

УДК 004.8:004.912:004.932
DOI 10.17513/snt.40903



CC BY 4.0

РОБАСТНАЯ КЛАССИФИКАЦИЯ ЭМОЦИЙ В КИТАЙСКИХ РУКОПИСНЫХ ТЕКСТАХ В УСЛОВИЯХ ШУМА ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ С ПРИМЕНЕНИЕМ ПРОФИЛЬНО-ОРИЕНТИРОВАННОЙ АУГМЕНТАЦИИ ДАННЫХ

Чжан Юйи ORCID ID 0009-0005-2435-9100

*Высшая школа технологий искусственного интеллекта
Института компьютерных наук и кибербезопасности федерального
государственного автономного образовательного учреждения высшего образования
«Санкт-Петербургский политехнический университет Петра Великого»,
Санкт-Петербург, Российская Федерация, e-mail: zyuyi2671@gmail.com*

В работе проверяется, что происходит с классификацией эмоций, когда на вход модели попадает не ручная транскрипция китайской рукописной строки, а результат PaddleOCR. Исходный корпус удобен для такой проверки: в нем есть изображения строк и их текстовые записи, но нет экспертной эмоциональной разметки. Поэтому сначала была зафиксирована слабая разметка, полученная по лексическим признакам и текстовому классификатору, а уже затем на ней обучались итоговые модели. Основным вариантом оставлена задача на семь эмоций. Одиннадцатиклассовая схема используется отдельно, как проверка поведения редких категорий. В режиме text-only лучшей оказалась RoBERTa: Accuracy 0,9304 и Macro-F1 0,8863. После замены ручных транскрипций OCR-выходом Macro-F1 снизилась до 0,8390. Для обучения с учетом таких ошибок был собран профиль фактических замен, удалений и вставок PaddleOCR; на его основе применялась профильно-ориентированная аугментация. В сквозном сценарии Macro-F1 выросла до 0,8636, а 95%-й доверительный интервал для разности не включал ноль. В одиннадцатиклассовом варианте Macro-F1 равна 0,8760, но редкие классы остаются ненадежными для самостоятельных выводов. Ручная проверка 480 объектов также показала, что псевдометки лишь частично совпадают с оценкой аннотатора. Поэтому приведенные метрики следует читать прежде всего как оценку устойчивости внутри выбранной слабой разметки, а не как точность относительно полноценного экспертного стандарта.

Ключевые слова: китайский рукописный текст, классификация эмоций, шум оптического распознавания символов, слабоконтролируемое обучение, профильно-ориентированная аугментация данных, надежность псевдометок

ROBUST EMOTION CLASSIFICATION OF CHINESE HANDWRITTEN TEXT UNDER OPTICAL CHARACTER RECOGNITION NOISE USING PROFILE-DRIVEN DATA AUGMENTATION

Zhang Yuyi ORCID ID 0009-0005-2435-9100

*Higher School of Artificial Intelligence Technologies, Institute of Computer Science
and Cybersecurity of the Federal State Autonomous Educational Institution
of Higher Education “Peter the Great Saint Petersburg Polytechnic University”,
Saint Petersburg, Russian Federation, e-mail: zyuyi2671@gmail.com*

This study examines what happens to emotion classification when the model receives not a manual transcription of a Chinese handwritten text line, but the output of PaddleOCR. The source corpus is suitable for this setting because it contains line images and their textual transcriptions, but it does not include expert emotion labels. A weak-label set was therefore fixed first, using lexical cues and a text classifier, and the final models were trained only after that step. The main setting uses seven emotion classes. The eleven-class scheme is treated separately as a test of rare-category behavior. In the text-only setting, RoBERTa achieved the best result, with an accuracy of 0.9304 and a Macro-F1 of 0.8863. After manual transcriptions were replaced with OCR output, Macro-F1 decreased to 0.8390. To train the model with such errors in mind, a profile of actual PaddleOCR substitutions, deletions, and insertions was collected and used for profile-driven data augmentation. In the end-to-end scenario, Macro-F1 increased to 0.8636, and the 95 % confidence interval for the difference excluded zero. In the eleven-class setting, Macro-F1 reached 0.8760, but rare classes remained unreliable for separate conclusions. A manual review of 480 items also showed that pseudo-labels only partly agreed with the annotator's judgments. The reported scores should therefore be read mainly as robustness estimates within the chosen weak-labeling scheme, not as accuracy against a full expert gold standard.

Keywords: Chinese handwritten text, emotion classification, optical character recognition noise, weakly supervised learning, profile-driven data augmentation, pseudo-label reliability

Введение

В готовом цифровом тексте классификатор получает уже проверенную строку. В случае рукописного китайского текста ситуация другая: сначала изображение проходит через OCR, и модель работает с тем, что распознаватель смог прочитать. Здесь ошибка даже в одном эмоционально значимом иероглифе может изменить итоговый класс. В используемом корпусе есть изображения и ручные транскрипции строк, но нет экспертной разметки эмоций. Поэтому весь эксперимент приходится рассматривать с двумя оговорками: качество зависит и от ошибок OCR, и от способа построения псевдометок.

В качестве основной схемы в статье оставлены семь эмоций: радость, грусть, гнев, страх, удивление, отвращение и тревога. По этой схеме дальше сравниваются два режима входа: ручная транскрипция и текст, полученный после PaddleOCR. На том же наборе проверяется и PDA, то есть обучение с шумом, похожим на реальные ошибки распознавания. Одиннадцатиклассовый вариант эксперимента вынесен за рамки основного результата – он выполняет функцию проверочного среза, показывающего, что происходит с моделью после расширения набора классами *neutral*, *calm*, *resignation* и *sarcasm*. Добавление этих категорий делает распределение классов еще менее сбалансированным, и именно это обстоятельство представляет отдельный интерес для анализа.

