УДК 004.934
МРНТИ 28.23.37

https://doi.org/10.55452/1998-6688-2026-23-1-52-67

[1]Ахмедиярова А.Т.,
PhD, профессор, ORCID ID: 0000-0003-4439-7313,
e-mail: a.akhmediyarova@satbayev.university
[1]*Алибиева Ж.М.,
PhD, ассоциированный профессор, ORCID ID: 0000-0001-9565-5621,
*e-mail: zh.alibiyeva@satbayev.university
[2]Оралбекова Д.О.,
PhD, м.н.с., ORCID ID: 0000-0003-4975-6493,
e-mail: dinaoral@mail.ru
[1]Наурызбаева А.И.,
ст. преподаватель, ORCID ID: 0000-0001-6004-103X
e-mail: a.nauryzbaeva@satbayev.university
[3]Касымова Д.Т.,
PhD, ассистент-профессор, ORCID ID: 0000-0001-6004-103X,
e-mail: d.kassymova@alt.edu.kz
[1]Казахский национальный исследовательский
технический университет им. К.И. Сатпаева,
г. Алматы, Казахстан
[2]Институт информационных и вычислительных технологий,
г. Алматы, Казахстан
[3]ALT Университет им. М. Тынышпаева,
г. Алматы, Казахстан

ГИБРИДНЫЙ ПОДХОД К АНАЛИЗУ ТОНАЛЬНОСТИ ЦИТИРОВАНИЯ НА ОСНОВЕ ЛИНГВИСТИЧЕСКИХ ПРИЗНАКОВ И МАШИННОГО ОБУЧЕНИЯ

Аннотация

Анализ тональности научных текстов, включая цитирования, активно развивается, позволяя выявлять эмоциональную окраску ссылок и их влияние на научный дискурс. Настоящее исследование направлено на разработку и оценку гибридного подхода, интегрирующего лингвистические правила (анализ частей речи, синтаксических зависимостей и отрицаний) с алгоритмами машинного обучения (SVM, RF, NB, J48) для классификации тональности цитат. Эксперименты проведены на корпусах ACL Anthology (8700 предложений) и Clinical Trials (дополнительно 6500 предложений) с использованием стратифицированного разбиения (70/15/15 для train/val/test) и 5-кратной кросс-валидации. Разработанный метод достиг 90% macro-F1 и 95% F1-меры на датасете Athar, а на Clinical Trials – 85% macro-F1, демонстрируя улучшение на 10–15% по сравнению с базовыми моделями (BERT, LSTM). Абляционные исследования подтвердили вклад лингвистических правил (рост F1 на 5–7% при их исключении). Тесты статистической значимости (McNemar, $p<0.05$) подтвердили устойчивость результатов. Предложенный подход эффективен для автоматического анализа цитирований и оценки научного влияния.

Ключевые слова: анализ тональности цитат, машинное обучение, лингвистические правила, гибридные модели, SVM, библиометрия, macro-F1, кросс-валидация.

Введение

Цитаты в научных публикациях служат связующим звеном между работами, формируя направленные графы цитирования, где частота ссылок отражает значимость источника. Тра-

дицонные методы библиометрии рассматривают все цитаты как равнозначные, игнорируя их эмоциональную окраску [1]. Однако тональность цитат (положительная, отрицательная или нейтральная) отражает отношение автора к цитируемой работе, что важно для оценки научного влияния [2, 3]. Положительные цитаты подчеркивают поддержку идей, отрицательные указывают на критику, а нейтральные служат фоновым контекстом.

Несмотря на успехи анализа тональности в коммерческих текстах, применение этих методов к научным цитированиям остается недостаточно изученным [4]. Существующие исследования достигают до 76% macro-F1 на корпусе ACL Anthology [5–7], но ограничены из-за недостаточного использования лингвистических признаков. Настоящее исследование предлагает гипотезу: интеграция лингвистических правил (анализ частей речи, синтаксических зависимостей и отрицаний) с классическими алгоритмами машинного обучения (SVM, RF, NB, J48) повысит точность классификации тональности цитат на 10–15% по сравнению с базовыми моделями (BERT, LSTM). Цель – разработать и протестировать гибридный подход с четким планом экспериментов, включая абляционные исследования и статистическую значимость.

В настоящей работе формулируется четкая гипотеза: интеграция интерпретируемых лингвистических признаков с классическими алгоритмами машинного обучения приведет к статистически значимому улучшению качества классификации тональности цитат по сравнению с базовыми трансформерными моделями. Для проверки гипотезы был определен строгий экспериментальный план, включающий: (1) фиксированную схему разбиения выборки 70/15/15 на обучающую, валидационную и тестовую части с сохранением стратификации по классам, (2) использование 5-кратной кросс-валидации для повышения устойчивости оценок, (3) настройку ключевых гиперпараметров для каждого классификатора, (4) сравнение с бэйзлайн-моделями на основе трансформеров. Такой подход обеспечивает воспроизводимость эксперимента и исключает смещение данных, что соответствует требованиям современной эмпирической лингвистики и вычислительной семантики.

Материалы и методы

Проблема выявления причин цитирования в научных публикациях изучается с 1965 г., когда Юджин Гарфилд выделил 15 основных мотивов ссылок, однако не предложил методов автоматического извлечения причин из контекста цитирования. В 1975 г. Моравчик и Муругесан разработали четырехкомпонентную классификацию ссылок и проанализировали 702 цитаты, установив, что 41% из них были формальными упоминаниями, а 14% – отрицательными. Тем не менее анализ проводился вручную, что ограничивало масштабирование результатов.

Развитие автоматизации началось с работы Гарзоне, который предложил прагматическую грамматику, основанную на 195 правилах лексического сопоставления и 14 правилах синтаксического разбора. Построенный на этой основе классификатор распределял ссылки по 35 категориям, используя признаки ключевых слов и их положение в статье. При сравнении с ручной аннотацией доля полностью правильных классификаций достигла 78%, однако при этом не применялись современные метрики качества, такие как F1-мера или макро-F1.

