

DOI: 10.22363/2313-1683-2025-22-1-123-143

EDN: UCTISS

УДК 159.9.07:004.8

Исследовательская статья

Проблема выделения текстовых маркеров депрессии / депрессивности при автоматическом анализе текста

Е.Н. Никитина^{id}, М.А. Станкевич^{id}, Н.В. Чудова^{id}✉

Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Российская Федерация
✉ nchudova@gmail.com

Аннотация. Работа посвящена актуальной междисциплинарной проблеме возможности выявления в текстах признаков депрессии. Цель исследования — определение тех параметров текста, которые могут при автоматическом анализе служить маркерами депрессии / депрессивности. Исследование двух корпусов текстов: эссе, написанных пациентами с депрессией и здоровыми испытуемыми, и постов в социальных сетях, написанных людьми с высокими и низкими показателями депрессивности по шкале Бека, — проводилось с помощью инструмента автоматического анализа текста TITANIS, разработанного в ФИЦ ИУ РАН. Получены данные анализа по 108 параметрам 557 текстов эссе (полученных от 110 пациентов с депрессией, и 447 здоровых испытуемых) и 224 текстов постов в социальных сетях (полученных от 89 испытуемых с признаками депрессивного состояния и 135 испытуемых без признаков депрессивности). Выявлены текстовые особенности, общие и специфичные для текстов больных депрессией и текстов людей с высокими показателями депрессивности по опроснику, представлена их психологическая интерпретация. Проведен лингвистический анализ специфики «депрессивного текста» в группе больных депрессией и в группе «депрессивных» по данным опросника. Показана роль не только частотно-лексических, но и грамматических в широком смысле признаков текста (части речи, морфемы, граммы, именные синтаксемы, признаки связности, членимость текста и др.). По результатам анализа предложены три комплексных показателя депрессии, включающих в себя ряд частных психолингвистических, лингвистических и психологических маркеров. Для текстов людей, имеющих признаки депрессивности по шкале Бека, выделены маркеры постов в соцсетях, объединяющиеся в два комплексных показателя; предложено рассматривать их в массовых обследованиях как показатели недовольства (враждебности), а не депрессии. Сделаны выводы по проблемам выделения текстовых маркеров депрессии и сформулированы предложения по вопросам методологии использования средств автоматического анализа текста в сетевой психодиагностике.

Ключевые слова: депрессия, депрессивность, текстовые маркеры, автоматический анализ текста, машинное обучение, социальные сети

© Никитина Е.Н., Станкевич М.А., Чудова Н.В., 2025



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License
<https://creativecommons.org/licenses/by-nc/4.0/legalcode>

Введение

С появлением социальных сетей исследовательское поле социогуманитарных наук резко расширилось и появилась возможность не задавать вопросы испытуемым/респондентам, а находить ответы в спонтанных сообщениях пользователей соцсетей. Для анализа сетевого контента были созданы различные инструменты, ориентированные почти исключительно на лексико-частотный анализ. Однако возможности автоматического анализа текста расширяют набор собственно лингвистических данных, принадлежащих разным уровням языковой системы, которые могли бы использоваться в сетевой психодиагностике (Смирнов, 2023).

Развитие методов автоматического анализа текста позволяет перейти к корпусным лингвостатистическим исследованиям текстовой продукции испытуемых. В таких междисциплинарных исследованиях свойства текстов сопоставляются со свойствами их авторов и обнаруженные связи далее рассматриваются как значимые для психодиагностики (Ениколопов и др., 2021; Литвинова, 2015; Tausczik, Pennebaker, 2010). Используются как так называемые психолингвистические показатели, введенные в науку еще психиатрами начала XX в., так и эмпирически выявленные параметры текстов, служащие разделителями корпусов при использовании методов математической статистики.

С развитием нейросетевого подхода в области анализа больших данных в отдельное направление выделились исследования, посвященные поискам текстовых маркеров психических нарушений, в первую очередь депрессии. Так, машинное обучение использовалось для построения модели, предсказывающей тяжесть депрессии на основе аудиоданных, видеоданных и текстовых данных англоязычных интервью (Gong, Poellabauer, 2017) и аудиоданных русскоязычных интервью (Шалилех и др., 2023). Во многих работах специалистов по машинному обучению материалом служат сообщения пользователей социальных сетей. Например, при отборе наиболее эффективных моделей машинного обучения для выявления депрессивного состояния по данным англоязычных социальных сетей использовались признаки, связанные с лингвистическим стилем, эмоциональной окрашенностью, категорией времени (Islam et al., 2018). Возможности применения семантического анализа постов в ходе машинного обучения также показаны для русскоязычной социальной сети ВКонтакте (Солохов, Кочкаров, 2024). В задаче выявления пользователей Twitter с признаками депрессии проверена эффективность различных нейросетевых архитектур (Orabi et al., 2018).

С появлением в начале 2020-х гг. технологии больших языковых моделей (LLM) и развитием методов обучения нейросетей проникновение программистов в сферу текстовой психодиагностики еще более усилилось, в связи с чем лингвистический анализ и психологическая интерпретация перестали рассматриваться как необходимые этапы в работе по созданию инструментов автоматического анализа текста в интересах психодиагностики. Сами алгоритмы распознавания текстовых «признаков» депрессии стали рассматриваться как необходимые и достаточные средства выявления среди пользова-

телей социальных сетей людей, страдающих депрессией. Так, в совсем свежей работе прикладных математиков подчеркивается, что «разработанный метод автоматической оценки депрессивного состояния обладает высокой точностью и не требует присутствия врача, что существенно повышает эффективность процесса выявления и оценки уровня депрессии» (Петраевский, Кравец, 2024, с. 95).

Однако с ростом числа работ в области автоматической текстовой психодиагностики растет внимание к проблеме значимости тех или иных текстовых маркеров, к проблеме валидности и надежности получаемых текстовых показателей (Литвинова, 2021). Накопленный в уже многочисленных исследованиях материал позволяет поставить вопрос о целесообразности продолжения использования данных поэлементного анализа, выделяющего в тексте отдельные лексемы, части речи, пунктуационные знаки и т. п. и о необходимости перехода к анализу комплексных показателей, характеризующих психологическую реальность, стоящую за высказыванием автора. Так, в метаанализе 84 публикаций, содержащих данные о связи параметров текстов с показателями опросника «Большой пятерки» (Moreno et al., 2021) вопрос поставлен так: «В литературе можно найти два основных метода анализа языка для изучения личности: методы, основанные на критериях экспертов... и методы, основанные на вычислительных моделях языка. Основное различие между ними заключается в том, что последние модели фокусируются на семантических отношениях между словами, фразами, абзацами и т. д., в то время как первые модели предполагают анализ, который направлен на обнаружение шаблонов (обычно определяемых как правила в регулярных выражениях или грамматике и даже поддерживаемых синтаксическими анализаторами)» (Moreno et al., 2021, p. 2). Сделан вывод о том, что «комбинация синтаксической и семантической информации по сравнению с использованием только одного типа информации» обладает большей эффективностью (Moreno et al., 2021, p. 5). В более ранней работе (Devyatkin, 2019) аналогичный вывод делается на основе анализа опыта работы специалиста по ИИ с лингвистами при анализе научных публикаций.

Вопрос о том, что именно показывают используемые «показатели» и в какой мере маркеры психического нездоровья, имеющие давнюю традицию применения в клинике и хорошо обоснованные теоретически, могут быть дополнены маркерами, полученными в работах, выполненных специалистами по автоматическому анализу текстов, — это вопрос, встающий в последние годы с особой остротой и перед психологами.