Цель исследования – оценка надежности классификации эмоций в китайских рукописных текстах в условиях шума оптического распознавания символов и проверка того, повышает ли аугментация, построенная на реальном профиле ошибок распознавания, устойчивость модели. Дополнительная цель состоит в явном описании надежности псевдометок и границ интерпретации результатов, полученных в слабоконтролируемой постановке.

Для достижения цели решались следующие задачи: сформировать слабоконтролируемые метки; сравнить текстовые, визуальные и мультимодальные модели; оценить влияние сквозного распознавания; проверить эффективность профильно-ориентированной аугментации; провести статистическое сравнение моделей; оценить надежность псевдометок и определить границы их интерпретации.

Отдельно проверяется гипотеза о полезности аугментации, построенной на основе таких ошибок. Интересует не абстрактная эффективность метода, а его работоспособность в полном конвейере обработки: от изображения строки через распознавание текста и вплоть до итогового эмоционального класса.

PDA формировалась на основе реальных ошибок распознавания, а не за счет произвольной замены символов – это принципиальный момент, отличающий подход от простого шумового искажения данных. Выход PaddleOCR сопоставлялся с ручной транскрипцией, после чего отдельно фиксировались замены, удаления и вставки символов. Полученные распределения ошибок затем применялись для зашумления обучающих строк.

В рамках того же эксперимента проводилось сравнение текстовых, визуальных и мультимодальных моделей. Ограничения, связанные с редкими классами и псевдоразметкой, обсуждаются отдельно, поскольку они относятся к устройству данных, а не к специфике OCR-шума как такового.

Экспериментальная часть исследования опирается на четыре текстовые модели, и каждая из них закрывает отдельный участок сравнения. TextCNN [1] выполняет роль простого символьного ориентира – с его помощью проверяется, хватает ли локальных сочетаний символов для разграничения эмоциональных классов. BiLSTM [2] проверяет схожую идею иначе, учитывая порядок символов в строке, но без предобученного трансформерного контекста, что позволяет отделить эффект последовательной обработки от эффекта глубокого предобучения. BERT [3] и RoBERTa, реализованная средствами библиотеки Transformers [4], представляют сильные контекстные энкодеры, задающие верхнюю планку качества в этом сопоставлении.

Для китайского языка вопрос сегментации стоит особенно остро. Отдельный иероглиф далеко не всегда совпадает с границей смысловой единицы, поэтому в работе учтены подходы к целнословному маскированию [5] – без их учета качество представления текста заметно снижается. Корпус GoEmotions [6] использован как ориентир для многоклассовой разметки эмоций. SemEval-2018 Task 1 [7] взят как дополнительный методологический фон для задач классификации эмоциональных состояний. Работы по китайской эмоциональной классификации и аффективным лексиконам [8–10] учитывались при формировании категориального и лексического основания исследования.

Ключевое отличие настоящей работы от типичных задач *emotion classification* состоит в отсутствии ручной эмоциональной разметки исходных данных. CASIA-HWDB содержит изображение строки и ее транскрипцию, но не содержит указания на эмоцию текста. По этой причине метки формировались слабым способом, через лексические признаки и предвари-

тельный текстовый классификатор. Такое решение неизбежно ставит вопрос о качестве полученных меток, и здесь стоит обратиться к литературе по псевдометкам [11], Co-teaching [12] и Noisy Student [13]. Эти работы предупреждают: автоматическая разметка расширяет обучающую выборку, однако вместе с ней в модель проникают и закрепляются ошибки. Соответственно, результаты, полученные на псевдометках, не следует приравнивать к результатам экспертной оценки – это разные по надежности источники истины.

Отдельного внимания заслуживает OCR-шум. Источником изображений выступает корпус CASIA-HWDB [14], а основным распознавателем выбран PaddleOCR [15] как современный и практически применимый инструмент для китайского текста; Tesseract [16] сохранен в качестве классического открытого ориентира для сравнения. Величина CER сама по себе мало что говорит о реальном ущербе. Ошибка распознавания становится проблемой лишь тогда, когда она меняет вход классификатора и тем самым сдвигает предсказанный эмоциональный класс. Перенос шума в последующие задачи уже обсуждался в работах [17, 18], и полученные в настоящем исследовании наблюдения в целом согласуются с этими выводами. Наконец, для проверки устойчивости моделей применялись две стратегии аугментации: EDA [19] как простая случайная аугментация и PDA, построенная на основе реально наблюдаемых ошибок PaddleOCR – заменах, удалениях и вставках символов.

Материалы и методы исследования

Формирование слабоконтролируемых меток

Каждая рабочая запись в исследовательской таблице содержала три обязательные позиции: идентификатор изображения, само изображение рукописной строки и ее ручную транскрипцию. Эмоциональная метка в этот набор изначально не входила – она формировалась отдельным этапом, предшествующим обучению итоговых моделей, что принципиально отличает предложенный подход от прямого заимствования разметки из корпуса.

Процедура присвоения метки была двухступенчатой. Сначала транскрипция сверялась с подготовленным списком лексических признаков, затем в дело вступал текстовый классификатор, предлагавший возможные категории. Полученная псевдоразметка фиксировалась и далее в процессе обучения не пересматривалась. Итоговые

модели работали уже с готовыми метками, не порождая целевой класс самостоятельно во время собственной оценки – это важное методологическое решение, исключающее циркулярность в проверке результатов.

Выбор псевдометки y_i производился по принципу максимальной апостериорной вероятности среди классов C :

$$y_i = \arg \max_{c \in C} p(c|x_i).$$

Наряду с выбранной категорией сохранялось и наибольшее значение вероятности (confidence) – оно применялось как диагностический признак при анализе распределения уверенности и при проверке варианта со взвешиванием функции потерь. Здесь необходимо оговориться: confidence нельзя трактовать как прямое доказательство совпадения псевдометки с решением человека, это разные вещи по своей природе.

Порог $\tau = 0,7$ использовался исключительно на предварительном этапе анализа распределения уверенности; никакого отсева объектов из основной обучающей выборки по этому порогу не производилось. Все строки, отнесенные к целевым категориям, были сохранены, поскольку словарное совпадение и вероятность классификатора имеют разную природу и без калибровки не могут считаться единой мерой правильности эмоциональной метки.

