что модели способны достигать высокой корреляции с самооценками, особенно для черт экстраверсии (до 0.57) и открытости опыту (до 0.48), подтверждая ценность пассивно собранных данных.

2.2 Векторные представления текстов и трансформеры

Революционным прорывом в области обработки естественного языка стало появление архитектуры трансформеров [15]. В работе Devlin et al. (2019) [16] рассматривалась задача создания универсальной предобученной языковой модели, применимой к широкому спектру задач обработки естественного языка. Ключевой инновацией BERT стало использование механизма внимания (attention) для учета контекста слов одновременно в обоих направлениях — как слева направо, так и справа налево, в отличие от предыдущих моделей, обрабатывающих текст последовательно. В качестве метода использовался двунаправленный трансформер (bidirectional transformer encoder), обучаемый на двух задачах: маскированное языковое моделирование (masked language modeling), где модель предсказывает случайно скрытые слова в предложении на основе окружающего контекста, и предсказание следующего предложения (next sentence prediction), где определяется, является ли одно предложение логическим продолжением другого. Обучение проводилось на масштабном корпусе текстов, включающем BooksCorpus (800 миллионов слов) и английскую Википедию (2.5 миллиарда слов). При тестировании на 11 стандартных бенчмарках NLP, включая GLUE и SQuAD, модель установила новые рекорды качества в большинстве задач, особенно в вопросно-ответных системах и анализе тональности текста.

В работе Reimers и Gurevych (2019) [9] рассматривалась задача создания эффективных семантических представлений целых предложений, поскольку оригинальная модель BERT требует попарного сравнения текстов, что вычислительно неэффективно. Авторы модифицировали архитектуру BERT, применив метод дообучения (fine-tuning) с использованием сиамских и триплетных нейронных сетей. Сиамская архитектура обрабатывает пары предложений через общие веса для определения их семантической близости, а триплетная сеть учится различать похожие и непохожие предложения, используя тройки примеров (якорь, положительный и отрицательный примеры). Обучение проводилось на задачах определения логического следования (Natural Language Inference, NLI) с использованием датасета SNLI, содержащего 570 тысяч пар предложений, и оценки семантического сходства тек-

стов (Semantic Textual Similarity, STS). Качество полученных векторных представлений оценивалось через косинусное сходство и корреляцию Спирмена. Результаты показали, что Sentence-BERT ускоряет вычисление схожести текстов в 65 000 раз по сравнению с оригинальным BERT, при этом достигая корреляции Спирмена 0.85 на бенчмарке STS, что сопоставимо с качеством попарного сравнения.

В последующей работе Reimers и Gurevych (2020) [17] решалась задача создания многоязычных векторных представлений предложений, способных работать с текстами на разных языках в едином семантическом пространстве. Для этого авторы применили метод дистилляции знаний (knowledge distillation), при котором предобученная «учительская» модель на английском языке передает свои знания «студенческой» многоязычной модели. Этот подход позволяет студенческой модели научиться создавать схожие векторные представления для семантически эквивалентных предложений на разных языках, даже без прямого обучения на параллельных текстах. Метод был протестирован на параллельных корпусах, охватывающих более 50 языков, включая русский. Качество моделей оценивалось на задачах кросс-языкового семантического поиска (поиск похожих по смыслу текстов на разных языках) и выравнивания параллельных текстов (bitext mining). Результаты продемонстрировали, что многоязычные версии Sentence-BERT достигают качества, сопоставимого с монолингвальными моделями для большинства языков, что делает их особенно ценными для работы с текстами на языках с ограниченными обучающими данными.

2.3 Методы снижения размерности и визуализации

Векторные представления текстов, получаемые с помощью трансформерных моделей, обычно имеют высокую размерность (например, 768 для базовых моделей SBERT [9]), что затрудняет их визуализацию и может приводить к проблемам при кластеризации, известным как «проклятие размерности».

В работе McInnes et al. (2018) [10] был представлен UMAP (Uniform Manifold Approximation and Projection) — современный метод нелинейного снижения размерности. В рамках исследования решалась задача эффективного преобразования данных в пространство низкой размерности при сохранении как локальной, так и глобальной структуры. Метод основан на построении топологического представления данных с помощью нечетких симплициальных комплексов и последующей оптимизации координат в целевом пространстве с помощью кросс-энтропии. Тестирование на стандартных датасетах, таких как MNIST и Fashion-MNIST, показало, что UMAP работает значительно быстрее своего популярного аналога t-SNE и лучше сохраняет глобальную структуру данных.

Преимущества UMAP были подтверждены в сравнительном исследовании Becht et al. (2018) [18], где рассматривалась задача визуализации высокоразмерных биологических данных. Авторы сравнили UMAP с методами t-SNE, PCA и FIt-SNE на данных секвенирования РНК единичных клеток (более 1 миллиона наблюдений). Оценка производительности по критериям сохранения структуры и вычислительной эффективности показала, что UMAP превосходит t-SNE в сохранении кластерной структуры, особенно при наличии иерархических связей между группами данных.

2.4 Современные алгоритмы кластеризации

Для выявления латентных групп в данных без предварительной разметки применяются методы обучения без учителя, в частности алгоритмы кластеризации.

В работе McInnes et al. (2017) [11] был предложен алгоритм HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Здесь решалась задача класте-

ризации данных произвольной формы, при которой количество кластеров определяется автоматически. Метод строит иерархию кластеров на основе расстояния взаимной достижимости (mutual reachability distance) и затем извлекает наиболее стабильные кластеры. Такой подход позволяет алгоритму самостоятельно определять оптимальное количество кластеров и эффективно обнаруживать выбросы (шум), что было продемонстрировано на синтетических и реальных наборах данных.

Теоретические основы и преимущества HDBSCAN были заложены в работе Campello et al. (2015) [19], где рассматривалась задача усовершенствования кластеризации на основе плотности. Авторы предложили иерархический подход, который позволяет алгоритму адаптироваться к кластерам с различной плотностью. Сравнительное исследование на 54 наборах данных с использованием метрики Adjusted Rand Index показало, что способность автоматически выбирать параметр плотности для каждого кластера делает HDBSCAN значительно более гибким и эффективным по сравнению с его предшественником DBSCAN.

2.5 Методы обучения без учителя в анализе личности

Применение методов обучения без учителя для анализа личности представляет растущий исследовательский интерес, о чем свидетельствует увеличение числа публикаций в этой области за последние годы [20].

Так, в работе Rastegari et al. (2023) [21] рассматривалась задача выявления латентных поведенческих кластеров в текстовых данных без какой-либо предварительной разметки личностных черт. Для этого авторы предложили двухэтапный метод: сначала для получения векторных представлений текста использовалась модель RoBERTa, дообученная на контрастных задачах для повышения семантической различимости векторов, а затем полученные эмбеддинги кластеризовались с помощью алгоритма HDBSCAN. На материале 150 000 текстовых записей с платформы Reddit авторы показали, что выявленные кластеры демонстрируют высокую внутреннюю согласованность и значимые различия в распределении черт «Большой пятерки», что подтверждает эффективность подхода для обнаружения естественных коммуникативных стилей.