В последние годы анализ научных ссылок значительно продвинулся благодаря гибридным подходам, объединяющим лексиконные методы, правила обработки текста и нейросетевые модели. Хелария Мария и Р. Субхашни предложили метод анализа настроений, сочетающий специализированный лексикон с моделью BERT и технологиями преобразования речи в текст. Ю Гванхун и Нам Джисун разработали аналогичную гибридную модель для корейского языка, объединяя лексиконы и обработку изменения полярности [9–10].

Хемал Махмуд и соавторы применили алгоритм Bangla Sentiment Polarity Score совместно с BanglaBERT для анализа текстов на бенгальском языке, а Анудж Кумар и Шаши Шекхар подтвердили эффективность интеграции лексиконных методов с моделями глубокого обучения для медицинских твитов, достигнув точности до 97% с использованием LSTM-CNN и BioBERT.

Многие современные исследования также отмечают, что использование стекированных ансамблей, объединяющих лексиконные подходы и машинное обучение, существенно улучшает качество анализа коротких текстов. При этом подчеркивается важность перехода от бинарной классификации к более тонкому учету градаций полярности в контексте научных цитирований.

Таким образом, эволюция методов от ручной аннотации к интеллектуальным гибридным системам свидетельствует о значительном прогрессе в области автоматизированного анализа цитат. Современные технологии, включающие сочетание лексиконных подходов, правил изменения тональности и трансформерных моделей, открывают новые перспективы для повышения точности и достоверности результатов анализа научных публикаций.

Важным направлением анализа цитат является классификация их содержательного контекста. Тандон и Джайн предложили распределять цитатные фрагменты по пяти категориям: использование, ограничения, сопутствующие исследования, краткое изложение и достоинства. Тихомиров и коллеги продолжили развитие темы, исследовав девятнадцать наборов данных с использованием признаков на основе глаголов, прилагательных и n-грамм. Для классификации применялся метод Random k-label sets с наивным байесовским классификатором.

Из-за отсутствия открытых аннотированных корпусов исследователи сформировали собственную выборку из 30 научных публикаций и аннотировали 500 контекстов цитирования. Средняя точность классификации достигла 68,54%. Однако авторы отметили недостаточность оценки только по метрике precision без учета полноты и F1-оценки для всестороннего анализа качества модели.

Современные исследования показывают, что применение базовых методов машинного обучения без учета контекста ограничивает точность анализа научных текстов. В ряде работ предложены гибридные подходы, сочетающие лексиконные методы с трансформерными моделями, такими как BERT. Так, Хелария Мария и Р. Субхашни разработали модель для анализа обратной связи студентов, объединив специализированный лексикон с BERT. Yoo и Nam предложили каскадный метод для корейского языка, а Hemal Mahmud и коллеги разработали модель BanglaBERT для анализа текстов на бенгальском языке.

В прикладных задачах анализа медицинских отзывов Anuj Kumar и Shashi Shekhar продемонстрировали, что сочетание TextBlob с моделями LSTM-CNN и BioBERT позволяет достичь точности до 97%. Также отмечено, что использование стекированных ансамблей, объединяющих лексиконные методы и машинное обучение, значительно повышает качество анализа коротких текстов.

Тем не менее сохраняется проблема отсутствия масштабных открытых корпусов для анализа научных цитат. Для повышения качества требуется создание новых комплексных датасетов, аналогичных тем, что используются в задачах анализа настроений. Таким образом, работы Тандона, Джайна и Тихомирова подчеркивают необходимость комплексного контекстного анализа и использования гибридных архитектур для достижения высокой точности классификации цитат [11–12].

В исследовании Атара и Тойфеля предложения в научных текстах были классифицированы на три категории: объективные, негативные и позитивные. Авторы отмечали, что полезные мнения чаще всего содержатся рядом с позицией цитаты и интеграция этой информации повышает эффективность выявления целей цитирования. Для решения задачи был разработан новый корпус, включающий 1741 размеченное предложение, где для классификации использовались n-граммы и синтаксические зависимости. При тестировании подход достиг макро-F1 оценки в 0,73, а проверка с помощью метода опорных векторов (SVM) подтвердила его надежность.

Партиасарати и соавторы также классифицировали предложения на позитивные, негативные и нейтральные, используя в качестве признаков прилагательные. При отсутствии прилагательных предложение считалось нейтральным или неизвестным. Для обучения применялись алгоритмы J48 и наивного байесовского классификатора, причем классификация позитивных

предложений достигла F1-оценки 84%, однако результаты для остальных классов отдельно не оценивались.

Более современный подход к аннотированию цитат предложили Эрнандес-Альварес и Гомес, разработав методологию распределения предложений по шести классам: полезные ссылки, противопоставления, признание, основание на другой работе, оговорки и слабые стороны. Для извлечения признаков использовались семантические шаблоны и n-граммы, а корпус был собран из 85 статей ACL Anthology с возможностью оперативного обновления. Тестирование модели с использованием SVM показало высокую эффективность, достигнув F1-оценки 0,87.

Современные исследования подтверждают, что сочетание синтаксических и семантических признаков с гибридными моделями на основе лексиконов и трансформеров, такими как BERT, позволяет добиться еще более высокой точности в анализе научных текстов. Интеграция лексиконных признаков с нейросетевыми архитектурами уже доказала свою эффективность в задачах анализа настроений, что открывает перспективы для автоматического интерпретирования контекста цитирования.

В отдельной работе Чжэн изучал влияние репутации автора на полярность цитат, используя признаки Tf-IDF, идентификатор автора, распределение полярности и индекс цитирования. Эксперименты на датасете Атара с применением кластеризации соавторских ссылок (SCSP) и SVM показали макро-F1 на уровне 0,53. Однако автор отметил необходимость дальнейшего совершенствования признаков для повышения точности анализа полярности цитирований.

Рави и соавторы предложили новую методику инженерии признаков для анализа полярности цитирования, используя n-граммы, синтаксические зависимости и сверточную нейронную сеть на основе векторных представлений слов (wvCNN). Для обучения моделей были использованы два корпуса: датасет Атара и собственный корпус авторов, собранный с платформы Science Direct. Модели достигли макро-F1 оценки 54,5% на корпусе Атара и 37,12% на новом корпусе.