Эмоциональный лексикон собирался автором применительно к категориям, задействованным в эксперименте, и включал китайские слова, устойчивые выражения и кандидатные фрагменты. Для семиклассовой постановки задачи применялись файлы модельной разметки по восьми исходным группам, а при расширении до одиннадцати классов добавилось 1951 правило переопределения для категорий neutral, calm, resignation и sarcasm. При пересечении строки с несколькими категориями итоговый класс определялся либо правилом переопределения, либо по наибольшей модельной уверенности. Сам лексикон остается рабочим инструментом построения псевдоразметки, а не внешним психологическим золотым стандартом.

Данные и распределение классов

Анализируемая семиклассовая выборка насчитывает 52 001 наблюдение. Баланс классов оставляет желать лучшего: радость и грусть доминируют по частоте встречаемости, тогда как страх, отвращение и тревога представлены существенно меньшим числом строк. Это обстоятельство напрямую влияет на интерпретацию метрики Macro-F1, которая особенно чувствительна к редким категориям.

Таблица 1

Число объектов по классам в основной семиклассовой выборке

Категория	Обучение	Валидация	Тест	Всего
радость	12 808	3332	4149	20 289
грусть	17 850	4372	5472	27 694
гнев	777	169	239	1185
страх	120	24	34	178
удивление	1360	296	382	2038
отвращение	261	67	92	420
тревога	126	31	40	197
Итого	33 302	8291	10 408	52 001

Примечание: подсчет автора по данным семиклассового эксперимента.

Таблица 2

Число объектов по категориям в одиннадцатиклассовой постановке

Категория	Обучение	Валидация	Тест	Всего
Нейтральность	1112	234	336	1682
Радость	12 412	3241	4026	19 679
Грусть	17 068	4213	5240	26 521
Гнев	763	169	237	1169
Страх	120	24	34	178
Удивление	1338	290	379	2007
Отвращение	261	67	92	420
Спокойствие	113	24	30	167
Тревога	126	31	40	197
Смирение	71	25	23	119
Сарказм	17	0	4	21
Итого	33 401	8318	10 441	52 160

Примечание: обработка автором результатов одиннадцатиклассового эксперимента.

Показательно, что для таких классов буквально несколько ошибок OCR или неудачных примеров при разбиении данных способны заметно исказить итоговую оценку – и это следует учитывать при любой содержательной интерпретации результатов. Распределение объектов по семи классам представлено в табл. 1.

Переход к одиннадцати категориям увеличил выборку до 52 160 объектов, однако проблему дисбаланса это не решило. Напротив, длинный хвост распределения стал более явным. Наиболее показательный пример здесь – sarcasm: для этой категории найден всего 21 объект. При этом в validation-выборке строк с sarcasm нет вообще, а в тестовой их обнаружено лишь 4.

Ввиду перечисленного одиннадцатиклассовая схема должна рассматриваться исключительно как дополнительный ин-

струмент анализа редких эмоций, но отнюдь не как надежный источник оценки качества по каждой отдельной новой категории. Распределение по одиннадцати категориям приведено в табл. 2.

Конфигурации моделей и профильно-ориентированная аугментация

В текстовой ветви эксперимента прошли проверку модели BERT, RoBERTa, TextCNN и BiLSTM – выбор, продиктованный необходимостью сопоставить архитектуры разной сложности и природы. Для обработки изображений использовалась Swin-Tiny (microsoft/swin-tiny-patch4-window7-224): вход приводился к размеру 224×224, представления патчей усреднялись и передавались классификатору. В мультимодальных конфигурациях текстовые и визуальные

признаки проецировались в единое пространство размерности 512, что позволило реализовать три схемы слияния данных – позднее слияние вероятностей, перекрестное внимание и вентильное слияние. Размер моделей закономерно варьировался: 102 млн параметров для варианта text-only, 28 млн для image-only и 131–134 млн для мультимодальных решений.

В сквозном варианте OCR сначала формировал строку текста, которая затем передавалась в классификатор эмоций. Профиль OCR строился на 500 строках из валидационной части данных: выход PaddleOCR сопоставлялся с ручной транскрипцией на уровне символов, после чего проводилось выравнивание и подсчет замен, удалений и вставок. Распознаватель не дообучался, использовались PaddleOCR 3.6.0, PaddlePaddle 3.3.1, режим text_recognition и настройка ch. CER на данном наборе составила 0,1181, а доли замен, удалений и вставок – 0,0463; 0,0157 и 0,0560 соответственно. Искажения вносились исключительно в обучающие тексты. Поскольку профиль OCR получен на валидационной выборке, это следует признать ограничением исследования, требующим отдельной калибровочной выборки в дальнейшем:

$$\Theta_{\text{OCR}} = \left\{ P_{\text{sub}}(b | a), P_{\text{del}}(a), P_{\text{ins}}(b | a) \right\}.$$

Здесь $P_{\text{sub}}(b | a)$ – вероятность распознать b вместо a , $P_{\text{del}}(a)$ – вероятность удаления a , а $P_{\text{ins}}(b | a)$ – вероятность вставки b в контексте a .

При подготовке обучающего примера оператор A искажал исходную строку x_i в соответствии с этим профилем:

$$\tilde{x}_i = A(x_i; \Theta_{\text{OCR}}).$$

План эксперимента

Классификацию оценивали по Accurasy, Macro-F1 и Weighted-F1. Для OCR вычислялась CER как нормированное расстояние Левенштейна; дополнительно считалась доля строк, полностью совпавших с ручной транскрипцией. Знак « $\pm$ » в таблицах относится к стандартному отклонению по трем запускам с разными случайными зёрнами, если в подписи не указано другое.

Для распознанной строки CER вычислялась следующим образом:

$$\text{CER} = \frac{S + D + I}{N}.$$

В формуле S , D и I обозначают число замен, удалений и вставок; N – число символов в ручной транскрипции.