В другом исследовании Guntuku et al. (2017) [22] решали задачу выявления связей между визуальным контентом в Twitter и личностными чертами. С помощью тематического моделирования (Latent Dirichlet Allocation) и последующей кластеризации методом K-means на данных от 3200 пользователей авторы анализировали паттерны в публикуемых изображениях. Результаты показали, что подходы без учителя способны выявлять нетривиальные корреляции между темами визуального контента и личностными чертами, которые не всегда обнаруживаются при использовании методов обучения с учителем.

Важность исследовательских подходов, не ограниченных заранее заданными категориями, была подчеркнута в обзорной работе Boyd и Pennebaker (2018) [23]. Авторы утверждают, что именно такие методы анализа языка позволяют обнаруживать новые, ранее неизвестные лингвистические маркеры психологических состояний и личностных черт.

2.6 Статистические методы анализа различий между группами

Для проверки гипотез о статистически значимых различиях в личностных чертах между выявленными кластерами используются методы статистического вывода. В работе Kim (2015) [24] представлен обзор методов множественных сравнений, таких как дисперсионный анализ (ANOVA) и апостериорные (post-hoc) тесты. Автор подчеркивает важность контроля ошибки первого рода и рекомендует использовать поправку Бонферрони в качестве консервативного метода, особенно в исследованиях с большим количеством срав-

нений, что является типичной ситуацией при анализе различий между несколькими кластерами.

Наконец, для правильной интерпретации результатов важно учитывать не только их статистическую значимость. В работе Sullivan и Feinn (2012) [25] подчеркивается необходимость оценки размера эффекта (effect size). Авторы объясняют, что при использовании больших выборок, что характерно для анализа данных из социальных сетей, даже тривиальные и практически не значимые различия между группами могут оказаться статистически значимыми. В таких случаях именно размер эффекта позволяет оценить реальную, или практическую, значимость выявленных различий, дополняя результаты статистических тестов.

2.7 Выводы по обзору литературы

Анализ современной литературы показывает, что:

1. Цифровые следы в социальных сетях содержат богатую информацию о личностных чертах пользователей, которая может быть извлечена с помощью методов машинного обучения;
2. Современные трансформерные модели, позволяют получать высококачественные векторные представления текстов, сохраняющие семантическую информацию;
3. UMAP демонстрирует преимущества перед классическими методами снижения размерности для визуализации и сохранения структуры данных;
4. HDBSCAN обеспечивает автоматическое определение количества кластеров и робастность к шуму, что критически важно для исследовательского анализа без априорных предположений о структуре данных;
5. Применение методов обучения без учителя для выявления стилей коммуникации с последующим анализом их связи с большой пятеркой является перспективным направлением, позволяющим обнаруживать новые закономерности;

Таким образом, предлагаемый в данной работе подход, объединяющий современные методы NLP, обучения без учителя и статистического анализа, имеет прочную методологическую основу и позволяет внести вклад в понимание связи между стилями коммуникации и оценками выраженности психологических особенностей пользователя.

3 Данные и методы

В этом разделе описываются исходные данные, этапы их подготовки, а также алгоритмы и метрики, использованные для выявления латентных стилей письменной коммуникации пользователей социальной сети ВКонтакте.

3.1 Предварительная обработка данных

Исходный датасет содержал сведения о 500 пользователях ВКонтакте: для каждого были доступны идентификатор профиля и оценки выраженности личностных черт по модели Big Five. Данные были предоставлены в формате CSV.

На первом этапе выполнялась очистка и нормализация идентификаторов. В строках с ID пользователей удалялись префиксы вида `https://vk.com/`, `http://vk.com/`,

www.vk.com/, а также параметры запросов и служебные символы. Это позволило привести все идентификаторы к единому формату и устранить дубликаты записей по одному и тому же пользователю.

Далее с помощью метода «users.get» VK API (версия 5.131) осуществлялась проверка профилей пользователей на доступность: определялось, являются ли страницы открытыми или закрытыми. Из анализа исключались удалённые и заблокированные аккаунты, а также полностью закрытые профили, к содержимому которых невозможно получить доступ программно. После этого для пользователей с открытыми профилями выполнялся сбор текстового контента.

Сбор текстов осуществлялся методом «wall.get» с параметром «filter = owner», то есть извлекались только собственные записи пользователя. Для каждого профиля запрашивалось не более 300 последних постов; между запросами выдерживалась пауза 0.34 секунды для соблюдения ограничений VK API. На данном этапе часть профилей была исключена из дальнейшего анализа из-за полного отсутствия текстовых публикаций.

Сводка по числу пользователей на разных этапах обработки приведена в табл. 1.

Таблица 1: Этапы обработки данных и количество оставшихся записей

Этап обработки	Количество	% от исходных
Исходный датасет	500	100.0%
После удаления дубликатов	457	91.4%
Открытые профили	312	62.4%

Из табл. 1 видно, что от исходного набора данных, составляющего 500 пользователей, после обработки осталось 312 респондентов. Основные причины сокращения выборки — дубликаты, закрытые или удалённые профили, а также отсутствие текстового контента.

После сбора данных выполнялась предобработка текстов постов. Этот этап включал:

- удаление ссылок, упоминаний, хэштегов, HTML-тегов и спецсимволов;
- оставление только буквенно-цифровых токенов (кириллица и латиница);
- токенизацию с помощью библиотеки razdel;
- лемматизацию с использованием pymorphy2.MorphAnalyzer и мемоизации (LRU-cache);
- удаление стоп-слов (стандартный список + технические слова {http, https, vk, com, www});
- фильтрацию токенов длиной ≤ 1 символа и чисто числовых токенов.

После предобработки для каждого пользователя формировался корпус очищенных и лемматизированных текстов. Среднее число токенов на пользователя составило около 2000, однако медиана была существенно ниже, что говорит либо о сильной неравномерной активности, либо о сильном неравномерной распределении активности: небольшое число очень «говорливых» аккаунтов сосуществует с большим количеством пользователей с малым числом постов.

3.2 Используемые алгоритмы

Цель формирования алгоритма состояла в том, чтобы::

1. получить компактные и информативные векторные представления текстов (эмбеддинги);
2. проверить наличие кластерной структуры в пространстве этих представлений;
3. выделить устойчивые кластеры пользователей, интерпретируемые как стили письменной коммуникации;
4. описать и интерпретировать найденные стили с помощью лингвистических признаков.
5. применить методы машинного обучения для предсказания оценки выраженности психологических особенностей пользователей на основе исходных и латентных признаков.