В связи с тем что методы машинного обучения работают как «черный ящик» и в принципе не предполагают интерпретации полученных результатов, опять актуально звучит анализ проблем методологии психологического исследования, предложенный Куртом Левином в программной статье «Закон и эксперимент в психологии» 1927 г.: «Проведение экспериментов и анализ их результатов были в огромной степени подчинены духу статистики, основывающемуся на частоте повторения одного и того же опыта и вычислении средних значений. При такого рода статистическом подходе отдельные и даже многочисленные исключения играют весьма незначительную роль. Методы

обоснования (доказательства) определяются идеей приведения возможно большего количества данных» (Левин, 2001, с. 26). В настоящий момент главная проблема применения методов машинного обучения к анализу текстов в интересах психодиагностики заключается в том, что в рамках нейросетевого подхода в качестве текстовых показателей выступают не разноуровневые лингвистические параметры, а лишь сочетания букв и слов (так называемые токены), поскольку текст рассматривается как более или менее вероятное сочетание символов. Таким образом, системность языка и речи не принимается в расчет, так что и никакая психологическая интерпретация, выявляющая закономерные особенности речи автора, обусловленные его психическим статусом, личностными особенностями и коммуникативными задачами, оказывается невозможна в принципе. При этом сами по себе методы машинного обучения могут работать с любым набором параметров, так что перед психолого-лингвистическим исследованием может быть поставлена задача создания осмысленных текстовых показателей. Проще говоря, проблема дальнейшего продвижения в области текстовой психодиагностики видится нам в распространении и на эту область исследований принципа, предложенного Л.С. Выготским, — анализ не «элементов», а далее неразложимых «единиц».

В последние годы мы предпринимаем попытки выйти за пределы вероятностного подхода, в рамках которого нами ранее были получены вполне приличные с точки зрения формальной полноты и точности результаты (Ениколопов и др., 2019; Ковалёв и др., 2019). Неудовлетворенность простой логикой «совместной встречаемости» элементарных текстовых параметров и психологических особенностей авторов текстов подтолкнула нас к созданию лингвистически, психологически и математически обоснованных инструментов — предикатно-эмотивного анализа (Григорьев и др., 2022) и алгоритма распознавания реакций на фрустрацию (Devyatkin et al., 2021). Применялся и математический аппарат, позволяющий искать в данных текстовой психодиагностики причины текстовых «событий», — метод каузальных связей AQ+JSM (Smirnov et al., 2020). В последнем исследовании в двух корпусах текстов, созданных больными с диагностированной депрессией и пользователями соцсетей с высокими показателями по шкале депрессии Бека, методом AQ+JSM искали текстовые параметры, причиной которых (не содержательно, а формально-логически) являлась бы клинически выраженная депрессия или депрессивное состояние, выявленное опросником. Результаты оказались обескураживающими: «Депрессия как заболевание выступила единственной причиной 27 текстовых „событий“, а депрессивность как черта личности — 40 текстовых „событий“; клиническая депрессия имеет 16 текстовых маркеров, депрессивность как черта личности — 27 маркеров; среди текстовых „событий“, причинами которых выступили депрессия и депрессивность, есть только два общих „события“: высокая доля уникальных слов и редкая встречаемость лексики уголовного жаргона» (Smirnov et al., 2020). Это заставило нас изменить тактику — перейти от изучения «элементарных» текстовых показателей к изучению «комплексных», лингвистически и психологически обоснованных показателей, инструменты для измерения которых нами были созданы в последующие два года.

Цель исследования — определение текстовых параметров, способных при автоматическом анализе текстов, написанных на русском языке, выступать маркерами депрессии / депрессивности.

Задачи исследования:

- 1) обработка текстов инструментом корпусного лингвостатистического анализа TITANIS;
- 2) статистический анализ различий текстовых показателей в подкорпусах, написанных авторами с признаками депрессии / депрессивности и с их отсутствием;
- 3) психологический и лингвистический анализ общих и различающихся маркеров депрессивности в обоих корпусах;
- 4) определение психолого-лингвистического радикала «депрессивного» текста.

Процедура и методы исследования

Материалы исследования. Исследование проводилось на двух корпусах текстов.

Первый корпус составили эссе, написанные людьми, страдающими депрессией, и здоровыми. Тексты пациентов с депрессией (110 человек, в том числе 66 женщин, 44 мужчин, средний возраст 28 лет) были предоставлены Научным центром психического здоровья (НЦПЗ). Тексты здоровых людей (447 человек, в том числе 298 женщин, 149 мужчин, средний возраст 23 года) были собраны авторами в ходе психодиагностического обследования студентов очного и заочного отделений нескольких вузов (г. Москва, г. Курган). Испытуемые писали эссе на тему «Я, другие, мир» объемом 1–2 страницы (2–3 тыс. знаков) (Stankevich et al., 2019).

Второй корпус включал в себя тексты, собранные в ходе исследования, посвященного проблеме автоматического выявления депрессивности пользователей социальных сетей по данным их персональных профилей в социальных сетях (Ignatiev, Smirnov, Stankevich, 2022). Пользователи социальной сети ВКонтакте заполняли опросник депрессии Бека и предоставляли доступ к данным своих аккаунтов. В используемый в настоящей работе корпус вошли только те посты, которые были написаны в период, не превышающий 6 месяцев до прохождения психодиагностического обследования. Посты каждого испытуемого за этот период склеивались, и совокупный текст не должен был превышать 5 тыс. знаков. Тексты испытуемых, получивших средние баллы по шкале депрессивности, не вошли в корпус; корпус был составлен из текстов 89 испытуемых (71 женщина, 18 мужчин, средний возраст 23 года), получивших высокие баллы (>29), и текстов 135 испытуемых (80 женщин, 55 мужчин, средний возраст 25 лет), получивших низкие баллы (<11).

В первом корпусе средний объем текста составил 1665 символов, медиана равна 1549 символам, а среднеквадратическое отклонение — 655. Во втором корпусе эти показатели составили 669, 1495 и 1618 соответственно.

Измеряемые показатели. Созданный в ФИЦ ИУ РАН лингвистический анализатор позволяет проводить морфологический, синтаксический и семан-

тический анализ текстов на русском языке. Спецификой этого анализатора является применение метода реляционно-ситуационного анализа Г.С. Осипова (Осипов, Смирнов, Тихомиров, 2008), который базируется на теории именных синтаксем Г.А. Золотовой (Золотова, 1988). На основе данных полного лингвистического анализа в настоящем исследовании вычислялись психолингвистические маркеры, оценивалась частота встречаемости предикатов-эмотивов (с учетом отрицания), проводился частеречный анализ. С опорой на данные нашего лингвистического анализатора работает и созданный в 2021 г. алгоритм распознавания типа фрустрационного реагирования (Девяткин и др., 2021; Devyatkin et al., 2021), что позволило в автоматическом режиме определять наличие в тексте определенного типа реакции на фрустрацию. Благодаря интеграции с лингвистическим анализатором разработанных в ФИЦ ИУ РАН словарей эмоционального реагирования (тематических групп слов, ТГС), включающих около 50 тыс. лексем (Кузнецова, 2020) проводился и лексико-частотный анализ. Общий функционал психодиагностического инструмента автоматического анализа текста TITANIS представлен в работе И.В. Смирнова с коллегами (Smirnov et al., 2021).

В данном исследовании анализ текстов проводился по 108 параметрам, обеспечивающим выявление частоты представленности в текстах следующих характеристик: психолингвистических маркеров возбуждения (коэффициент Трейгера, коэффициент опредмеченности действия, коэффициент логической связности), эмотивных предикатов (позитивы, негативы, амбиваленты, деэмотивы; эмотивы базовых эмоций), направленности фрустрационного реагирования (экстрапунитивная, интропунитивная, импунитивная), тематических групп слов (ТГС), связанных с состоянием стресса (эмоциональный, рациональный и предметно-социальный аспекты, всего 21 ТГС), элементарных синтаксических характеристик, позволяющих проводить комплексный лингвистический анализ.

