УДК 004.8

Исследование влияния предобработки текста на качество тематической классификации

Д. Ю. Подзол, И. А. Коломойцева
ФГБОУ ВО «Донецкий национальный технический университет» (г. Донецк)
кафедра «Программная инженерия» им. Л.П,Фельдмана
E-mail: david.podzol@mail.ru, bolatiger@mail.ru.

Аннотация
В работе исследуется влияние различных стратегий предобработки текста на качество тематической классификации русскоязычных документов. Сравниваются модели SVM, LSTM и RuBERT при использовании трех уровней очистки данных. Показано, что умеренная предобработка повышает точность классических и рекуррентных моделей, тогда как чрезмерная фильтрация снижает качество трансформерных архитектур. На основе результатов предложена адаптивная стратегия предобработки, учитывающая особенности каждой модели.

Введение

Автоматическая тематическая классификация является одной из базовых задач в области обработки естественного языка. Её успешное решение позволяет эффективно структурировать большие текстовые массивы, обеспечивая работу интеллектуальных систем анализа данных. В то же время, качество классификации зависит не только от архитектуры модели, но и от корректности этапа предобработки данных.

Процедуры нормализации, токенизации и лемматизации формируют основу дальнейшего анализа текста, устраняя шум и приводя слова к единой форме. Однако, как показывают исследования [1], [2], чрезмерная фильтрация может приводить к потере смысловых зависимостей, что особенно важно для современных контекстных моделей. Так же полагает, что предобработка является не просто техническим этапом, а ключевым фактором, определяющим способность модели извлекать значимые признаки из текстов.

Целью исследования является комплексное изучение влияния степени предобработки текста на качество тематической классификации русскоязычных документов и проведение эксперимента на корпусе Lenta.ru с использованием моделей SVM, LSTM и RuBERT. Гипотеза заключается в том, что между уровнем очистки и качеством классификации существует нелинейная зависимость, при которой оптимальный результат достигается при умеренной глубине предобработки.

Для достижения поставленной цели в работе решаются следующие задачи:

– провести сравнительный анализ существующих подходов к предобработке текстов;

– реализовать эксперимент на корпусе Lenta.ru с применением моделей SVM, LSTM и RuBERT;

– оценить влияние каждого сценария очистки на точность и устойчивость моделей;

– определить оптимальную стратегию предобработки для разных архитектур.

Помимо общепринятых задач автоматической тематической классификации, в современных прикладных сценариях (медиаресурсы, мониторинг новостей, автоматическая кодировка тематик социальных сообщений) предобработка текста задаёт не только качество модели, но и её способность к быстрому развёртыванию и поддержке на потоковых данных. Практическая потребность в решениях, работающих в реальном времени, диктует требования к эффективности предобработки: методы должны быть не только информативными, но и вычислительно экономичными.

Анализ существующих исследований

Роль предобработки текста в машинном обучении активно изучается на протяжении последних двух десятилетий. В работах Mikolov и соавторов [3] было доказано, что корректная сегментация токенов и нормализация данных улучшают устойчивость обучаемых моделей. Исследования Vaswani и коллег [1] показали, что даже в архитектуре Transformer, основанной на механизме самовнимания, корректная токенизация определяет успех классификации. Devlin и соавторы [2], создавшие модель BERT, отмечают, что для языков со сложной

морфологией (в том числе русского) выбор токенизатора оказывает значительное влияние на итоговую точность.

Российские исследования также подтверждают данную зависимость. Так, Стрелец и Иванников [5] показали, что применение лемматизации и фильтрации стоп-слов существенно повышает точность классификации текстов. Мотовских [5] пришёл к выводу, что избыточная очистка данных, напротив, снижает эффективность нейросетевых моделей. Kuratov и Архипов [4], адаптируя модель BERT для русского языка (RuBERT), подчёркивают, что сохранение синтаксических и морфологических элементов является необходимым условием сохранения контекста.

Таким образом, проблема предобработки остаётся актуальной и требует комплексного анализа с учётом архитектурных особенностей моделей.