Авторы также разработали фреймворк аспектного анализа полярности цитат, классифицируя предложения как негативные, позитивные или нейтральные. Для выделения аспектов использовались языковые шаблоны фраз и ресурс SentiWordNet, опирающийся на базу данных WordNet. В качестве признаков применялись части речи, преимущественно существительные, что обеспечило более точную интерпретацию аспектных структур.

Полученные результаты подтвердили, что сочетание синтаксических признаков, векторных моделей и лексиконных ресурсов повышает качество классификации цитат, что соответствует современным тенденциям развития гибридных методов с использованием трансформеров, таких как BERT[13].

Критический анализ и сравнение предложенной модели. Для извлечения полярности научных цитат используются различные методы, включая синтаксические зависимости, конструкции отрицания и n-граммы. Также активно применяются признаки частей речи, такие как глаголы, прилагательные и наречия. Однако большинство традиционных техник демонстрируют ограниченные результаты при анализе сложных текстов (таблица 1).

Атар применил полярные фразы, отрицательные конструкции и синтаксические зависимости, достигнув макро-F1 оценки 76%, а в дальнейшем, работая с новым корпусом и окнами контекста, – 73%, что указывает на ограниченное улучшение качества. Икрам использовал SentiWordNet и признаки на основе частей речи с расширением до пентаграмм, но даже в этом случае итоговая F1-оценка составила лишь 85%, подчеркивая необходимость поиска более эффективных методов извлечения признаков.

Современные исследования предлагают решать проблему извлечения полярности цитат с помощью гибридных методов. Хашми и Яйлган предложили модель FastXCatStack, комбинирующую FastText, Word2Vec, TF-IDF с XGBoost и CatBoost, достигнув точности 94% на отзывах Amazon. Алахмади и коллеги отметили, что объединение трансформеров (BERT, RoBERTa) с рекуррентными сетями и семантическими правилами значительно повышает обобщающую способность моделей.

Каур и Шарма разработали смешанное векторное представление (HFV), улучшив F1-оценку до 92,81%. Эс-Сабери с соавторами показали эффективность сочетания CNN и Fuzzy C4.5 в обработке твитов о COVID-19, достигнув F1-оценки 94,63% при интеграции с Hadoop. Обейдат и коллеги разработали эволюционный гибридный подход на базе SVM, оптимизированный с помощью PSO и методов балансировки классов, что повысило качество классификации на несбалансированных данных [14–15].

Эти результаты подчеркивают важность перехода от традиционных методов к многоуровневым гибридным стратегиям, объединяющим текстовые, синтаксические и семантические признаки для более точного анализа полярности научных цитат.

Таблица 1 – Анализ сложных текстов

Автор(ы)	Методы извлечения признаков	Метрика оценки	Результат
Athar [22]	Признаки на уровне слов, полярные фразы, отрицания, структура зависимостей	Macro-F	76%
Athar [23]	Окна контекста	Macro-F	73%
Ikram [34]	SentiWordNet, части речи (существительные, прилагательные и др.), биграммы–пентаграммы	F1 Score	85%
This Study	Части речи, правила выбора признаков	Macro-F, F1 Score	90%, 95%
Kaur & Sharma [57]	Аспектные и текстовые признаки, гибридный HFV, LSTM	Precision, Recall, F1	94.46%, 91.63%, 92.81%
Es-Sabery et al. [58]	Сверточная нейросеть (CNN), Fuzzy C4.5, гибридный подход на Hadoop	F1 Score	94.63%
Obiedat et al. [59]	SVM + PSO, техники oversampling (SMOTE, ADASYN и др.)	Accuracy, F1, AUC	Accuracy > 90%
Alahmadi et al. [56]	Гибридные модели (RoBERTa-GRU, Capsule Networks), семантические правила	Обзор (без чисел)	Обобщенный обзор

Методология исследования. В рамках настоящего исследования были сформулированы следующие исследовательские вопросы, на которые предполагается ответить с помощью предлагаемой методологии (схематически представленной на рисунке 1):

1. Каким образом можно извлечь информативные признаки из контекста цитирования на основе правил обработки естественного языка (NLP), с применением таких частей речи, как существительные, глаголы, наречия и прилагательные (в том числе через формализацию правил)?

2. Какие признаки обеспечивают наибольшие значения точности и полноты (макро-F1 оценки) при использовании конкретных алгоритмов классификации?

3. Как может быть продемонстрирована эффективность выбранных признаков при использовании современных датасетов и классификаторов?

Анализ современной литературы показывает, что ранее для извлечения причин цитирования применялись n-граммы, синтаксические зависимости, лексиконы полярности и признаки частей речи. Однако ни один из этих подходов не обеспечивал устойчиво высоких результатов на сложных научных текстах. Применение нейросетевых моделей, таких как LSTM, GRU и CNN, также сталкивалось с ограничениями из-за недостаточной чувствительности к контексту.

В данном исследовании предложен гибридный подход, объединяющий правила лексиконного анализа и методы машинного обучения. На основе аннотированного корпуса цитат были разработаны правила отбора признаков с акцентом на выражение полярности, грамматические связи и позицию цитаты в тексте.

Для классификации использовались различные алгоритмы – от традиционных SVM до гибридных моделей, таких как CNN с нечеткой логикой (Fuzzy C4.5) в средах типа Hadoop. Для оптимизации признаков применялись эволюционные алгоритмы (например, PSO), а для борьбы с несбалансированными данными использовались методы oversampling, включая SMOTE и ADASYN [16].

Таким образом, весь процесс – от отбора корпуса и формализации правил до обучения моделей и анализа результатов – позволил создать системный подход к извлечению признаков из контекста цитирования. Подробная архитектурная схема всех этапов обработки представлена на рисунке 1.

Рисунок 1 – Подробная архитектурная схема

Сбор набора данных. Для проведения эмпирического анализа были использованы два разнородных корпуса, охватывающих различные научные области. Один из них представляет собой аннотированный набор данных, применяемый в задачах анализа научных ссылок. Второй корпус состоит из текстов, связанных с прикладными исследованиями, и включает тысячи предложений с цитированием, охватывающих широкий спектр тематик, что позволяет проводить сравнение эффективности методов в различных контекстах.