Для каждого класса с отдельно рассчитывались Precision, Recall и F1:

$$P_c = \frac{TP_c}{TP_c + FP_c},$$

$$R_c = \frac{TP_c}{TP_c + FN_c},$$

$$F1_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}.$$

Затем показатели по классам объединялись с помощью макроусреднения и взвешивания по числу объектов:

$$M_{F1} = \frac{1}{|C|} \sum_{c \in C} F1_c, \quad W_{F1} = \sum_{c \in C} \frac{n_c}{N} F1_c.$$

В расчетах приняты обозначения: M_{F1} обозначает Macro-F1, W_{F1} – Weighted-F1, n_c – число тестовых объектов класса c , N – размер всей тестовой выборки. Сопоставление PDA с базовой моделью проводилось на идентичном наборе тестовых объектов, что обеспечивало корректность сравнения. Парный bootstrap включал 10 000 повторов, по итогам которых строился 95 %-й доверительный интервал разности Macro-F1; для b -й bootstrap-выборки фиксировалась соответствующая разность между моделью с PDA и базовой моделью. В одиннадцатиклассовой задаче перечень категорий оставался фиксированным при вычислении метрик. Когда класс отсутствовал в валидационной или тестовой выборке, его Precision, Recall и F1 приравнивались нулю и учитывались при расчете Macro-F1 – отдельных выводов по классам с нулевой или крайне малой поддержкой при этом не формулировалось:

$$\Delta_b = M_A(D_b^*) - M_0(D_b^*),$$

M_A и M_0 обозначают Macro-F1 модели с PDA и исходной модели. Границы доверительного интервала брались по 2,5-му и 97,5-му процентилем распределения Δ :

$$CI_{95\%} = [q_{0.025}(\Delta), q_{0.975}(\Delta)].$$

Анализ распределения данных по выборкам показал следующее соотношение долей: обучающая выборка составила около 64,0 %, валидационная – 15,9 %, тестовая – 20,0 %. При этом проверка на пересечение image_id между частями не выявила ни одного совпадения, что говорит о корректности разбиения на уровне изображений.

Однако здесь обнаруживается существенная проблема. Отдельные транскрип-

ции встречаются одновременно и в обучающей, и в тестовой выборках. Это обстоятельство способно искусственно завышать качество text-only моделей относительно результатов, полученных после проведения дедупликации. Иными словами, модель может демонстрировать хорошую метрику не за счет обобщающей способности, а благодаря простому запоминанию текстов.

Еще одно ограничение связано с отсутствием в метаданных поля writer_id. Из-за этого в рамках данной версии данных невозможно оценить, насколько модель способна обобщаться на почерк ранее не встречавшихся авторов – а ведь именно это качество представляет наибольшую практическую ценность.

Что касается параметров обучения, для BERT и RoBERTa использовалась длина входной последовательности в 64 токена, размер батча 8 и не более 20 эпох. Оптимизатор AdamW применялся со скоростью обучения $1e-5$ для предобученного энкодера и $2e-5$ для классификационной головы и слоев слияния; weight decay составил 0,01, warmup – 10 % от общего числа шагов. Для TextCNN и BiLSTM символьный вход имел длину 64, batch size – 128, обучение продолжалось до 15 эпох. Выбор итоговой контрольной точки во всех случаях осуществлялся по значению Macro-F1 на валидационной выборке.

Результаты исследования и их обсуждение

Сравнение моделей в семиклассовой задаче

При семиклассовой классификации лидирующую позицию заняла RoBERTa: Macro-F1 составил $0,8863 \pm 0,0044$. BERT показала сопоставимый, хотя и чуть более скромный результат – $0,8854 \pm 0,0083$. Модели BiLSTM и TextCNN уступили обеим трансформерным архитектурам, продемонстрировав Macro-F1 0,8523 и 0,8366 соответственно.

Вентильное мультимодальное слияние дало значение 0,8470 – результат промежуточный, не превзошедший текстовые модели. Показательная ситуация с моделью, опирающейся исключительно на изображение: Macro-F1 упал до 0,1532. В рамках использованной архитектуры и примененных псевдометок визуальная ветвь так и не смогла улучшить показатели text-only решений.

Однако делать вывод о бесполезности признаков почерка преждевременно: при иной архитектуре или наличии ручной эмоциональной разметки ситуация может измениться существенным образом.

Сводное сравнение моделей приведено в табл. 3.

Влияние сквозного распознавания

На тестовых изображениях CER PadleOCR составила 0,1563, а доля полностью совпавших строк – 0,1177. С ручными транскрипциями классификатор получил Accuracy 0,9286 и Macro-F1 0,8889. При подаче OCR-выхода значения снизились до 0,8961 и 0,8390. Потеря 0,0499 по Macro-F1 показывает, что распознавательные ошибки затрагивают не только общую точность, но и качество по малочисленным категориям. Результаты сквозной классификации после OCR представлены в табл. 4.

Эффект профильно-ориентированной аугментации

На чистых транскрипциях исходная RoBERTa получила Macro-F1 $0,8863 \pm 0,0044$. Для случайной аугментации значение было $0,8858 \pm 0,0022$, для взвешивания по confidence – $0,8887 \pm 0,0059$, для PDA – $0,8919 \pm 0,0026$. Различия невелики. Поэтому этот эксперимент рассматривается как дополнительное наблюдение и не подтверждает универсальное улучшение классификации чистого текста. Сравнение стратегий робастного обучения приведено в табл. 5.