Векторные представления текстов. Для кодирования текстов использовалась многоязычная модель Sentence-BERT `paraphrase-multilingual-MiniLM-L12-v2` из библиотеки `sentence-transformers` [26]. Каждый пост пользователя преобразовывался в эмбеддинг размерности 384. Данная размерность обусловлена архитектурой модели MiniLM-L12, в которой используется компактное скрытое представление, обеспечивающее баланс между качеством семантического кодирования и вычислительной эффективностью.

Затем для каждого пользователя вычислялся усреднённый вектор по всем его постам с последующей L2-нормализацией. Таким образом, один вектор отражает усреднённый стиль письменной коммуникации конкретного пользователя.

Снижение размерности и визуализация. Для анализа структуры пространства эмбеддингов и их наглядной визуализации использовались два метода снижения размерности:

- *UMAP* (*Uniform Manifold Approximation and Projection*) — для построения двумерных проекций, на которых затем отображались результаты кластеризации [10];
- *t-SNE* — как альтернативный метод визуализации, применяемый после предварительного сжатия данных методом PCA до 50 компонент [27].

Использование двух разных методов визуализации позволило проверить устойчивость наблюдаемой структуры данных. Сходная форма и расположение групп пользователей в проекциях UMAP и t-SNE указывает на то, что выявленные кластеры отражают реальные различия в стилях письменной коммуникации, а не являются следствием особенностей конкретного метода проекции.

Алгоритмы кластеризации. Для выделения групп пользователей с похожими стилями общения были применены три алгоритма:

- *k-means* — классический алгоритм кластеризации с фиксированным числом кластеров k. Рассматривались значения k от 2 до 8. Преимуществом данного метода является простая интерпретация кластеров через центроиды и невысокая вычислительная сложность [28].
- *HDBSCAN* — плотностной иерархический метод, автоматически определяющий число кластеров и выделяющий точки шума. Алгоритм позволяет обнаруживать кластеры произвольной формы и различной плотности, что делает его удобным для исследовательского анализа текстовых данных [11].
- *Gaussian Mixture Models (GMM)* — вероятностный алгоритм, аппроксимирующий распределение данных смесью многомерных нормальных распределений. Число компонент подбиралось по информационному критерию [29].

Интерпретация кластеров. После выделения кластеров их содержательная интерпретация выполнялась двумя способами:

- TF–IDF-анализом наиболее характерной лексики для каждого кластера;
- сравнением базовых текстовых характеристик пользователей, включая объём текста, среднюю длину слова, коэффициент уникальности лексики, количество символов и частоту использования знаков препинания.

В результате для каждого кластера были получены наборы лексических и стилистических признаков, позволяющие описать различия между группами пользователей и использовать их для дальнейшего анализа и интерпретации.

3.3 Используемые метрики качества

Для оценки качества эмбеддингов и работы алгоритмов кластеризации применялся набор количественных метрик.

Кластеризуемость данных. На матрице пользовательских эмбеддингов вычислялась статистика Хопкинса H, которая показывает, насколько распределение точек в пространстве отличается от случайного. Значения H около 0,5 соответствуют почти однородным данным, тогда как $H \to 1$ указывает на наличие выраженной кластерной структуры. Величина H используется в работе как глобальный индикатор того, имеет ли смысл применять методы кластеризации к данному набору эмбеддингов.

Дополнительно анализировались:

- распределения косинусного сходства между постами одного пользователя и постами разных пользователей (для оценки качества эмбеддингов);
- распределение попарных косинусных сходств между пользователями;
- график расстояний до k-го ближайшего соседа (k-NN) для диагностики плотностной структуры.

Численные значения этих метрик и их интерпретации приводятся в разделе с результатами.

Внутренние метрики качества кластеризации. Для сравнения различных алгоритмов и настроек использовались:

- коэффициент силуэта (Silhouette Score), характеризующий отношение расстояния до собственного кластера к расстоянию до ближайшего соседнего [30];
- индекс Дэвиса–Болдина (Davies–Bouldin Index), измеряющий степень перекрытия кластеров [31];
- информационный критерий (BIC) при выборе числа компонент для Gaussian Mixture Models [32];
- стабильность кластеров при повторных запусках, оцениваемая, например, с помощью Adjusted Rand Index (ARI) между различными разбиениями [33].

Метрики качества предсказания оценок выраженности психологических особенностей пользователей. Для оценки качества моделей, предсказывающих выраженность психологических особенностей пользователей по шкалам Big Five, использовались стандартные метрики регрессии:

- средняя абсолютная ошибка (MAE), характеризующая среднее отклонение предсказанных значений от истинных оценок;
- среднеквадратичная ошибка (RMSE), более чувствительная к крупным ошибкам предсказания и позволяющая оценить устойчивость модели;
- коэффициент детерминации (R^2), отражающий долю дисперсии целевой переменной, объясняемую моделью.

Использование нескольких метрик позволило комплексно оценить качество предсказания и сравнить различные модели и наборы признаков как по средней точности, так и по устойчивости к отдельным ошибочным предсказаниям.

Совместное использование этих метрик позволило не только выбрать рабочую конфигурацию алгоритмов кластеризации, но и убедиться, что выделенные стили коммуникации являются статистически обоснованными и содержательно интерпретируемыми.

4 Результаты

4.1 Активность пользователей и структура корпуса

Перед построением эмбеддингов была проведена первичная оценка распределения пользовательской активности. На рисунке 1 представлено распределение количества постов на одного автора.

Большинство пользователей публикуют сравнительно небольшое количество сообщений (менее 50 постов), при этом присутствует небольшая группа пользователей с высокой активностью (до 300 постов). Такое распределение характеризуется выраженной асимметрией и указывает на неоднородность текстовых данных: у более активных пользователей стиль письменной коммуникации выражен стабильнее, тогда как у пользователей с малым числом сообщений векторные представления могут обладать большей вариативностью, поскольку формируются на основе ограниченного объёма текста.

После очистки и лемматизации среднее число токенов на пользователя составило около 2000, медианное значение — около 80, что дополнительно подтверждает неравномерность распределения текстовой активности. У небольшой части пользователей накапливаются значительные объёмы текста, в то время как у большинства пользователей данные представлены короткими текстовыми фрагментами. Данная особенность учитывалась при интерпретации результатов и обучении моделей, которые должны корректно работать как для пользователей с высокой, так и с низкой текстовой активностью.