Следует отметить влияние субсловных методов представления слов, таких как FastText, которые учитывают внутреннюю морфологию слова и тем самым устойчивее к вариативности словоформ и опечаткам по сравнению с чисто словными представлениями. Эти методы особенно полезны в задачах с ограниченным объёмом размеченных данных, поскольку они позволяют обобщать информацию по морфемам и суффиксам. [6]

При выборе инструментов для предобработки важно учитывать доступность и качество морфологического анализа для русского языка: библиотеки вроде pymorphy2 предоставляют стабильные средства для лемматизации и разбора форм, но их применение требует аккуратной настройки фильтров стоп-слов и правил нормализации, чтобы не избавиться от информативных форм. [7]

Современные подходы к морфологическому разбиению и сегментации слов (нейросетевые модели сегментации и BPE/SentencePiece) позволяют строить токенизацию, адаптированную под корпус, что сокращает необходимость ручной лемматизации и делает систему более гибкой для языков с богатой морфологией. Это направление даёт основу для разработки динамических токенизаторов, которые подстраиваются под статистику корпуса. [8]

Методология исследования и проведение эксперимента

В качестве экспериментальной базы использовался русскоязычный корпус Lenta.ru, включающий около восьми тысяч документов, сгруппированных по семи тематическим категориям: политика, экономика, наука, культура, спорт, технологии и происшествия. Для обеспечения сбалансированности данных была выполнена стратифицированная выборка, что позволяло равномерно распределить документы между категориями и снизить влияние дисбаланса классов на результаты классификации [5].

Исходные тексты проходили три уровня предобработки. Минимальная очистка включала удаление пунктуации, приведение текста к нижнему регистру и удаление лишних пробелов, при этом сохранялось максимальное количество информации, включая служебные слова и морфологические особенности, что особенно важно для контекстных моделей [4]. Стандартный уровень предобработки дополнительно включал лемматизацию с использованием инструментов морфологического анализа русского языка, таких как pymorphy2, и фильтрацию стоп-слов, что позволяло устранить шумовые и редкие токены и улучшить качество работы статистических и рекуррентных моделей [5]. Контекстная предобработка была ориентирована на модели типа Transformer и включала сохранение служебных слов и синтаксических конструкций при минимальной очистке пунктуации, что позволяло сохранять грамматические и контекстные связи в тексте [1],[2],[4].

Для анализа применялись три класса моделей. Метод опорных векторов (SVM) с признаковым представлением TF-IDF обучался с настройкой гиперпараметров с помощью сеточного поиска, что обеспечивало оптимальный баланс между переобучением и недообучением [5]. Рекуррентная нейросеть LSTM использовала эмбеддинги Word2Vec, обученные на корпусе Lenta.ru, при этом применялось регуляризованное обучение с Dropout и батч-нормализацией для повышения устойчивости модели и предотвращения переобучения [3]. Контекстная модель RuBERT, дообученная на корпусе Lenta.ru для задачи классификации, использовала оптимизатор AdamW с небольшим шагом обучения и стратегией ранней остановки для предотвращения переобучения [2],[4]

Качество классификации оценивалось с помощью метрик Accuracy и F1-macro, что позволяло учитывать баланс между точностью и полнотой для всех классов. Для повышения устойчивости оценки результатов проводилась кросс-валидация с пятью фолдами, что обеспечивало более надёжное сравнение эффективности различных стратегий предобработки текста [5].

Чтобы объективно оценить влияние различных методов предобработки, в исследование были включены несколько контрольных сценариев, каждый из которых отражал свой подход к обработке исходных текстов. Один из них предусматривал орфографическую нормализацию, направленную на приведение словоформ к единому стандарту и исправление наиболее распространённых

опечаток. Другой сценарий был основан на применении субсловной токенизации с использованием методов BPE или SentencePiece, что позволяло исключить необходимость лемматизации и обеспечивало устойчивость к вариативности словоформ. Третий сценарий представлял собой гибридный подход, в котором лемматизация сочеталась с субсловной сегментацией для обработки редких токенов, сохраняя при этом баланс между нормализацией и сохранением морфологических особенностей.