Таблица 3

Сравнение моделей в основной семиклассовой задаче

Модель	Точность	Макро-F1	Взвешенная F1
RoBERTa	$0,9304 \pm 0,0014$	$0,8863 \pm 0,0044$	$0,9304 \pm 0,0014$
BERT	$0,9282 \pm 0,0019$	$0,8854 \pm 0,0083$	$0,9282 \pm 0,0020$
BiLSTM	$0,9045 \pm 0,0010$	$0,8523 \pm 0,0044$	$0,9043 \pm 0,0009$
TextCNN	$0,8878 \pm 0,0060$	$0,8366 \pm 0,0072$	$0,8875 \pm 0,0062$
Вентильное слияние	$0,9049 \pm 0,0018$	$0,8470 \pm 0,0056$	$0,9049 \pm 0,0018$
Только изображение	0,4840	0,1532	0,4701

Примечание: результаты трех запусков; «±» обозначает стандартное отклонение

Таблица 4

Результаты сквозной классификации после PaddleOCR
в семиклассовой задаче

Метрика	Значение
Посимвольная ошибка распознавания	0,1563
Точное совпадение распознавания	0,1177
Точность при эталонном тексте	0,9286
Макро-F1 при эталонном тексте	0,8889
Сквозная точность	0,8961
Падение точности	0,0325
Сквозная макро-F1	0,8390
Падение макро-F1	0,0499

Примечание: составлена автором на основе полученных данных в ходе исследования

Таблица 5

Сравнение стратегий робастного обучения на чистом тексте

Настройка	Точность	Макро-F1	Взвешенная F1
Исходная RoBERTa	$0,9304 \pm 0,0014$	$0,8863 \pm 0,0044$	$0,9304 \pm 0,0014$
Случайная аугментация	$0,9290 \pm 0,0008$	$0,8858 \pm 0,0022$	$0,9290 \pm 0,0008$
Взвешивание по уверенности	$0,9311 \pm 0,0018$	$0,8887 \pm 0,0059$	$0,9311 \pm 0,0018$
Профильно-ориентированная аугментация	$0,9313 \pm 0,0012$	$0,8919 \pm 0,0026$	$0,9313 \pm 0,0012$

Примечание: составлена автором на основе полученных данных в ходе исследования

Таблица 6

Влияние профильно-ориентированной аугментации в сквозном конвейере

Метрика	Базовая модель	Аугментация	Δ (95 % ДИ)
Точность при эталонном тексте	0,9286	0,9305	+0,0019 [−0,0020; 0,0054]
Макро-F1 при эталонном тексте	0,8889	0,8928	+0,0039 [−0,0084; 0,0153]
Сквозная точность	0,8961	0,9102	+0,0140 [0,0093; 0,0192]
Сквозная макро-F1	0,8390	0,8636	+0,0246 [0,0102; 0,0407]

Примечание: составлена автором на основе полученных данных в ходе исследования

Парный анализ дал разные результаты для двух входов. На эталонном тексте прирост Macro-F1 равен 0,0039; интервал [−0,0084; 0,0153] пересекает ноль. После OCR прирост Assurasy составил 0,0140, а Macro-F1 – 0,0246. Для Macro-F1 получен интервал [0,0102; 0,0407], полностью расположенный выше нуля. Статистически подтверждается именно эффект в условиях распознавательного шума. Сводные значения показаны в табл. 6, а изменение F1-мер по классам – на рис. 1.

Одиннадцатиклассовое расширение и длинный хвост

После расширения системы меток до одиннадцати категорий RoBERTa получила Assurasy $0,9297 \pm 0,0007$ и Macro-F1 $0,8760 \pm 0,0150$. По сравнению с семиклассовым экспериментом среднее значение Macro-F1 ниже, а разброс между запусками заметно выше. Наибольшую неопределенность создают категории с несколькими десятками примеров.

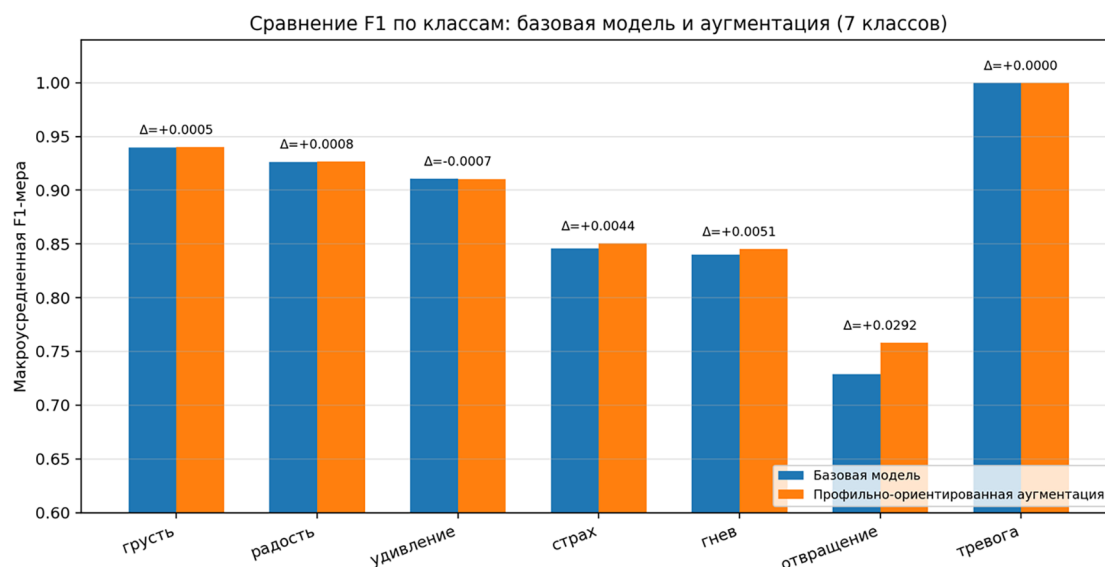


Рис. 1. Сравнение F1-меры по классам между базовой моделью и профильно-ориентированной аугментацией в семиклассовой задаче
Примечание: составлен автором по результатам исследования