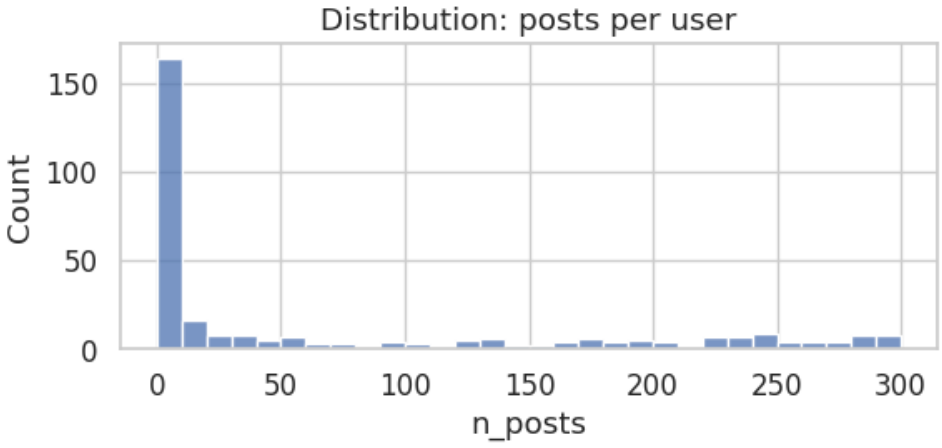

Рис. 1: Распределение количества постов на пользователя

4.2 Качество эмбеддингов Sentence–BERT

Для того чтобы обосновать использование Sentence–BERT в качестве инструмента для анализа стиля письменной коммуникации, необходимо проверить, отражают ли полученные эмбеддинги устойчивые индивидуальные особенности пользователей, а не только тематическое сходство отдельных текстов. С этой целью были вычислены распределения косинусного сходства между постами одного автора (intra-user) и между постами разных авторов (inter-user).

Если эмбеддинги действительно фиксируют индивидуальный стиль, то тексты одного пользователя должны быть в среднем более близки друг к другу, чем тексты, принадлежащие разным пользователям. Наличие такого различия является необходимым условием для последующего усреднения эмбеддингов по постам и применения методов кластеризации на уровне пользователей.

Результаты сравнения распределений представлены на рисунке 2 и в таблице 2.

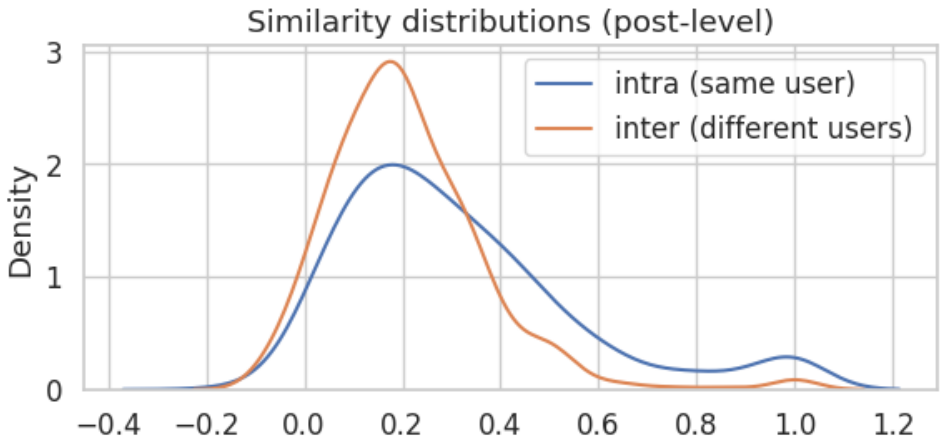

Рис. 2: Сходство постов одного пользователя и разных пользователей

Таблица 2: Статистики семантического сходства постов

Тип сходства	Среднее ± std	Медиана
Intra-user	0.311 ± 0.242	0.298
Inter-user	0.208 ± 0.162	0.192

Как видно из таблицы, среднее сходство текстов одного пользователя заметно выше, чем среднее сходство между разными пользователями. Аналогичный сдвиг наблюдается и для медианных значений. График на рис. 2 показывает, что кривая intra-user в целом смещена вправо относительно inter-user. Это означает, что эмбеддинги фиксируют устойчивые индивидуальные особенности письменной речи: типичные темы, эмоциональность, характерные формулировки и структуру высказываний. Именно на этом эффекте далее строится идея стиля, который используется и в кластеризации, и в построении новых признаков.

4.3 Выявление латентных признаков и кластеризация

Для перехода от индивидуальных эмбеддингов к признакам был реализован следующий алгоритм:

- построение эмбеддингов Sentence-BERT размерности 384 для каждого пользователя (усреднение по его постам);
- понижение размерности методом анализа главных компонент (*Principal Component Analysis, PCA*) до 10 компонент, отражающих основные направления вариации в стилях [34];
- кластеризация KMeans и Gaussian Mixture Models (GMM) для k в диапазоне от 2 до 8;
- выбор финальной модели по внутренним метрикам качества и интерпретируемости результатов.

С точки зрения метрик качества наилучшие значения силуэта давали модели KMeans при $k = 2$ и $k = 3$. Однако при $k = 2$ кластеры получаются слишком крупными и гетерогенными, а при $k = 3$ удаётся выделить более содержательные группы. В итоге в качестве итогового решения был выбран KMeans с $k = 3$ кластерами.

Итоговая двухмерная проекция пользовательских эмбеддингов с раскраской по кластерам представлена на рисунке 3.

