УДК 004

ПОВЫШЕНИЕ УСТОЙЧИВОСТИ КЛАССИФИКАЦИИ КОРОТКИХ ТЕКСТОВ К СТОХАСТИЧЕСКОМУ ШУМУ НА ОСНОВЕ ПЛОТНОСТНОЙ ОЧИСТКИ ОБУЧАЮЩИХ ВЫБОРОК

Б. Б. Баишев[1] [0009-0007-9287-4248], А. П. Халов[2] [0009-0005-4584-8245]

[1]*Назарбаев Университет, г. Астана, Казахстан*

[2]*Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Россия*

[1]baishevbasar@gmail.com, [2]khalov.a@phystech.edu

Аннотация

Рассмотрена задача классификации коротких текстовых заявок в условиях значительного дисбаланса классов и зашумленности реальных потоков обращений. Показана ограниченная эффективность методов синтетического расширения выборки при работе с зашумленной разметкой. Предложен гибридный метод, сочетающий предварительную плотностную очистку данных и многоуровневое ансамблирование моделей. Применение алгоритма плотностной кластеризации позволило исключить 16.5% информационного шума от общего объема выборки. Финальная модель представлена двухуровневой архитектурой и оптимизирована с помощью байесовского поиска гиперпараметров. На отложенной тестовой выборке достигнуто значение метрики R@3, равное 97.4%. Предложенный метод позволяет автоматизировать процесс распределения заявок, существенно снижая нагрузку на операторов и сокращая время диспетчеризации обращений.

Ключевые слова: обработка естественного языка, зашумленные текстовые данные, ансамблевое обучение, робастная классификация, фильтрация шума.

ВВЕДЕНИЕ

Классификация коротких текстов является одной из фундаментальных задач современной обработки естественного языка, находящей широкое применение в автоматизации систем управления ИТ-услугами. В качестве ключевого

инструмента такой автоматизации используются методы машинного обучения, обеспечивающие интеллектуальную диспетчеризацию обращений и минимизирующие долю рутинных операций при первичной обработке данных. Несмотря на значительные успехи использования моделей глубокого обучения, задача обработки текстов из реальных пользовательских сценариев остается крайне актуальной. Для таких данных характерна высокая степень зашумленности: наличие опечаток, использование неформальной или узкоспециализированной технической лексики, а также нестандартные синтаксические структуры предложений. Эти факторы в совокупности приводят к существенной деградации метрик классификации стандартных моделей.

Особую сложность представляет сценарий, при котором обучающая выборка характеризуется распределением с «длинным хвостом» [1], что создает выраженный дисбаланс классов в сочетании с высоким уровнем шума. Эта проблема отчетливо проявляется в системах управления ИТ-услугами, где входящие обращения содержат зашумленный текст (например, *«fw:re: нужен обмен м/у базами urk/cnt/alm»*) и специфическую техническую номенклатуру (например, *«Cisco Catalyst»*, *«1C:Enterprise»*). Ввиду постоянного обновления подобных сущностей модель сталкивается с множеством внесловарных токенов, отсутствовавших на этапе обучения, что ведет к ошибкам семантического анализа. Непропорциональность классов также является критическим фактором: количество массовых типовых заявок (например, *«Сброс пароля»*) может на несколько порядков превышать число критических инцидентов в «хвосте» распределения (например, *«нет счф на номер оргтехники»*).

Известные методы борьбы с дисбалансом, такие как алгоритмы синтетического расширения выборки, например SMOTE [2], эффективно работают на искусственно созданных данных, однако демонстрируют ограниченную производительность в условиях зашумленной разметки. Во-первых, алгоритмы генерации новых примеров чувствительны к качеству исходных объектов, что ведет к тиражированию выбросов и формированию на их основе ложных искусственных кластеров. Во-вторых, искусственное выравнивание баланса искажает априорное распределение классов, приводя к общему снижению точности и полноты классификации.

В настоящей работе предложен альтернативный подход, основанный на концепции приоритета качества обучающих данных [3]. Смещая фокус с усложнения архитектур нейросетевых моделей на повышение репрезентативности исходной выборки, мы используем стратегию плотностной фильтрации. Этот метод позволяет эффективно исключать шумовые объекты, находящиеся в разреженных областях признакового пространства и не формирующие устойчивых семантических кластеров. Научная новизна исследования заключается в разработке комплексной методологии совместного применения плотностной очистки данных и ансамблевых методов, что обеспечивает экспериментально подтвержденный прирост точности классификации по сравнению с базовыми подходами.

ФОРМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ

Задача формулируется как многоклассовая классификация текстовых сообщений в условиях зашумленности разметки и признакового пространства.

