МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

Классификация текстов разных жанров с использованием ритмических признаков

АВТОРЕФЕРАТБАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 «Математическое обеспечение и администрирование
информационных систем»
факультета компьютерных наук и информационных технологий

Подкидышевой Дарьи Александровны

Научный руководитель:
Зав. кафедрой к.ф.-м.н, доцент _____ Огнева М. В.

Зав. кафедрой:
к.ф.-м.н, доцент _____ Огнева М. В.

Саратов 2026

ВВЕДЕНИЕ

Согласно данным аналитического блога TechBlog, каждую секунду в Google делается более 40 тысяч поисковых запросов. В интернете хранятся миллионы терабайтов данных, среди которых значительную часть занимают текстовые материалы самой разной природы – новостные статьи, художественная литература, научные тексты, песенная лирика, диалоги. Такой объём текстовых данных требует серьёзной систематизации и автоматической обработки, поэтому активно развивается область обработки естественного языка – NLP (Natural Language Processing). Это важное направление искусственного интеллекта, занимающееся разработкой методов и алгоритмов для анализа текстов на естественных языках.

NLP применяется повсеместно: поиск в интернете, анализ тональности отзывов, обнаружение плагиата, машинный перевод, автоматическое реферирование. Одной из фундаментальных задач NLP является классификация текстов – автоматическое отнесение текстового документа к одной из заранее определённых категорий. Классификация по жанрам представляет особый интерес, поскольку жанровая принадлежность текста определяет не только его тематику, но и способ изложения, синтаксическую организацию, пунктуационный профиль и стилистические особенности. Автоматическое распознавание жанра востребовано в системах поиска и рекомендаций, при организации электронных библиотек, в задачах анализа медиаконтента.

Традиционные подходы к жанровой классификации опираются на лексические признаки – частотные представления слов и семантические векторы. Вместе с тем ряд исследований показывает, что дополнение лексических методов структурными и ритмическими характеристиками текста способно улучшить качество классификации, особенно когда жанры различаются не только по тематике, но и по форме изложения. Настоящая

работа посвящена исследованию именно этого направления: изучению вклада ритмических признаков текста в задачу автоматической жанровой классификации и оценке их эффективности в сочетании с традиционными лексическими методами.

Цель бакалаврской работы – реализация и сравнительный анализ методов классификации текстов различных жанров с использованием ритмических признаков.

Поставленная цель определила **следующие задачи**:

1. Провести обзор существующих подходов к классификации жанров текстов и к использованию ритмических признаков в задаче классификации;

2. Рассмотреть методы предобработки текстов, подходы к их векторному представлению и алгоритмы машинного обучения, применяемые в задачах классификации;

3. Подготовить корпусы текстов различных жанров, осуществить разметку жанров;

4. Определить и формализовать набор ритмических признаков текста;

5. Реализовать модели классификации текстов на основе трёх типов лексического представления и оценить их точность на каждом корпусе;

6. Исследовать эффект от добавления ритмических признаков к лексическим представлениям и провести сравнительный анализ результатов;

7. Проанализировать важность отдельных ритмических признаков и сформулировать рекомендации по их применению в задачах жанровой классификации.

Методологические основы исследования в области классификации текстов и ритмических признаков представлены в работах К. Лагутиной, Н. Лагутиной, Е. Бойчук, а также T. Mikolov, C. Manning, L. Breiman, C. Cortes, V. Vapnik.

Практическая значимость бакалаврской работы. Полученные

результаты демонстрируют, что включение ритмических признаков целесообразно при построении систем жанровой классификации, особенно для жанров с выраженными структурными различиями. Разработанный набор из 14 ритмических признаков, и методика их применения могут быть использованы при создании систем автоматической обработки текстов, организации электронных библиотек и анализа медиаконтента.

Структура и объём работы. Бакалаврская работа состоит из введения, 6 разделов, заключения, списка использованных источников и 4 приложений. Общий объём работы – 86 страниц, из них 67 страниц – основное содержание, включая 18 рисунков и 24 таблицы, список использованных источников информации – 20 наименований.

КРАТКОЕСОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Обзор источников»

Посвящён анализу существующих подходов к жанровой классификации текстов. Рассматриваются базовые методы векторного представления: Bag of Words, TF-IDF и Word2Vec. Систематизируются исследования, в которых ритмические и стилистические признаки применялись для классификации текстов. Показано, что ритмические признаки формируют дополнительный источник информации, не зависящий от тематического содержания текста: классические алгоритмы на ритмических признаках достигают F-меры 94.8% (RandomForest) и точности 84–99% при верификации авторства (SVM). Установлено, что систематическое сравнение ритмических признаков с тремя базовыми типами лексического представления на нескольких корпусах остаётся малоизученным, что и определяет вклад настоящей работы.

