УДК 004.03.032.26 ***Самойлова Л., Даниловский В.М.***

Самойлова Л.

Сибирский государственный индустриальный университет

(г. Новокузнецк, Россия)

Даниловский В.М.

Сибирский государственный индустриальный университет

(г. Новокузнецк, Россия)

ПРОГНОЗИРОВАНИЕ ЛИЧНОСТНЫХ ХАРАКТЕРИСТИК MBTI С ИСПОЛЬЗОВАНИЕМ РЕКУРРЕНТНОЙ НЕЙРОННОЙ СЕТИ LSTM И ТЕКСТОВЫХ ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ

Аннотация: *в данной статье рассматривается задача прогнозирования типов личности по системе Майерс-Бриггс (MBTI) на основе текстовых данных с использованием алгоритмов машинного обучения. Исследование фокусируется на применении модели Long Short-Term Memory (LSTM) в сочетании с техникой Random Oversampling для достижения высокой точности предсказаний. Используется публичный набор данных из Kaggle, содержащий тексты постов пользователей социальных сетей. Результаты показывают, что подход с использованием LSTM и RMSprop оптимизатора с learning rate 10^-3 обеспечивает наивысшую точность в 86,31%. Эта работа подчеркивает важность предобработки данных и балансировки классов для улучшения производительности моделей машинного обучения.*

Ключевые слова: *нейронная сеть, прогнозирование личности, машинное обучение, текстовые данные, социальные сети, анализ данных.*

Личностные характеристики играют важную роль в различных аспектах жизни, включая карьеру, здоровье и личные отношения. Методология Майерс-Бриггс (MBTI) является одним из наиболее популярных инструментов для оценки личности. Основанная на учениях Карла Юнга в 1943 году, она

расширяет его типологию о темпераментах до 16 типов личности, комбинируя дихотомии в различные сочетания [1].

В последние годы, с развитием технологий и социальных сетей, автоматическое прогнозирование личности на основе текстовых данных стало актуальной задачей. На рисунке 1 сможем увидеть, какие предикторы за какую характеристику отвечают.

Для данного исследования используется публичный набор данных из Kaggle [2], содержащий 8675 постов пользователей, каждый из которых имеет метку MBTI типа. Датасет включает в себя 16 типов личности, таких как INTJ, ESFP и другие. Каждый пользователь имеет 50 образцов текстов, что в сумме составляет 433750 комментариев.

Можем увидеть на рис.2, что данные несбалансированные - типы INFP, INFJ, INTP и INTJ оставляли в разы больше записей, а ESFP, ESFJ, ESTJ минимальное количество. Чтобы результаты исследования вышли объективными, потребуется балансировка данных.

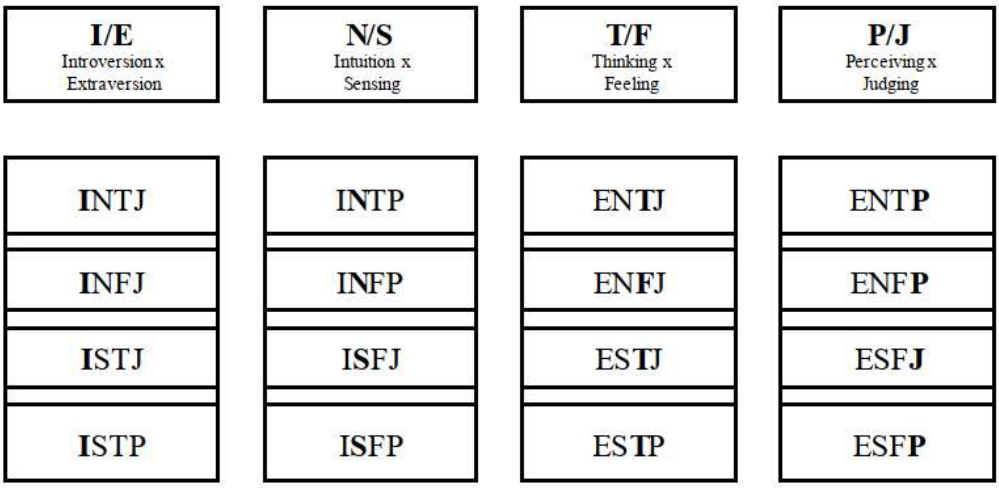

Рисунок 1. Схематическое отображение четырех дихотомических шкал.

Предобработка данных включает несколько этапов:

- очистка данных - удаление ссылок, приведение текста к нижнему регистру, удаление стоп-слов и пунктуации,

- токенизация - разделение текста на отдельные слова,

- лемматизация - приведение слов к их начальной форме.

