© 2022 г. Ю.Ю. ДЮЛИЧЕВА, канд. физ.-мат. наук
(dyulichevayuyu@cfuv.ru)
(ФГАОУ ВО «Крымский федеральный университет им. В.И. Вернадского», Симферополь)

ВЫЯВЛЕНИЕ АФФЕКТИВНЫХ СОСТОЯНИЙ НА ОСНОВЕ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ КОММЕНТАРИЕВ В СОЦИАЛЬНЫХ СЕТЯХ

В статье рассмотрена задача классификации 3553 англоязычных комментариев из социальной сети Reddit на основе различных подходов к векторизации текстов комментариев: мешок слов, TF-IDF, анализ биграмм на основе точечной взаимной информации PMI и сентимента, глубокая модель представления языка BERT. Применение гибридного подхода на основе векторизации текстов с помощью BERT и анализа биграмм позволило повысить качество классификации комментариев до 91%. На основе кластерного анализа 1857 англоязычных комментариев, содержащих описание тревожностей, с помощью BERT+k-Means были выделены кластеры. В исследовании предложен гибридный подход, основанный на применении метода тематического моделирования LDA, метода анализа тональности VADER, точечной взаимной информации, анализа частей речи и позволяющий выделять биграммы и триграммы для описания кластеров комментариев. Для визуализации извлеченных закономерностей в виде триграмм был построен граф знаний, описывающий предметную область, а сопоставление слов выделенных целевых триграмм со словами кастомного словаря, описывающего различные аффективные расстройства, позволило определить типы психосоциологических стрессоров, с которыми связаны аффективные расстройства.

Ключевые слова: биграммы, сентиментный анализ, LDA, BERT, VADER, BoW, TF-IDF, граф знаний, ментальное здоровье.

DOI: 10.31857/S0005231022120029, **EDN:** KRIUIZ

1. Введение

Анализ негативных настроений на основе текстов комментариев в социальных медиа, связанных с проявлением страха, тревожности, скуки, печали и т.п. является перспективным направлением для оценивания состояния ментального здоровья, в целом, и выявления различных аффективных состояний, в частности. Некоторые исследователи отмечают, что в социальных сетях люди описывают проблемы, симптомы и проявление своей ментальной болезни более свободно, чем на приеме у врача [1, 2]. По этой причине наблюдается рост интереса со стороны исследователей к применению методов обработки естественного языка для выявления закономерностей в текстах

комментариев, характерных для различных типов расстройств и их диагностики. Копперсмит и др. выявили преобладание личных местоимений первого лица в депрессивных комментариях на основе анализа частей речи [3]; Сарсам и др. отмечают преобладание эмоциональных состояний, связанных с выражением печали в сообщениях суицидального характера [4]. Некоторые исследователи отмечают, что измененное эмоциональное состояние и желание намеренно исказить смысл влияют на лингвистические показатели текста, которые могут быть использованы на этапе векторизации текстов комментариев для улучшения качества классификации [5, 6].

Пандемия COVID-19 и, в частности, самоизоляция, масочный режим и вакцинация привели к росту проявления аффективных состояний в комментариях социальных сетей. Так, Зенг и др. исследовали влияние пандемии COVID-19 на выражение депрессивных эмоций в твиттах [7]; Саифуллах и др. продемонстрировали эффективность применения случайного леса совместно с подходом к векторизации на основе TF-IDF для классификации тревожных комментариев на Youtube, связанных с COVID-19 [8].

Разработка новых методов анализа текстов комментариев в области исследования ментального здоровья направлена не только на выявление комментариев, относящихся к различным видам аффективных расстройств, но и на создание систем поддержки принятия решений по оказанию персонализированной помощи людям, страдающим такими расстройствами. Интерес представляет и задача определения той точки невозврата в сообщениях, когда негативное эмоциональное состояние и негативное отношение ко всем аспектам жизни приводят к суицидальной идеации [4].

Целью работы является исследование эффективности применения различных подходов к векторизации текстов комментариев и, в частности, на основе анализа биграмм для решения задач классификации и кластеризации комментариев с описанием различных аффективных расстройств, а также выявление закономерностей, способствующих пониманию психосоциологических стрессоров, с которыми связаны аффективные расстройства.