Методы обработки данных. Оба корпуса текстов проходили процедуру подсчета текстовых показателей при помощи инструмента TITANIS. После выделения показателей в рамках каждого корпуса проводилось сравнение текстовых показателей между выборкой испытуемых без признаков депрессии и выборкой испытуемых с депрессией или депрессивностью по опроснику Бека. Для показателей с нормальным распределением использовался тест Стьюдента для независимых выборок, а для остальных показателей применялся критерий хи-квадрат, с приведением к бинарному формату значений показателей: 1, если значение показателя больше нуля, и 0 во всех остальных случаях.

Результаты исследования

Полученные результаты статистического анализа обоих корпусов с указанием среднего значения маркера в группе «депрессия / здоровые» и в группе «депрессивные / недепрессивные», представлены в табл. 1 и 2 (включены только показатели, которые значимо различают тексты людей с признаками депрессивного состояния и людей, не имеющих его признаков).

Таблица 1 / Table 1

Результаты статистического анализа первого корпуса / Results of statistical analysis of the first corpus

Признак / Marker *	Депрессия, сред. / Subjects with depression, mean	Здоровые, сред. / Healthy subjects, mean
Доля местоим. 1-го лица ед. числа / Pronouns: 1 st person singular rate	0,5272	0,2747
Местоимения-прилагательные / Pronouns-adjectives	0,0552	0,0827
ТГС: Эмоц. и телесн. сост. / Dict.: bodily states and emotions	0,0135	0,006
Доля местоимений 1-го лица / Pronouns: 1 st person singular rate	0,5543	0,3967
Эмотивы: печаль / Emotives: sadness	0,0063	0,0021
Доля глаголов 3-го лица / Verbs: 3 rd person rate	0,1844	0,2973
Эмотивы-негативы / Emotives: negative	0,0133	0,0061
Доля глаголов 1-го лица / Verbs: 1 st person rate	0,1959	0,1057
Доля местоим. 1-го лица мн. числа / Pronouns: 1 st person plural rate	0,0149	0,1116
Доля глаголов прош. врем. / Verbs: past tense rate	0,2467	0,1408
ТГС: социальность предметная / Dict.: healthcare	0,0082	0,0029
Коэффициент Трейгера / Trager coef.	1,2745	0,9879
ТГС: социальность полож. / Dict.: positive social	0,026	0,0177
Доля инфинитивов / Infinitives rate	0,225	0,2753
ТГС: Когнит. деят. и коммун. / Dict.: cognitive communication	0,1266	0,1445
Тональность текста / Sentiment rate	0,0935	3,6104
Местоимения-существительные / Noun-pronouns	0,1241	0,1093
Предлоги / Prepositions	0,1086	0,0993
Частицы / Particles	0,0465	0,0548
Доля местоимений 2-го лица / Pronouns: 2 nd person rate	0,0065	0,0374
ТГС: аффектогенная лексика / Dict.: affective lexis	0,042	0,0483
Союзы / Conjunctions	0,0944	0,1027
Фрустрац. реакция: I группа / Frustration reaction: Group I	0,0317	0,0125
ТГС: рационал. оценка отриц. / Dict.: rational negative assessment	0,0122	0,0093
Число уник. слов / Число слов / Unique words rate	0,5793	0,5528
Эмотивы: удивление / Emotions: wonder	0,0004	0,0015
Наречия / Adverbs	0,0707	0,0631
Доля глаголов 2-го лица / Verbs: 2 nd person rate	0,0042	0,0201
Прич. + дееприч. / кол. предлож. / Participles-gerunds coef.	0,1224	0,2228
ТГС: рационал. оценка полож. / Dict.: rational positive assessment	0,0338	0,0392
Глаголы / Verbs	0,1717	0,164
Число отриц. приставок / Negative prefixes rate	0,0481	0,0434
Фрустрац. реакция: M группа / Frustration reaction: M group	0,0368	0,0537
Доля местоимений 3-го лица / Pronouns: 3 rd person rate	0,0641	0,086
Эмотивы-амбиваленты / Emotives: ambivalents	0,0022	0,0014
Кол. знаков пункт. / Кол. слов / Punctuation per word rate	0,2054	0,2158
Коэф. опредмеченности дейст. / Verbs per nouns coef.	0,8032	0,7455
ТГС: инвективы / Dict.: invectives	0,0029	0,0019
Числительные / Numerals	0,0023	0,0016
ТГС: мягкие инвективы / Dict.: soft invectives	0,004	0,0031
Эмотивы: страх / Emotives: fear	0,0017	0,0011
ТГС: деструктивная активность / Dict.: destructive activity	0,0055	0,0041
ТГС: тематическая экономика / Dict.: topical economics	0,0005	0,0014
Коэфф. логической связности / Logical coherence coef.	0,8997	1,0062
Сред. длина предл. (в кол. слов) / Words per sentence rate	13,3939	14,6944

Примечание: представлены только данные, значимо различающиеся ($p < 0,05$) для больных депрессией и здоровых / *Note:* only markers that differ significantly ($p < 0.05$) between the depressed and healthy subjects are presented.

Таблица 2 / Table 2

**Результаты статистического анализа второго корпуса /
Results of statistical analysis of the second corpus**

Признак / Marker *	Депрессив- ные сред. / Depressive mean	Недепрессив- ные сред. / Non-depressive mean
Фрустрац. реакция: E группа / Frustration reaction: E group	0,575	0,4456
Фрустрац. реакция: inf группа / Frustration reaction: inf group	0,3616	0,4789
Коэффициент Трейгера / Trager coef.	1,34	1,0694
Тональность текста / Sentiment rate	1,1666	4,2826
Глаголы / Verbs	0,1663	0,1526
Эмотивы-негативы / Emotives: negatives	0,0064	0,0038
Наречия / Adverbs	0,0549	0,0638
Предлоги /Prepositions	0,102	0,1118
ТГС: когнит. деятельн. и коммун. / Dict.: cognitive activity and communication	0,0975	0,1083
ТГС: катастрофы / Dict.: catastrophes	0	0,0006
Местоименные наречия / Pronominal adverbs	0,0137	0,0177
Прилагательные / Adjectives	0,0709	0,0792
ТГС: инвективы / Dict.: invectives	0,0052	0,0027
Коэфф. опредмеченности дейст. / Verbs per nouns coef.	0,5893	0,5006
Доля глаголов 2-го лица / Verbs: 2 nd person rate	0,0829	0,0586
Фрустрац. реакция: M группа / Frustration reaction: M group	0,0144	0,0365
Доля местоим. 1-го лица мн. числа / Pronouns: 1 st person plural rate	0,0314	0,0509
Доля прописных символов / Uppercase rate	0,0283	0,0338
Числительные-прилагательные / Numerals-adjectives	0,0002	0,001
Эмотивы-амбиваленты / Emotives: ambivalents	0,0001	0,0006
ТГС: тематическая экономика / Dict.: topical economics	0,0002	0,0007

Примечание: представлены только данные, значимо различающиеся ($p < 0,05$) для испытуемых с высокими и низкими показателями по шкале депрессии / *Note:* only markers that differ significantly ($p < 0.05$) between the subjects with high and low scores on the depression scale are presented.

В первом корпусе из 108 текстовых показателей дифференцирующими оказались 62 показателя; во втором корпусе таких показателей оказалось 29. Из этих показателей в обоих корпусах 11 совпадают, а 5 имеют обратное значение (в одном корпусе показатель выше в группе «депрессивные», в другом — в группе «недепрессивные»).

Обсуждение результатов

Первый корпус текстов

Рассмотрим более подробно данные, полученные в процессе анализа эссе на тему «Я, другие, мир», которое писали пациенты НЦПЗ с диагнозом «депрессия» и здоровые испытуемые.