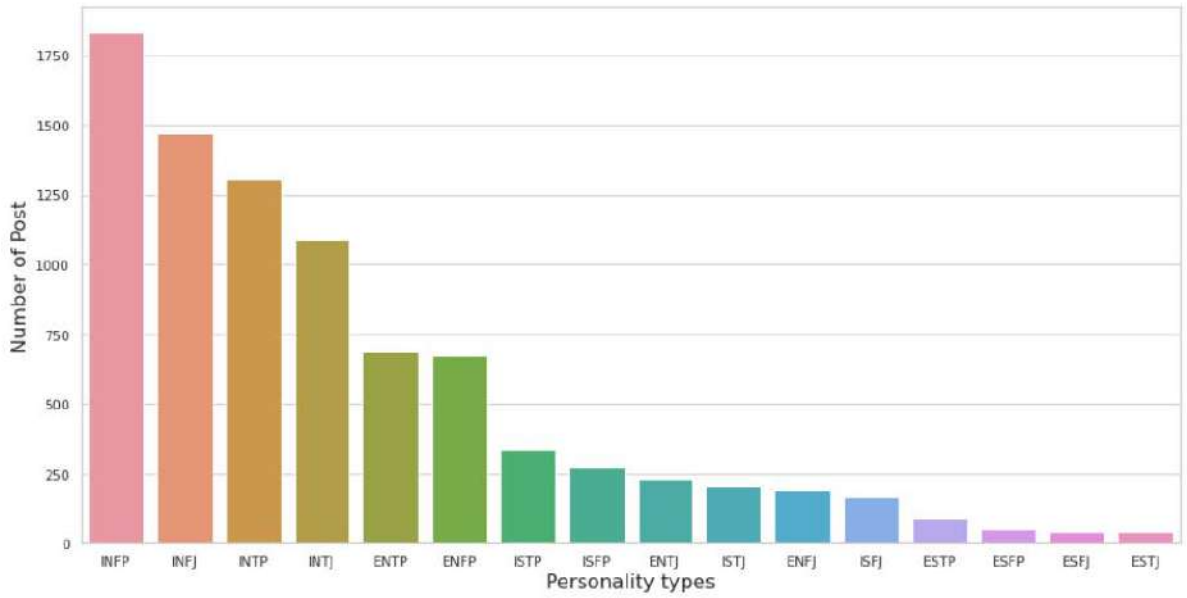

Рисунок 2. Диаграмма количества постов каждого типа.

Проблема дисбаланса данных решается с помощью техники Random Oversampling, которая увеличивает количество примеров для меньшинств, сохраняя информацию из большинства классов. Это особенно важно для алгоритмов машинного обучения, так как дисбаланс данных может привести к смещенным и неэффективным моделям. Формула для определения точности модели может быть представлена следующим образом:

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

где TP – количество истинно положительных предсказаний,
TN – количество истинно отрицательных предсказаний,
FP – количество ложноположительных предсказаний,
FN – количество ложноотрицательных предсказаний.

Искусственные нейронные сети представляют собой математическую модель, которая воспроизводит принципы функционирования биологических нейронных сетей. Они состоят из искусственных нейронов, подобных биологическим, и способны обрабатывать информацию. Структура искусственной нейронной сети включает входной слой, скрытые слои (при наличии) и выходной слой (Рисунок 3). Входной слой получает внешние данные, а затем сигналы передаются через синапсы нейронам следующего слоя для обработки. Скрытые слои выполняют сложные логические преобразования, а выходной слой формирует окончательный результат. Искусственные нейронные сети могут быть простыми или глубокими, в зависимости от количества скрытых слоев, что позволяет им находить сложные взаимосвязи в данных и выполнять глубокое обучение [1].

Рассмотрим модель простой трехслойной искусственной нейронной сети, представленной на рисунке 3. Она содержит:

Векторы входного слоя (input layer), принимают данные и передают их в нейронную сеть для последующей обработки, представлены как $I = [\, I_1, I_2, I_3, \dots, I_M]^T$, где I_i – это i-ый элемент вектора входных данных, а M – количество входных нейронов.

Скрытый слой (hidden layer), в который поступают данные для обработки, вектор скрытого слоя представлен $H = [\, H_1, H_2, H_3, \dots, H_O]^T$, где H_i – это i-ый элемент вектора входных данных, а O – количество нейронов.

Векторы выходного слоя (output layer) $O = [\, O_1, O_2, O_3, \dots, O_N]^T$, где O_i – это i-ый элемент вектора входных данных, а N – количество выходных нейронов.