2. Обзор литературы по тематике исследования

Наиболее исследованной платформой социальных медиа с точки зрения выявления аффективных расстройств на основе анализа текстов комментариев является Твиттер. В табл. 1 приведены некоторые исследования, направленные на извлечение закономерностей из твиттов, способствующих улучшению качества классификации аффективных комментариев.

Волк и др. продемонстрировали эффективность классификации текстов комментариев с учетом сентимента и выявления депрессивных состояний на основе анализа колл-грамм и модели представления языка BERT [5], а Мойин и др. отмечают, что векторизация на основе биграмм существенно увеличивает качество классификации в отличие от использования триграмм [11], поэтому исследование авторов было направлено на выявление би-

Таблица 1. Анализ твиттов для выявления некоторых типов аффективных расстройств

Авторы	Тип аффективного расстройства	Датасет	Особенности	Методы анализа текстов	Результаты
1	2	3	4	5	6
Бирджали и др. [1]	суицидальное расстройство	892 твитта со словами из словаря суицидальных слов	Авторский словарь слов, связанных с суицидальным настроением (мыслями) и оценка семантической схожести на основе WordNet	Векторизация текстов на основе признаков частотности, анализа n-грамм и распознавания частей речи, классификация твиттов на основе SVM, ME и NB	Наибольшая точность классификации (precision) 89,5% достигнута на основе SMO
Рабани и др. [2]	суицидальное расстройство	4266 твиттов	Применение ансамблей методов машинного обучения	Векторизация на основе BoW, TF-IDF, классификация твиттов на основе беггинга, ансамбля голосования, AdaBoost, случайного леса, стекинга	Наибольшая точность классификации (accuracy) 98,5% достигнута на основе случайного леса
Сарсам и др. [4]	суицидальное расстройство	4987 твиттов, из них 1000 твиттов (для обучения модели) из двух классов: твитты с проявлением суицидальных мыслей и обычные твитты	Использование сентиментных признаков, выявленных на основе NRC Affect Intensity Lexicon и SentiStrength	Векторизация на основе мешка слов (BoW), методы на основе лексикона, алгоритмы классификации YATSI и LLGC	Наибольшая точность классификации (accuracy) 86,97% достигнута на основе YATSI+ сентиментные признаки

Таблица 1. (окончание)

1	2	3	4	5	6
Пиллай и др. [9]	стрессовое/ расслабленное состояние	1000 твиттов	Введение шкалы баллов для оценивания экспертами степени стрессового/расслабленного состояния и исследование неоднозначности смысла слов	Предобработка текста твиттов на основе анализа повторяющихся букв, эмотиконов, пунктуации, сентимента и теггирование твиттов с учетом смысла слов	Улучшение точности алгоритмов классификации за счет исследования неоднозначности смысла слов
Ораби и др. [10]	депрессивное состояние	1145 пользователей с исследованием слов твиттов этих пользователей	Оптимизация векторного представления с учетом специфичных слов для предметной области	Векторизация на основе скип-грамм и непрерывного мешка слов	Наилучший результат достигнут на основе однослойной сверточной нейронной сети со слоем глобального максимального пуллинга и оптимизацией векторного представления и составляет 86,967% (AUC-оценка)

грамм и триграмм, описывающих предметную область, и рассмотрение подходов к векторизации текстов комментариев на основе анализа биграмм и их характеристик для решения задач классификации и кластеризации комментариев, содержащих описание аффективных расстройств.

Несмотря на разработку эффективных подходов для предобработки, векторизации и классификации твиттов и текстов комментариев в социальных сетях по классам, соответствующим различным аффективным состояниям, исследования, направленные на извлечение закономерностей, описывающих причины таких состояний, остаются актуальными. Следует отметить, что

комментарии в социальных сетях существенно отличаются от твиттов, поскольку позволяют более подробно описывать мысли и чувства. В частности, людям, страдающим депрессиями, свойственна смена настроения при написании комментариев в социальных сетях [12]. Кроме того, в комментариях людей, страдающих депрессией, отмечается описание позитивного прошлого, которое сменяется описанием негативного настоящего, поэтому требуется разработка специальных методов от этапа предобработки текстов комментариев до этапа классификации, кластеризации и извлечения закономерностей из них с учетом особенностей аффективных расстройств.