Тексты пациентов с депрессией. Первое, на что следует обратить внимание, — это отсутствие различий между текстами здоровых и больных по таким показателям, которые из соображений здравого смысла в первую очередь считаются маркерами сниженного настроения и «депрессивного» состояния. В текстах здоровых людей лексика положительных эмоций, эмотивы-

позитивы, лексика мотивации и лексика стенических реакций не повышена по сравнению с текстами больных депрессией. И наоборот, тематика негативной социальности, как и тематика катастроф и всевозможных неприятностей, связанных с проблематикой ЖКХ и власти, не представлена в текстах больных шире, чем в текстах здоровых. Также в текстах больных нет повышенной встречаемости эмотивов стыда и отвращения. Таким образом, лексико-частотный сравнительный анализ, при котором учитываются слова, означающие негативные и позитивные переживания, и слова, связанные в житейском сознании с бытовыми и социальными проблемами, не дает значимых различий. Эти данные указывают на то, что далеко не все текстовые маркеры, «очевидные» для житейского сознания, в частности сознания программистов, создающих методики автоматической диагностики депрессии, являются релевантными задаче распознавания «депрессивных текстов».

Обратимся теперь к обнаруженным статистически значимым различиям.

Во-первых, бросается в глаза неравномерность направленности фрустрационного реагирования — в группе пациентов с депрессией значимо чаще встречаются реакции, направленные на себя (интропунитивное реагирование). При этом различий со здоровыми по экстрапунитивным реакциям (реакциям возмущения, обвинения и требований) не наблюдается, а вот реагирование импунитивного типа, отрицающее значимость неприятностей, встречается в текстах больных реже, чем в текстах здоровых. Действительно, чувство вины и ощущение собственной малоценности при депрессии роднит это состояние с самообвинительной реакцией на фрустрацию (I-реакция) и препятствует самоуспокоению в духе «ничего страшного» (M'-реакция), «никто не виноват» (M-реакция), «как-то обойдётся» (m-реакция).

Во-вторых, обнаружено вполне объяснимое своеобразие текстов больных в части использования эмотивных предикатов. Как уже было сказано, в этих текстах нет повышенной представленности эмотивов таких негативных переживаний, как стыд и отвращение, но зато есть значимое повышение встречаемости эмотивов, описывающих состояние печали и страха. За счет этого и вся группа эмотивов-негативов ярче представлена в текстах больных, а печаль и страх оказываются эмблематичны для текста депрессивного больного. При этом представленность эмотивов удивления значимо снижена, что вполне согласуется с клиническим представлением о депрессии — происходящее вокруг видится больному лишенным красок, серым, как бы покрытым пылью. Таким образом, тонкий анализ лексики — выявление предикатов-эмотивов — дает вполне ясную картину происходящего: при депрессии снижается не столько способность радоваться, сколько способность удивляться, поскольку человек концентрируется на себе и своей немощи, так что печаль-страх превращается в универсальное эмоциональное действие.

Третье, что следует отметить, это то, что анализ лексики, осуществляемый лингвистически корректно, на основе тематических групп слов, содержащих по несколько сотен или тысяч лексем, позволяет описать семантическую специфику текстов пациентов с депрессией. Можно видеть, что в текстах пациентов психиатрического стационара значимо чаще встречаются слова, свя-

занные с темой здравоохранения и темой телесных состояний. Кроме того, употребление лексики рациональности (как позитивной, так и негативной) и лексики коммуникации снижено, что вполне ожидаемо для страдания такого рода. Также снижена и тональность текста (единственный показатель, который в системе TITANIS работает не на основе авторских словарей, а на основе используемого в маркетинге словаря *linis-crowd*).

В завершение психологического анализа текстов пациентов с депрессией остановимся на результатах оценки психолингвистических показателей. Для текстов больных характерно повышение показателей коэффициента Трейгера (отношение количества глаголов к количеству прилагательных) и коэффициента опредмеченности действия (отношение количества глаголов к количеству существительных), а также снижение показателя коэффициента логической связности (отношение общего количества служебных слов (союзов, предлогов, частиц) к числу предложений), что свидетельствует о большем возбуждении авторов этих текстов, об их готовности к общей ненаправленной активности. Связано ли это собственно с депрессией или является отражением текущего статуса больных (выход из фазы депрессии, а не сам эпизод депрессии, когда больные просто не в состоянии написать эссе длиной в целую страницу) либо случайным попаданием в экспериментальную группу преимущественно людей с такой разновидностью депрессии, еще предстоит выяснить.

Лингвистический комментарий к статистически значимым признакам, отличающим тексты эссе пациентов с депрессией, может стать содержательным дополнением, подтверждающим и развивающим психологические трактовки, приведенные выше.

Во-первых, отмеченный выше рост числа эмотивных предикатов печали и страха в эссе пациентов с депрессией сопровождается общим негативизмом, о чем свидетельствует увеличение числа отрицательных приставок и форм (*не, без*).

Обращает на себя внимание уменьшение числа знаков пунктуации в текстах больных. Причиной может быть игнорирование правил пунктуации (своеобразный «поток сознания») либо избегание графических средств в качестве «строительного материала» смайликов и средств интонационной выразительности (восклицательные и вопросительные знаки, особенно в сериях). Если так, то незначительная доля пунктуационных знаков может свидетельствовать о малой эмоциональности текста, проявляемой графическими средствами. Предположение о снижении эмоционального фона в депрессивных текстах находит подтверждение и в уменьшении количества модальных частиц как выразителей эмоциональности — наряду с падением количества пунктуационных знаков как «графических» выразителей эмоциональности.

Во-вторых, среди текстовых параметров, значимо различающих корпуса, есть лингвистические признаки, определяющие связность текста. Это коэффициент логической связности, количество причастий / деепричастий на количество предложений, средняя длина предложения, часть речи *союз*. Уменьшение количества этих признаков в депрессивных эссе проявляет струк-

турно-семантическую дробность письма и мышления, короткую речевую стратегию. Сюда же косвенно можно отнести и рост местоимений 1-го лица ед. ч.: если с точки зрения смысла и тематического разворачивания пишущий в депрессии сосредоточен на собственном Я и утверждении собственного Я, то в структурном плане рост этого показателя может быть приемом, вытесняющим текстовую технику связности, обеспечиваемую «синтаксическими нулями» (Никитина, Онипенко 2019). Сосредоточенность на собственном Я в депрессивных эссе подтверждается уменьшением числа местоимений 2-го лица (в адресатном и обобщенно-личном значении).

В-третьих, важным аспектом интерпретации статистически значимых различий становится разграничение признаков жанра и признаков темы. Поясним на примере роста глаголов в форме прошедшего времени в депрессивных эссе. Форма прошедшего времени проявилась на фоне темы «Я, другие, мир», так как группа «здоровых» развивала рассуждения в форме настоящего неактуального времени, показывая общие закономерности соотношения между личностью и обществом, в то время как авторы в депрессии смещались в область собственного прошлого, личных воспоминаний. Однако при разработке других тем, склоняющих авторов к анализу собственного опыта («Мой прошлогодний отпуск», «Как я был маленьким» и т.п.), выбор глагольных форм здоровыми и больными авторами мог бы и не отличаться. При этом можно полагать, что в целом депрессия склонна «скатываться» к возвращению в прошлое, что маркируется формой прош. и Я-субъектом, но контрастировать это с текстами здоровых будет только на определенном тематическом фоне (скорее всего, каузирующем общие рассуждения).

Еще одна ремарка, касающаяся взаимодействия психологических и лингвистических признаков. В наш список параметров депрессии не попала пока такая категория, как Я-каузатор негативных переживаний или действий других (*Я виноват, Они поссорились из-за меня, Я не обрадовал их*), однако она обсуждается в другой работе (Никитина, Онипенко, 2022) и коррелирует с психологическим показателем — см. выше об интропунитивных реакциях при депрессии.