При использовании субсловных эмбеддингов применялись предобученные представления, которые затем до обучались на корпусе Lenta.ru и интегрировались в архитектуру LSTM. Аналогичные эмбеддинги использовались и в SVM, но в этом случае предварительно агрегировали в единый вектор документа. В случае RuBERT применялась стратегия частичного сохранения пунктуации и служебных слов при работе токенизатора SentencePiece, поскольку эти элементы оказывают значимое влияние на устойчивость механизма самовнимания и формирование контекстных зависимостей [8].

Результаты исследований и экспериментов

Результаты экспериментов демонстрируют, что влияние предобработки текста на качество тематической классификации сильно зависит от архитектуры используемой модели. Для SVM, основанной на признаковом представлении TF-IDF, стандартная очистка текста, включающая лемматизацию и фильтрацию стоп-слов, привела к заметному улучшению показателей Accuracy с 0.78 до 0.85 и F1-макро с 0.76 до 0.84. Это объясняется тем, что классические алгоритмы чувствительны к "шумам" в виде различных форм слова и служебных слов, и их удаление упрощает пространство признаков, повышая стабильность классификации.

При сравнении моделей также было отмечено различие в характере ошибок. Для SVM типичными были ошибки, связанные с путаницей между тематически близкими категориями, такими как «политика» и «экономика», что обусловлено отсутствием глубокого контекстного анализа. LSTM лучше справлялась с подобными случаями, однако испытывала сложности при работе с длинными текстами, где важные признаки распределены по всему документу. RuBERT показал наименьшее количество ошибок, особенно в случаях скрытых тематических связей, что подтверждает его способность учитывать более сложные языковые зависимости.

Для модели LSTM с эмбеддингами Word2Vec умеренная очистка текста также оказалась полезной. При стандартной предобработке Accuracy составила 0.88, а F1-макро — 0.87. Расширенная предобработка (с более глубокой фильтрацией и нормализацией) позволила улучшить показатели до 0.90 и 0.89 соответственно. Это говорит о том, что рекуррентные нейросети извлекают контекстные зависимости между словами, и частичное удаление нерелевантных токенов помогает модели лучше концентрироваться на значимых признаках, не разрушая при этом семантические связи.

Контекстная модель RuBERT продемонстрировала иную тенденцию. Для стандартной предобработки Accuracy и F1-макро достигли 0.94, что является наивысшим результатом среди всех моделей. При применении контекстной предобработки, которая предполагает удаление части служебных слов и более сильную фильтрацию, показатели слегка снизились до 0.93. Это свидетельствует о том, что Transformer-модели полагаются на сохранение синтаксических и морфологических элементов для механизма самовнимания, и чрезмерная очистка текста может нарушить внутренние представления модели, снижая её способность корректно различать темы. Влияние уровня предобработки на качество тематической классификации приведено в таблице 1.

Таблица 1 – Влияние уровня предобработки на качество тематической классификации

Модель	**Уровень предобработки**	**Точность**	**F1**	**Изменение к базовой**
SVM (TF-IDF)	Без очистки	0.78	0.76	—
SVM (TF-IDF)	Стандартная	0.85	0.84	+8%
LSTM (Word2Vec)	Стандартная	0.88	0.87	—
LSTM (Word2Vec)	Расширенная	0.90	0.89	+2%
RuBERT	Стандартная	0.94	0.94	—
RuBERT	Контекстная	0.93	0.93	−1%

Дополнительно следует отметить, что зависимость качества классификации от степени предобработки носит нелинейный характер. Для SVM и LSTM постепенное увеличение степени очистки повышает качество, пока не достигается оптимальный баланс между удалением шума и сохранением информативных элементов текста. Для RuBERT наблюдается противоположная тенденция: слишком сильная очистка ведёт к потере ключевых контекстных сигналов. Это подчёркивает необходимость выбора стратегии

предобработки с учётом архитектуры модели и специфики корпуса.

Анализ результатов

Анализ результатов подтверждает, что предобработка является контекстно-зависимым этапом, эффективность которого определяется природой модели. Классические алгоритмы, работающие с частотными признаками, нуждаются в строгой нормализации текста, поскольку их качество зависит от статистической чистоты данных. Нейросетевые модели, напротив, учатся на контексте и способны компенсировать часть лексического шума за счёт внутренних представлений. Следовательно, подход к предобработке должен быть адаптивным.