3. Датасет и методология исследования

3.1. Классификация текстовых сообщений из двух классов: класс с описанием аффективных состояний и класс обычных комментариев

В исследовании использовался сбалансированный датасет из 3553 комментариев: 1857 комментариев с описанием тревожностей и 1696 обычных комментариев из социальной сети Reddit. Разметка комментариев по двум классам выполнялась вручную с привлечением двух практикующих экспертов, оказывающих помощь людям с различными типами аффективных расстройств. Рассмотренный датасет является частью датасета, описанного в работе [13] и представленного на платформе Kaggle.

В качестве базового алгоритма классификации рассматривался алгоритм случайного леса. Для повышения качества алгоритма классификации изучались различные подходы к векторизации текста:

1) применение мешка слов (BoW);

2) применение меры TF-IDF;

3) применение глубокой модели представления языка BERT;

4) применение анализа биграмм на основе точечной взаимной информации, а также числовых оценок сентимента, полученных с помощью метода VADER, реализованного в python-библиотеке vaderSentiment.

Кратко опишем методы, перечисленные выше. Модель «мешок слов» (BoW) основана на извлечении всех слов из текстов комментариев и сопоставлении им частотности их появления в комментариях. Мера TF-IDF (TF — term frequency, IDF — inverse document frequency) вычисляется как произведение отношения числа вхождений выбранного слова к общему количеству слов в комментарии и инверсии частоты, с которой некоторое слово встречается в комментариях корпуса [14]. Глубокая модель представления языка BERT (Bidirectional Encoder Representations from Transformers) реализует архитектуру трансформер и позволяет учитывать контекст и представление токена, а также его положение внутри предложения и номер предложения в корпусе [15].

Для оценки качества классификации исследуемых подходов на сбалансированном датасете использовались показатели (1)–(4), приведенные ниже, и 5-кратная перекрестная проверка.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{4}$$

где TP, TN, FP, FN — истинно положительные, истинно отрицательные, ложно положительные, ложно отрицательные значения соответственно.

3.2. Кластеризация текстовых сообщений, содержащих описания аффективных состояний

На этапе кластеризации был использован датасет из 1857 сообщений с описанием различных тревожностей. Приведем пример случайного комментария с авторской орфографией из исследуемого датасета: «Приступ длился несколько часов. Это было похоже на проблемы с кровообращением, и я запаниковал и, конечно же, снова оказался в отделении неотложной помощи. На этот раз ко мне сразу же приехал врач. Он хотел поговорить о моем тревожном состоянии. Он сказал, что может провести еще несколько тестов, но не думает, что это поможет.»

Следуя методологии для выявления математической тревожности на основе анализа комментариев MOOK, изложенной в работе [16], для выделения кластеров использовалась векторизация текстов комментариев на основе глубокой модели представления языка BERT и алгоритма кластеризации k-Means.

Рассмотрим гибридный подход, основанный на применении метода тематического моделирования LDA, метода анализа тональности VADER, точечной взаимной информации, анализа частей речи, и позволяющий выделять биграммы и триграммы для описания кластеров комментариев.

Алгоритм анализа биграмм и построение на их основе триграмм основан на следующих основных этапах:

1) извлечение M ключевых слов с наибольшей частотой из тем, определенных на основе латентного размещения Дирихле (LDA) и выделение на основе анализа частей речи существительных или глаголов. Латентное размещение Дирихле направлено на извлечение скрытых (латентных) тем из документов, причем при построении тематической модели и определении количества

тем учитывался показатель когерентности для обеспечения схожести термов в рамках одной темы;

2) извлечение ключевых биграмм кластера, у которых левый и/или правый токен является одним из M ключевых слов тем;

3) во множестве всех биграмм для ключевой биграммы извлекается левая и правая соседняя биграмма и осуществляется склеивание по общим словам для получения триграммы;

4) на основе анализа частей речи удаляются триграммы, содержащие MD (модальные глаголы), более двух наречий или прилагательных (RB, JJ) и т.п.;

5) множество целевых триграмм формируется на основе редких триграмм, имеющих негативную тональность. Редкие триграммы извлекаются на основе значений pPMI согласно (5), а негативная тональность определяется с помощью метода сентиментного анализа VADER (Valence Aware Dictionary and sentiment Reasoner). Метод VADER основан на правилах и словарях, в которых словам из словаря экспертами сопоставлены оценки полярности [17].