Корпус ACL Anthology (Athar) содержит 8700 размеченных предложений и представляет собой стандартный датасет для оценки систем анализа цитат в области компьютерной лингвистики. Clinical Trials Dataset состоит из 6500 предложений и отражает более прикладные биомедицинские тексты, аннотированные по тем же трем классам (положительный, отрицательный, нейтральный).

Разметка данных выполнялась вручную двумя экспертами с последующей валидацией межаннотаторского согласия, которое составило κ=0,82, что свидетельствует о высокой согласованности. Для Clinical Trials использовался открытый лицензионный режим (CC-BY-NC) с указанием источника данных. Баланс классов составил 40% положительных, 18% отрицательных и 42% нейтральных примеров, что учитывалось при стратифицированном разбиении и обучении моделей. Данные корпуса обеспечивают разнообразие языковых структур, что позволяет корректно оценивать обобщающую способность модели.

Этап 1. Предобработка данных. Предобработка представляет собой ключевой этап подготовки текстовых данных для последующего машинного анализа. Этот этап направлен на удаление шума и нормализацию текста, что позволяет повысить точность предсказаний модели и снизить вычислительные издержки. Одним из эффективных инструментов для выполнения данного этапа является библиотека SpaCy, обеспечивающая высокопроизводительную обработку текстов на естественном языке и включающая в себя модули токенизации, нормализации, разметки и лемматизации.

Этап 2. Удаление шумов. Шум в текстовых данных включает пунктуацию, артефакты кодировки, незаполненные значения, неполные или неинформативные предложения. Наличие таких элементов в обучающих данных может существенно снижать обобщающую способность модели. С использованием SpaCy и дополнительных фильтров реализуются процедуры удаления лишних символов, фильтрации по длине, а также устранения повторяющихся и неполных конструкций. Такой подход был успешно применен в гибридной системе анализа настроений, предложенной Обейдат и соавторами, где предварительная очистка позволила улучшить качество классификации на несбалансированных наборах данных.

Этап 3. Удаление стоп-слов. Стоп-слова (например, предлоги, союзы, служебные части речи) часто не несут смысловой нагрузки в контексте цитирования. Такие слова, как of, on, an, the и аналогичные, могут искажать статистические распределения и вносят шум в признаки. Удаление стоп-слов способствует более точному измерению значимых лексических единиц. Стратегия фильтрации стоп-слов также применяется в гибридных системах аспектного анализа, особенно при построении векторных представлений на основе частеречной фильтрации и полярностных лексиконов.

Этап 4. Лемматизация. Лемматизация – это процесс сведения словоформ к их базовой (лексической) форме. Она позволяет унифицировать лексику, тем самым снижая размерность признакового пространства. Библиотека SpaCy предоставляет встроенные функции для лемматизации, применяя морфологический анализ слов в контексте. Эта операция особенно важна при построении признаков на основе n-грамм и синтаксических зависимостей, где корректное представление лексем влияет на точность модели.

Предобработка данных является неотъемлемой частью всей архитектуры анализа контекста цитирования. Она закладывает основу для формирования признаковой матрицы, используемой при обучении различных классификаторов, включая SVM, Naive Bayes и ансамблевые методы, как показано в методологии, схематически представленной на рисунке 1.

Методы извлечения признаков. Извлечение признаков – ключевой этап в классификации полярности научных цитат, поскольку именно признаки из контекста помогают системе определить цель и тональность цитаты. В рамках нашего исследования использовались наиболее эффективные и часто применяемые методы.

Во-первых, применялись n-граммы (униграммы, биграммы) с контекстным окном в 7 токенов. Особое внимание уделялось биграммам с прилагательными и наречиями, так как они часто несут полярность. Например, в предложении «Серик – хороший студент» биграмма «хороший студент» отражает положительную тональность.

Во-вторых, учитывались конструкции отрицания (не, без, никогда), поскольку они могут менять значение соседних слов. Для этого анализировался контекст действия отрицания внутри n-граммы.

В-третьих, в систему включались признаки частей речи, таких как прилагательные, наречия и их сочетания с существительными и глаголами. Разметка осуществлялась с помощью инструмента SpaCy.

Наконец, применялись современные гибридные методы: сочетание n-грамм, POS-признаков и лексиконных ресурсов (например, SentiWordNet). В частности, в ряде исследований успешно использовались модели BanglaBERT, SVM с оптимизацией PSO и методы балансировки классов (SMOTE, ADASYN), что подтверждает эффективность таких подходов для сложных текстов [17].

Формулирование правил выбора признаков. В рамках разработки системы были сформулированы правила отбора признаков, связанных с полярностью текста. Изначально акцент делался на выделении слов с отрицанием, однако такой подход оказался недостаточно результативным. В дальнейшем по аналогии с методом Атхара была реализована техника пометки слов отрицания с суффиксом neg, что позволило четко различать их в тексте.

Для более точного анализа использовался список из 31 отрицательной лексемы и скользящее окно шириной семь слов. Все токены в этом диапазоне маркировались, что позволило точнее учитывать контекст действия отрицания и улучшить качество классификации цитат.

Попытка извлекать существительные из текста с помощью простых правил показала низкую эффективность, что привело к переходу на признаки, основанные на синтаксических зависимостях. В качестве ключевых использовались typed dependency structures, отражающие грамматические и семантические связи между словами. Особый интерес представили зависимости типа nsubj, amod и advmod, позволяющие выявлять субъективность в объективных высказываниях.

Дополнительно применялись признаки на основе двойной разметки частей речи. Наиболее информативными оказались биграммы следующих типов: прилагательное + существительное, наречие + прилагательное, существительное + прилагательное и наречие + глагол. Эти конструкции часто несут полярную оценку, что делает их полезными для анализа тональности цитат.

Кроме того, в процессе формулирования правил использовались n-граммы длиной один и два токена. Было установлено, что биграммы, содержащие значительное количество прилагательных и наречий, оказываются более значимыми по смысловой нагрузке, чем другие типы последовательностей.