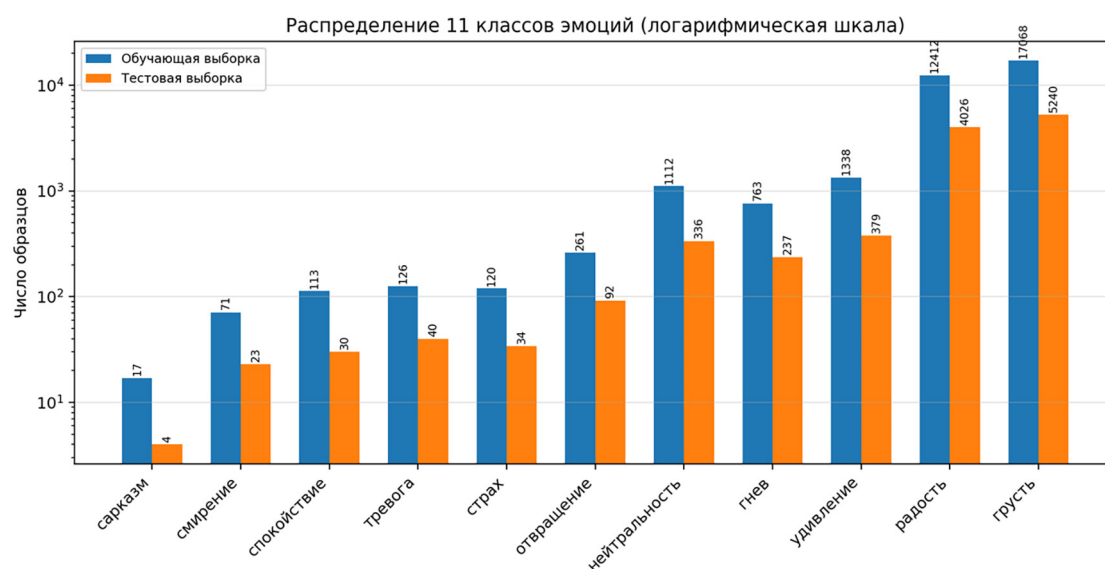


Рис. 2. Распределение одиннадцати категорий эмоций в логарифмической шкале
Примечание: составлен автором по статистике выборки

Семь и одиннадцать классов используют разные системы меток и разные правила псевдоразметки, поэтому сравнение показывает только общую тенденцию. Особенно мало данных для сарказма, спокойствия и смирения. У сарказма нет валидационных примеров и имеется лишь четыре тестовые строки. Результаты по этим категориям следует считать предварительными. Распределение одиннадцати категорий показано на рис. 2.

Человеческая проверка надежности псевдометок

Для выборочной ручной проверки из тестовой части задачи, включающей семь классов, было отобрано 480 объектов. Аннотатор, не имея доступа к псевдометкам, определял эмоцию исключительно по ручной транскрипции и изображению строки. Полученное совпадение аннотатора с автоматической разметкой оказалось невысоким и составило 0,3854, при этом 95 %-й

интервал Уилсона для данного показателя расположился в границах $[0,3430; 0,4297]$. Коэффициент к Коэна принял значение $0,2763$ – величину, которую принято трактовать как слабое согласие.

Важная оговорка: поскольку проверку осуществлял единственный аннотатор, коэффициент характеризует лишь степень совпадения человеческой оценки с автоматической меткой, но никак не межэкспертное согласие как таковое. Это существенное ограничение методики.

На тех же 480 объектах модель продемонстрировала Ассигасу $0,8771$ относительно псевдометок – показатель, на первый взгляд, впечатляющий. Однако Ассигасу относительно ручных меток составила лишь $0,3833$. Разрыв между этими цифрами по-

казателен: высокое совпадение с псевдоразметкой отнюдь не свидетельствует о таком же соответствии реальному человеческому восприятию эмоций. Модель, судя по всему, успешно воспроизводит внутреннюю логику принятой системы псевдоразметки, но точность с позиции человека при этом не гарантируется.

Отсюда следует практический вывод. Дальнейшая проверка требует привлечения нескольких независимых аннотаторов, разработки единой инструкции для их работы и обязательного расчета межаннотаторского согласия – только так можно получить достоверную картину качества разметки. Итоги ручной проверки приведены в табл. 7, а соотношение оценок показано на рис. 3.

Таблица 7

Человеческая проверка и оценка модели в семиклассовой задаче

Метрика	Значение
Число проверенных образцов	480
Согласие псевдометок с человеком	$0,3854$
95 %-й доверительный интервал Уилсона	$[0,3430; 0,4297]$
к Коэна	$0,2763$
Точность модели относительно человека	$0,3833$
Точность модели относительно псевдометок	$0,8771$