Пусть $D = \{(x_i, y_i)\}_{i=1}^{N}$ — обучающая выборка, где объект $x_i = (t_i, m_i)$ включает неструктурированное текстовое описание t_i и вектор метаданных m_i .

Целевая переменная $y_i \in C = \{c_1, \ldots, c_k\}$ соответствует одной из K групп поддержки. Специфика предметной области (домена) накладывает на множество D три ключевых ограничения.

Сверхкраткость векторов: средняя длина $|t_i| \leq 7$ токенов, что приводит к высокой разреженности признакового пространства и дефициту контекстной информации.

Классовый дисбаланс: распределение классов $P(y)$ имеет «тяжелый хвост». Коэффициент дисбаланса, определяемый как $\rho = \frac{\max(N_c)}{\min(N_c)}$, достигает значений $\rho > 1000$. Значительная часть классов представлена малым количеством примеров ($n_c < 10$), недостаточным для обучения параметрических моделей без предварительной аугментации или регуляризации.

Стохастический шум: существует подмножество $D_{\text{noise}} \subset D$, для которого истинная метка y_i присвоена ошибочно вследствие человеческого фактора, либо текст t_i не содержит семантической информации, релевантной для задачи классификации.

Целью настоящей работы является построение отображения $f: X \to C$, которое минимизирует функцию потерь на тестовой выборке. Основной метрикой качества выбрана $\mathrm{R@k}$, так как в прикладном сценарии критически важно наличие истинного класса в списке из K наиболее вероятных рекомендаций системы.

ОБЗОР АНАЛОГИЧНЫХ ИССЛЕДОВАННИЙ

Современные подходы к классификации коротких текстов эволюционировали от базовых частотных методов, таких как TF-IDF [4], к использованию плотных векторных представлений в сочетании с ансамблевыми алгоритмами. Исследования подтверждают высокую эффективность комбинации нейросетевых кодировщиков и алгоритмов градиентного усиления [5], а также многоуровневой композиции классификаторов на базе решающих деревьев [6]. Однако в домене систем управления ИТ-услугами итоговая точность глубоких моделей критически зависит от качества предобработки и устойчивости к «шумным» классам [7].

Проблема дисбаланса, типичная для журналов регистрации событий [8], существенно ограничивает применимость стандартных подходов. В работе [9] показано, что генерация искусственных примеров на коротких зашумленных текстах часто искажает семантику и не превосходит тривиальное дублирование [9].

В качестве альтернативы активно исследуется алгоритмическая фильтрация данных, в частности выявление структурных аномалий с помощью плотностной кластеризации HDBSCAN [10, 11] поверх стандартных векторных представлений. Существенным ограничением таких решений является использование «замороженных» общелексических моделей, не способных формировать корректные векторные представления для специфического ИТ-сленга. В настоящей работе этот пробел устраняется за счет интеграции модели [12], адаптированной к предметной области (доменно-адаптированной), что обеспечивает семантически корректную фильтрацию шума и формирование качественного признакового пространства.

МЕТОДЫ

Для решения формализованной задачи классификации в условиях «длинного хвоста» и шума разработан многоступенчатый конвейер, основанный на концепции приоритета качества данных. Архитектура включает три последовательных этапа: предварительную подготовку и квотирование данных, плотностную фильтрацию признакового пространства и многоуровневое ансамблирование моделей.

Подготовка и балансировка данных

Для формирования обучающей выборки из исходного корпуса со значительным дисбалансом был применен алгоритм квотирования, основанный на степенном сглаживании частот. Целевой размер выборки Q_c для класса C рассчитывался следующим образом:

$$Q_c = \text{clip}\left(N_{\text{total}} \frac{(N_c)^{\alpha}}{\sum_j (N_j)^{\alpha}}, L_{\min}, Rm_{\text{ref}}\right),$$

где N_c – исходное количество примеров класса, clip – функция усечения, ограничивающая вычисляемое значение заданным диапазоном, α – коэффициент сглаживания, увеличивающий вес редких классов, $L_{\min}$ – нижний порог, гарантирующий корректность перекрестной проверки, Rm_{ref} – верхний порог, в котором R ограничивает отношение преобладающего класса к медианному значению m_{ref}. Для исключения смещения в сторону крупных клиентов квота заполнялась стратифицированно. Количество примеров $n_{c,s}$, отбираемых от конкретного источника s для класса c, определялось пропорционально его доле в исходных данных:

$$n_{c,s} \propto Q_c \ \frac{N_{c,s}}{N_c},$$

где $N_{c,s}$ – исходное количество заявок класса c от источника s. Это обеспечивает репрезентативность выборки и предотвращает переобучение модели на специфической лексике одного заказчика.