Второй корпус текстов

Рассмотрим теперь особенности текстов пользователей соцсетей, показавших высокий уровень депрессивности по шкале Бека, и сравним эти особенности с тем, что были обнаружены для текстов пациентов с депрессией.

Тексты депрессивных. Первое, на что можно обратить внимание, это изменение в распространенности различных типов фрустрационного реагирования. Как и в текстах больных депрессией, в текстах депрессивных снижена встречаемость импунитивных реакций. Однако ведущим типом реагирования оказываются не самообвинения, а обвинения, т. е. не чувство вины и готовность справляться с проблемами самому, а агрессия и требовательность к другим. Энергичность обвинительного настроя, отличающая депрессивных пользователей соцсетей от недепрессивных, может быть рассмотрена скорее как признак наличия враждебной картины мира (Ениколопов, Чудова, 2017).

Вторая особенность текстов депрессивных пользователей соцсетей, так же как и больных депрессией, связана с использованием предикатов-эмотивов, описывающих негативные переживания. Число эмотивов-негативов значительно повышено, однако в текстах этих испытуемых нет какого-то общего переживания — представленность классов предикатов, презентирующих печаль, страх, стыд, отвращение, гнев, не разделяют достоверно подкорпуса текстов депрессивных и недепрессивных пользователей. Можно предположить, что среди тех, кто получил высокие баллы по шкале Бека, действительно есть те, кто, как и люди с диагнозом «депрессия», часто использует эмотивы печали и страха, но кроме них в эту группу попали, очевидно, и те, для кого характерны другие негативные переживания.

Наконец, можно отметить, что психолингвистические маркеры и характер лексической составляющей (определяемый по ТГС) дают картину, схожую с тем, что наблюдается для текстов больных депрессией, но картина эта оказывается смазанной и малоинформативной. Из трех психолингвистических показателей возбуждения сработали два (коэффициент Трейгера и коэффициент опредмеченности действия), а лексическое своеобразие подкорпуса постов депрессивных пользователей соцсетей сводится к редкому обсуждению катастроф, экономики и коммуникации и к частому использованию инвектив.

Лингвистический анализ данных по второму корпусу дополняет и подтверждает результаты психологического анализа.

Доля глаголов 2-го лица у депрессивных растет, что является свидетельством диалогизма (необязательно вежливого), обращенности к адресату, и это согласуется с психологической характеристикой пишущих как людей агрессивного склада, склонных к обвинению. На этом фоне понятно уменьшение местоимений 1-го лица мн.ч. — признака человеческого единения.

Общее увеличение доли глаголов сочетается с уменьшением неглагольных частей речи (предлог, прилагательное, наречие). Уменьшение части речи *предлог* свидетельствует о том, что депрессивные пишущие опускают в своих высказываниях пространственно-временные, целевые и причинные синтаксемы, организуемые предложно-падежными формами, — т. е. исключают из высказывания координаты мира и причинно-следственные отношения; эти синтаксемы составляют одну семантическую категорию с местоименными наречиями, количество которых также падает. Связав уменьшение именных и местоименно-наречных компонентов с ростом глагольных высказываний, получаем неполные глагольные предложения, которые должны восполняться из ситуации общения и носить энергичный, категоричный характер. Дополняет поведенческую картину уменьшение в текстах депрессивных количества частей речи *прилагательное* и *наречие*, что может свидетельствовать об отсутствии рефлексии, приоритете действия над наблюдением и размышлением: прилагательные (качественные) отвечают за детализированность, конкретизацию описания предметов внешнего мира, оценочность, наречия несут значения признака действия и предикативного статуюального признака (состояния человека и пространства).

Доля прописных символов в депрессивных постах снижается по сравнению с недепрессивными, что может быть связано с неряшливостью, торопливостью пишущих или с тем, что они мало прибегают к capslock для интонационного выделения и ранжирования информации.

Подведём итоги. Текстовые признаки больных депрессией и лиц с высокими показателями депрессивности не показали существенного сходства. Очевидно, что лингвистически вторая группа — это не «ослабленный вариант» картины, представленный в первой группе, а самостоятельное явление. В целом складывается впечатление, что если в текстах пациентов с депрессией мы видим закономерные речевые проявления состояния и проблем больного человека, то в текстах людей с высокими баллами по опроснику Бека мы встречаемся с разнонаправленными тенденциями, со взаимоослабляющими проявлениями речевой системности, присущей текстам людей с разными психологическими особенностями. Для психодиагностики эта ситуация хорошо знакома: если болезнь проявляет себя во всем и всегда, то для здоровой психики характерно большое разнообразие факторов, оказывающих влияние на конечный продукт деятельности. В конце концов по одному опроснику, даже многошкальному, не говоря уже о специализированных одношкальных, никто не станет давать психологическое заключение. В нашем случае второй корпус составлен из текстов людей, которые характеризуются не только высокими баллами по опроснику депрессивности, но и тем, что они дали согласие на прохождение онлайн-тестирования, что они достаточно часто пишут посты в соцсетях, что они быстро откликнулись на призыв поработать добровольцами для научного исследования. Какой из этих (и множества других, неизвестных нам) факторов дал больший, чем депрессивность, вклад в речевые проявления, обнаруженные в этом корпусе текстов, мы не знаем. В связи с этим нам представляется разумным при решении задачи выделения «депрессивного радикала» в тексте ориентироваться на характеристики, специфичные для текстов пациентов с депрессией.

Выводы

Полученные данные позволяют характеризовать «депрессивный радикал» для развернутого высказывания о себе и своем отношении к миру следующим образом:

— негативизм: эмотивные предикаты негативного характера в целом; эмотивные предикаты печали и страха в частности; отрицательные приставки и формы; самообвинительные реакции;

— фиксация на себе и собственном прошлом: местоимения 1-го лица ед. ч.; интропунитивные реакции; склонность к анализу собственного прошлого, что маркируется формой прошедшего времени и Я-субъектом; сосредоточенность на конкретно-собственном Я, подтверждаемая уменьшением числа местоимений 2-го лица (в адресатном и обобщенно-личном значении); лексика болезни и телесных состояний;

— утомляемость и снижение интереса: снижение эмотивных предикатов удивления; дробность письма и мышления, короткая речевая стратегия,

проявляющаяся в уменьшении лингвистических признаков, определяющих связность текста (коэффициент логической связности, количество причастий / деепричастий на количество предложений, средняя длина предложения, часть речи *союз*); уменьшение представленности в текстах лексики рациональности и лексики коммуникации.

Эти три комплексных текстовых показателя могут выступать «кандидатами» на роль маркера депрессии, в том числе и при автоматическом анализе текстов социальных сетей в задачах изучения динамики массового психического состояния. Разумеется, для подтверждения этого предположения требуется эмпирическая проверка на корпусе текстов сетевых дискуссий.

Текстовые особенности постов в соцсетях людей, имеющих повышенные показатели по шкале депрессии Бека, сводятся:

- к негативизму (рост числа эмотивных предикатов негативного характера; преобладание экстрапунитивных реакций; рост числа инвектив), проявляемому непосредственно в коммуникации (доля глаголов 2-го лица растет, что является свидетельством диалогизма; уже упомянутая склонность к обвинительным реакциям и повышенная требовательность);

- энергичности и категоричности, т. е. сниженной рефлексии и недостатку внимания к деталям, свойствам, обстоятельствам (повышение доли глаголов, не только в отношении к существительным (коэффициент опредмеченности действия) и прилагательным (коэффициент Трейгера), но и к наречиям и предлогам).

Как можно видеть, тексты пользователей соцсетей с повышенными показателями по шкале депрессивности Бека близки текстам эссе пациентов с депрессией только в плане общего снижения настроения и недовольства происходящим. В остальном такие тексты трудно назвать собственно депрессивными. Скорее, эти два комплексных показателя могут использоваться для выявления роста недовольства и агрессии пользователей социальных сетей; опираться при этом следует на такие частные текстовые показатели, как повышение числа эмотивов-негативов, инвектив, глагольности, экстрапунитивных реакций.