Второй раздел «Теоретические основы классификации текстов»

Содержит описание методов, применяемых в работе. Формализована задача текстовой классификации. Описаны три метода векторного

представления текста: модель «мешка слов» (BoW), метод TF-IDF и Word2Vec. Определён набор из 14 ритмических признаков, сгруппированных по четырём аспектам: синтаксическая структура (средняя и стандартная длина предложения, средняя и стандартная длина слова), лексические характеристики (индекс разнообразия словаря, доля повторов), пунктуационный профиль (частоты запятых, восклицательных и вопросительных знаков, многоточий) и морфологический состав (доли прилагательных, наречий, глаголов и существительных).

Третий раздел «Методы машинного обучения для классификации текстов»

Посвящён теоретическому описанию шести классификационных алгоритмов: логистической регрессии, метода опорных векторов (SVM), случайного леса (Random Forest), наивного байесовского классификатора (GaussianNB / MultinomialNB), решающего дерева и метода k ближайших соседей (KNN). Для каждого алгоритма приведены математическое обоснование, принцип работы и особенности применения к задачам текстовой классификации. Обоснован выбор GaussianNB для непрерывных ритмических признаков и MultinomialNB для частотных лексических представлений.

Четвёртый раздел «Классификация на корпусе книжных текстов шести жанров»

Описывает первый эксперимент. Корпус сформирован из датасета Project Gutenberg (Kaggle) и включает шесть жанров: Culture/Civilization/Society, Poetry, Philosophy & Ethics, Travel & Geography, History – General, Nature/Gardening/Animals – около 1000 текстов на каждый жанр. В корпус включались только книги с однозначной жанровой принадлежностью на английском языке. Предобработка включала лемматизацию и удаление стоп-слов для лексических признаков;

ритмические признаки извлекались из оригинальных текстов с помощью библиотеки spaCy из первых 20 000 символов каждого текста.

На ритмических признаках лучший результат – SVC 51.3%, что существенно превышает случайное угадывание (16.7%), однако указывает на недостаточность ритмики для надёжного разграничения тематически близких книжных жанров. Поэзия классифицируется лучше всего (до 79% F-мера) благодаря выраженному структурному профилю. Travel & Geography практически неотличима от History и Culture на уровне ритмических параметров: все три жанра написаны в описательно-повествовательном регистре со схожей длиной предложений. Доминирующие признаки: std_word_length (0.1088), lexical_diversity (0.0907), avg_word_length (0.0836); наименее информативен ellipsis (0.0165), что логично для книжных текстов.

На лексических признаках TF-IDF обеспечил наилучшие результаты: SVC – 81.3%, LogisticRegression – 81.0%. Word2Vec занял промежуточное положение (SVC – 80.9%, LogisticRegression – 80.3%), BoW уступил из-за отсутствия нормировки по длине документа (лучший – LogisticRegression 76.4%; SVC просел до 66.3%). Труднее всего классифицируется жанр Culture/Civilization/Society, тематически пересекающийся сразу с несколькими соседними категориями.

При объединении ритмических признаков с лексическими картина неоднородна. Наибольший прирост показал KNN (+1.19% на TF-IDF): добавление ритмических измерений уплотняет жанровые кластеры в расширенном пространстве признаков. RandomForest и MultinomialNB незначительно ухудшаются, что объясняется устойчивостью ансамблевых методов к добавлению коррелированных признаков. Лучший абсолютный результат по корпусу: SVC + TF-IDF + ритмика – 81.86%.

Пятый раздел «Классификация на корпусе из пяти жанров»

Представляет второй эксперимент на смешанном корпусе: к трём

книжным жанрам (Poetry, Philosophy & Ethics, Nature/Gardening/Animals) добавлены новостные статьи CNN/DailyMail и энциклопедические статьи Википедии (по 1000 текстов каждый). Добавление структурно выраженных нелитературных жанров кардинально изменяет информативность ритмических признаков: RandomForest достигает 81.1% только на ритмике – против 51.3% на книжном корпусе. Энциклопедические и новостные тексты классифицируются ритмикой практически безошибочно (179–180/200 верно).

Ключевой новый результат – стабильный положительный эффект ритмики при комбинировании с Word2Vec: все шесть моделей без исключения демонстрируют неотрицательный прирост. Это объясняется ортогональностью семантической и структурной информации. Лучший результат на TF-IDF: RandomForest – 97.2%. Наиболее информативный признак на данном корпусе: std_sentence_length (0.1099).

Шестой раздел «Классификация на корпусе из трёх жанров»

описывает третий эксперимент на корпусе с максимальными структурными различиями между жанрами: кинодиалоги (Cornell Movie-Dialog Corpus, 603 документа), тексты песен (500K+ Spotify Songs, 1000) и народные сказки (Fairy Tales from Around the World, 1000). Жанры принципиально различаются по синтаксической организации: диалоги характеризуются короткими репликами и высокой долей вопросительных предложений, песни – строфической структурой и повторами, сказки – нарративными формульными конструкциями.

На ритмических признаках RandomForest достигает 98.5%, GaussianNB – 98.1%, DecisionTree – 97.3%. Разрыв с лучшим лексическим методом (SVC+TF-IDF, 98.7%) статистически несущественен, при этом RandomForest на ритмике обучается в 27 раз быстрее: 1.27 с против 33.91 с. DecisionTree на ритмических признаках (97.3%) превосходит тот же классификатор на TF-IDF (92.7%) и BoW (95.4%).

Доминирующие признаки: std_sentence_length (0.2432) и avg_sentence_length (0.2429) вместе дают почти половину суммарной важности. Вопросительные знаки (0.1612) впервые оказываются на третьем месте, ellipsis впервые входит в топ-5 (0.0840). Наибольший прирост от добавления ритмики к TF-IDF показал DecisionTree: +5.37% – один из крупнейших приростов по всем экспериментам.

ЗАКЛЮЧЕНИЕ

В настоящей работе проведено исследование применимости ритмических признаков текста в задаче автоматической классификации по жанрам. Эксперименты выполнялись на трёх корпусах различной природы с использованием шести классификационных алгоритмов и трёх типов лексического представления – BoW, TF-IDF и Word2Vec.

Основным результатом работы является установленная прямая зависимость информативности ритмических признаков от природы разграничиваемых жанров. Точность классификации исключительно на 14 ритмических признаках варьировала от 51.3% на корпусе шести книжных жанров до 98.5% на корпусе диалогов, песен и сказок при полностью идентичном наборе признаков и алгоритмов. Ритмические признаки фиксируют структурную организацию текста: жанры, различающиеся по форме, разделяются ритмикой значительно эффективнее, чем жанры, различающиеся преимущественно по тематике.

Добавление ритмических признаков к лексическим давало стабильный положительный эффект в большинстве случаев. Наиболее выраженный и воспроизводимый результат наблюдался для DecisionTree – прирост достигал 5.37% при добавлении ритмики к TF-IDF на корпусе диалогов, песен и сказок. Этот паттерн воспроизводился на всех трёх корпусах и при всех типах лексического представления.

На корпусе диалогов, песен и сказок ритмические признаки оказались сопоставимы с лексическими методами по точности: RandomForest на ритмике достиг 98.5% при времени обучения 1.27 с против 98.7% у SVC+TF-IDF при 33.91 с. При достаточной структурной различимости жанров ритмический подход обеспечивает сопоставимое качество при кратно меньших вычислительных затратах.

Набор наиболее информативных ритмических характеристик менялся в зависимости от корпуса: на книжных жанрах доминировали признаки длины и разнообразия слов, на смешанном корпусе – дисперсия длины предложений, на корпусе диалогов и песен – средняя и стандартная длина предложения совместно с долей вопросительных знаков.

Таким образом, поставленные задачи выполнены в полном объёме. Результаты подтверждают целесообразность включения ритмических признаков в системы жанровой классификации текстов в качестве дополнения к лексическим методам, особенно для жанров с выраженными структурными различиями. Направлениями дальнейшей работы могут служить расширение набора ритмических признаков за счёт фонетических характеристик, применение методов отбора признаков для оптимизации их состава под конкретный корпус, а также проверка подхода на многоязычных данных.

Основные источники информации

1. Лагутина К. В. Классификация текстов по жанрам на основе ритмических признаков / К. В. Лагутина, Н. С. Лагутина, Е. И. Бойчук // Моделирование и анализ информационных систем. – 2021. – Т. 28, № 3. – С. 280–291.

2. Лагутина К. В. Обзор моделей построения текстовых признаков для классификации текстов на естественном языке / К. В. Лагутина, Н. С. Лагутина // Труды 29-й конференции FRUCT. – 2021.

3. Lagutina K., Lagutina N., Boychuk E., Larionov V., Paramonov I. Authorship verification of literary texts with rhythm features // Proceedings of the 28th Conference FRUCT. – IEEE, 2021.

4. Mikolov T. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv. – 2013.

5. Manning C. D. Introduction to Information Retrieval / C. D. Manning, P. Raghavan, H. Schütze. – Cambridge University Press, 2008. – 544 p.

6. Breiman L. Random Forests / L. Breiman // Machine Learning. – 2001. – Vol. 45, № 1. – P. 5–32.

7. Cortes C. Support-Vector Networks / C. Cortes, V. Vapnik // Machine Learning. – 1995. – Vol. 20, № 3. – P. 273–297.

8. Pedregosa F. Scikit-learn: Machine Learning in Python / F. Pedregosa [et al.] // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.

9. Vanetik N. Genre Classification of Books in Russian with Stylometric Features / N. Vanetik [et al.] // Information. – 2024. – Т. 15, № 340.

10. Хобсон Л. Обработка естественного языка в действии / Л. Хобсон, Х. Ханнес, Х. Коул. – 3-е изд. – СПб. : Питер, 2020. – 576 с.