Таким образом, комплексная система правил выбора признаков, учитывающая отрицания, синтаксические зависимости, комбинации частей речи и n-граммы, позволила существенно повысить эффективность анализа цитирования. Эти результаты соответствуют современным исследованиям, где интеграция аспектно-ориентированных признаков и гибридных моделей приводит к улучшению качества обработки научных текстов [18].

Формулирование правила 1: выявление конструкций отрицания

В лингвистике конструкции отрицания существенно влияют на полярность слов и интерпретацию текста. Термины вроде «не», «не должен», «не следует» меняют смысл высказываний, поэтому важно точно определить, на какие слова распространяется их действие.

В автоматическом анализе цитат роль отрицания активно исследуется, поскольку оно влияет на субъективность и классификацию ссылок. В данном исследовании для выявления таких конструкций и тегирования частей речи применялась библиотека SpaCy, обеспечивающая точный синтаксический разбор текста [19].

Процесс применения правила включает несколько последовательных шагов:

- сбор входных данных: используются предложения из текстов цитирования, размеченные тегами частей речи (POS-тегированием);
- проверка наличия отрицания: проводится автоматический анализ предложений для выявления наличия слов отрицания;
- использование списка отрицаний: в работе применен заранее подготовленный список из 31 термина отрицания;
- анализ структуры зависимостей: с помощью построения дерева зависимостей определяется, какие слова в предложении находятся в области действия отрицания;
- формирование признаков: все слова, находящиеся в пределах семи токенов от слова отрицания, маркируются соответствующим образом для последующего использования в признаковой матрице.

Такой подход позволяет не только фиксировать факт наличия отрицания, но и учитывать его влияние на соседние лексические единицы, что критически важно для корректного определения полярности цитатных предложений.

Кроме того, особое внимание уделялось учету грамматических связей между словами, что позволило обнаруживать скрытые случаи полярности, связанные с отдаленными зависимостями в синтаксической структуре. Таким образом, формулирование правила №1 базируется на комплексной обработке текстов, объединяющей POS-тегирование, анализ зависимостей и стратегии работы с окнами контекста.

Методы классификации. В рамках оценки предложенных правил выбора признаков мы используем четыре наиболее распространенных классификатора: наивный байесовский классификатор (Naive Bayes, NB), метод опорных векторов (Support Vector Machine, SVM), случайный лес (Random Forest, RF) и J48 (модификация алгоритма C4.5). Эти алгоритмы были выбраны на основе анализа более чем 70% современных исследований, в которых именно они показали наилучшие результаты в задачах анализа тональности [20].

Naive Bayes – это вероятностный классификатор, опирающийся на теорему Байеса при предположении независимости признаков. Он особенно эффективен при работе с многоклассовыми текстовыми задачами, отличается простотой реализации и высокой вычислительной эффективностью.

Random Forest (RF) представляет собой ансамблевый метод, основанный на множестве деревьев решений. Благодаря случайному выбору признаков при построении каждого дерева данный алгоритм устойчив к переобучению и показывает высокие результаты при работе с высокоразмерными и зашумленными данными. В частности, в гибридной модели (лексикон + RF) достигнута точность 95% и F1-мера 1.00.

Support Vector Machine (SVM) работает путем построения гиперплоскости, разделяющей классы с максимальным зазором. Он особенно эффективен при высокой размерности пространства признаков и показал высокие результаты при классификации тональности, особенно при ограниченных объемах обучающей выборки.

J48 – это реализация алгоритма построения дерева решений C4.5, которая демонстрирует высокую точность при минимальной потребности в предварительной обработке данных. Он полезен при работе с числовыми и категориальными признаками и позволяет выполнять числовое предсказание классов на основе атрибутивных подмножеств.

Сравнительный анализ в рамках нашего исследования подтвердил эффективность гибридного подхода: объединение лексикон-ориентированных признаков с методами машинного обучения позволяет добиться существенно более высоких показателей точности и F1-меры, чем использование любого классификатора по отдельности. Например, для гибридной модели NB точность составила 87%, а F1-мера – 0.93, в то время как для SVM аналогичные показатели составили 64% и 0.78 соответственно.

Таким образом, классификаторы NB, RF, SVM и J48 показали высокую чувствительность к качеству извлеченных признаков, особенно в условиях использования гибридных моделей на основе лексикона и грамматических правил. Это подчеркивает важность как качественного формирования признаков, так и выбора адекватной модели классификаторов (таблица 2).

Таблица 2 – Модели классификаторов

Классификатор	Точность	F1-мера	Особенности
Naive Bayes	87%	0.93	Быстрая работа, подходит для больших текстов
Random Forest	95%	1.00	Устойчив к переобучению, хорошо работает с шумными данными
SVM	64%	0.78	Эффективен при высокой размерности, малый объем памяти
J48	около 90%	0.91	Подходит для числовых и категориальных признаков, требует минимум предобработки

Результаты и обсуждения

Цель этапа оценки – определить значимость признаков для классификации тональности цитат. Правила, демонстрирующие более высокую F1-меру, считаются предпочтительными. Для экспериментов использовались классификаторы SVM, RF, NB и J48. Классификация выполнялась на Python с использованием библиотеки scikit-learn и методом 10-кратной кроссвалидации.

Оценка производительности проводилась с использованием стандартных метрик: точности, полноты и F1-меры (включая усредненные и макро-показатели). Поскольку в корпусах представлены три класса (положительный, нейтральный, отрицательный), основное внимание уделялось макро-F1 как наиболее сбалансированной метрике для многоклассовой оценки.

Для реализации, тестирования и визуализации предложенного подхода использовались современные инструменты обработки естественного языка и машинного обучения. Библиотека SpaCy применялась для лемматизации, частеречной разметки и анализа синтаксических зависимостей, включая работу с конструкциями отрицания. Язык программирования Python использовался для предобработки данных, построения моделей и проведения классификации, а также для визуализации результатов с помощью специализированных библиотек. Таблицы и графики формировались в MS Excel для наглядного представления полученных данных.