Проведенное исследование показало, что, хотя лексические показатели текста широко применяются в психодиагностике, их возможности ограничены, поскольку лексика обусловлена тематикой текста. Лексические показатели могут быть значимы в той части, в какой являются атематичными (например, лексика эмоций). Различия между нормой и патологией в большей мере проявляются в грамматическом обустройстве текста (хотя и грамматика не полностью свободна от тематики).

За рамками настоящего исследования остались вопросы влияния жанра текста и его темы на конкретные текстовые признаки депрессивного состояния его автора. Жанр спонтанного высказывания в соцсетях отличается от жанра эссе, написанного по просьбе экспериментатора. Однако совокупность записанных в свободной форме соображений о различных аспектах жизни, каковой и является «текст», собранный из постов испытуемого за полгода, довольно близок по широте тематического охвата к эссе на тему «Я, другие,

мир». Отметим в связи с этим, что более детализированный учет обстоятельств создания текста — это путь, оправданный как с лингвистической, так и с психологической точки зрения. Как писал К. Левин: «В эксперименте речь идет не о воспроизведении возможно большего числа одинаковых случаев, но о систематическом варьировании условий — т. е. об анализе условий путем осуществления совокупности различных случаев» (Левин, 2001, с. 27).

Проведенное исследование показывает, что усилия при создании средств автоматического анализа текста в интересах психодиагностики можно вкладывать в создание комплексных текстовых маркеров, имеющих как психологическое, так и лингвистическое обоснование. Данный путь, требующий совместной работы психологов, лингвистов и специалистов по искусственному интеллекту, обладает несомненным достоинством в сравнении с путем автоматизации на основе лишь процедур машинного обучения — получаемые таким способом текстовые показатели психического нездоровья и психологического неблагополучия интерпретируемы.

Список литературы

- Григорьев О.Г., Кузнецова Ю.М., Никитина Е.Н., Смирнов И.В., Чудова Н.В. Каузативно-эмотивный анализ. Часть I : Методика изучения эмоциональных реакций пользователей социальных сетей // Психологический журнал. 2022. Т. 43. № 3. С. 114–121. <https://doi.org/10.31857/S020595920020501-7> EDN: CDCABF
- Девяткин Д.А., Ениколопов С.Н., Салимовский В.А., Чудова Н.В. Речевые реакции на фрустрацию: автоматическая категоризация // Психологические исследования. 2021. Т. 14. № 78. <https://doi.org/10.54359/ps.v14i78.160> EDN: NZDKDO
- Ениколопов С., Чудова Н. Проблема проявлений враждебной установки // Психологические исследования. 2017. Т. 10. № 54. <https://doi.org/10.54359/ps.v10i54.360>
- Ениколопов С.Н., Ковалёв А.К., Кузнецова Ю.М., Старостина Е.В., Чудова Н.В. Признаки, характерные для текстов, написанных в состоянии фрустрации // Вестник Московского университета. Серия 14. Психология. 2019. №. 3 С. 66-85. <https://doi.org/10.11621/vsp.2019.03.66> EDN: XDBMYJ
- Ениколопов С.Н., Кузнецова Ю.М., Осипов Г.С., Смирнов И.В., Чудова Н.В. Метод реляционно-ситуационного анализа текста в психологических исследованиях // Психология. Журнал Высшей школы экономики. 2021. Т. 18. № 4. С. 748-769. <http://doi.org/10.17323/1813-8918-2021-4-748-769> EDN: KKDCCT
- Золотова Г.А. Синтаксический словарь : Репертуар элементарных единиц русского синтаксиса / отв. ред. Ю.Н. Караулов. Москва : Наука, 1988. 439 с.
- Ковалёв А.К., Кузнецова Ю.М., Минин А.Н., Пенкина М.Ю., Смирнов И.В., Станкевич М.А., Чудова Н.В. Методы выявления по тексту психологических характеристик автора (на примере агрессивности) // Вопросы кибербезопасности. 2019. № 4 (32). С. 72-79. <https://doi.org/10.21681/2311-3456-2019-4-72-79> EDN: DYMSAU
- Кузнецова Ю.М. Опыт применения тематических групп слов как средства компьютерного анализа текстов // Безопасность: Информация, Техника, Управление : сборник избранных статей по материалам Международной научной конференции. Санкт-Петербург : НАЦРАЗВИТИЕ, 2020. С. 28-36. <https://doi.org/10.37539/SITB294.2020.37.95.003> EDN: FABJYR
- Левин К. Динамическая психология : избранные труды. Москва : Смысл, 2001. 572 с.
- Литвинова Т.А., Литвинова О.А. Идентификация и диагностирование личности автора письменного текста. Воронеж : Воронежский государственный педагогический университет, 2015. 332 с. EDN: YTAXFL

- Литвинова Т. А., Бирюкова Е.Д., Загоровская О.В. Диагностирование пола, возраста и психологических характеристик автора письменного текста с использованием методов корпусной и компьютерной лингвистики: возможности и ограничения // Вопросы криминологии, криминалистики и судебной экспертизы. 2021. № 1. С. 105–111. EDN: JVTTCW
- Никитина Е.Н., Онипенко Н.К. Когнитивно-лингвистическая интерпретация результатов автоматического анализа текстов психически больных // Искусственный интеллект и принятие решений. 2019. Т. 1. № 3. С. 60–69. <https://doi.org/10.14357/20718594190307> EDN: EGPAAJ
- Никитина Е.Н., Онипенко Н.К. Синтаксис и семантика эмотивных глаголов (к проблеме лингвистической интерпретации текстов, написанных авторами с психическими расстройствами) // Вестник Томского университета. 2022. Филология. № 79. С. 109–130. <https://doi.org/10.17223/19986645/79/6> EDN: ZWVCPE
- Осинов Г.С., Смирнов И.В., Тихомиров И.А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. 2008. № 2. С. 3–10. EDN: JYAFSD
- Петраевский В.А., Кравец А.Г. Метод интеллектуального анализа текстовой информации для психиатрической диагностики // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2024. № 2. С. 95–104. <https://doi.org/10.24143/2072-9502-2024-2-95-104> EDN: JCMQGR
- Смирнов И.В. Интеллектуальный анализ текстов на основе методов разноуровневой обработки естественного языка. Москва : ФИЦ ИУ РАН, 2023. 356 с. EDN: VCDQEW
- Солохов Т.Д., Кочкаров А.А. Прогнозирование депрессии на основе пользовательских данных русскоязычной социальной сети // Моделирование, оптимизация и информационные технологии. 2024. Т. 12. № 2. <https://doi.org/10.26102/2310-6018/2024.45.2.016> EDN: HIPMUD
- Шалилех С., Концева А.О., Шишкова Т.И., Худякова М.В., Драгой О.В. Диагностика тяжести симптомов депрессии при помощи объяснимого искусственного интеллекта // Доклады Российской академии наук. Математика, информатика, процессы управления. 2023. Т. 514. № 2. С. 242–249. <https://doi.org/10.31857/S26869543236009> EDN: GOCKCK
- Devyatkin D. Extraction of cognitive operations from scientific texts // Communications in Computer and Information Science. 2019. Vol. 1093. P. 189–200. https://doi.org/10.1007/978-3-030-30763-9_16 EDN: ZPRKDQ
- Devyatkin D., Chudova N., Chuganskaya A., Sharypina D. Methods for recognition of frustration-derived reactions on social media // Lecture Notes in Computer Science. 2021. Vol. 12948 LNAI. P. 17–30. https://doi.org/10.1007/978-3-030-86855-0_2 EDN: OSDDAH
- Gong Y., Poellabauer C. Topic modeling based multi-modal depression detection // Proceedings of the 7th Annual Workshop on Audio/Visual emotion challenge (AVEC '17). NY : Association for Computing Machinery, 2017. P. 69–76. <https://doi.org/10.1145/3133944.3133945>
- Ignatiev N., Smirnov I., Stankevich M. Predicting depression with text, image, and profile data from social media // Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods – ICPRAM. SciTePress, 2022. P. 753–760. <https://doi.org/10.5220/0010986100003122>
- Islam M.R., Kabir M. A., Ahmed A., Kamal A. R. M., Wang H., Ulhaq A. Depression detection from social network data using machine learning techniques // Health Information Science and Systems. 2018. Vol. 6. Article 8. <https://doi.org/10.1007/s13755-018-0046-0> EDN: ANHIXD
- Moreno J.D., Martinez-Huertas J.A., Olmos R., Jorge-Botana G., Botella J. Can personality traits be measured analyzing written language? A meta-analytic study on computational