Примечание: составлена автором на основе полученных данных в ходе исследования

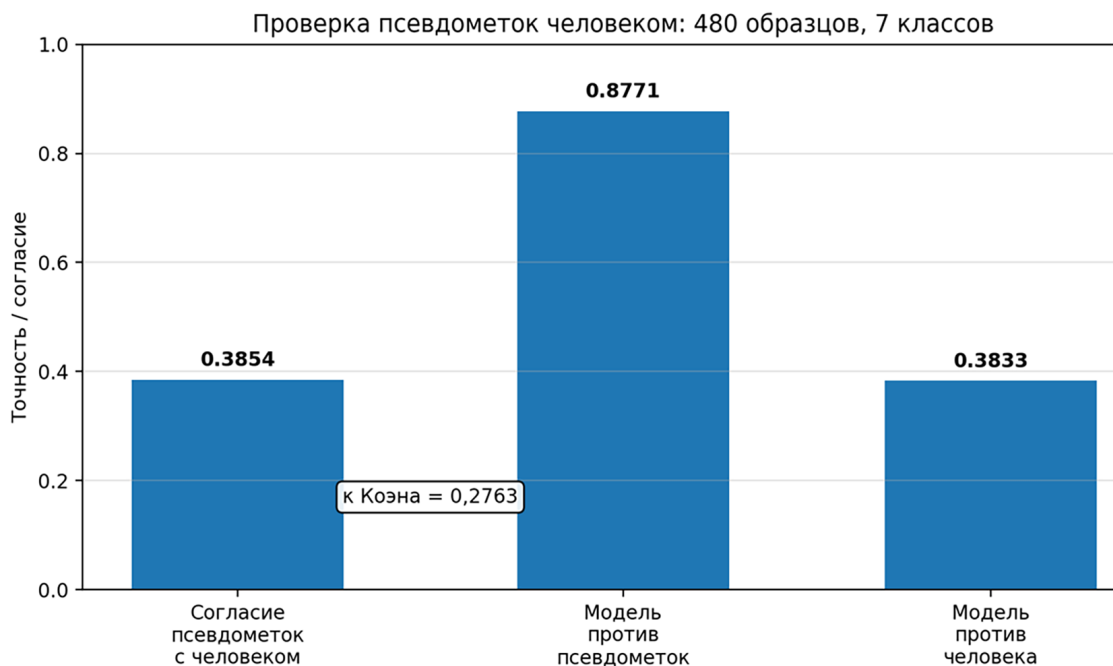


Рис. 3. Результаты ручной проверки надежности псевдометок
Примечание: построен автором по данным проверки 480 объектов

Анализ результатов позволяет утверждать: основной выигрыш от применения PDA проявляется именно при работе с OCR-входом. Причина проста – ручная транскрипция уже содержит полный, корректный текст, и добавление дополнительного символа шума практически не дает эффекта. Иная ситуация складывается с распознанным текстом. В OCR-строках регулярно встречаются замены, удаления и вставки символов, характерные для конкретного движка распознавания. PDA воспроизводит именно эти типы ошибок на этапе обучения, благодаря чему обучающий шум приближается к тестовому, а не имитирует произвольные искажения.

Случайная аугментация подобного результата не дает. Она не учитывает специфику того, какие символы чаще всего искажает OCR, и способна затронуть как второстепенный фрагмент текста, так и эмоционально ключевое слово – с равной вероятностью. Сравнение с EDA подтверждает важный тезис: недостаточно просто добавлять шум в текст. Куда существеннее приближать характер этого шума к реальным ошибкам конкретного OCR-движка на конкретном корпусе данных.

Отдельного внимания заслуживает одиннадцатиклассовый эксперимент. Он выявил проблему длинного хвоста распределения классов: редкие категории оказывают непропорционально сильное влияние на Macro-F1 и заметно увеличивают разброс результатов между отдельными запусками. При нулевой или минимальной поддержке того или иного класса содержательный анализ его качества классификации становится попросту невозможным.

Основная доказательная часть работы опирается на семиклассовый сквозной сценарий. Результаты, полученные для категорий sarcasm, calm и resignation, следует рассматривать как предварительные, не более того. Ручная проверка продемонстрировала: псевдометки пригодны для воспроизводимого сравнения моделей внутри единой процедуры, однако не могут заменить экспертный золотой стандарт разметки. Высокая Ассигасу относительно автоматических меток отражает лишь согласованность модели с конкретной системой разметки, а не качество классификации в подлинно человеческом понимании. Для полноценной оценки эмоциональной классификации требуется расширенная независимая разметка с расчетом межаннотаторского согласия – без этого шага любые выводы остаются неполными.

В семиклассовом эксперименте лучший текстовый результат продемонстрирова-

ла модель RoBERTa. При замене ручных транскрипций выходом PaddleOCR значение Macro-F1 снизилось с 0,8889 до 0,8390. Эта цифра количественно подтверждает существенное влияние ошибок распознавания на качество классификации эмоций. Применение PDA в OCR-сценарии повысило Macro-F1 с 0,8390 до 0,8636. Разность с базовой моделью составила 0,0246, а 95 %-й доверительный интервал [0,0102; 0,0407] не содержит нуля – следовательно, прирост статистически значим. Для чистых транскрипций подобного, статистически подтвержденного прироста от PDA получить не удалось.

Отсюда следует практический вывод: смысл предложенного метода состоит в адаптации к реальным ошибкам OCR, а не в универсальном улучшении любого текстового входа независимо от его происхождения.

В одиннадцатиклассовой задаче оценки редких категорий оказались нестабильными – что закономерно, учитывая проблему длинного хвоста. Ручная проверка 480 объектов зафиксировала лишь ограниченное совпадение псевдометок с решением отдельного аннотатора. Высокие метрики по автоматическим меткам нельзя без оговорок интерпретировать как аналогичную точность относительно независимой человеческой разметки – это принципиальное ограничение всего подхода.

Работа имеет и другие ограничения: обнаружены повторяющиеся тексты между обучающей и тестовой выборками, отсутствует writer_id, а OCR-профиль построен по части валидационных данных. Эти факторы способны исказить итоговые оценки.

В дальнейшей работе необходимы дедупликация данных, writer-disjoint разбиение выборок и формирование отдельного calibration set. Ручную разметку целесообразно расширить за счет привлечения нескольких независимых аннотаторов, а метод PDA – проверить на других OCR-движках, чтобы установить границы его применимости. До проведения перечисленных дополнительных проверок все выводы настоящей работы следует относить исключительно к описанной слабоконтролируемой схеме эксперимента.

Список литературы

1. Kim Y. Convolutional neural networks for sentence classification // Proceedings of EMNLP. 2014. P. 1746–1751. URL: <https://aclanthology.org/D14-1181/> (дата обращения: 01.07.2026). DOI: 10.3115/v1/D14-1181.
2. Hochreiter S., Schmidhuber J. Long short-term memory // Neural Computation. 1997. Vol. 9. No. 8. P. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
3. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language