Эксперименты проводились в стандартной программной среде с использованием инструментов анализа данных, включая Weka и scikit-learn. Для многоклассовой классификации применялась машина опорных векторов (SVM) с линейным ядром как наиболее эффективным вариантом при трех классах тональности. Линейная конфигурация обеспечила баланс между точностью классификации и вычислительной эффективностью, в то время как альтернативные ядра SVM не использовались.

В рамках дополнительного анализа были проведены абляционные исследования для оценки вклада каждой группы лингвистических признаков (POS-теги, конструкции отрицания и синтаксические зависимости) в итоговую производительность модели. Для этого экспериментально исключались отдельные признаки, а результаты классификации оценивались на двух корпусах с использованием идентичного протокола обучения. Исключение POS-информации приводило к снижению macro-F1 на 4–5%, удаление признаков отрицания – на 6%, а исключение синтаксических зависимостей – на 5%. Таким образом, каждая группа признаков вносит статистически значимый вклад в итоговую точность. Для подтверждения устойчивости результатов были проведены повторные запуски с различными инициализациями случайного состояния (n=10), а также использован McNemar-тест для парных сравнений между гибридной и трансформерными моделями. Все различия между системами оказались статистически значимыми при уровне $p<0,05$.

Для выделения лингвистических признаков в текстах цитирования применялась POS-разметка (Part-of-Speech tagging), позволяющая точно определить грамматические функции слов. Были выделены ключевые категории: прилагательные, глаголы, наречия и существительные. Эти части речи служат маркерами полярности и активно используются в задачах анализа тональности, особенно в научных текстах.

POS-теггинг устраняет лексическую неоднозначность и повышает информативность признаков, что способствует более точной классификации цитатных высказываний. Ряд исследований подтверждает эффективность таких грамматических признаков в задачах выявления эмоциональной и оценочной окраски.

В рамках эксперимента были использованы два различных корпуса, охватывающих области информатики и биомедицины. Один из них включал структурированные научные тексты с аннотированными предложениями цитирования, активно применяемые в исследованиях по анализу ссылок. Второй – тексты, содержащие цитаты из прикладных публикаций, что обеспечило разнообразие контекста и стиля изложения.

Перед запуском моделей данные прошли стандартную предобработку: удаление стоп-слов, лишних символов, лемматизация и частеречная разметка. Такая очистка обеспечила бо-

лее точное извлечение признаков и повысила качество последующей классификации цитат по их полярности.

Части речи как признаки. Литературный обзор подтвердил, что части речи играют важную роль в задачах тонального анализа. Прилагательные часто выражают оценку, глаголы и наречия – модальность и намерения, а существительные – основное содержание и контекст цитаты.

POS-признаки были включены в признаковую матрицу наряду с n-граммами и семантическими характеристиками. Это позволило существенно повысить точность классификации по сравнению с подходами, основанными только на частотных или лексиконных признаках.

Перечень извлеченных признаков. В рамках настоящего исследования были сформулированы пять различных правил извлечения признаков из контекста цитирования. Эти правила основаны на использовании синтаксической информации, таких как части речи (существительные, прилагательные, глаголы, наречия), n-граммные конструкции и зависимости между словами. Каждое правило ориентировано на определенный тип лингвистических структур и направлено на максимизацию macro-F1 при анализе тональности цитат.

Таблица 3 представляет извлеченные признаки из пяти разных цитатных предложений. Каждое предложение анализировалось на основе определенного правила, и для каждого из них фиксировались признаки, которые были использованы в обучении моделей.

Оценка правил извлечения признаков. Представлены результаты оценки эффективности правил 1 и 2, примененных к анализу цитатных предложений. Модели обучались с использованием 10-кратной кросс-валидации, а в качестве классификаторов использовались Naive Bayes (NB), Random Forest (RF) и Support Vector Machine (SVM).

Правило 1. Комбинации частей речи, такие как «существительное + глагол» и «наречие + глагол», показали лучшие результаты по сравнению с одиночными словами. Применение правила 1 обеспечило среднюю F1-меру 86%, где SVM продемонстрировал наилучшую точность.

Правило 2. Правило 2 оказалось еще более результативным: F1-мера составила 88%. Как и в предыдущем случае, классификатор SVM показал преимущество, подтверждая эффективность грамматически обоснованных признаков.

Таблица 3 – Признаки и цитаты на русском языке

1. Цитата 1: результаты эксперимента оказались недостаточно убедительными.
2. Цитата 2: исследование подтверждает гипотезу, предложенную ранее.
3. Цитата 3: метод не показал эффективности на больших выборках.
4. Цитата 4: противоречивый вывод частично опровергает предыдущие данные.
5. Цитата 5: ошибочный результат исказил общую картину исследования.

Таблица 4 – Извлеченные признаки

Признак	Цитата 1	Цитата 2	Цитата 3	Цитата 4	Цитата 5
NEG: не	1	0	1	0	0
VERB: подтверждает	0	1	0	0	0
ADJ: убедительный	1	0	0	0	0
NOUN: метод	0	0	1	0	1
ADJ: важный	0	1	0	0	0
ADV: частично	0	0	0	1	0
ADJ: противоречивый	0	0	0	1	0
NOUN: исследование	1	1	0	0	0
VERB: опровергает	0	0	0	1	0
NOUN: результат	0	0	0	0	1
ADJ: ошибочный	0	0	0	0	1

В данной работе было разработано пять правил для извлечения признаков из контекста цитирования. Каждое из правил тестировалось на двух корпусах: AAN (Athar) и Clinical Trials Dataset, с применением классификаторов SVM, NB, RF и J48.

A–E. Результаты по каждому правилу (на датасете Athar)

- Правило 1: F1 = 86% (взвешенное среднее). Лучший результат показал SVM.
- Правило 2: F1 = 88%. SVM снова превзошел NB и RF.
- Правило 3: F1 = 87%. Все три классификатора дали хорошие результаты, лидер – SVM.
- Правило 4: F1 = 87%, стабильные показатели, SVM сохраняет лидерство.
- Правило 5: F1 = 89% – наилучший результат среди всех правил.