- methods // *Personality and Individual Differences*. 2021. Vol. 177. Article 110818. <https://doi.org/10.1016/j.paid.2021.110818> EDN: WADDNS
- Orabi A.H., Buddhitha P., Orabi M. H., Inkpen D. Deep learning for depression detection of twitter users // *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. New Orleans, LA : Association for Computational Linguistics, 2018. P. 88-97. <https://doi.org/10.18653/v1/W18-0609>
- Smirnov I. V., Ushakova A. V., Chudova N. V. Method for detecting text markers of depression and depressiveness // *Lecture Notes in Computer Science*. 2020. Vol. 12412 LNAI. P. 325–337. https://doi.org/10.1007/978-3-030-59535-7_24 EDN: YIMANP
- Smirnov I., Stankevich M., Kuznetsova Y., Suvorova M., Larionov D., Nikitina E., Grigoriev O., Savelov M. TITANIS: A tool for intelligent text analysis in social media // *Lecture Notes in Computer Science*. 2021. Vol. 12948 LNAI. P. 232-247. https://doi.org/10.1007/978-3-030-86855-0_16 EDN: DQVODN
- Stankevich M., Kuznetsova Y., Smirnov I., Kiselnikova N., Enikolopov S. Predicting depression from essays in Russian // *Computational Linguistics and Intellectual Technologies*. 2019. Vol. 18. No. 25. P. 637-647. EDN: LQTSTX
- Tausczik Y. R., Pennebaker J. W. The psychological meaning of words: LIWC and computerized text analysis methods // *Journal of Language and Social Psychology*. 2010. Vol. 29. No. 1. P. 24-54. <https://doi.org/10.1177/0261927X09351676>

История статьи:

Поступила в редакцию 5 марта 2024 г.

Доработана после рецензирования 20 сентября 2024 г.

Принята к печати 24 сентября 2024 г.

Для цитирования:

Никитина Е.Н., Станкевич М.А., Чудова Н.В. Проблема выделения текстовых маркеров депрессии / депрессивности при автоматическом анализе текста // Вестник Российского университета дружбы народов. Серия: Психология и педагогика. 2025. Т. 22. № 1. С. 123–143. <http://doi.org/10.22363/2313-1683-2025-22-1-123-143>

Вклад авторов:

Е.Н. Никитина — лингвистический анализ данных, написание и редактирование текста. М.А. Станкевич — сбор и обработка данных, редактирование текста. Н.В. Чудова — концепция и дизайн исследования, психологический анализ данных, написание и редактирование текста.

Заявление о конфликте интересов:

Авторы заявляют об отсутствии конфликта интересов.

Сведения об авторах:

Никитина Елена Николаевна, кандидат филологических наук, старший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Российская Федерация, 117321, Москва, пр-т 60-летия Октября, д. 9. ORCID: 0000-0002-6207-8693; eLibrary SPIN-код: 6989-9498. E-mail: yelenon@mail.ru

Станкевич Максим Алексеевич, младший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Российская Федерация, 117321, Москва, пр-т 60-летия Октября, д. 9. ORCID: 0000-0003-0705-5832; eLibrary SPIN-код: 1916-7298. E-mail: stankevich@isa.ru

Чудова Наталья Владимировна, кандидат психологических наук, старший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Российская Федерация, 117321, Москва, пр-т 60-летия Октября, д. 9. ORCID: 0000-0001-9306-1280; eLibrary SPIN-код: 3421-2959. E-mail: nchudova@gmail.com

DOI: 10.22363/2313-1683-2025-22-1-123-143

EDN: UCTISS

UDC 159.9.07:004.8

Research article

The Problem of Identifying Text Markers of Depression and Depressiveness in Automatic Text Analysis

Elena N. Nikitina^{id}, Maksim A. Stankevich^{id}, Natalia V. Chudova^{id}✉

Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, Moscow, Russian Federation
✉nchudova@gmail.com

Abstract. The paper examines the interdisciplinary topic of the possibility of determining the psychological characteristics of authors from their texts, which may be useful for artificial intelligence methods. The aim of the study was to identify textual markers of depression and depressiveness. For this purpose, a study of two corpora of texts was carried out using a linguistic analyzer developed at the Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences (FRC CSC RAS). One corpus consisted of 557 essays written by patients with clinical depression ($N = 110$) and healthy subjects ($N = 447$), and the other was formed by 224 social media posts written by people with high (89) and low (135) scores of depressiveness on the Beck Depression Inventory (BDI). In total, data on 108 text parameters were obtained for both corpora. The authors identified textual features common and specific to the texts of the depressed patients and the texts of those with a high level of depressiveness according to the questionnaire data, and provided their psychological and linguistic interpretations. At the same time, not only lexical features were taken into account, but also grammatical ones (in the broad sense), such as parts of speech, morphemes, grammemes, locative, temporal and causal noun phrases, indicators of text segmentation and text coherence, etc. Based on the results of the analyses, three complex indicators of depression were proposed, including a number of specific psycholinguistic, linguistic and psychological markers. For the texts of the subjects with signs of depression according to the BDI, markers were selected from social media messages, which were combined into two complex indicators. They are proposed to be considered in mass surveys as indicators of dissatisfaction (hostility) rather than depression. The authors also discuss the theoretically and experimentally identified problem of identifying text markers of depression and formulate proposals on the methodology of using AI tools in network psychodiagnostics.

Key words: depression, depressiveness, text markers, automatic text analysis, machine learning, social networks

References

Devyatkin, D. (2019). Extraction of cognitive operations from scientific texts. *Communications in Computer and Information Science*, 1093, 189–200. https://doi.org/10.1007/978-3-030-30763-9_16