Точечная взаимная информация (PMI) вычисляется по формуле

$$PMI(w_1, w_2, w_3) = \log_2 \left(\frac{P(w_1, w_2, w_3)}{P(w_1)P(w_2)P(w_3)} \right),$$

где $P(w_1)$, $P(w_2)$, $P(w_3)$ — вероятность появления токена (слова) w_1, w_2, w_3 соответственно в тексте комментария, $P(w_1, w_2, w_3)$ — вероятность появления тройки слов (w_1, w_2, w_3) – триграммы в тексте комментария.

Для выявления редких триграмм использовалась модификация pPMI, вычисляемая по формуле

$$pPMI(w_1, w_2, w_3) = \max(0, PMI(w_1, w_2, w_3)). \tag{5}$$

Алгоритм анализа триграмм основан на построении всех триграмм для каждого кластера и выделении редких триграмм с негативной тональностью с последующим анализом частей речи на основе шаблонов: (JJ, VB, NN), (NN(P), VBD, NN(S)), (NN, VBN, NN), (NNP, VBG, NNP), (JJ, VBG, NN), (JJ, VB + VB, NN), (NN, JJ, NN) и т.п. или центральное слово триграммы имеет тег зависимости ROOT, где JJ — прилагательное, NN(NNP, NNS) — существительное множественного или единственного числа, VBG — герундий или простое причастие, VB — глагол, VBN — причастие прошедшего времени. Например, на основе шаблона (NN|JJ, VB, NN(S)|JJ) — (существительное | прилагательное, глагол, существительное (множественное число) | прилагательное) были извлечены триграммы: (паническая, произойти, атака), (кататония, обнаружить, симптомы), (нападение, влиять, травматический) и т.п.

Для сопоставления ключевых слов триграмм каждого кластера с типами психосоциологических стрессоров использовался психолингвистический словарь LIWC [19] и построенный на его основе кастомный словарь, содержащий

различные синонимы слов «тревожность», «страх», «одиночество», а также слова, описывающие различные родственные отношения и социальные связи (для сопоставления с социологическим стрессором, связанным с построением отношений); слова, описывающие различные типы боли, части тела, учреждения здравоохранения (для сопоставления с социологическим стрессором, связанным с состоянием здоровья и здравоохранением) и т.п.

4. Результаты исследования

4.1. Классификация текстовых сообщений из двух классов: класс с описанием аффективных состояний и класс обычных комментариев

Результаты 5-кратной перекрестной проверки для оценки качества алгоритма классификации (случайный лес (RF)) в зависимости от различных подходов к векторизации текстов комментариев представлены в табл. 2.

Как видно из табл. 2, наилучшая точность классификации была достигнута за счет расширения векторного пространства на основе анализа биграмм и составила 91,1%.

4.2. Кластеризация текстовых сообщений, содержащих описания аффективных состояний

Предварительный этап обработки комментариев включал удаление пунктуации, стоп-слов, токенизацию, приведение к нормальной форме. Оптимальное количество кластеров для решения задачи кластерного анализа 1857 сообщений, содержащих описание аффективных состояний, оценивалось на основе голосования различных методов определения числа кластеров (силуэтный метод, метод локтя и т.д.), реализованного в R-пакете NbClust [18], и равно 7.

В табл. 3 приведен фрагмент результатов кластерного анализа на основе BERT+k-Means с выделением биграмм и триграмм с негативным сентиментом, содержащих ключевые слова, определенные на основе тематического

Таблица 2. Оценка эффективности различных подходов к векторизации текстов комментариев

Подходы	Accuracy	Precision	Recall	F1
BoW + RF	$0{,}696 \pm 0{,}018$	$0{,}679 \pm 0{,}022$	$0{,}848 \pm 0{,}014$	$0{,}760 \pm 0{,}013$
TF-IDF + RF	$0{,}712 \pm 0{,}019$	$0{,}700 \pm 0{,}014$	$0{,}837 \pm 0{,}034$	$0{,}763 \pm 0{,}015$
BERT + RF	$0{,}723 \pm 0{,}018$	$0{,}736 \pm 0{,}017$	$0{,}770 \pm 0{,}018$	$0{,}750 \pm 0{,}015$
Биграммы (pPMI + + сентимент) + RF	$0{,}714 \pm 0{,}016$	$0{,}725 \pm 0{,}020$	$0{,}796 \pm 0{,}026$	$0{,}759 \pm 0{,}013$
TF-IDF + + Биграммы + RF	$\mathbf{0{,}824 \pm 0{,}012}$	$\mathbf{0{,}832 \pm 0{,}013}$	$\mathbf{0{,}876 \pm 0{,}020}$	$\mathbf{0{,}854 \pm 0{,}011}$
BERT + + Биграммы + RF	$\mathbf{0{,}911 \pm 0{,}018}$	$\mathbf{0{,}926 \pm 0{,}019}$	$\mathbf{0{,}934 \pm 0{,}014}$	$\mathbf{0{,}928 \pm 0{,}015}$