На этапе лексической очистки из текстов заявок удалялась нерелевантная информация: IP-адреса и URL-ссылки заменялись на специальные токены <ip>

и <url> соответственно. Было также произведено удаление неразрывных пробелов, невидимых символов Unicode и приведение текста к нижнему регистру. Для снижения размерности словаря и исключения утечки данных между обучающей и валидационной выборками реализован двухэтапный поиск дубликатов. Поиск **точных дубликатов** осуществлялся путем удаления записей с идентичным хеш-значением алгоритма SHA-1, вычисленным от нормализованного текста. Для выявления **нечетких дубликатов** (семантически близких заявок, отличающихся опечатками или автогенерируемыми метками времени) применялся алгоритм SimHash [15]. Вектор признаков для хеширования формировался на основе символьных n-грамм ($n \in \{3, 4, 5\}$), что обеспечивает устойчивость к незначительным изменениям в тексте. Пороговое значение расстояния Хэмминга для определения дубликата было установлено на уровне $d \leq 3$ бит.

Плотностная фильтрация

Процедура очистки данных реализуется через последовательность трех шагов.

Векторизация. Для преобразования текстов в векторное пространство использован доменно-адаптированный кодировщик на базе архитектуры XLM-RoBERTa Large [12]. Модель принимает на вход нормализованный текст t_i, а в качестве векторного представления заявки v_i используется скрытое состояние специального токена [CLS] последнего скрытого слоя:

$$v_i = h_{[\mathrm{CLS}]}\,.$$

Снижение размерности. Для повышения плотности кластеров перед подачей в алгоритм кластеризации размерность векторов была снижена с 1024 до 64 компонент методом главных компонент.

Плотностная кластеризация. Разделение на семантические группы производилось алгоритмом HDBSCAN [11]. Данный метод позволяет автоматически определять количество кластеров на основе плотности распределения и явно выделять объекты, находящиеся в разреженных областях.

По итогам работы алгоритма объекты, получившие метку шума, интерпретировались как стохастические аномалии (нерелевантные заявки, спам, редкие выбросы) и исключались из обучающей выборки.

Ансамблевая классификация

Финальный этап конвейера отвечает за построение пространства признаков и обучение двухуровневого ансамбля моделей.

Конструирование признаков. Процедура включает в себя семантическое обогащение коротких текстов путем конкатенации структурированных метаданных (тегов иерархии, меток длины и флагов детализации). Для повышения разделяющей способности сгенерированы доменные ключевые слова. С помощью критерия χ^2 выделены токены с максимальной предсказательной силой для каждого класса. Специфика источников данных учитывается через формирование профиля клиента с использованием кодирования средним значением целевой переменной [15]. Профиль включает вектор априорных вероятностей, коэффициент доминирования преобладающего класса и информационную энтропию распределения заявок $H = -\sum p_i \log p_i$. Дополнительно генерируется вектор базовых инженерных метрик: логарифмированные длины, коэффициент лексического разнообразия.

Базовые модели первого уровня. Для формирования семантического пространства решений использован гибридный подход. За фиксацию лексических паттернов отвечают две модели логистической регрессии, обученные на символьных и словных n-граммах с TF-IDF векторизацией (с применением сублинейного масштабирования). Для извлечения глубоких семантических признаков использован трансформер XLM-RoBERTa Large, предварительно адаптированный на корпусе технических текстов [12]. С учетом наличия субъективного шума в разметке инцидентов обучение нейросетевой модели производилось с применением сглаживания меток для предотвращения переобучения на ошибочных примерах:

$$y_{\text{target}} = (1 - \varepsilon) y_{\text{true}} + \frac{\varepsilon}{K},$$

где ε – коэффициент сглаживания, K – общее количество классов.

Архитектура мета-классификатора. Итоговое решение формируется алгоритмом градиентного усиления XGBoost [16] в парадигме многоуровневого ансамблирования [13]. Для предотвращения утечки данных вероятностные прогнозы базовых моделей генерируются строго методом перекрестного прогнози-

рования на отложенных блоках выборки. Поскольку алгоритмы на основе деревьев решений эффективнее работают с признаками, линейно разделяемыми на интервале $(-\infty, +\infty)$, было применено обратное сигмоидальное преобразование вероятностей p в пространство логарифмов отношения шансов:

$$z = \ln \frac{p}{1-p}.$$

Итоговый вектор признаков x_{meta} для обучения финальной мета-модели формируется путем объединения:

$$x_{\text{meta}} = [z_{\text{RoBERTa}} \oplus z_{\text{CharLR}} \oplus z_{\text{WordLR}} \oplus x_{\text{client}} \oplus x_{\text{domain}} \oplus x_{\text{extra}}],$$

где $\oplus$ обозначает операцию векторного сцепления преобразованных прогнозов базовых моделей (z), профиля клиента (x_{client}), доменных (x_{domain}) и инженерных (x_{extra}) признаков.