Процесс передачи данных от входного слоя к скрытому заключается в использовании матрицы весов $W^{[1]}$. Вычисление активации скрытого слоя происходит по формуле:

$$H = f_1\left(W^{[1]}I + b^{[1]}\right), (2)$$

где $W^{[1]} = [w_{qm}]_{Q\times M}$, это матрица весов от входного слоя к скрытому,

$b^{[1]} \in R^{Q\times 1}$ – матрица смещения скрытого слоя,

$f_1(*)$ – функция активации скрытого слоя.

Аналогично происходит вычисление передачи данных от скрытого слоя к выходному:

$$O = f_2\left(W^{[2]}H + b^{[2]}\right), (3)$$

где $W^{[2]} = [w_{nq}]_{N\times Q}$, это матрица весов от скрытого слоя к выходному,

$b^{[2]} \in R^{N\times 1}$ – матрица смещения выходного слоя,

$f_2(*)$ – функция активации выходного слоя.

Более сложные искусственные нейронные сети, содержащие большее количество скрытых слоев, имеют сильную вычислительную мощность и могут находить сложные взаимосвязи в данных [4]. Это делает их мощным инструментом для анализа большого объема информации [5].

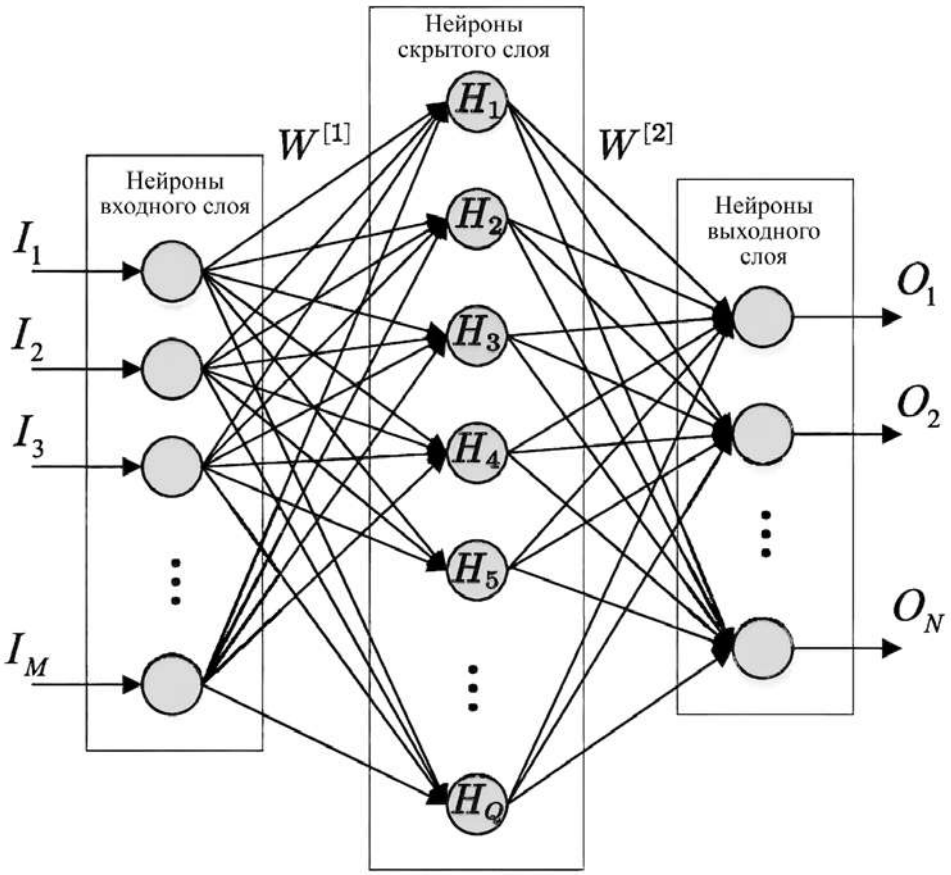

Рисунок 3. Структурная схема модели трехслойной ИНС.

Для прогнозирования MBTI типов личности используется модель Long Short-Term Memory (LSTM). Архитектура сети включает следующие слои:

- входной слой Embedding,
- два слоя LSTM с 128 и 64 нейронами соответственно и Dropout 0.2,
- полносвязный слой с 64 нейронами и активацией ReLU,
- выходной слой с 4 нейронами и активацией Softmax.

Оптимизаторы, используемые в модели: Adadelta, Adam, RMSprop и SGD с разными значениями learning rate (10^-2, 10^-3, 10^-4). Также исследование включает применение следующих алгоритмов: Naive Bayes, Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN).

Программная часть. Для реализации модели использовались следующие инструменты и библиотеки: Python, основной язык программирования для разработки модели. Pandas, для работы с табличными данными. NumPy, для

численных вычислений. Scikit-learn, для реализации различных алгоритмов машинного обучения. Keras, для создания и обучения модели LSTM. NLTK, для обработки текстов на этапе предобработки данных. Imbalanced-learn, для балансировки данных с помощью техники Random Oversampling.

Решение задач классификации, присутствующих в исследовании, используется кросс-энтропия или функция потерь. Определяется по формуле:

$$-\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)], (4)$$

где: N — количество классов (в нашем случае N = 16, так как у нас 16 типов личности MBTI),

y_i — истинная метка класса для i-го примера (принимает значение 1, если пример принадлежит к данному классу, и 0 в противном случае),

p_i — предсказанная вероятность того, что i-й пример принадлежит к данному классу.

Результаты. Тестирование модели показало, что использование RMSprop оптимизатора и learning rate 10^-3 обеспечивает наивысшую точность 86,31%. Результаты модели с и без Random Oversampling представлены в таблице ниже.

Полученные результаты подтверждают, что модель LSTM с Random Oversampling и оптимизатором RMSprop является эффективным подходом для прогнозирования типов личности MBTI.

Таблица 1. Результаты точности моделей LSTM с и без Random Oversampling для различных значений learning rate.

Learning Rate	**Accuracy (без ROS)**	**Accuracy (с ROS)**
10^-2	22,28%	82,28%
10^-3	35,72%	86,31%
10^-4	42,38%	83,50%

Это демонстрирует важность правильной предобработки данных и использования методов балансировки для улучшения производительности модели. Дальнейшие исследования могут включать более сложные методы обработки текста и использование более крупных и разнообразных датасетов.

Заключение. Настоящее исследование демонстрирует, что применение LSTM в сочетании с техникой Random Oversampling позволяет достичь высокой точности в прогнозировании MBTI типов личности на основе текстовых данных. Будущие работы могут быть направлены на улучшение методов обработки данных и использование дополнительных источников текстовой информации для повышения точности предсказаний.

СПИСОК ЛИТЕРАТУРЫ:

1. Geyer, Peter. Psychological Type and the MBTI ®: Past, Present and Future? // Научное издание. – 2013. – DOI: 10.13140/2.1.1088.1928. – Режим доступа: https://www.researchgate.net/publication/264789741_Psychological_Type_and_the_MBTI_R_Past_Present_and_Future (дата обращения: 06.05.2024);
2. Datasnaek. MBTI Type Dataset [Электронный ресурс]. – URL: https://www.kaggle.com/datasets/datasnaek/mbti-type?resource=download (дата обращения: 06.05.2024);
3. Ростовцев, В.С. Искусственные нейронные сети: учебник для вузов / В. С. Ростовцев. — 2-е изд., стер. — Санкт-Петербург: Лань, 2021. — 216 с. — ISBN 978-5-8114-7462-2. — Текст: электронный // Лань: электронно-библиотечная система. — URL: https://e.lanbook.com/book/160142 (дата обращения: 24.03.2024);
4. Başaran, S., & Ejimogu, O. H. (2021). A Neural Network Approach for Predicting Personality from Data. Sage Open, 11(3). https://doi.org/10.1177/21582440211032156 (дата обращения: 26.03.2024);
5. Зинченко Алексей Алексеевич Количественное моделирование процесса подбора персонала // Управленческие науки. 2015. №3. URL:

https://cyberleninka.ru/article/n/kolichestvennoe-modelirovanie-protsessa-podbora-personala (дата обращения: 27.03.2024);

6. Ontoum, Sakdipat, Chan, Jonathan H. Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning. – 2021

Samoilova L., Danilovsky V.M.

Samoilova L.

Siberian State Industrial University

(Novokuznetsk, Russia)

Danilovsky V.M.

Siberian State Industrial University

(Novokuznetsk, Russia)

PREDICTION OF PERSONAL CHARACTERISTICS MBTI USING RECURRENT NEURAL NETWORK LSTM AND TEXT DATA OF SOCIAL NETWORKS

Abstract: *this article discusses the problem of predicting personality types according to the Myers-Briggs system (MBTI) based on text data using machine learning algorithms. The research focuses on the application of the Long Short-Term Memory (LSTM) model in combination with the Random Oversampling technique to achieve high prediction accuracy. A public dataset from Kaggle is used, containing texts of posts from social network users. The results show that the approach using LSTM and RMSprop optimizer with a learning rate of 10^-3 provides the highest accuracy of 86.31%. This work highlights the importance of data preprocessing and class balancing to improve the performance of machine learning models.*

Keywords: *neural network, MBTI, personality prediction, machine learning, LSTM, Random Oversampling, text data, social networks, data analysis.*