Таблица 3. Пример трех выделенных кластеров и построение их описаний на основе анализа биграмм и триграмм

Кластеры	Мощность	Примеры триграмм, полученных слиянием биграмм, построенных на основе LDA, pPMI, VADER	Примеры триграмм на основе VADER и pPMI
Кластер 1	304	(паническая, происходить, атака), (стресс, не верить, личность), (кататония, обнаружить, симптомы), (класс, напугать, наркотик), (болезнь, притворяющийся, парень), (старая, вращающаяся, дверь), (заснуть, падать, поворачиваться), (выпускной, травмирующий, школа), (маленький, сжимающий, паника), (жилье, плохой, школа)	(прикрытие, вынужденный, отсутствие), (плохой, получить, место), (работы, побочная, зарплата), (работы, обслуживать, оплата), (площадь, сказать, покидать), (книги, не являться, событие), (расходы, вынудить, покрывать)
Кластер 2	289	(желудок, физически, больной), (тревожность, влиять, страх), (свадьба, хронический, тревожность), (мысль, относить, ипохондрия), (терапевт, кликнуть, игра), (мчащийся, постоянно, тревожность), (эмоциональный, чувствовать, сексуальный), (гадость, облажаться, симптомы), (гадость, облажаться, игра), (злой, чувствовать, учащенно дышащий), (тревожность, числа, драма), (инцест, начать, жестокое обращение), (боль, сердце, вдохновение), (сердце, чувствовать, общение), (сердце, отслеживание, частота), (парень, чувствовать, обманутый), (нападение, повлиять, травматический)	(злой, говорить, ревнивый), (чувство, не являться, что-нибудь)
Кластер 3	251	(продукты, бороться, явный), (клинический, ноги, депрессия), (депрессия, посттравматический стрессовый синдром, гнев), (объект, проблемы, плохое обращение), (диагностируемый, тревожность, депрессия), (наркотики, проблемы, злоупотребление), (диета, подразумевать, ненависть), (дети, доставлять боль, нам), (жестокое обращение, инвалидность, ребенок), (квартира, проблемы, жестокое обращение)	(демоны, уничтожить, жизнь), (болезнь, битва, посттравматический стрессовый синдром), (дыхание, мысли, страх), (страхи, исказить, реальность)

моделирования с помощью LDA и ранжирования с помощью точечной взаимной информации.

Например, из триграмм кластера 1 можно выделить закономерности, описывающие панические приступы, состояние кататонии, бессоницы, тревожностей, связанных со школой, наркотиками, условиями проживания и поиском работы.

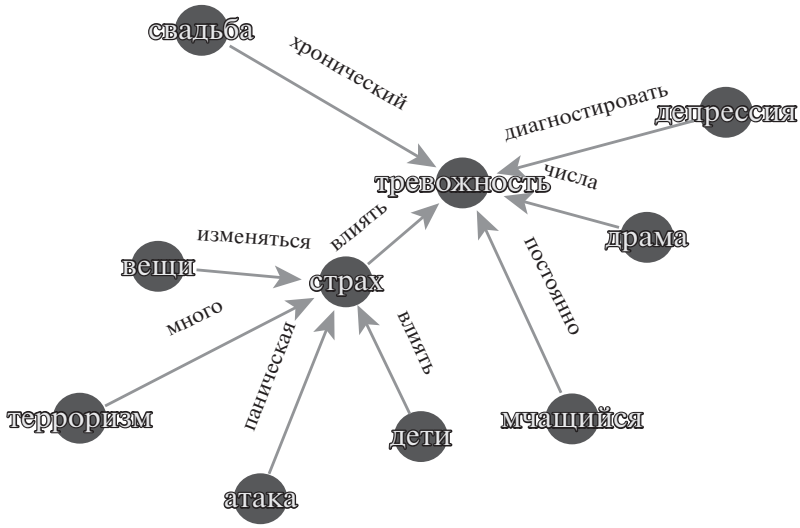
Для реализации алгоритма анализа биграмм использовались следующие python-библиотеки анализа данных: nltk, gensim, spacy, sentence-transformers.

