УДК 004.85
2.3.1. Системный анализ, управление и обработка информации




А.Д. Бухаров, Д.Е. Жаравин
Вологодский государственный университет

МЕТОДЫ СИНТЕТИЧЕСКОЙ АУГМЕНТАЦИИ ДАННЫХ ДЛЯ УЛУЧШЕНИЯ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ЗАПРОСОВ НА МАЛЫХ ВЫБОРКАХ

В данной статье рассматриваются методы синтетической аугментации данных для повышения точности классификации текстовых запросов при малых объемах обучающей выборки. Основное внимание уделяется трем ключевым методам аугментации: синонимизации, перефразированию и бэктрансляции. Экспериментальная часть исследования демонстрирует влияние каждого метода на качество обучаемой модели. Полученные результаты показывают, что аугментация данных позволяет повысить F1-меру системы и может быть полезна специалистам в области NLP при ограниченных данных.

Машинное обучение, аугментация данных, классификация текста, обработка естественного языка, F1-мера.

На сегодняшний день во многих сферах деятельности современные технологии стали неотъемлемой частью рабочих процессов. Компании стремятся автоматизировать рутинные операции, повысить качество обслуживания и ускорить обработку запросов. В связи с этим анализ текстовых данных становится одной из основных задач в области искусственного интеллекта и обработки естественного языка, поскольку является ключевым инструментом для повышения эффективности и удобства работы с информацией.

Использование машинного обучения для классификации и анализа текстовых запросов уже давно стало частью многих бизнес-процессов и технологических решений [1]. Однако внедрение таких систем осложняется тем, что объем размеченных данных ограничен, это характерно для многих прикладных задач, таких как классификация запросов в системах технической поддержки, анализ отзывов клиентов или обработка обращений. Это также относится к ряду специализированных задач, где сбор достаточного объема информации затруднен из-за специфики предметной области или ограниченности ресурсов.

В таких условиях традиционные методы машинного обучения сталкиваются с целым рядом сложностей, к которым относится дисбаланс классов, недостаточная репрезентативность и шум в данных. Однако самой распространенной проблемой является переобучение модели.

Переобучение возникает в том случае, когда нейронная сеть вместо поиска закономерностей и умения обобщать, начинает слишком точно запоминать обучающие данные [2], в результате модель плохо справляется с реальными примерами и начинает ошибаться. На примере автоматической классификации текстовых запросов, модель может научиться распознавать конкретные формулировки и фразы, но не их общий смысл, но если одно и то же предложение будет сформулировано иначе, система будет неверно его интерпретировать.

В большинстве случаев переобучение возникает из-за малого количества примеров для каждого класса. Если в обучающем датасете, например, 1000 записей, которые необходимо распределить по 5 категориям, то в таком случае, на каждый класс будет приходиться всего по 200 примеров, что недостаточно для обучения сложной модели. Для создания хорошей системы, при условии, что нейронная сеть глубокая, обучается с нуля, и не имеет заранее выученных признаков, потребуется от 10 000 до 100 000 примеров на один класс. В связи с этим основные методы борьбы с переобучением направлены либо на уменьшение склонности модели запоминать данные, либо на уменьшение ее сложности [3], так как слишком мощная модель на малом объеме данных легко переобучается, а изменение ее архитектуры помогает избавиться от излишней гибкости.

Первый метод получил название «регуляризация», который на практике подразделяется на два основных вида: L1/L2-регуляризация, когда в функцию потерь добавляется некоторый штраф за слишком большие веса, и Dropout, когда во время обучения случайно отключается некоторая часть нейронов. Первый способ делает одну часть весов равными нулю, а вторую сглаживает, не давая им чрезмерно увеличиваться. Второй же способ направлен на предотвращение избыточной зависимости от отдельных признаков и заставляет сеть распределять информацию между разными путями. В конечном итоге под действием регуляризации модель становится более разряженной, менее чувствительной к шуму, а также устойчивой на новых данных.

К отдельному классу методов борьбы с переобучением стоит отнести «аугментацию данных», поскольку она работает не с разбиением или перераспределением выборки, а с преобразованием непосредственно самих исходных данных [4]. Аугментация создает новые варианты текстов и тем самым расширяет обучающий датасет, искусственно увеличивая количество примеров.

В рамках исследования рассматриваются три основных метода аугментации текстовых данных [5]:

1. Синонимизация, которая направлена на замену отдельных слов на подходящие синонимы, что повышает разнообразие, сохраняя при этом общий смысл предложения.

2. Перефразирование, которое основано на переписывании исходного текста, другими словами, делая выборку богаче за счет изменения структуры, что в свою очередь помогает модели лучше понимать разные варианты формулировок.

3. Бэктрансляция, являющаяся одним из самых надежных способов создавать естественные вариации текста без потери смысла, который основан на переводе фразы на другой язык и обратно.

Оценка качества модели является критически важным этапом в процессе разработки и внедрения систем машинного обучения [6]. Это особенно актуально в условиях ограниченного объема данных, так как неправильно выбранные метрики могу привести к ложному ощущению высокого качества моделей.