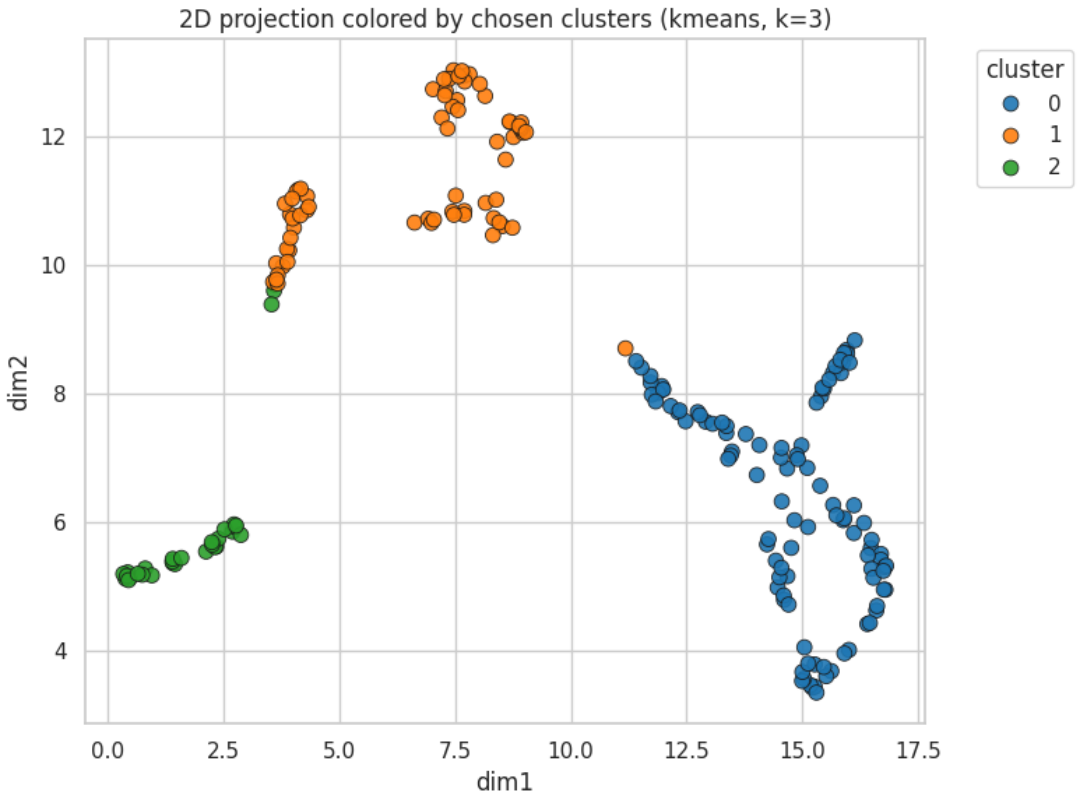

Рис. 3: UMAP-проекция пользовательских эмбеддингов, окрашенная по кластерам KMeans ($k = 3$)

На рисунке видно, что три кластера в пространстве UMAP образуют сравнительно компактные и разделённые области, между которыми имеются заметные «промежутки». Это говорит о наличии реальной структурированности в данных. Размеры кластеров приведены в таблице 3.

Таблица 3: Размеры финальных кластеров

Кластер	Количество пользователей	Доля
0	102	53,4%
1	69	36,1%
2	29	10,5%

Как видно из таблицы, наиболее распространённым является кластер 0, включающий более половины пользователей, тогда как кластер 2 заметно менее многочисленный. Такое распределение отражает то, что один тип стилевого поведения является «по умолчанию» в выборке, а другие встречаются реже и имеют более специфические характеристики.

4.4 Интерпретация кластеров

Для того чтобы понять, чем именно кластеры отличаются между собой, были использованы:

- лексические признаки (TF–IDF по леммам);
- стилистические признаки: объём текста (число слов), коэффициент уникальности слов, временная активность;
- содержательный анализ наиболее частотных и наиболее характерных слов внутри каждого кластера.

На основе этих признаков были получены следующие интерпретации:

- **Кластер 0** — деловой и структурированный стиль: преобладают формальные сообщения, рекламные и информационные посты, объявления, описание мероприятий и проектов; часто используются повторяющиеся шаблонные формулировки.
- **Кластер 1** — социальный и эмоциональный стиль: личные заметки о повседневной жизни, эмоциях и отношениях; больше обращений к другим людям, эмоциональных оценок и более разговорная лексика.
- **Кластер 2** — сокращённый стиль: короткие, часто фрагментарные записи; значительная часть постов удалена или содержит минимальное количество текста; низкая общая текстовая активность.

Ключевые различия по коэффициенту уникальности лексики показаны на рисунке 4, а различия по временной активности — на рисунке 5.

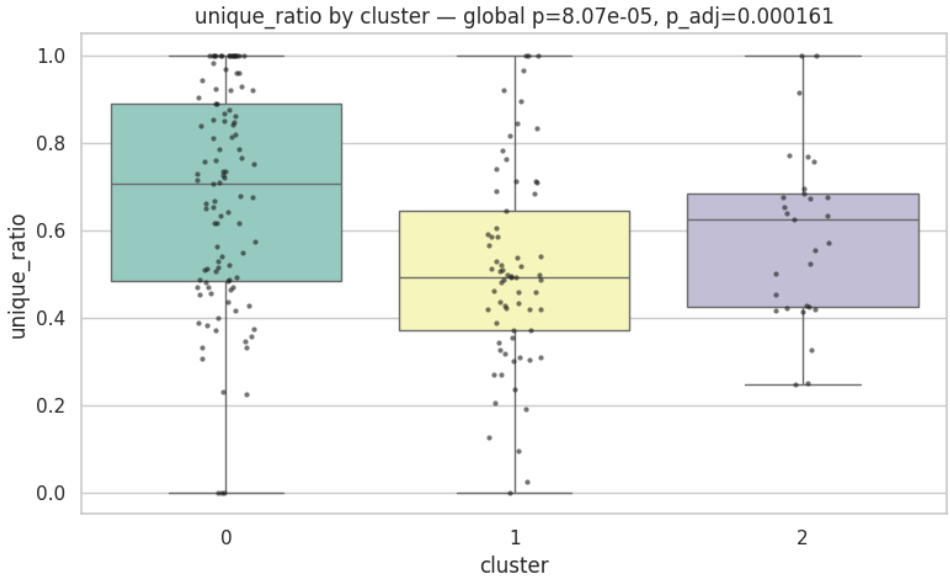

Рис. 4: Распределение коэффициента уникальности слов по кластерам

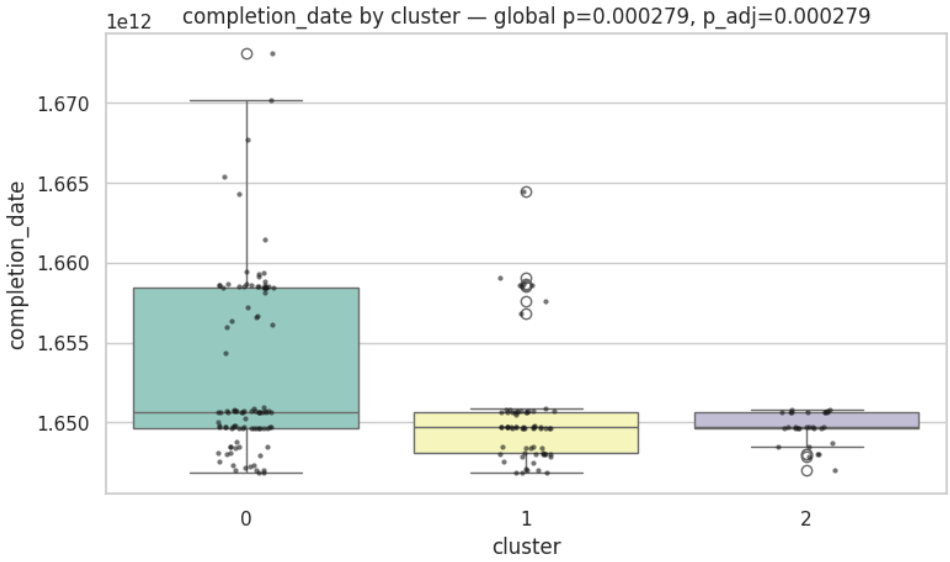

Рис. 5: Различия во временной активности пользователей по кластерам

Из рисунка 4 видно, что деловой кластер имеет более низкий коэффициент уникальности за счёт повторяемых формулировок (оформление акций, типовые объявления), в то время как социально-эмоциональный кластер демонстрирует более разнообразную лексику. Минималистичный кластер отличается большим разбросом, что объясняется малым количеством слов: даже небольшие изменения текста сильно влияют на показатель уникальности.