Для каждого кластера отдельно выполнялся частотный анализ слов триграмм на основе кастомного словаря, описывающего психосоциологические стрессоры. Например, частотный анализ триграмм кластера 2 позволил определить, что большая часть комментариев рассматриваемого кластера (48% всех триграмм кластера) описывает проблемы, связанные со здоровьем, например, (сердце, отслеживание, частота), (желудок, физически, больной), (мысль, относить, ипохондрия) и т.п. Значительная часть комментариев рассматриваемого кластера (21% всех триграмм кластера) описывает проблемы, связанные с построением социальных отношений, например, (свадьба, хронический, тревожность), (злой, говорить, ревнивый), (парень, чувствовать, обманутый), (инцест, начать, жестокое обращение) и т.п.

4.3. Построение графа знаний на основе анализа биграмм и триграмм

Графы знаний хорошо зарекомендовали себя в области визуализации извлеченных закономерностей, выделении основных характеристик и демонстрации взаимосвязей между ними. Графы знаний широко применяются в области медицины, например, для представления медицинских знаний по инсультам [20], для демонстрации персонализированных предложений по подбору рациона для людей, страдающих диабетом [21], однако исследования, связанные с представлением в виде графа знаний закономерностей, описывающих проблемы с ментальным здоровьем, автору неизвестны. Далее предлагается построение графа знаний для описания аффективных состояний на основе полученных триграмм.

Для каждой триграммы (w_{i-1}, w_i, w_{i+1}) строятся вершины графа с метками w_{i-1} и w_{i+1}, а также ребро с меткой w_i. Сначала выбирается целевое слово из словаря аффективных расстройств, например, «тревожность», извлекаются все триграммы, построенные на основе предложенных выше шаблонов со словом «тревожность», например, (свадьба, хронический, тревожность), (тревожность, влиять, страх), (мчащийся, постоянно, тревожность), (тревожность, числа, драма), (тревожность, диагностировать, депрессия) и т.п. Вершинами графа становится целевое слово, например, «тревожность», а также слова следующих частей речи, входящие в триграмму: существительное или любое слово с тегом ROOT (корневое слово). Ориентированным ребрам графа приписываются слова триграмм, связанные с описанием аффективного состояния и относящиеся к частям речи прилагательное, наречие, причастие,



Фрагмент графа знаний для слов «тревожность» и «страх», относящихся к проявлению аффективных состояний, составлено автором.

если слов таких частей речи нет в триграмме, то ребру приписывается оставшееся в триграмме слово. Для построения графа знаний использовалась библиотека для распознавания частей речи spacy и библиотека для построения графов PyViz. Фрагмент графа знаний, демонстрирующий построенные закономерности для слов «тревожность» и «страх», относящихся к аффективным состояниям, представлен на рисунке.

Из графа знаний, представленного на рисунке, видны тревожности, вызванные депрессией, свадьбой, драмой, а также описание тревожности, как хронической, связанной с числами, или возникающей под влиянием страха.

5. Заключение

Активное использование социальных сетей привело к накапливанию огромного количества комментариев, которые оставляют пользователи. Методы обработки естественного языка совместно с алгоритмами машинного обучения позволили получить интересные результаты в области оценки эмоционального состояния, как отдельных групп пользователей социальных сетей, так и общества в целом. В последнее время активно развивается такое направление киберпсихологии как оценивание психического состояния на основе анализа комментариев в социальных сетях и влияние различного контента на физическое и ментальное здоровье человека.

В работе продемонстрирована эффективность применения биграмм для повышения качества классификации комментариев, содержащих описание аффективных расстройств, и возможности извлечения биграмм и триграмм для описания предметной области. Дальнейшие исследования автора будут

направлены на улучшение качества извлекаемых закономерностей для выявления причин различных типов психосоциологических стрессоров, приводящих к проявлению тревожных расстройств в текстах комментариев социальных медиа.

СПИСОК ЛИТЕРАТУРЫ

1. *Birjali M., Beni-Hssane A., Erritali M.* Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks // 8th Int. Conf. Emerging Ubiquitous Systems and Pervasive Networks, Procedia Computer Science. 2017. V. 113. P. 65–72.
2. *Rabani S.T., Khan O.R., Khanday Akib Mohi U.D.* Detection of Suicidal Ideation on Twitter using Machine Learning & Ensemble Approaches // Baghdad Sci. J. 2020. V. 17. No. 4. P. 1328–1339.
3. *Coppersmith G., Dredze M., Harman C.* Quantifying Mental Health Signals in Twitter // Proc. Workshop Comput. Linguist. Clinical Psychol.: From Linguist. Signal Clinical Reality. Associat. Comput. Linguist. 2014. P. 51–60.
4. *Sarsam S.M., Al-Samarraie H.A., Ahmed I., Alnumay A., Smith A.P.* A Lexicon-based Approach to Detecting Suicide-related Text on Twitter // Biomed. Signal Proc. Control. 2021. V. 65. No. 102355.
5. *Wolk A., Chlasta K., Holas P.* Hybrid Approach to Detecting Symptoms of Depression in Social Media Entries // Twenty-Fifth Pacific Asia Conf. Information Systems. 2021. arXiv:2106.10485.
6. *Gillam L., Tariq M., Ahmad K.* Terminology and the Construction of Ontology // Terminology. 2005. V. 11. No. 1. P. 55–81.
7. *Zhang Y., Lyu H., Liu Y., Zhang X., Wang Yu., Luo J.* Monitoring Depression Trend on Twitter during the COVID-19 Pandemic: Observational Study // JMIR Format. Res. 2020. 39 p.
8. *Saifullah S., Fauziah Yu., Aribowo A.S.* Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety based on Social Media Data // arXiv:2101.06353, 2021.
9. *Pillai R.G., Thelwall M., Orasan C.* Detection of Stress and Relaxation Magnitudes for Tweets // WWW '18: Compan. Proc. Web Conf. 2018. P. 1677–1684.
10. *Orabi A.H., Buddhitha P., Orabi M.H., Inkpen D.* Deep Learning for Depression Detection of Twitter Users // Proc. Fifth Workshop Comput. Linguist. Clinical Psychol.: From Keyboard Clinic. 2018. P. 88–97.
11. *Moyeen S.I., Mabud Md.S.R, Nayem Z., Mamun Md.Al.* Sentiment Analysis of English Tweets using Bigram Collocations // EPRA Int. J. Res. Development (IJRD). 2021. V. 6. I. 9. P. 220–227.
12. *Величко А.Н., Карпов А.А.* Аналитический обзор систем автоматического определения депрессии по речи // Artificial Intellig., Knowledge and Data Engineer. 2021. No. 3. P. 497–529.
13. *Turcan E., McKeown K.* Dreddit: A Reddit Datasets for Stress Analysis in Social Media // arXiv: 1911.00133v1. 2019.
14. *Jones K.S.* A Statistical Interpretation of Term Specificity and Its Application in Retrieval // J. Document. 2004. V. 60. No. 5. P. 493–502.

15. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv: 1810.04805. 2018. https://doi.org/10.48550/arXiv.1810.04805

16. *Дюличева Ю.Ю.* Учебная аналитика МООК как инструмент анализа математической тревожности // Вопросы образования (Educat. Stud. Moscow). 2021. No. 4. С. 243–265.

17. *Hutto C., Gilbert E.* VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text // Eight Int. AAAI Conf. Weblogs and Social Media. 2014. V. 8. No. 1. P. 216–225.

18. *Charrad M., Ghazzali N., Boiteau V., Niknafs A.* NbClust: An R Package for Determining the Relevant Number of Clusters in DataSet // J. Statist. Software. 2014. V. 61. No. 6. P. 1–36. https://doi.org/10.18637/jss.v061.i06

19. *Tausczik Y.R., Pennebaker J.W.* The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods // J. Lang Soc Psychol. 2010. V. 29. No. 1. P. 24–54.

20. *Cheng B., Zhang J., Liu H., Cai M., Wang Y.* Research on Medical Knowledge Graph for Stroke // J. Healthcare Engineer. 2021. V. 2021 (5531327).

21. *Haussmann S., Seneviratne O., Chen Y. et al.* FoodKG: A Semantics-Driven Knowledge Graph for Food Recommendation // Semant.-Web – ISWC. 2019. P. 146–162.

Статья представлена к публикации членом редколлегии А.А. Лазаревым.

Поступила в редакцию 31.01.2022

После доработки 28.05.2022

Принята к публикации 29.06.2022