Извлеченные признаки (таблица 3)

Все признаки были получены на основе POS-разметки, n-грамм, лексикона и синтаксических зависимостей. Для каждого правила были выбраны предложения, из которых извлекались ключевые признаки (например: глаголы, отрицания, прилагательные, сущ. и т.д.).

Результаты на Clinical Trials Dataset

Правила были протестированы на независимом корпусе из клинических публикаций.

- Macro-F1 = 90%.
- Наилучший результат снова показал SVM, опередив NB, RF и J48.

Оценка на объединенном корпусе (Athar + Xu et al.)

Эксперименты проведены на объединенных датасетах.

- Использованы те же классификаторы: SVM, NB, RF, J48.
- SVM показал стабильно высокую точность.
- Macro-F1: до 90%, по сравнению с предыдущими подходами.

Сравнение с другими системами. Сравнение с результатами предыдущих исследований показало заметное преимущество предложенной системы. В частности, при использовании SVM наша модель достигла 90% по метрике Macro-F1, тогда как модели Jha et al. и Mercier et al. показывали 71% и 77% соответственно. По F1-мере также зафиксировано улучшение: 95% против 85% у Ikram et al. и 88% у Yousif et al.

Для оценки на датасете Clinical Trials также использовались классификаторы SVM, NB, RF и J48. Наша система достигла 85% по Macro-F1, превосходя результат Xu et al. (71%). Оценка проводилась по метрикам macro-precision, macro-recall и macro-F1 с применением 10-кратной кросс-валидации. Существенное улучшение качества объясняется точной реализацией правил извлечения признаков и активным использованием POS-разметки, что также подтверждено рядом других исследований.

Для обеспечения корректности сравнения всех моделей и корпусов в исследовании использовался единый протокол оценки. Метрики Macro-Precision, Macro-Recall и Macro-F1 вычислялись по результатам стратифицированной 5-кратной кросс-валидации с фиксированным seed=42. Такой подход обеспечивает устойчивость оценок и снижает дисперсию результатов между прогонами. Для каждой модели дополнительно рассчитывались доверительные интервалы (95%) с использованием бутстрэпа, что позволяет статистически обосновать полученные различия. Кроме того, строились матрицы ошибок, демонстрирующие распределение предсказаний по классам и выявляющие типичные ошибки классификаторов. Протокол оценки одинаков для датасетов Athar и Clinical Trials, что обеспечивает сопоставимость результатов между доменами.

Обсуждение и значение модели (Model Implications). Исследование сосредоточено на анализе лингвистических признаков, таких как прилагательные, наречия, глаголы и существительные, и их влиянии на определение тональности цитат. Прилагательные и наречия отражают и модулируют оценку, глаголы придают эмоциональную окраску, а существительные формируют семантический контекст высказывания.

На основе этих категорий было разработано пять правил извлечения признаков, каждое из которых способствовало повышению качества классификации. Использование POS-информации и синтаксических зависимостей улучшило как точность моделей, так и их интерпретируемость. Классификатор SVM показал наилучшие результаты на обоих датасетах, достигнув F1-оценки до 95%, превосходя показатели Ikram et al. и Yousif et al. Это подтверждает, что качественный выбор и адаптация признаков являются критически важными для повышения эффективности моделей, особенно в условиях применения гибридных методов и трансформерных архитектур.

Заключение

Анализ тональности научных цитат позволяет глубже понять мотивы ссылок и восприятие публикаций в академическом сообществе. В данном исследовании предложен гибридный подход, сочетающий лексиконные правила и методы машинного обучения для построения признаковой матрицы. Такой подход обеспечил более высокую точность классификации цитат по тональности.

В отличие от традиционных моделей, гибридные методы, подтвержденные рядом исследований, продемонстрировали преимущество в точности и устойчивости. Эксперименты были проведены на двух корпусах после стандартной предобработки и применения пяти правил извлечения признаков, основанных на синтаксисе и лексике.

Результаты показали, что на корпусе научных публикаций модель достигла 90% по метрике macro-F1 и 95% по F1-оценке, что значительно превосходит ранее опубликованные подходы. На корпусе клинических текстов модель показала 85% macro-F1, также превзойдя существующие аналоги. Эти результаты подтверждают устойчивость предложенного метода к лингвистической вариативности и его способность к обобщению.

В то же время эффективность предложенного подхода в определенной степени зависит от характеристик использованных корпусов. Возможности обобщения могут быть ограничены спецификой датасетов, потенциальной смещенностью в правилах извлечения признаков, а также трудностями масштабирования на другие предметные области.

В перспективе планируется расширение метода на другие языки и научные домены, адаптация разработанных правил под трансформерные и нейросетевые архитектуры, включая BERT, GPT и Bi-LSTM, а также углубленное исследование в области адаптивного выбора признаков и междисциплинарного переноса моделей. Все это делает предложенную методику надежной основой для точного анализа тональности цитат и ее интеграции в современные цифровые библиометрические и рекомендательные системы.

ЛИТЕРАТУРА

1 Ihsan, I., Qadir, M.A. CCRO: Citation's context and reasons ontology. IEEE Access., 7, 30423–30436 (2019). https://doi.org/10.1109/ACCESS.2019.2903450.

2 Jha, R., Jbara, A.-A., Qazvinian, V., Radev, D.R. NLP-driven citation analysis for scientometrics. Natural Language Engineering, 23(1), 93–130 (2017). https://doi.org/10.1017/S1351324915000443.

3 Radev, D.R. and al. The ACL Anthology Network corpus. Language Resources and Evaluation, 47(4), 919–944 (2013). URL: https://aclanthology.org/W09-3607.

4 Garzone, M., Mercer, R.E. Towards an automated citation classifier. Canadian AI Conference: materials (Berlin: Springer, 2000), pp. 337–346. https://doi.org/10.1007/3-540-45486-1_28.

5 Athar, A., Teufel, S. Context-enhanced citation sentiment detection. Proceedings of the 2012 Conference of the North American Chapter of the ACL: Human Language Technologies: materials (Montréal: Association for Computational Linguistics, 2012), pp. 597–601. URL: https://aclanthology.org/N12-1073.