4.5 Распределение оценок выраженности психологических особенностей пользователей

На следующем этапе анализировались уже не только текстовые, но и психологические данные. Распределение пяти оценок выраженности психологических особенностей пользователей социальной сети на основе теста "BigFive"по выборке представлены на рисунке 6.

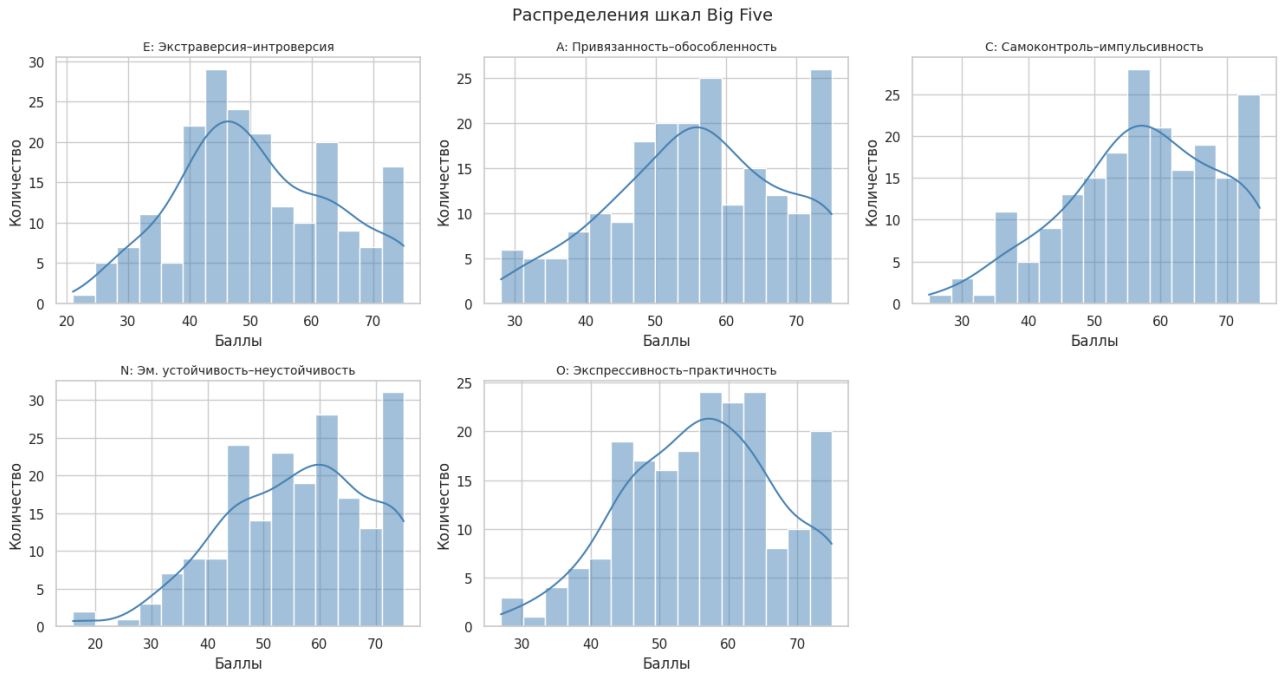

Рис. 6: Распределения шкал Big Five по выборке

Все шкалы имеют относительно плавные распределения и с заметной вариативностью, что позволяет корректно ставить задачу регрессии: модель может учиться не на двух-трёх классах, а на непрерывном диапазоне значений, а ошибки удобно интерпретировать в абсолютных баллах.

4.6 Предсказание оценок выраженности психологических особенностей пользователей

Для предсказания оценок выраженности пяти психологических особенностей по модели Big Five были сформированы два варианта набора признаков, отличающиеся уровнем использования текстовой информации.

- **Baseline** — признаки исходного обработанного датасета, включающие анкетные данные и базовые статистические характеристики текстов пользователей (объём текста, средняя длина слова, коэффициент уникальности лексики и показатели активности).
- **Enhanced** — расширенный набор признаков, включающий все baseline-признаки, а также дополнительные латентные и стилистические характеристики, извлечённые из текстов пользователей.

Дополнительные признаки в расширенном наборе формировались следующим образом. Для каждого пользователя были построены усреднённые эмбеддинги Sentence-BERT на основе его постов, после чего пространство эмбеддингов было понижено в размерности методом анализа главных компонент (PCA). В расширенный набор вошли первые 10 PCA-компонент, отражающие основные направления вариации в стилях письменной коммуникации.

Кроме того, в качестве признаков использовались индикаторы кластерной принадлежности пользователя, полученные в результате кластеризации в пространстве эмбеддингов, а также набор стилистических текстовых характеристик, описывающих объём, разнообразие и интенсивность текстовой активности.

Таким образом, исходный набор содержал $N_{\text{base}} = 8$ признаков, а после добавления латентных и кластерных характеристик размерность признакового пространства увеличилась до $N_{\text{enhanced}} = 22$. Это позволило явно включить в модель информацию о скрытой структуре стиля письменной коммуникации пользователей и оценить вклад новых признаков в качество предсказания.

Таким образом, в enhanced-модель явным образом добавлялись признаки, описывающие скрытую структуру стиля письменной коммуникации, а baseline-модель служила отправной точкой для сравнения.

В качестве алгоритмов использовались Random Forest и HistGradientBoosting в обёртке `MultiOutputRegressor`, что позволяет одновременно предсказывать пять шкал.

Результаты baseline-модели для Random Forest представлены в таблице ниже.

Таблица 4: Baseline (Random Forest): ошибки предсказания оценок выраженности психологических особенностей

Психологическая особенность	MAE	RMSE
Экстраверсия	10,82	13,20
Привязанность	9,94	12,20
Самоконтроль	9,22	11,59
Эмоциональная устойчивость	10,77	13,39
Экспрессивность	9,41	11,39

Аналогичные результаты для HistGradientBoosting приведены в табл. 5.

Таблица 5: Baseline (HistGradientBoosting): ошибки предсказания оценок выраженности психологических особенностей

Психологическая особенность	MAE	RMSE
Экстраверсия	12,05	14,85
Привязанность	10,55	13,17
Самоконтроль	10,48	13,24
Эмоциональная устойчивость	11,43	14,42
Экспрессивность	10,35	12,62

Видно, что baseline-модели дают ошибки порядка 9–12 баллов MAE и ещё более высокие значения RMSE. Это можно считать рабочим, но не очень точным уровнем, особенно если речь идёт об индивидуальной оценке.