ЭКСПЕРИМЕНТЫ

Набор данных и базовые стратегии

Настоящее исследование проводилось на закрытом корпоративном корпусе системы технической поддержки, содержащем 570 тыс. текстовых записей. Исходный дисбаланс классов достигал значения коэффициента 220. Применение разработанного алгоритма адаптивного квотирования сократило обучающую выборку до 55 тыс. объектов, снизив дисбаланс до 14.6 при сохранении репрезентативности редких классов. Распределение классов до и после квотирования представлено на рис. 1.

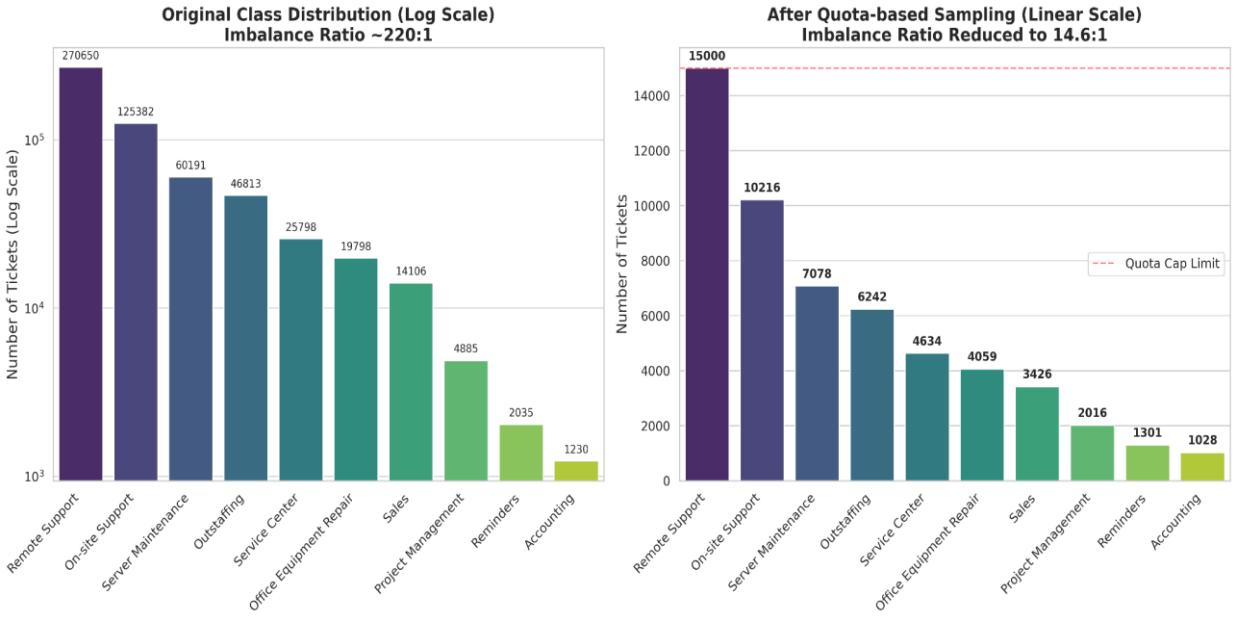

Рис. 1. Распределение классов до и после квотирования.

Далее обоснуем стратегию очистки**.** Для определения оптимального метода работы с шумом сравнивались три подхода к подготовке данных: синтетическое расширение (SMOTE), базовый подход (обучение на данных, сбалансированных методом квотирования, без удаления шума) и предложенная плотностная фильтрация (HDBSCAN). Базовым классификатором выступал алгоритм градиентного усиления. Сравнительный анализ (рис. 2) подтвердил негативное влияние синтетического расширения на зашумленных данных: метрика точности снизилась из-за тиражирования ошибок разметки. Напротив, плотностная фильтрация повысила точность на 1.99% и F1-меру на 2.57% относительно базового подхода, доказав, что удаление стохастического шума эффективнее искусственного увеличения выборки. Необходимо отметить, что плотностная кластеризация является наиболее ресурсоемким этапом конвейера, однако она проводится однократно на этапе подготовки данных и выполняется на центральном процессоре. Это исключает аппаратную зависимость от графических ускорителей и обеспечивает высокую универсальность предложенного решения при внедрении.

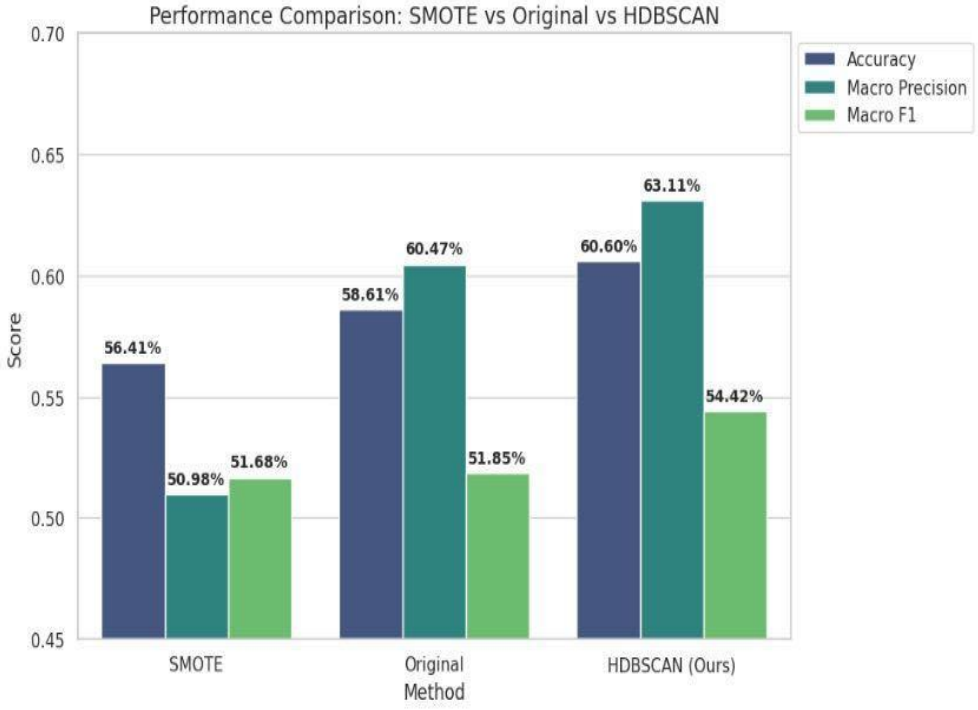

Рис. 2. Сравнение влияния различных методов предобработки на метрики классификации.

Оценка ансамбля и оптимизация

В качестве базовых моделей первого уровня обучались логистические регрессии на символьных и словных n-граммах, а также доменно-адаптированный трансформер. Оценка производилась методом пятикратной перекрестной проверки. Как видно из табл. 1, линейные модели продемонстрировали качество, сопоставимое с нейросетевым подходом, обеспечив при этом необходимую независимость предсказаний для успешного ансамблирования.

Табл. 1. Результаты базовых моделей первого уровня.

Режим	Признаки	Точность
A	Char-level TF-IDF + LogReg	0.654 ± 0.005
B	Word-level TF-IDF + LogReg	0.645 ± 0.005
C	XLM-RoBERTa Large	0.691 ± 0.006

Оптимизация мета-классификатора. Векторы вероятностей базовых моделей агрегировались с мета-признаками для обучения финальной мета-модели. Применение алгоритма байесовской оптимизации гиперпараметров [17] позволило улучшить обобщающую способность ансамбля (табл. 2).

Табл. 2. Сравнение производительности ансамбля

Конфигурация	Точность	Макро-F1	ROC-AUC
Default XGBoost	0.765 ± 0.004	0.701	0.953
Optuna Tuned	0.767 ± 0.004	0.706	0.954

График функции потерь на валидационных выборках демонстрирует высокую устойчивость алгоритма. Выход кривой на асимптотическое плато без последующего роста свидетельствует об эффективной работе механизма ранней остановки (рис. 3).

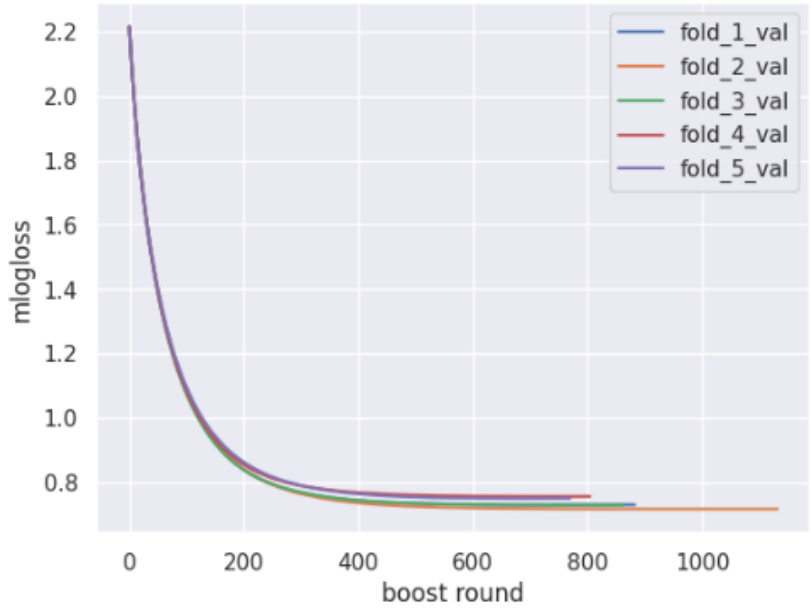

Рис. 3. График функции потерь на валидационных выборках.

Итоговая оценка качества

Для валидации применимости решения в промышленной эксплуатации проводилось тестирование на отложенной выборке из 15 тыс. заявок, имитирующей реальный поток обращений. В качестве ключевых метрик использовались точность первого выбора (R@1) и точность по трем лучшим предсказаниям (R@3), отражающая эффективность системы как помощника оператора. Сравнительный анализ сценариев эксплуатации представлен в табл. 3. В идеальных условиях (клиенты, известные системе, очищенные данные) метрика R@1 составила 81.7%, а R@3 достигла 97.4%. Важно отметить, что даже в наиболее сложном сценарии, включающем зашумленные данные и запросы от новых заказчиков, метрики сохранили высокие значения: R@1 = 72.5%, R@3 = 91.5%. Это подтверждает применимость модели в промышленной эксплуатации: более чем в 90% случаев верное решение находится среди трех предложенных рекомендаций.

Табл. 3. Сравнительный анализ производительности.

Сценарий	Покрытие	Полнота R@1	Полнота R@3	Взвешенная F1-мера	ROC-AUC
Известные клиенты, без шума	84.31%	0.817	0.974	0.821	0.961
Известные/новые клиенты, без шума	84.0%	0.773	0.942	0.794	0.951
Известные клиенты, с шумом	100%	0.764	0.943	0.762	0.923
Известные/новые клиенты, с шумом	100%	0.725	**0.915**	0.735	0.908

ЗАКЛЮЧЕНИЕ

Предложен комплексный метод автоматической классификации коротких текстов для систем технической поддержки, функционирующих в условиях силь-

ного классового дисбаланса и высокого уровня стохастического шума. Экспериментально доказано преимущество алгоритмов плотностной фильтрации над традиционными методами синтетического расширения выборки. В частности, применение кластеризации HDBSCAN позволило выявить и исключить 16.5% зашумленных объектов, что предотвратило тиражирование ошибок разметки и обеспечило значимый прирост метрик качества по сравнению с алгоритмом SMOTE. Полученные результаты подтверждают фундаментальную гипотезу: при работе с реальными корпоративными данными качество обучающей выборки приоритетнее ее объема.

Итоговая архитектура решения, реализованная в парадигме многоуровневого ансамблирования и настроенная с помощью байесовской оптимизации гиперпараметров, объединила глубокие семантические, лексические и интерпретируемые мета-признаки. Ансамблевая модель продемонстрировала высокую надежность: на отложенной тестовой выборке метрика R@3 достигла 97.4%. Данный результат позволяет эффективно использовать разработанный классификатор в контуре промышленной эксплуатации как систему поддержки принятия решений, где вероятность ошибки в рекомендациях составляет менее 3%. Разработанный вычислительный конвейер отличается высокой степенью универсальности и не имеет жесткой привязки к специфике исходной выборки. В связи с этим перспективным направлением для дальнейших исследований является валидация предложенной методологии на текстовых корпусах из смежных технических предметных областей.

СПИСОК ЛИТЕРАТУРЫ

1. *Zhang Y. et al.* Deep Long-Tailed Learning: A Survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023. Vol. 45, No. 3. P. 3079–3099. https://doi.org/10.1109/TPAMI.2021.3114116

2. *Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P.* SMOTE: synthetic minority over-sampling technique // Journal of Artificial Intelligence Research. 2002. Vol. 16. P. 321–357. https://doi.org/10.1613/jair.953

3. *Zha D. et al.* Data-centric Artificial Intelligence: A Survey // ACM Computing Surveys. 2025. Vol. 57, No. 5. Article 129. https://doi.org/10.1145/3711118

4. *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval // Information Processing & Management. 1988. Vol. 24, No. 5. P. 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

5. *Batiuk T., Dosyn D.* Intellectual analysis of textual data in social networks using BERT and XGBOOST // Vìsnik Nacìonal'nogo Unìversitetu L'vìvs'ka Polìtehnìka Serìâ Ìnformacìjnì Sistemi Ta Merežì. 2025. Vol. 17. P. 44–60. https://doi.org/10.23939/sisn2025.17.044

6. *Parmar M., Tiwari A.* Enhancing text classification performance using stacking ensemble method with TF-IDF feature extraction // Proceedings of the 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI). Kathmandu, Nepal. 2024. P. 166–174. https://doi.org/10.1109/ICMCSI61480.2024.10493890

7. *Zemp M.* Text classification of service desk tickets. Master's thesis. Winterthur, Zurich University of Applied Sciences. 2021. https://www.zhaw.ch/storage/shared/upload/MAS21_Ticket_Classification_Zemp.pdf (дата обращения: 12.02.2026)

8. *Akhbardeh F., Alm C.O., Zampieri M., Desell T.* Handling extreme class imbalance in technical logbook datasets // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP). Online. 2021. P. 4034–4045. https://doi.org/10.18653/v1/2021.acl-long.312

9. *Padurariu C., Breaban M.E.* Dealing with data imbalance in text classification // Procedia Computer Science. 2019. Vol. 159. P. 736–745. https://doi.org/10.1016/j.procs.2019.09.229

10. *Asyaky M.S., Mandala R.* Improving the performance of HDBSCAN on short text clustering by using word embedding and UMAP // 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA). Bandung, Indonesia. 2021. P. 1–6. https://doi.org/10.1109/ICAICTA53211.2021.9640285

11. *McInnes L., Healy J., Astels S.* hdbscan: Hierarchical density based clustering // Journal of Open Source Software. 2017. Vol. 2, No. 11. P. 205. https://doi.org/10.21105/joss.00205

12. *Халов А.П., Атаева О.М.* Автоматические и полуавтоматические методы построения графа знаний предметной области и расширения онтологии // Электронные библиотеки. 2025. Т. 28, № 6. С. 1481–1519. https://doi.org/10.26907/1562-5419-2025-28-6-1481-1519

13. *Wolpert D.H.* Stacked generalization // Neural Networks. 1992. Vol. 5, No. 2. P. 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

14. *Charikar M.S.* Similarity estimation techniques from rounding algorithms // Proceedings of the thirtieth annual ACM symposium on Theory of computing (STOC). 2002. P. 380–388. https://doi.org/10.1145/509907.509965

15. *Micci-Barreca D.* A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems // SIGKDD Explorations Newsletter. 2001. Vol. 3, No. 1. P. 27–32. https://doi.org/10.1145/507533.507538

16. *Chen T., Guestrin C.* XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). San Francisco, USA. 2016. P. 785–794. https://doi.org/10.1145/2939672.2939785

17. *Akiba T., Sano S., Yanase T., Ohta T., Koyama M.* Optuna: A next-generation hyperparameter optimization framework // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). Anchorage, USA. 2019. P. 2623–2631. https://doi.org/10.1145/3292500.3330701

IMPROVING SHORT TEXT CLASSIFICATION ROBUSTNESS TO STOCHASTIC NOISE BASED ON DENSITY-DRIVEN TRAINING DATA CLEANING

B. B. Baishev[1] [0009-0007-9287-4248], A. P. Khalov[2] [0009-0005-4584-8245]

[1]*Nazarbayev University, Astana, Kazakhstan*

[2]*Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia*

[1]baishevbasar@gmail.com, [2]khalov.a@phystech.edu

Abstract

The paper addresses the problem of short text request classification under conditions of significant class imbalance and high noise levels in real-world communication flows. The limited effectiveness of synthetic oversampling techniques when dealing with noisy labeling is demonstrated. A hybrid method is proposed, combining preliminary density-based data cleaning and multi-level model ensembling. The application of a density-based clustering algorithm enabled the exclusion of 16.5% of informational noise from the total sample volume. The final model features a two-level architecture and is optimized using Bayesian hyperparameter search. A Recall@3 (R@3) metric of 97.4% was achieved on a hold-out test set. The proposed method allows for the automation of the request distribution process, significantly reducing operator workload and decreasing dispatch time.

Keywords*: natural language processing, noisy text data, ensemble learning, robust classification, noise filtering.*

REFERENCES

1. *Zhang Y. et al.* Deep Long-Tailed Learning: A Survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023. Vol. 45, No. 3. P. 3079–3099. https://doi.org/10.1109/TPAMI.2021.3114116

2. *Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P.* SMOTE: synthetic minority over-sampling technique // Journal of Artificial Intelligence Research. 2002. Vol. 16. P. 321–357. https://doi.org/10.1613/jair.953

3. *Zha D. et al.* Data-centric Artificial Intelligence: A Survey // ACM Computing Surveys. 2025. Vol. 57, No. 5. Article 129.
https://doi.org/10.1145/3711118

4. *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval // Information Processing & Management. 1988. Vol. 24, No. 5. P. 513–523.
https://doi.org/10.1016/0306-4573(88)90021-0

5. *Batiuk T., Dosyn D.* Intellectual analysis of textual data in social networks using BERT and XGBOOST // Vìsnik Nacìonal'nogo Unìversitetu L'vìvs'ka Polìtehnìka Serìâ Ìnformacìjnì Sistemi Ta Merežì. 2025. Vol. 17. P. 44–60.
https://doi.org/10.23939/sisn2025.17.044

6. *Parmar M., Tiwari A.* Enhancing text classification performance using stacking ensemble method with TF-IDF feature extraction // Proceedings of the 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI). Kathmandu, Nepal. 2024. P. 166–174.
https://doi.org/10.1109/ICMCSI61480.2024.10493890

7. *Zemp M.* Text classification of service desk tickets. Master's thesis. Winterthur, Zurich University of Applied Sciences. 2021.
https://www.zhaw.ch/storage/shared/upload/MAS21_Ticket_Classification_Zemp.pdf

8. *Akhbardeh F., Alm C.O., Zampieri M., Desell T.* Handling extreme class imbalance in technical logbook datasets // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP). Online. 2021. P. 4034–4045.
https://doi.org/10.18653/v1/2021.acl-long.312

9. *Padurariu C., Breaban M.E.* Dealing with data imbalance in text classification // Procedia Computer Science. 2019. Vol. 159. P. 736–745.
https://doi.org/10.1016/j.procs.2019.09.229

10. *Asyaky M.S., Mandala R.* Improving the performance of HDBSCAN on short text clustering by using word embedding and UMAP // 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA). Bandung, Indonesia. 2021. P. 1–6.
https://doi.org/10.1109/ICAICTA53211.2021.9640285

11. *McInnes L., Healy J., Astels S.* hdbscan: Hierarchical density based clustering // Journal of Open Source Software. 2017. Vol. 2, No. 11. P. 205. https://doi.org/10.21105/joss.00205

12. *Khalov A.P., Ataeva O.M.* Automatic and semi-automatic methods for constructing a domain knowledge graph and ontology expansion // Russian Digital Libraries Journal. 2025. Vol. 28, No. 6. P. 1481–1519 (in Russian). https://doi.org/10.26907/1562-5419-2025-28-6-1481-1519

13. *Wolpert D.H.* Stacked generalization // Neural Networks. 1992. Vol. 5, No. 2. P. 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

14. *Charikar M.S.* Similarity estimation techniques from rounding algorithms // Proceedings of the thirtieth annual ACM symposium on Theory of computing (STOC). 2002. P. 380–388. https://doi.org/10.1145/509907.509965

15. *Micci-Barreca D.* A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems // SIGKDD Explorations Newsletter. 2001. Vol. 3, No. 1. P. 27–32. https://doi.org/10.1145/507533.507538

16. *Chen T., Guestrin C.* XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). San Francisco, USA. 2016. P. 785–794. https://doi.org/10.1145/2939672.2939785

17. *Akiba T., Sano S., Yanase T., Ohta T., Koyama M.* Optuna: A next-generation hyperparameter optimization framework // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). Anchorage, USA. 2019. P. 2623–2631. https://doi.org/10.1145/3292500.3330701

СВЕДЕНИЯ ОБ АВТОРАХ

БАИШЕВ Басар Бауржанович – студент 2-го курса, ассистент-исследователь, Назарбаев Университет, кафедра «Компьютерные науки». Область научных интересов: обработка естественного языка (NLP), машинное обучение, нейронные сети, анализ несбалансированных данных.

Bassar Baurzhanovich BAISHEV – second-year undergraduate student, research assistant at the Department of Computer Science Nazarbayev University. Research interests: natural language processing (NLP), machine learning, neural networks, imbalanced data analysis.

email: baishevbasar@gmail.com

ORCID: 0009-0007-9287-4248

***ХАЛОВ* Андрей Петрович** – аспирант МФТИ (ФПМИ), кафедра «Интеллектуальные системы». Область научных интересов: онтологическое моделирование, графы знаний, извлечение знаний из текстов (NER/RE, RAG), многоагентные системы и планирование, применение LLM в корпоративных ИС.

Andrey Petrovich KHALOV – PhD student at the Moscow Institute of Physics and Technology (MIPT), Phystech School of Applied Mathematics and Informatics, Department of Intelligent Systems. Research interests: ontological modeling, knowledge graphs, information extraction from text (NER/RE, RAG), multi-agent systems and planning, application of LLMs in enterprise information systems.

email: khalov.a@phystech.edu

ORCID: 0009-0005-4584-8245

Материал поступил в редакцию 24 марта 2026 года