- understanding // Proceedings of NAACL-HLT. 2019. P. 4171–4186. URL: <https://aclanthology.org/N19-1423/> (дата обращения: 01.07.2026). DOI: 10.18653/v1/N19-1423.
4. Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., von Platen P., Ma C., Jernite Y., Plu J., Xu C., Le Scao T., Gugger S., Drame M., Lhoest Q., Rush A. M. Transformers: State-of-the-Art Natural Language Processing // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020. P. 38–45. URL: <https://aclanthology.org/2020.emnlp-demos.6/> (дата обращения: 01.07.2026). DOI: 10.18653/v1/2020.emnlp-demos.6.
 5. Cui Y., Che W., Liu T., Qin B., Yang Z. Pre-training with whole word masking for Chinese BERT // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. Vol. 29. P. 3504–3514. DOI: 10.1109/TASLP.2021.3124365.
 6. Demszky D., Movshovitz-Attias D., Ko J., Cowen A., Nemade G., Ravi S. GoEmotions: A dataset of fine-grained emotions // Proceedings of ACL. 2020. P. 4040–4054. URL: <https://aclanthology.org/2020.acl-main.372/> (дата обращения: 01.07.2026). DOI: 10.18653/v1/2020.acl-main.372.
 7. Mohammad S., Bravo-Marquez F., Salameh M., Kiritchenko S. SemEval-2018 task 1: Affect in tweets // Proceedings of SemEval. 2018. P. 1–17. URL: <https://aclanthology.org/S18-1001/> (дата обращения: 01.07.2026). DOI: 10.18653/v1/S18-1001.
 8. Wen S., Wan X. Emotion Classification in Microblog Texts Using Class Sequential Rules // Proceedings of the AAAI Conference on Artificial Intelligence. 2014. Vol. 28. Is. 1. DOI: 10.1609/aaai.v28i1.8709.
 9. Odbal, Wang Z. Segment-based fine-grained emotion detection for Chinese text // Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing. 2014. P. 52–60. URL: <https://aclanthology.org/W14-6809/> (дата обращения: 23.07.2026). DOI: 10.3115/v1/W14-6809.
 10. Xu L., Lin H., Pan Y., Ren H., Chen J. Constructing the affective lexicon ontology // Journal of the China Society for Scientific and Technical Information. 2008. Is. 2. P. 180–185.
 11. Lee D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks // ICML Workshop on Challenges in Representation Learning. 2013. URL: https://www.researchgate.net/publication/280581078_Pseudo-Label_The_Simple_and_Efficient_Semi-Supervised_Learning_Method_for_Deep_Neural_Networks (дата обращения: 01.07.2026).
 12. Han B., Yao Q., Yu X., Niu G., Xu M., Hu W., Tsang I., Sugiyama M. Co-teaching: Robust training of deep neural networks with extremely noisy labels // Proceedings of NeurIPS. 2018. P. 8527–8537. URL: <https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html> (дата обращения: 01.07.2026).
 13. Xie Q., Luong M.-T., Hovy E., Le Q. V. Self-training with noisy student improves ImageNet classification // Proceedings of CVPR. 2020. P. 10687–10698. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Xie_Self-Training_With_Noisy_Student_Improves_ImageNet_Classification_CVPR_2020_paper.html (дата обращения: 01.07.2026).
 14. Liu C.-L., Yin F., Wang D.-H., Wang Q.-F. CASIA online and offline Chinese handwriting databases // Proceedings of ICDAR. 2011. P. 37–41. DOI: 10.1109/ICDAR.2011.17.
 15. PaddleOCR contributors. PaddleOCR: Global Leading OCR Toolkit and Document AI Engine. GitHub repository. URL: <https://github.com/PaddlePaddle/PaddleOCR> (дата обращения: 23.07.2026).
 16. Smith R. An overview of the Tesseract OCR engine // Proceedings of ICDAR. 2007. P. 629–633. DOI: 10.1109/ICDAR.2007.4376991.
 17. Zhang J., Zhang Q., Wang B., Ouyang L., Wen Z., Li Y., Chow K.-H., He C., Zhang W. OCR hinders RAG: Evaluating the cascading impact of OCR on retrieval-augmented generation // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2025. P. 17443–17453. URL: https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_OCR_Hinders_RAG_Evaluating_the_Cascading_Impact_of_OCR_on_ICCV_2025_paper.html (дата обращения: 01.07.2026).
 18. Vitman O., Kostiuk Y., Plachinda P., Zhila A., Sidorov G., Gelbukh A. Evaluating the impact of OCR quality on short texts classification task // Advances in Computational Intelligence: MICAI 2022, Lecture Notes in Computer Science. Cham: Springer, 2022. Vol. 13613. P. 163–177. DOI: 10.1007/978-3-031-19496-2_13.
 19. Wei J., Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks // Proceedings of EMNLP-IJCNLP. 2019. P. 6382–6388. URL: <https://aclanthology.org/D19-1670/> (дата обращения: 01.07.2026). DOI: 10.18653/v1/D19-1670.
 20. Tsai Y.-H. H., Bai S., Liang P. P., Kolter J. Z., Morency L.-P., Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences // Proceedings of ACL. 2019. P. 6558–6569. URL: <https://aclanthology.org/P19-1656/> (дата обращения: 01.07.2026). DOI: 10.18653/v1/P19-1656.
 21. Rahman W., Hasan M. K., Lee S., Zadeh A. B., Mao C., Morency L.-P., Hoque E. Integrating multimodal information in large pretrained transformers // Proceedings of ACL. 2020. P. 2359–2369. URL: <https://aclanthology.org/2020.acl-main.214/> (дата обращения: 01.07.2026). DOI: 10.18653/v1/2020.acl-main.214.
 22. Baltrušaitis T., Ahuja C., Morency L.-P. Multimodal machine learning: A survey and taxonomy // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019. Vol. 41. Is. 2. P. 423–443. DOI: 10.1109/TPAMI.2018.2798607.
 23. Wang H., Mao Y., Miao Q., Xiao Q., Zhang Y. Dual-branch multitask fusion network for offline Chinese writer identification // ACM Transactions on Asian and Low-Resource Language Information Processing. 2024. Vol. 23. Is. 2. Art. 24. P. 24:1–24:22. DOI: 10.1145/3638554.

Конфликт интересов: Автор заявляет об отсутствии конфликта интересов.

Conflict of interest: The author declares that there is no conflict of interest.

Финансирование: Авторы заявляют об отсутствии внешнего финансирования.

Financing: The research was performed without external funding.