При добавлении скрытых латентных признаков из текстов пользователей качество существенно улучшается. Сводное сравнение baseline и enhanced-модели приведено в таблице 6.

Таблица 6: Сравнение MAE и RMSE baseline и расширенной модели

Психологическая особенность	MAE(base)	MAE(enhanced)	RMSE(base)	RMSE(enhanced)
Экстраверсия	10,82	6,45	13,20	8,80
Привязанность	9,94	6,21	12,20	8,73
Самоконтроль	9,22	6,09	11,59	8,66
Эмоциональная устойчивость	10,77	6,55	13,39	8,95
Экспрессивность	9,41	6,02	11,39	8,51

Ошибка MAE снижается примерно на 25–45% по всем психологическим особенностям, а RMSE уменьшается в среднем на 4–5 баллов. Это означает, что добавленные латентные признаки, извлечённые из текстов пользователей, действительно несут полезную информацию о психологических особенностях пользователей и существенно повышают точность предсказаний.

5 Обсуждение результатов

Полученные результаты демонстрируют, что эмбеддинги Sentence–BERT в сочетании с методами снижения размерности и кластеризации позволяют выделять устойчивые стилистические паттерны в текстах пользователей. Эти паттерны оказываются информативными как для анализа письменной коммуникации, так и для построения моделей предсказания личностных характеристик.

Во-первых, анализ внутрипользовательского и межпользовательского сходства (рис. 2, табл. 2) показал, что посты одного автора в среднем заметно ближе друг к другу, чем к постам других пользователей. Это означает, что Sentence–BERT фиксирует индивидуальные особенности письменной речи: характерный набор тем, типичные речевые конструкции, стиль выражения эмоций. Наличие такого устойчивого стилевого профиля является ключевым условием для дальнейшей кластеризации и построения латентных признаков.

Во-вторых, полученные эмбеддинги демонстрируют выраженную кластерную структуру. Визуализация в UMAP-пространстве (рис. 3) показывает, что пользователи образуют три относительно компактные и разделённые группы, а распределение размеров кластеров (табл. 3) указывает на естественную неоднородность стилей поведения. Интерпретация кластеров на основе лексических и стилистических признаков (рис. 4, 5) позволила выделить деловой, социально-эмоциональный и минималистичный типы текста. Таким образом, кластеризация не только формально делит пользователей на группы, но и соответствует понятным с содержательной точки зрения стилям.

В-третьих, анализ распределений шкал Big Five (рис. 6) показал, что выборка обладает достаточной вариативностью по каждому из пяти измерений личности. Это позволяет рассматривать задачу предсказания как регрессию по непрерывным шкалам, а не как грубое деление на несколько категорий.

Особенно важным является сравнение baseline и расширенной модели предсказания психологических характеристик (табл. 4, 5, 6). Baseline-вариант, основанный только на исходных признаках, показывает ограниченную точность (ошибки порядка 9–12 баллов MAE), что делает его пригодным скорее для анализа тенденций на уровне группы, чем для точного индивидуального прогноза. При добавлении скрытых текстовых признаков — PCA-компонент эмбеддингов, кластерных меток и стилистических характеристик — ошибки заметно уменьшаются. Это означает, что индивидуальный стиль письменной коммуникации несёт существенную информацию о личностных особенностях и может быть эффективно интегрирован в табличные модели.

Важно подчеркнуть, что в работе последовательно реализованы оба шага, на которые обращал внимание научный руководитель: сначала были найдены скрытые латентные признаки из постов (эмбеддинги SBERT, их проекции PCA, кластерные и стилистические характеристики), затем эти признаки были добавлены к исходному датасету, и уже на расширенном наборе признаков были построены модели, предсказывающие оценки выраженности психологических особенностей пользователя. Отдельно были обучены и проанализированы baseline-модели на исходных признаках, что позволяет корректно сравнить вклад новых латентных признаков.

В то же время полученные результаты следует интерпретировать с осторожностью. Выборка по пользователям относительно невелика, оценки Big Five могут содержать су-

щественный индивидуальный шум, а активность по текстам сильно различается между пользователями. Усреднение эмбеддингов по постам приводит к потере части внутриличностной вариативности: автор, совмещающий деловые и личные тексты, может оказаться в промежуточной позиции. Кроме того, алгоритм KMeans предполагает сферические границы кластеров и может не полностью раскрывать сложные структуры в данных.

Несмотря на эти ограничения, работа демонстрирует, что даже относительно простая комбинация современных текстовых эмбеддингов, кластеризации и классических регрессионных моделей позволяет получить содержательную карту стилевых типов пользователей и существенно улучшить качество предсказания психологических шкал. Полученные латентные текстовые признаки могут быть использованы в дальнейшем в более сложных моделях, на расширенных выборках и в практических задачах персонализации контента и анализа аудитории.

Заключение

В данной работе была поставлена и решена задача выявления латентных стилей письменнои коммуникации пользователеи социальной сети ВКонтакте с помощью методов обучения без учителя и анализа связи этих стилеи с личностными характеристиками по модели Большои пятёрки. Отдельной целью было показать, что скрытые признаки, извлечённые из текстов, могут улучшать качество предсказания психологических шкал по сравнению с моделями, основанными только на исходных табличных данных.

Для этого был собран и подготовлен корпус текстов 200 пользователь, прошедших психологическое тестирование. На этапе предварительнои обработки были нормализованы идентификаторы профилеи, отфильтрованы закрытые и неактивные аккаунты, выполнена очистка текстов, токенизация и лемматизация. Сводная статистика показала, что существенная часть исходнои выборки не пригодна для анализа из-за отсутствия открытого текстового контента, что является важным практическим ограничением при работе с реальными социальными сетями. В итоговый датасет были включены только пользователи с достаточным объёмом доступных записей.

Для представления текстов в векторной форме использовалась многоязычная модель Sentence–BERT paraphrase-multilingual-MiniLM-L12-v2. Анализ внутрипользовательского и межпользовательского сходства (рис. 2, табл. 2) показал, что эмбеддинги устойчиво отражают индивидуальные особенности стиля: посты одного пользователя в среднем заметно ближе друг к другу, чем к постам других пользователей. Это подтверждает применимость современных трансформерных моделей для задач анализа стиля письменной коммуникации и создания признаков, описывающих индивидуальный языковой профиль.

Далее эмбеддинги пользователей были понижены в размерности до 10 компонент с помощью PCA, после чего была выполнена кластеризация в этом скрытом пространстве. На основе сравнения нескольких вариантов было выбрано решение KMeans с тремя кластерами, которое обеспечило баланс между устойчивостью, отделимостью групп и их содержательной интерпретируемостью. В итоге были выделены три устойчивых стиля коммуникации:

- деловой и структурированный стиль — объёмные, формальные и менее разнообразные по лексике публикации, связанные с работой, проектами, рекламой и официальными сообщениями;
- социально-эмоциональный стиль — преимущественно личные сообщения о повседневной жизни, эмоциях и отношениях, с более разговорной и разнообразной лексикой;

- минималистичный стиль — редкие и краткие записи, фрагментарные сообщения и посты с минимальным текстом или удалённым содержимым.