Оптимальным является использование гибридной стратегии, при которой система автоматически подбирает глубину очистки в зависимости от архитектуры и языковой специфики корпуса.

Есть предположения, что интеграция обучаемых механизмов предобработки в модели глубокого обучения станет следующим этапом развития систем тематической классификации. Более глубокий анализ показал, что предобработка влияет не только на итоговые метрики, но и на внутренние представления моделей. В частности, применение лемматизации приводило к сглаживанию семантического пространства Word2Vec, что упрощало задачу классификации для LSTM, но нарушало более тонкие контекстные зависимости, необходимые для RuBERT.

Кроме того, модели по-разному реагировали на удаление пунктуации: для SVM это сокращало разреженность признакового пространства, тогда как для RuBERT частичное сохранение пунктуации помогало более точному учёту синтаксических структур. Таким образом, влияние очистки является многоуровневым и требует более детального изучения.

Дополнительно стоит отметить, что выбор стратегии предобработки напрямую влияет на способность модели выявлять тонкие тематические связи между словами и фразами. Например, для контекстных моделей сохранение служебных слов и синтаксических конструкций позволяет лучше улавливать зависимые связи и полисемию, что критично для точной классификации новостных и экспертных текстов.

В то же время для классических моделей чрезмерная детализация может создавать разреженное признаковое пространство и ухудшать обобщающую способность алгоритма. Это подчёркивает необходимость комплексного подхода: не только экспериментальной оптимизации предобработки для каждой архитектуры, но и разработки методов динамического подбора параметров очистки текста на этапе обучения, что может стать перспективным направлением будущих исследований.

Интерпретация моделей с использованием методов объяснимости дала дополнительные инсайты. Применение LIME позволило выявить, какие токены (или их подчасти в случае субсловной токенизации) вносили наибольший вклад в решение классификатора; это позволило уточнить список стоп-слов и исключаемых конструкций для классических моделей [9].

Аналогично, использование SHAP обнаружило взаимодействия между токенами, которые не очевидны при простом подсчёте частот; для RuBERT это позволило показать, что наличие сочетаний служебных слов и пунктуации часто усиливает вклад соседних лексических единиц при принятии решения трансформером [10]. Эти результаты подтверждают мысль о том, что инструменты интерпретации должны стать стандартной частью пайплайна предобработки, так как они позволяют адаптивно корректировать очистку на основе вкладов токенов в итоговую метрику.

Практические рекомендации

На основании результатов эксперимента можно сформулировать ряд практических рекомендаций, ориентированных на применение различных стратегий предобработки в зависимости от типа используемой модели. Для классических алгоритмов машинного обучения, таких как SVM и логистическая регрессия, наилучших результатов удаётся достичь при использовании лемматизации в сочетании с фильтрацией стоп-слов и применением субсловной агрегации для редко встречающихся токенов. Такой подход уменьшает разреженность признакового пространства и обеспечивает более устойчивую классификацию. [7], [6]

Для рекуррентных нейронных сетей предпочтительно сохранять часть орфографических особенностей текста, поскольку это способствует более точному моделированию контекста. Наилучшие результаты достигаются при использовании субсловных эмбеддингов, таких как FastText, которые позволяют учитывать внутреннюю структуру слова и обеспечивают высокую устойчивость к редким словоформам. [6]

При работе с трансформными моделями, включая RuBERT, оптимальной стратегией является минимальная очистка текста, исключающая удаление служебных слов и пунктуации. Если дополнительная фильтрация всё же применяется, она должна быть предварительно проверена на влияние как на итоговые метрики, так и на интерпретируемость модели. [2], [4]

Во всех рассматриваемых сценариях значимую роль играет анализ вкладов токенов в итоговое решение классификатора, поэтому использование методов объяснимости LIME и SHAP позволяет выявлять ключевые элементы текста и корректировать стратегии очистки на основе их важности. [9], [10]

Для систем, работающих в условиях ограниченных вычислительных ресурсов или в реальном времени, компромиссным и наиболее эффективным подходом является применение субсловной токенизации без дорогостоящей лемматизации, поскольку она обеспечивает баланс между скоростью работы и качеством классификации.

Ограничения исследования

Несмотря на полученные результаты, исследование имеет ряд ограничений, которые необходимо учитывать при интерпретации выводов.

Корпус Lenta.ru, используемый в качестве экспериментальной базы, обладает определённой тематической однородностью, что ограничивает переносимость результатов на другие типы текстов, такие как форумы, микроблоги или специализированные документы. Для формирования более универсальных рекомендаций требуется проведение аналогичных экспериментов на мультидоменных корпусах.

Дополнительным ограничением является использование фиксированного набора гиперпараметров для всех моделей. Полноценная оптимизация параметров под каждый сценарий предобработки могла бы существенно повлиять на итоговые значения метрик и изменить сравнение стратегий.

Ограничением также является способ анализа интерпретируемости: в работе применялись только LIME и SHAP, тогда как современные трансформные архитектуры требуют инструментов, способных учитывать структуру многоголового механизма внимания. Расширение набора инструментов могло бы дать более глубокое понимание влияния предобработки на внутренние представления модели.

Кроме того, некоторые методы обработки текста, включая сложные орфографические корректоры или контекстные нормализаторы, намеренно не включались в исследование из-за ограничений объёма работы. Однако именно эти методы способны оказать значительное влияние на качество классификации в реальных прикладных системах, что делает их включение важным направлением дальнейших экспериментов.

Перспективные направления развития исследований

Современные тенденции развития методов обработки естественного языка указывают на необходимость расширения исследований, связанных с адаптивной предобработкой текста. Одним из ключевых направлений является интеграция модулей интеллектуальной очистки, способных автоматически подстраиваться под характеристики корпуса, структуру предложений и архитектуру используемой модели. Такие модули могут использовать эвристики, статистический анализ, а также механизмы внимания для определения оптимального уровня фильтрации.

Кроме того, перспективным является использование нейросетевых токенизаторов нового поколения, которые способны динамически выявлять морфемные структуры слов, что особенно актуально для языков со сложной морфологией, таких как русский. Это позволит уменьшить зависимость от стандартных инструментов лемматизации и повысить точность моделей, работающих с контекстом.

Важным направлением является интеграция методов интерпретируемости, таких как LIME и SHAP, которые позволяют анализировать влияние конкретных токенов на итоговое решение модели. Это открывает возможность формировать предобработку, учитывающую влияние отдельных элементов текста на качество классификации.

Также следует учитывать развитие мультимодальных моделей, способных одновременно анализировать текст, изображения и метаданные. Для таких моделей требуется новая стратегия предобработки, ориентированная на согласование разных типов данных. В дальнейшем возможно создание единой универсальной системы предобработки, способной адаптироваться к многозадачным архитектурам.

Таким образом, дальнейшее развитие исследований в области предобработки текста будет направлено на автоматизацию, адаптивность и интеграцию новых методов интерпретируемости и мультимодального анализа.

В числе приоритетных направлений — разработка обучаемых (end-to-end) модулей предобработки, которые интегрируются в архитектуру модели и обучаются совместно с задачей классификации; такие модули могут динамически решать, какие токены сохранять, а какие — сглаживать или агрегировать.

Кроме того, актуальным является исследование адаптивных токенизаторов, которые используют статистику корпуса и сигналы обратной связи от интерпретируемости

модели для автоматической корректировки уровня очистки [8],[11].

Также перспективно исследование мультидоменных и мультиязычных пайплайнов предобработки, где на этапе обучения модель получает стимулы для различной обработки в зависимости от жанра текста и ожиданий по длине документа.

Для практического использования важна автоматизация выбора режима предобработки на основании простых метрик корпуса (средняя длина документа, доля не словарных токенов, частота опечаток).

Заключение

Проведённое исследование подтверждает, что предобработка текста является ключевым фактором, влияющим на качество тематической классификации. Для классических моделей оптимальной является стандартная очистка текста, включающая лемматизацию и фильтрацию стоп-слов. Для моделей на основе нейронных сетей целесообразно ограничиваться минимальной очисткой, тогда как для Transformer-моделей необходимо сохранять грамматические и контекстные элементы. Кроме того, следует подчеркнуть, что предобработка играет важную роль в обеспечении устойчивости моделей к распределительным сдвигам. Модели, обученные на данных с оптимальной очисткой, демонстрируют более стабильные результаты при применении к текстам, отличающимся по стилю, лексике или структуре. Это особенно важно в практических сценариях, где данные могут быть неоднородными.

Также важным выводом является то, что предобработка способствует улучшению интерпретируемости моделей. Благодаря снижению уровня шума и устранению неинформативных элементов текста повышается прозрачность принятия решений классификаторами, что особенно актуально для ответственных областей.

Таким образом, предобработка является неотъемлемой частью современного конвейера анализа текстов и требует дальнейших исследований в направлении автоматизации, адаптивности и интеграции методов объяснимого искусственного интеллекта.

Результаты работы позволяют сделать вывод о необходимости разработки адаптивных алгоритмов предобработки, учитывающих архитектуру модели и особенности корпуса. Дальнейшие исследования будут направлены на интеграцию методов объяснимого машинного обучения (LIME, SHAP) и адаптацию подходов для мультиязычных корпусов.

Литература

1. Vaswani, A. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar [и др.] // Advances in Neural Information Processing Systems. – 2017. – DOI: 10.48550/arXiv.1706.03762.

2. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // NAACL-HLT. – 2019. – DOI: 10.18653/v1/N19-1423.

3. Mikolov, T. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen, G. Corrado, J. Dean. – arXiv preprint, 2013. – DOI: 10.48550/arXiv.1301.3781.

4. Kuratov, Y. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language (RuBERT) / Y. Kuratov, M. Arkhipov. – arXiv preprint, 2019. – DOI: 10.48550/arXiv.1905.07213.

5. Стрелец, А. И. Методы классификации текстовых данных по темам / А. И. Стрелец, В. С. Иванников, А. А. Орлов, А. В. Атавина // Международный журнал гуманитарных и естественных наук. – 2019. – № 6(1). – С. 74–76. – DOI: 10.24411/2500-1000-2019-11252.

6. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. Enriching Word Vectors with Subword Information. arXiv preprint, 2017. (FastText).

7. Zharkov, D., & Korobov, M. pymorphy2: Open-source morphological analyzer for Russian and Ukrainian. (Описание инструмента pymorphy2).

8. Kudo, T., & Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. arXiv preprint, 2018.

9. Ribeiro, M. T., Singh, S., & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. (LIME)

10. Lundberg, S. M., & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 (NIPS 2017). (SHAP)

11. Sennrich, R., Haddow, B., & Birch, A. Neural Machine Translation of Rare Words with Subword Units. Proceedings of ACL, 2016. (BPE / subword methods)

Подзол Д.Ю., Коломойцева И.А. Исследование влияния предобработки текста на качество тематической классификации. *В работе исследуется влияние различных стратегий предобработки текста на качество тематической классификации русскоязычных документов. Сравниваются модели SVM, LSTM и RuBERT при использовании трех уровней очистки данных. Показано, что умеренная предобработка повышает точность классических и рекуррентных моделей, тогда как чрезмерная фильтрация снижает качество трансформерных архитектур. На основе результатов предложена адаптивная стратегия предобработки, учитывающая особенности каждой модели.*

Ключевые слова: *тематическая классификация, предобработка текста, машинное обучение, нейросетевые модели, RuBERT.*

Podzol D.Yu.. Kolomoitseva I.A. Research on the Impact of Text Preprocessing on the Quality of Topic Classification. *The study examines the impact of various text preprocessing strategies on the quality of topic classification for Russian-language documents. The SVM, LSTM, and RuBERT models are compared under three levels of data cleaning. The results show that moderate preprocessing improves the accuracy of classical and recurrent models, while excessive filtering reduces the performance of transformer-based architectures. Based on the findings, an adaptive preprocessing strategy tailored to the characteristics of each model is proposed.*
eywords: *topic classification, text preprocessing, machine learning, neural models, RuBERT.*

Статья поступила в редакцию 02.12.2025
Рекомендована к публикации профессором Федяевым О. И.