Для оценки качества моделей классификации используется специальная метрика F1-мера, она сочетает в себе два других показателя: точность и полноту. Точность определяет долю правильно предсказанных положительных примеров среди всех примеров, которые модель предсказала как положительные. Полнота представляет собой долю правильно предсказанных положительных примеров среди всех действительно положительных примеров. Расчет F1 производится по формуле (1):

$$F1_i = 2 \cdot \frac{\frac{TP_i}{TP_i + FP_i} \cdot \frac{TP_i}{TP_i + FN_i}}{\frac{TP_i}{TP_i + FP_i} + \frac{TP_i}{TP_i + FN_i}}, \quad (1)$$

где TP_i – количество правильно классифицированных примеров в класс i;

FP_i – количество ошибочно классифицированных примеров в класс i;

FN_i – количество примеров класса i, ошибочно классифицированных как другие классы.

В задачах, когда все классы одинаково важны, а данные не сбалансированы, для оценки качества системы целесообразнее использовать Macro-F1 – среднее арифметическое значение F1 по всем классам, которое рассчитывается по формуле (2):

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=1}^{N} F1_i, \quad (2)$$

где N – общее количество классов в задаче;

$F1_i$ – F1-мера для i-го класса.

Показатель Macro-F1 удобен тем, что подчеркивает слабые места модели, например если F1 для одного класса значительно ниже, чем для других, это может сигнализировать о необходимости дополнительной разметки данных или улучшения аугментации для этого класса.

В рамках исследования используется реальный датасет, состоящий из 1000 писем, разделенных на пять основных классов, которые соответствуют типичным запросам в техническую поддержку.

Важным и обязательным этапом перед обучением модели является предобработка данных, так как именно на этом этапе формируется качественное входное представление текста [7]. К основным этапам подготовки данных относятся следующие шаги:

1. Сначала текст разбивается на отдельные токены, что позволяет модели воспринимать текст как последовательность элементов и анализировать его структуру.

2. Затем удаляются слова, не несущие полезной информации для классификации. Такой шаг снижает размерность данных, уменьшает шум и помогает повысить точность модели.

3. После этого выполняется приведение слов к их нормальным формам. Это сокращает количество вариантов одного и того же слова и помогает модели лучше понимать общий смысл текста.

4. Далее текст векторизуется и преобразуется в числовое представление. На этом этапе оценивается значимость слов, а результат его работы можно напрямую подать в линейные модели.

Когда подготовка данных завершена, выбирается и настраивается базовая модель нейронной сети. Разные архитектуры обеспечивают разные уровни обработки текста [8], а уже поверх обученной модели при необходимости можно применять аугментацию. После завершения обучения проводится тестирование на отдельной выборке, которая не использовалась в обучении, что позволяет объективно оценить качество модели по F1-метрике.

В таблице представлено сравнение результатов экспериментов с различными методами аугментации данных на модели DoBERT, которая лучше всего подходит для поставленной задачи, так как заранее предобучена на большой выборке текстовых данных и требует меньше вычислительных ресурсов по сравнению с полноценной BERT [9].

Визуализация полученных экспериментальных данных представлена на графике (рис.), что позволяет наглядно оценить динамику изменений.

Таблица

Сравнение результатов экспериментов с аугментацией данных

Метод аугментации	Macro-F1 (среднее ± σ)	Прирост F1	Преимущества	Недостатки
Без аугментации	0,71 ± 0,03	–	Простота реализации	Низкое разнообразие данных
Синонимизация	0,74 ± 0,03	3 %	Сохранение смысла, простота	Риск неестественных формулировок
Парафразирование	0,75 ± 0,02	5 %	Естественные вариации текста	Высокие вычислительные затраты
Бэктрансляция	0,77 ± 0,02	8 %	Наиболее естественные вариации	Возможная потеря нюансов смысла