Стилистические и лингвистические признаки (объём текста, коэффициент уникальности слов, темп и характер активности) показали значимые и устойчивые различия между кластерами (рис. 4, 5), что подтверждает, что выделенные группы отражают не случайные флуктуации, а устойчивые паттерны поведения. Визуализация в пространстве UMAP (рис. 3) показала, что кластеры занимают компактные и хорошо отделённые области, что дополнительно подтверждает наличие структуры в данных.

Ключевым шагом работы стало использование выявленных латентных признаков, полученных из постов, для предсказания оценок по шкалам Big Five. Были сформированы два набора признаков: baseline, включающий исходные табличные и простые текстовые признаки, и расширенный (enhanced), в который дополнительно вошли PCA-компоненты SBERT-эмбеддингов, индикаторы кластерной принадлежности и стилистические характеристики текстов. Для обоих вариантов наборов признаков были обучены модели многоцелевой регрессии на основе Random Forest и HistGradientBoosting.

Сравнение результатов показало, что baseline-модели обеспечивают лишь умеренную точность предсказания шкал Big Five (табл. 4, 5). При переходе к расширенному набору признаков (табл. 6) средние ошибки MAE снижаются примерно на 25–45% по всем пяти шкалам, а RMSE уменьшается в среднем на 4–5 баллов. Это даёт основание заключить, что скрытые латентные признаки, извлечённые из текстов пользователей (стили, кластеры, компоненты векторных представлений), существенно повышают предсказательную способность моделей и действительно несут информацию, связанную с личностными особенностями.

С теоретической точки зрения работа подтверждает, что:

1. современные эмбеддинги предложений на основе трансформеров способны улавливать устойчивый индивидуальный стиль письменной коммуникации;

2. методы обучения без учителя (кластеризация в скрытом пространстве эмбеддингов) позволяют выделять осмысленные стили коммуникации даже при отсутствии явной разметки;

3. базовые лингвистические признаки (объём текста, длина слова, уникальность лексики, характеристики активности) хорошо согласуются с полученными кластерами и служат дополнительным инструментом их интерпретации;

4. стилистические и латентные текстовые признаки могут использоваться как важное дополнение к табличным данным при моделировании личностных характеристик.

Практическая значимость результатов заключается в том, что выявленные стили и полученные латентные признаки могут использоваться как высокоуровневые характеристики пользователей в системах персонализации контента, рекомендательных сервисах, маркетинговой аналитике и HR-подборе. Построенный в работе пайплайн (сбор данных, предобработка, построение эмбеддингов, понижение размерности, кластеризация, извлечение текстовых признаков и обучение моделей предсказания) является модульным и может быть адаптирован к другим платформам и языкам.

Вместе с тем исследование имеет ряд ограничений. Во-первых, относительно небольшой объём выборки и сильный дисбаланс по активности снижают статистическую мощность выводов, особенно при анализе связи с личностными чертами Big Five. Во-вторых, усреднение эмбеддингов по постам сглаживает внутриличностную вариативность стиля, а удаление эмодзи и части пунктуации уменьшает количество эмоциональных сигналов в

данных. В-третьих, использование KMeans накладывает предположение о примерно сферической форме кластеров и может не полностью соответствовать реальной структуре стилевых распределений.

Перспективами дальнейшей работы являются:

- расширение выборки пользователей и включение дополнительных источников текстов;
- использование более богатых признаков, включающих эмодзи, типы пунктуации, временную динамику и параметры сетевой активности;
- применение более гибких методов агрегирования эмбеддингов (взвешивание по TF–IDF, специализированные модели на уровне пользователя);
- углублённое исследование связи выделенных стилевых кластеров и латентных факторов с личностными оценками по модели Big Five на более крупных данных.

В целом полученные результаты демонстрируют, что методы обучения без учителя в сочетании с современными моделями представления текста и простыми регрессионными моделями являются перспективным инструментом для анализа психологических особенностей пользователей по их цифровым следам в социальных сетях и создают основу для дальнейших, более детальных исследований в этом направлении.

Список литературы

[1] Perrin, A., & Anderson, M. (2021). Social media use in 2021. *Pew Research Center*, 7(4), 1–4.

[2] Tkalčič, M., Kunaver, M., Košir, A., & Tasič, J. (2011). Personality based user similarity measure for a collaborative recommender system. *Proceedings of the 5th ACM Conference on Recommender Systems: Workshop on Human Decision Making in Recommender Systems*, 30–37.

[3] Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48), 12714–12719.

[4] Roulin, N., & Levashina, J. (2019). LinkedIn as a new selection method: Psychometric properties and assessment approach. *Personnel Psychology*, 72(2), 187–211.

[5] Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press.

[6] Ashton, M. C., & Lee, K. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research*, 39(2), 329–358.

[7] Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.

[8] Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036–1040.

[9] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3982–3992.

[10] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

[11] McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11), 205.

[12] Fernández-Tobías, I., Cantador, I., Kaminskas, M., & Ricci, F. (2012). Cross-domain recommender systems: A survey of the state of the art. *Spanish Conference on Information Retrieval*, 24.

[13] Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6), 934–952.

[14] Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30), 17680–17687.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

[16] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.

[17] Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4512–4525.

[18] Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1), 38–44.

[19] Campello, R. J., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1), 1–51.

[20] Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big Five personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, 124, 150–159.

[21] Rastegari, S., & Ghafari, H. (2023). Unsupervised Discovery of Behavioral Clusters using Contrastive Language Models and HDBSCAN. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, 1121–1132.

[22] Guntuku, S. C., Lin, W., Carpenter, J., Ng, W. K., Ungar, L. H., & Preoţiuc-Pietro, D. (2017). Studying personality through the content of posted and liked images on Twitter. *Proceedings of the 2017 International Conference on Web Science (WebSci '17)*, 223–227.

[23] Boyd, R. L., & Pennebaker, J. W. (2018). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, 18, 63–68.

[24] Kim, H. Y. (2015). Statistical notes for clinical researchers: post-hoc multiple comparisons. *Restorative Dentistry & Endodontics*, 42(2), 172–176.

[25] Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education*, 4(3), 279–282.

[26] Reimers, N. (2021). *Sentence-Transformers: Multilingual Sentence Embeddings*. Software (Python library). Available via the `sentence-transformers` project.

[27] van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.

[28] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137.

[29] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[30] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.

[31] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224–227.

[32] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

[33] Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218.

[34] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics.