УДК 004.8:004.912

О СООТВЕТСТВИИ АВТОМАТИЧЕСКОЙ ПРОЦЕДУРЫ РАЗМЕТКИ МЕТОДОМ ЛУЧШЕГО-ХУДШЕГО МАСШТАБИРОВАНИЯ РЕШЕНИЮ ЗАДАЧ АНАЛИЗА ТОНАЛЬНОСТИ ФИНАНСОВЫХ НОВОСТЕЙ

Г.Б. Никифоров
Ассистент отделения информационно-коммуникационных технологий образовательного департамента Передовой инженерной школы гибридных технологий в станкостроении Союзного государства, ПсковГУ
nikiforov.gb@pskgu.ru

В.И. Пименов
Заведующий кафедрой информационных технологий, СПбГУПТД, доктор технических наук, с.н.с.
v_pim@mail.ru

Д.А. Андреев
Заведующий отделением информационно-коммуникационных технологий образовательного департамента Передовой инженерной школы гибридных технологий в станкостроении Союзного государства, ПсковГУ, кандидат технических наук
d.andreev@pskgu.ru

Г.Б. Никифоров

Окончил Псковский государственный университет. Область научных интересов связана с разработкой методов и алгоритмов обработки естественного языка и машинного обучения в финансовой сфере.

В.И. Пименов

Окончил Балтийский государственный технический университет «Военмех» имени Д.Ф. Устинова. Область научных интересов связана с разработкой методов и алгоритмов интеллектуального анализа данных, когнитивных технологий, компьютерных систем обработки и анализа данных, компьютерной графики, распознавания образов, мультимедиа технологий и 3D-моделирования.

Д.А. Андреев

Окончил Псковский государственный политехнический институт. Область научных интересов связана с разработкой моделей, алгоритмов и показателей качества формализованного описания и анализа технологий производства продукции.

Аннотация. В работе рассматривается автоматическая процедура разметки методом лучшего-худшего масштабирования (BWS) для непрерывного анализа тональности финансовых новостных заголовков на русском языке. На наборе данных из 532 финансовых новостных заголовков выполнено сопоставление автоматической процедуры разметки относительно ручной. Показано, что автоматическая процедура разметки сохраняет межаннотаторскую согласованность ($к \approx 0,30$) *и внутреннюю надёжность при расщеплении* ($SHR \approx 0,87$), *обеспечивает сопоставимость итоговых оценок* ($R2 \approx 0,49$). *Результаты подтверждают применимость предобученной большой языковой модели (LLM) без дообучения на наборе данных для масштабируемой количественной разметки финансовых текстов в условиях ограниченного объёма данных.*

Ключевые слова: искусственный интеллект, машинное обучение, предобученные большие языковые модели, автоматическая разметка, анализ тональности.

Введение. Количественный анализ тональности финансовых новостей является значимым компонентом моделей прогнозирования рыночных показателей и систем поддержки инвестиционных решений [1, 2].

В отличие от задач классификации с дискретными метками, в прикладных финансовых исследованиях востребованы непрерывные оценки тональности, отражающие интенсивность и направленность информационного воздействия.

Формирование таких оценок осложняется ограниченным объёмом размеченных наборов данных и выраженной субъективностью интерпретации финансовых новостей [3].

Одним из подходов к получению непрерывных оценок является метод лучшего-худшего масштабирования (Best-Worst Scaling, BWS), основанный на относительном сравнении текстов и обеспечивающий устойчивость итоговых оценок при сравнении объектов.

Ручная процедура разметки методом лучшего-худшего масштабирования характеризуется высокой трудоёмкостью и низкой масштабируемостью, что затрудняет её применение при потоковой обработке данных.

Умеренная согласованность между аннотаторами дополнительно увеличивает требования к объёму разметки для обеспечения устойчивости оценок.

Развитие предобученных больших языковых моделей создаёт предпосылки для автоматизации сравнительной разметки, включая автоматическую процедуру разметки методом лучшего-худшего масштабирования без дополнительного обучения на специализированных наборах данных [4].

При этом в условиях отсутствия объективного эталона разметки оценка качества автоматической разметки требует сопоставления с ручными оценками и анализа метрик качества разметки [5].

Это обуславливает необходимость эмпирической проверки сопоставимости автоматической и ручной процедур разметки методом лучшего-худшего масштабирования по характеристикам согласованности и практической применимости в задачах количественного анализа тональности.

Постановка задачи. В исследовании использован набор данных FiNeS [6], содержащий финансовые новостные заголовки и их ручную разметку.

Ручная разметка выполнена методом лучшего-худшего масштабирования, предполагающего попарные сравнительные задачи выбора наиболее и наименее позитивного текста.

Пусть задано множество финансовых новостных заголовков:

$$T=t1,t2,...,tN, \quad (1)$$

где t_i – один финансовый новостной заголовок; i = 1, 2,…, N – индекс финансового новостного заголовка; N = 532 – количество финансовых новостных заголовков в наборе данных FiNeS.

Для каждого финансового новостного заголовка t_i определяется непрерывная оценка тональности $si \in [-1, 1]$.

В рамках исследования рассматриваются две её реализации: H – результат ручной процедуры разметки методом лучшего-худшего масштабирования;

- siA – результат автоматической процедуры разметки методом лучшего-худшего масштабирования с использованием предобученной большой языковой модели без дополнительного обучения на наборе данных.

Таким образом, задача исследования заключается в реализации проведения автоматической процедуры разметки методом лучшего-худшего масштабирования на основе предобученной большой языковой модели для построения непрерывных оценок тональности финансовых новостей в сравнении с полученными результатами ручной процедуры разметки.

Для этого требуется:

1. оценить межаннотаторскую согласованность решений и внутреннюю надёжность автоматической процедуры разметки при сопоставлении с ручной;
2. оценить степень количественного соответствия автоматических оценок ручным на непрерывной шкале;
3. оценить применимость автоматической разметки для использования в задачах количественного анализа тональности.

Важно отметить, что поставленная задача имеет ряд ограничений: малый объём размеченных данных, использование непрерывной шкалы оценивания и отсутствие этапа обучения модели на размеченных данных.

Методология исследования. Разметка реализована посредством решения попарных сравнительных задач.

Для каждого заголовка t_i сформировано множество из L = 8 других заголовков, с которыми он участвует в попарных сравнительных задачах.

Таким образом, общее число сформированных попарных сравнительных задач составило:

$$K=NL=532\cdot8=4256, \quad (2)$$

где K – число попарных сравнительных задач; N – число финансовых новостных заголовков в наборе данных; L – число назначенных попарных сравнительных задач для каждого заголовка.

Каждая попарная сравнительная задача имела вид:

$$Tk=ti,tj,ij, \quad (3)$$

где T_k – k-я попарная сравнительная задача двух заголовков, k = 1, 2, …, K – номер попарной сравнительной задачи; i, j = 1, 2, …, N – индексы финансовых новостных заголовков набора данных.

Формирование попарных сравнительных задач осуществлялось случайным образом без обучения или предварительного структурирования, что минимизировало систематические смещения и обеспечило равномерное покрытие набора данных.

В ручной процедуре разметки методом лучшего-худшего масштабирования для каждой попарной сравнительной задачи аннотатор выбирал более позитивный заголовок из пары, второй заголовок фиксировался как менее позитивный.

Вычисление ручной оценки тональности $s_i^{(H)}$ для финансового новостного заголовка t_i осуществлялось по формуле:

$$\text{siH}=\text{BiH-WiHCi}, \quad (4)$$

где $B_i^{(H)}$ – число выборов новостного финансового заголовка t_i как более позитивного, $W_i^{(H)}$ – число выборов новостного финансового заголовка t_i как менее позитивного, C_i – число попарных сравнительных задач, в которых участвовал заголовок t_i.

Автоматическая процедура разметки методом лучшего-худшего масштабирования реализована предобученной большой языковой моделью GigaChat 2 Lite [7], без дополнительного обучения на наборе данных. Для каждой попарной сравнительной задачи выполнено отображение:

$$\text{M:Tkt+,t-}, \quad (5)$$

где M – предобученная большая языковая модель, T_k – попарная сравнительная задача, t^+ – элемент, выбранный как более позитивный, t^- – элемент, выбранный как менее позитивный.

Вычисление автоматической оценки тональности $s_i^{(A)}$ осуществлено аналогично ручной процедуре (4):

$$\text{siA}=\text{BiA-WiACi}, \quad (6)$$

где $B_i^{(A)}$ – число выборов новостного финансового заголовка t_i как более позитивного, $W_i^{(A)}$ – число выборов новостного финансового заголовка t_i как менее позитивного, C_i – число попарных сравнительных задач, в которых участвовал заголовок t_i.

Для оценки качества разметок использовались:

- коэффициент межаннотаторской согласованности Флейсса κ, который характеризует межаннотаторскую согласованность и отражает степень совпадения решений разных участников [8];
- показатель надёжности при расщеплении *SHR* (split-half reliability) с поправкой Спирмена-Брауна, рассчитываемый как корреляция оценок, полученных на двух случайных подмножествах попарных сравнительных задач, который оценивает внутреннюю устойчивость оценок тональности при разбиении выборки попарных сравнительных задач на две части [9].

Количественное соответствие автоматических оценок ручным определялось коэффициентом детерминации R^2, который характеризует долю объяснённой дисперсии ручных оценок, объясняемую автоматической процедурой разметки.

Экспериментальные результаты. На рисунке 1 представлено распределение оценок тональности si для ручной и автоматической разметок набора данных FiNeS.

По оси абсцисс отложены значения оценок тональности в диапазоне $[-1;1]$, по оси ординат – плотность распределения.

Ручная разметка отображена штриховкой, автоматическая – сплошной заливкой.

Оба распределения охватывают полный интервал шкалы и характеризуются неоднородной плотностью по диапазону значений.

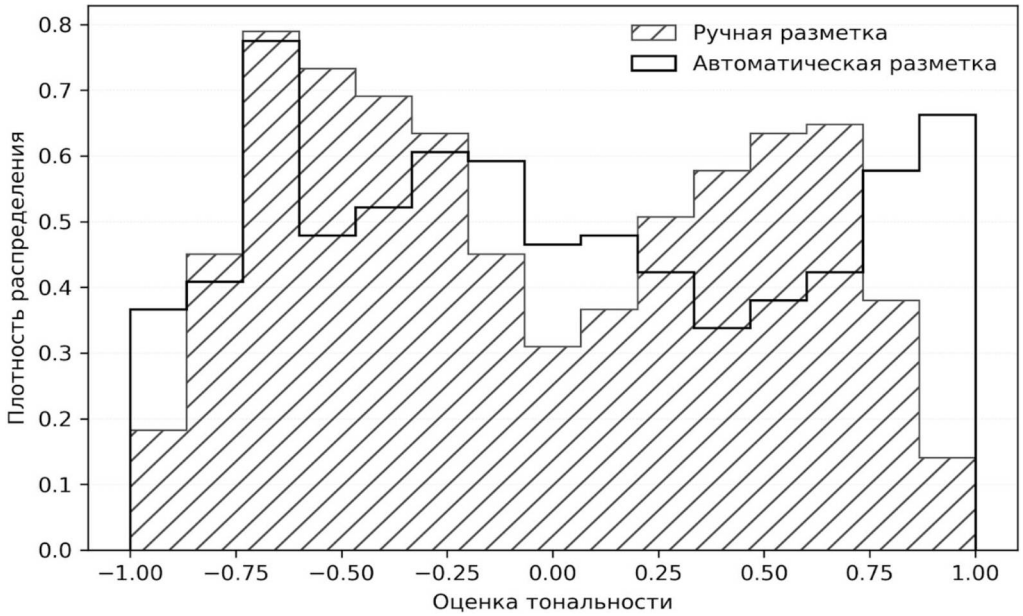

Рисунок 1. Распределение оценок тональности для ручной и автоматической разметок

Для ручной разметки коэффициент межаннотаторской согласованности Флейсса составил κ(H) ≈ 0,26, показатель надёжности при расщеплении – SHR(H) ≈ 0,85. Для автоматической разметки получены значения κ(A) ≈ 0,30 и SHR(A) ≈ 0,87 соответственно. Коэффициент детерминации между оценками ручной и автоматической разметок составил R2 ≈ 0,49.

Обсуждение результатов. Полученные значения коэффициента согласованности Флейсса κ и показателя надёжности при расщеплении SHR отражают различные характеристики процедуры разметки методом лучшего-худшего масштабирования. Значения κ ≈ 0,26-0,30 соответствуют умеренному уровню межаннотаторской согласованности, что свойственно для задач интерпретации тональности финансовых текстов, допускающих вариативность трактовок и контекстную неоднозначность. В то же время показатели SHR ≈ 0,85-0,87 отражают высокую внутреннюю устойчивость оценок тональности si при повторном случайном разбиении попарных сравнительных задач, что указывает на их стабильность.

Сопоставление ручной и автоматической разметок не выявило снижения межаннотаторской согласованности (κ) и внутренней надёжности (SHR) при переходе к автоматической процедуре разметки. Для автоматической процедуры разметки зафиксированы более высокие значения как κ, так и SHR, что позволяет говорить о сохранении структурных свойств разметки и отсутствии негативного влияния на воспроизводимость оценок. Распределения тональности (рисунок 1) демонстрируют сходную форму и сопоставимый диапазон значений, что подтверждает близость статистических характеристик двух процедур разметки.

Коэффициент детерминации R2 ≈ 0,49 указывает на значимую долю объяснённой дисперсии ручных оценок, объясняемую автоматической процедурой разметки, и свидетельствует об умеренной количественной сопоставимости оценок при сохранении естественных расхождений, обусловленных субъективностью задачи и случайным формированием попарных сравнительных задач.

В совокупности результаты показывают, что переход к автоматической процедуре разметки методом лучшего-худшего масштабирования не приводит к ухудшению показателей согласованности и обеспечивает сопоставимый уровень устойчивости оценок тональности s_i. Применение предобученной большой языковой модели в роли аннотатора

позволяет реализовать полностью автоматическую процедуру разметки методом лучшего-худшего масштабирования без дополнительного обучения, что потенциально расширяет возможности масштабирования и применения метода в условиях потокового поступления данных.

К дополнительным ограничениям исследования следует отнести использование одной языковой модели, относительно небольшой объём набора данных (532 заголовка) и анализ исключительно на уровне заголовков без учёта полного текста публикаций. Указанные факторы ограничивают обобщаемость результатов и требуют дополнительной валидации результатов на расширенных наборах данных и при использовании большего количества моделей.

Заключение. Проведённое исследование показало, что автоматическая процедура разметки методом лучшего-худшего масштабирования на основе предобученной большой языковой модели демонстрирует сопоставимые, а по отдельным метрикам – несколько более высокие значения межаннотаторской согласованности и надёжности при расщеплении в сравнении с ручной процедурой разметки. Кроме того, автоматической процедуре разметки указанным методом не свойственны проблемы, связанные с высокой трудоёмкостью и ограничениями масштабируемости в условиях потокового поступления данных. *Значение коэффициента детерминации $R2 \approx 0{,}49$ показывает, что автоматическая процедура разметки объясняет около 49% дисперсии ручных оценок тональности s_i*, и подтверждает возможность использования автоматической процедуры разметки в задачах количественного анализа тональности.

Таким образом, автоматическая процедура разметки методом лучшего-худшего масштабирования может рассматриваться как инструмент формирования и расширения специализированных финансовых наборов данных без этапа дополнительного обучения модели.

Дальнейшее повышение показателей качества разметки и коэффициента детерминации R^2 связано с оптимизацией стратегий формирования попарных сравнительных задач и проведением ансамблирования моделей. Последующее развитие работы связано с переходом от случайного формирования подмножеств попарных сравнительных задач к более структурированным стратегиям их конструирования.

В частности, возможно формирование групп на основе семантической близости заголовков либо с учётом текущего распределения полученных оценок тональности. Такой подход потенциально повысит информативность парных сравнений и устойчивость оценок тональности s_i.

Отдельным направлением является интеграция автоматической процедуры разметки методом лучшего-худшего масштабирования с динамическими моделями анализа информационных потоков, в которых оценки тональности могут использоваться в качестве количественного признака в задачах прогнозирования рыночных показателей.

Предложенная процедура разметки также может применяться для расширения специализированных финансовых наборов данных ограниченного объёма и построения регрессионных моделей тональности с непрерывной целевой переменной. Дополнительная проверка обобщаемости результатов требует экспериментов на более крупных выборках и при использовании дополнительных языковых моделей.

Список литературы

[1] Kirtac, K. Sentiment Trading with Large Language Models / K. Kirtac, G. Germano // Finance Research Letters. – 2024. – Vol. 62. – Part B. – Article 105227. – PP. 1–9. – DOI: 10.1016/j.frl.2024.105227.

[2] Zhang, Z. FinSentLLM: Multi-LLM and Structured Semantic Signals for Enhanced Financial Sentiment Forecasting / Z. Zhang, R. Fu, Y. He et al. // arXiv preprint. – 2025. – PP. 1–5. – DOI: 10.48550/arXiv.2509.12638.

[3] Gilardi, F. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks / F. Gilardi, M. Alizadeh, M. Kubli // Proceedings of the National Academy of Sciences. – 2023. – Vol. 120. – No. 30. – Article e2305016120. – PP. 1–3. – DOI: 10.1073/pnas.2305016120.

[4] Bagdon, C. «You are an expert annotator»: Automatic Best-Worst-Scaling Annotations for Emotion Intensity Modeling / C. Bagdon, P. Karmalker, H. Gurulingappa, R. Klinger // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Mexico: Association for Computational Linguistics, 2024. – Vol. 1. – PP. 7924–7936. – DOI: 10.18653/v1/2024.naacl-long.439.

[5] Dhurandhar, A. Ranking Large Language Models without Ground Truth / A. Dhurandhar, R. Nair, M. Singh et al. // Findings of the Association for Computational Linguistics (ACL). – Bangkok: Association for Computational Linguistics, 2024. – PP. 2431-2452. – DOI: 10.18653/v1/2024.findings-acl.143.

[6] Financial News Sentiment Dataset (FiNeS) [Электронный ресурс]. – URL: https://github.com/WebOfRussia/financial-news-sentiment (дата обращения: 23.02.2026).

[7] GigaChat 2 Lite | Документация для разработчиков [Электронный ресурс]. – URL: https://developers.sber.ru/docs/ru/gigachat/models/gigachat-2-lite (дата обращения: 23.02.2026).

[8] Zapf, A. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? / A. Zapf, S. Castell, L. Morawietz, A. Karch // BMC Medical Research Methodology. – 2016. – Vol. 16. – Article 93. – PP. 1–10. – DOI: 10.1186/s12874-016-0200-9.

[9] Pronk, T. Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment / T. Pronk, D. Molenaar, R. W. Wiers, J. Murre // Psychonomic Bulletin & Review. – 2021. – Vol. 29. – PP. 44–54. – DOI: 10.3758/s13423-021-01948-3.

Авторский вклад

Никифоров Глеб Борисович – постановка задачи исследования, разработка и реализация автоматической процедуры разметки методом лучшего-худшего масштабирования, проведение эксперимента, анализ результатов, подготовка текста статьи.

Пименов Виктор Игоревич – постановка задачи исследования, формирование научной концепции исследования, интерпретация результатов, редактирование статьи.

Андреев Дмитрий Анатольевич – постановка задачи исследования, методическое сопровождение исследования, экспертная оценка результатов, редактирование статьи.

ON THE APPLICABILITY OF AN AUTOMATIC ANNOTATION PROCEDURE BASED ON BEST–WORST SCALING FOR FINANCIAL NEWS SENTIMENT ANALYSIS

G.B. Nikiforov
Assistant, Department of Information and Communication Technologies, Educational Department, Advanced Engineering School of Hybrid Technologies in the Machine Tool Industry of the Union State, PskovSU

V.I. Pimenov
Head of the Department of Information Technologies, SPbSUITD, Doctor of Technical Sciences, Associate Professor

D.A. Andreev
Head of the Department of Information and Communication Technologies, Educational Department, Advanced Engineering School of Hybrid Technologies in the Machine Tool Industry of the Union State, PskovSU, PhD of Technical Sciences

Abstract. This paper examines an automatic annotation procedure based on Best-Worst Scaling (BWS) for continuous sentiment analysis of Russian-language financial news headlines. A comparison between automatic and manual annotation procedures was conducted on a dataset consisting of 532 financial news headlines. The results show that the automatic annotation procedure preserves inter-annotator agreement ($\kappa \approx 0.30$) and split-half reliability (SHR ≈ 0.87), while ensuring comparability of the resulting sentiment scores ($R2 \approx 0.49$). The findings confirm the applicability of a pretrained large language model (LLM), used without additional fine-tuning on the dataset, for scalable quantitative annotation of financial texts under limited data conditions.

Keywords. artificial intelligence, machine learning, pretrained large language models, automated annotation, sentiment analysis.