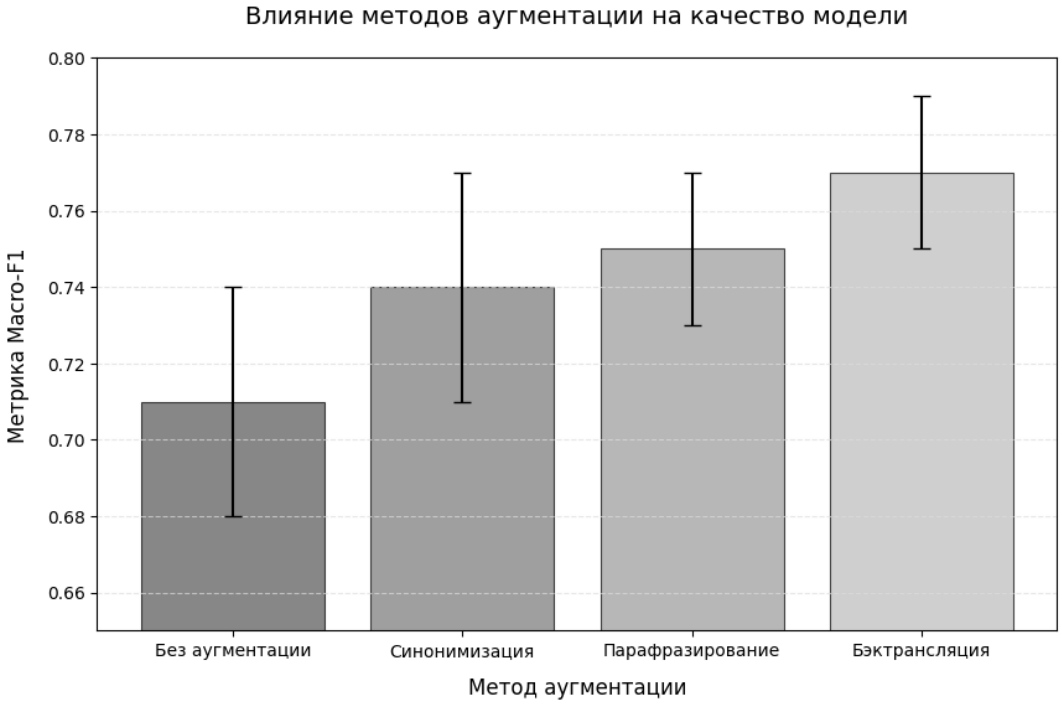


Рис. Оценка качества моделей по метрике Macro-F1

По результатам эксперимента установлено, что метод бэктрансляции показывает наилучшие результаты, прирост F1-меры на 8 % выше, чем у базовой модели. Это связано с тем, что в выборке появляются наиболее естественные вариации текстовых запросов, что улучшает обобщающую способность модели. Перефразирование также демонстрирует высокие результаты, но требует больше вычислительных ресурсов для генерации. Метод синонимизации является наиболее простым, однако его эффективность ниже.

Таким образом, использование методов аугментации данных по отдельности дает заметный положительный эффект и способствует повышению качества модели. Однако для более сложных задач целесообразно будет комбинировать несколько подходов аугментации, что обеспечивает более высокое разнообразие обучающей выборки [10].

Перспективным направлением дальнейшего развития является исследование аугментации в сочетании с другими техниками повышения качества моделей, такими как регуляризация, трансферное и активное обучение, методы балансировки и очистки данных от шума. Дополнительные преимущества может дать применение методов самообучения и использование гибридных моделей на базе генетических алгоритмов, которые могли бы объединить самые сильные стороны различных подходов.

Литература

1. Дейтел, П. Python: искусственный интеллект, большие данные и облачные вычисления / П. Дейтел, Х. Дейтел. – Санкт-Петербург : Питер, 2020. – 864 с.

2. Жаравин, Д. Е. Влияние стартовых параметров на эффективность генетических алгоритмов на примере многомерной тестовой функции / Д. Е. Жаравин // Системы управления и информационные технологии. – 2026. – № 1(103). – С. 8–11.

3. Брюс, П. Практическая статистика для специалистов Data Science / П. Брюс, Э. Брюс. – Санкт-Петербург : БХВ-Петербург, 2021. – 341 с.

4. Лейн, Х. Обработка естественного языка в действии / Х. Лейн, Х. Хапке, К. Ховард. – Санкт-Петербург : Питер, 2020. – 576 с.

5. Вандер, Д. Python для сложных задач. Наука о данных и машинное обучение / Д. Вандер. – Санкт-Петербург : Питер, 2020. – 576 с.

6. Ланц, Б. Машинное обучение на R. Экспертные техники для прогностического анализа / Б. Ланц. – Санкт-Петербург : Питер, 2020. – 464 с.

7. Ахмадуллин, Ф. Р. Основы анализа данных и интеллектуальные системы / Ф. Ахмадуллин, А. Л. Золкин, С. Н. Косников. – Санкт-Петербург : Лань, 2025. – 176 с.

8. Уайт, Э. Создание встраиваемых систем / Э. Уайт. – Астана : АЛИСТ, 2025. – 256 с.

9. Просиз, Д. Прикладное машинное обучение и искусственный интеллект для инженеров / Д. Просиз. – Астана : АЛИСТ, 2024. – 432 с.

10. Шалев-Шварц, Ш. Идеи машинного обучения. От теории к алгоритмам / Ш. Шалев-Шварц, Ш. Бен-Давид. – Москва : ДМК Пресс, 2019. – 432 с.

A.D. Bukharov, D.E. Zharavin
Vologda State University

METHODS OF SYNTHETIC DATA AUGMENTATION FOR IMPROVING TEXT QUERY CLASSIFICATION ON SMALL DATASETS

This article examines synthetic data augmentation methods for improving the accuracy of text query classification with small training datasets. The focus is on three key augmentation methods: synonymization, paraphrasing and backtranslation. The experimental part of the study demonstrates the influence of each method on the quality of the trained model. The results demonstrate that data augmentation can improve the system's F1 score and may be useful for NLP specialists with limited data.

Machine learning, data augmentation, text classification, natural language processing, F1 score.