- Devyatkin, D., Chudova, N., Chuganskaya, A., & Sharypina, D. (2021). Methods for recognition of frustration-derived reactions on social media. *Lecture Notes in Computer Science*, 12948 LNAI, 17–30. https://doi.org/10.1007/978-3-030-86855-0_2
- Devyatkin, D., Enikolopov, S., Salimovsky, V., & Chudova, N. (2021). Speech reactions to frustration: automatic categorization. *Psychological Studies*, 14(78). (In Russ.). <https://doi.org/10.54359/ps.v14i78.160>
- Enikolopov, S., & Chudova, N. (2017). The problem of a hostile attitude. *Psychological Studies*, 10(54). (In Russ.). <https://doi.org/10.54359/ps.v10i54.360>
- Enikolopov, S.N., Kovalev, A.K., Kiznetsova, J.M., Chudova, N.V., & Starostina, E.V. (2019). Features of texts written by a frustrated person. *Moscow University Psychology Bulletin*, (3), 66–85. (In Russ.). <http://doi.org/10.11621/vsp.2019.03.66>
- Enikolopov, S.N., Kuznetsova, Y.M., Osipov, G.S., Smirnov, I.V., & Chudova, N.V. (2021). The method of relational-situational analysis of text in psychological research. *Psychology. Journal of the Higher School of Economics*, 18(4), 748–769. (In Russ.). <http://doi.org/10.17323/1813-8918-2021-4-748-769>
- Gong, Y., & Poellabauer, C. (2017). Topic modeling based multi-modal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual emotion challenge (AVEC '17)* (pp. 69–76). NY: Association for Computing Machinery. <https://doi.org/10.1145/3133944.3133945>
- Grigoriev, O., Kuznetsova, Y., Nikitina, E., Smirnov, I., & Chudova, N. (2022). Causative-Emotive Analysis. Part I. Emotional reactions of social networks users research. *Psikhologicheskii Zhurnal*, 43(3), 114–121. (In Russ.). <https://doi.org/10.31857/S020595920020501-7>
- Ignatiev, N., Smirnov, I. V., & Stankevich, M. (2022). Predicting depression with text, image, and profile data from social media. In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods - ICPRAM* (pp. 753–760). SciTePress. <https://doi.org/10.5220/0010986100003122>
- Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. *Health Information Science and Systems*, 6, 8. <https://doi.org/10.1007/s13755-018-0046-0>
- Kovalev, A.K., Kuznetsova, Y.M., Minin, A.N., Penkina, M.Y., Smirnov, I.V., Stankevich, M.A., & Chudova, N.V. (2019). Text analysis approach for identifying psychological characteristics (with aggressiveness as an example). *Voprosy kiberbezopasnosti—Cybersecurity Issues*, 4(32), 72–79 (in Russ.). <https://doi.org/10.21681/2311-3456-2019-4-72-79>
- Kuznetsova, Y.M. (2020). A trial of the thematic groups of words for text mining. In *Security: Information, Technology, Behavior* (pp. 28–36). St. Petersburg: “NATSRZAVITIE” Publ. (In Russ.). <https://doi.org/10.37539/SITB294.2020.37.95.003>
- Levin, K. (2001). *Dynamic psychology: Selected works*. Moscow: Smysl Publ. (In Russ.).
- Litvinova, T. A., & Litvinova, O. A. (2015). *Identification and diagnosis of the personality of the author of a written text*. Voronezh: Voronezh State Pedagogical University Publ. (In Russ.).
- Litvinova, T.A., Biryukova, E.D., & Zagorovskaya, O.V. (2021). Predicting gender, age and psychological characteristics of the author of a written text using corpus and computer linguistics methods: possibilities and limitations. *Issues of Criminology, Criminalistics and Forensic Examination*, (1), 105–111. (In Russ.).
- Moreno, J. D., Martinez-Huertas, J. A., Olmos, R., Jorge-Botana, G., & Botella, J. (2021). Can personality traits be measured analyzing written language? A meta-analytic study on computational methods. *Personality and Individual Differences*, 177, 110818. <https://doi.org/10.1016/j.paid.2021.110818>
- Nikitina, E.N., & Onipenko, N.K. (2019). A cognitive linguistic interpretation of statistical analysis results based on texts by persons with mental disorder. *Artificial Intelligence and Decision Making*, 1(3), 60–69. (In Russ.). <https://doi.org/10.14357/20718594190307>

- Nikitina, E.N., & Onipenko, N.K. (2022). On linguistic component in a psycholinguistic research (based on the material of psych verbs in texts by persons with mental disorders). *Tomsk State University Journal of Philology* (79), 109–130. (In Russ.) <http://dx.doi.org/10.17223/19986645/79/6>
- Orabi, A.H., Buddhitha, P., Orabi, M.H., & Inkpen, D. (2018, June). Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 88–97). New Orleans, LA.: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0609>
- Osipov, G.S., Smirnov, I.V., & Tikhomirov, I.A. (2008). Relational-situational method of search and analysis of texts and its applications. *Artificial Intelligence and Decision Making*, (2), 3–10 (In Russ.).
- Petraevskiy, V.A., & Kravets, A.G. (2024). The textual information intellectual analysis method for psychiatric diagnosis. *Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and Informatics*, (2), 95–104. (In Russ.). <https://doi.org/10.24143/2072-9502-2024-2-95-104>
- Shalileh, S.A., Koptseva, A.O., Shishkovskaya, T.I., Khudyakova, M.V., & Dragoy, O.V. (2023). An explained artificial intelligence-based solution to identify depression severity symptoms using acoustic features. *Doklady Rossiiskoj Akademii Nauk. Matematika, Informatika, Processy Upravleniya*, 514(2), 242–249. (In Russ.) <https://doi.org/10.31857/S26869543236009>
- Smirnov, I. V. (2023). *Intelligent text analysis based on multi-level natural language processing methods*. Moscow: FRC CSC RAS. (In Russ.).
- Smirnov, I., Stankevich, M., Kuznetsova, Y., Suvorova, M., Larionov, D., Nikitina, E., & Grigoriev, O. (2021). TITANIS: A tool for intelligent text analysis in social media. *Lecture Notes in Computer Science*, 12948 *LNAI*, 232–247. https://doi.org/10.1007/978-3-030-86855-0_16
- Smirnov, I.V., Ushakova, A.V., & Chudova, N.V. (2020). Method for detecting text markers of depression and depressiveness. *Lecture Notes in Computer Science*, 12412 *LNAI*, 325–337. https://doi.org/10.1007/978-3-030-59535-7_24
- Solokhov, T.D., & Kochkarov A.A. (2024). Forecasting the depression with user data from Russian language social network. *Modeling, Optimization and Information Technology*, 12(2). (In Russ.). <https://doi.org/10.26102/2310-6018/2024.45.2.016>
- Stankevich, M.A., Kuznetsova, Y.M., Smirnov, I.V., Kiselnikova, N.V., & Enikolopov, S.N. (2019). Predicting depression from essays in Russian. *Computational Linguistics and Intellectual Technologies*, 18(25), 637–647.
- Tausczik, Y.R., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1), 24–54. <https://doi.org/10.1177/0261927X09351676>
- Zolotova, G.A. (1988). *Syntactic dictionary: Repertoire of elementary units of Russian syntax*. Moscow: Nauka Publ. (In Russ.)

Article history:

Received 5 March 2024

Revised 20 September 2024

Accepted 24 September 2024

For citation:

Nikitina, E.N., Stankevich, M.A., & Chudova, N.V. (2025). The problem of identifying text markers of depression and depressiveness in automatic text analysis. *RUDN Journal of Psychology and Pedagogics*, 22(1), 123–143. (In Russ.) <http://doi.org/10.22363/2313-1683-2025-22-1-123-143>

Authors' contribution:

Elena N. Nikitina — linguistic expertise of research materials, text writing, text editing. *Maksim A. Stankevich* — data collection, processing and analysis, text editing. *Natalia V. Chudova* — concept and design of the research, psychological expertise of research materials, text writing, text editing.

Conflicts of interest:

The authors declare that there is no conflict of interest.

Bio notes:

Elena N. Nikitina, PhD in Philology, Senior Researcher at the Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences (9 Ave. of the 60th Anniversary of October, 117321, Moscow, Russian Federation). ORCID: 0000-0002-6207-8693; eLibrary SPIN-code: 6989-9498. E-mail: yelenon@mail.ru

Maksim A. Stankevich, Junior Researcher at the Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences Mo (9 Ave. of the 60th Anniversary of October, 117321, Moscow, Russian Federation). ORCID: 0000-0003-0705-5832; eLibrary SPIN-code: 1916-7298. E-mail: stankevich@isa.ru

Natalia V. Chudova, PhD in Psychology, Senior Researcher at the Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences (9 Ave. of the 60th Anniversary of October, 117321, Moscow, Russian Federation). ORCID: 0000-0001-9306-1280; eLibrary SPIN-code: 3421-2959. E-mail: nchudova@gmail.com