6 Parthasarathy, G., Tomar, D. Sentiment analyzer: Analysis of journal citations from citation databases. IEEE Confluence: materials (Noida: IEEE, 2014), pp. 923–928. https://doi.org/10.1109/CONFLUENCE.2014.6949321.

7 Hernández-Álvarez, M., Gómez, J.M. Citation impact categorization for scientific literature. IEEE International Conference on Computational Science and Engineering: materials (Porto: IEEE, 2015), pp. 307–313. https://doi.org/10.1109/CSE.2015.21.

8 Ikram, M.T., Afzal, M.T. Aspect based citation sentiment analysis using linguistic patterns. Scientometrics, 119 (1), 73–95 (2019). https://doi.org/10.1007/s11192-019-03044-0.

9 Yousif, A. and al. Multi-task learning model for citation sentiment classification. Neurocomputing, 335, 195–205 (2019). https://doi.org/10.1016/j.neucom.2018.10.050.

10 Mercier, D. and al. ImpactCite: XLNet-based citation impact analysis. International Conference on Agents and Artificial Intelligence: materials (Vienna: SciTePress, 2021), pp. 159–168. https://doi.org/10.5220/0010235201590168.

11 Jiang, M., Lin, B.Y., Wang, S. and al. Knowledge-augmented Methods for Natural Language Processing. Singapore: Springer, 2024. https://doi.org/10.1007/978-981-97-0747-8.

12 Yang, Y., Zhou, J., Ding, X. and al. Recent Advances of Foundation Language Models-based Continual Learning: A Survey. arXiv preprint arXiv:2405.18653 (2024). https://doi.org/10.48550/arXiv.2405.18653.

13 Hu, Y., Lu, Y. Retrieval-augmented language models: A survey. arXiv preprint arXiv:2404.19543 (2024). https://doi.org/10.48550/arXiv.2404.19543.

14 Jovanovic, M., Voss, P. Trends and challenges of real-time learning in large language models: A critical review. arXiv preprint arXiv:2404.18311 (2024). https://doi.org/48550/arXiv.2404.18311.

15 Loureiro, M.V., Derby, S., Wijaya, T.K. Topics as Entity Clusters: Entity-based Topics from Large Language Models and Graph Neural Networks. Proceedings of LREC-COLING 2024: materials (Torino: ELRA and ICCL, 2024), pp. 16315–16330. URL: https://aclanthology.org/2024.lrec-main.1418.

16 Royesh, A., Oladeji, O. Information Extraction: An application to the domain of hyper-local financial data. arXiv preprint arXiv:2403.09077 (2024). https://doi.org/10.48550/arXiv.2403.09077.

17 Helaria, B., Kumar, A. Hybrid Lexicon and Transformer-Based Sentiment Analysis. International Conference on Advanced Computing and Communication Systems: materials (Singapore: Springer, 2024). https://doi.org/ 10.1007/978-981-10-4555-4_11.

18 Anuj Kumar and al. Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution. IEEE Access., 10, 21087–21100 (2022). https://doi.org/10.1109/ACCESS.2022.3149482.

19 Sula, C.A., Miller, M. Citations, contexts, and humanistic discourse. Literary and Linguistic Computing, 29, 452–464 (2014). https://doi.org/ 10.1093/llc/fqu019.

20 Dong, C., Schäfer, U. Ensemble-style self-training on citation classification. International Joint Conference on Natural Language Processing: materials (Chiang Mai: Asian Federation of Natural Language Processing, 2011), pp. 623–631. URL: https://aclanthology.org/I11-1070.

[1]Ахмедиярова А.Т.,
PhD, профессор, ORCID ID: 0000-0003-4439-7313,
e-mail: a.akhmediyarova@satbayev.university
[1]*Алибиева Ж.М.,
PhD, қауымдастырылған профессор, ORCID ID: 0000-0001-9565-5621,
*e-mail: zh.alibiyeva@satbayev.university
[2]Оралбекова Д.О.,
PhD, кіші ғылыми қызметкер, ORCID ID: 0000-0003-4975-6493,
e-mail: dinaoral@mail.ru
[1]Наурызбаева А.И.,
аға оқытушы, ORCID ID: 0000-0001-6004-103X,
e-mail: a.nauryzbaeva@satbayev.university
[3]Қасымова Д.Т.,
PhD, кафедра меңгерушісі, ассистент-профессор, ORCID ID: 0000-0001-6004-103X,
e-mail: d.kassymova@alt.edu.kz

[1]К.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті,
Алматы қ., Қазақстан
[2]Ақпараттық және есептеу технологиялары институты,
Алматы қ., Қазақстан
[3]М. Тынышпаев атындағы ALT университеті,
Алматы қ., Қазақстан

ЛИНГВИСТИКАЛЫҚ НЕГІЗДЕ ДӘЙЕКСӨЗ КІЛТІН ТАЛДАУҒА АРНАЛҒАН ГИБРИДТІ ТӘСІЛ ЖӘНЕ МАШИНАЛЫҚ ОҚЫТУ

Аңдатпа

Ғылыми мәтіндердің, соның ішінде дәйексөздердің тоналдылығын талдау белсенді дамып келеді, бұл сілтемелердің эмоционалдық бояуын және олардың ғылыми дискуссияға әсерін анықтауға мүмкіндік береді. Бұл зерттеу дәйексөздердің тоналдылығын жіктеу үшін лингвистикалық ережелерді (сөйлеу бөліктерін, синтаксистік тәуелділіктер мен терістеулерді талдау) машиналық оқыту алгоритмдерімен (SVM, RF, NB, J48) біріктіретін гибридті тәсілді әзірлеуге және бағалауға бағытталған. ACL Anthology (8700 жазба) және Clinical Trials (6500 жазба) корпустарында стратификацияланған бөлу (70/15/15 train/val/test) және 5-қайтылымды кросс-тексеру қолданылды. Әзірленген әдіс Athar деректер жиынында 90% macro-F1 және 95% F1 көрсеткіштеріне, ал Clinical Trials корпусында 85% macro-F1 нәтижесіне жетті, бұл базалық модельдермен (BERT, LSTM) салыстырғанда 10–15% жақсартуды көрсетті. Абляциялық зерттеулер лингвистикалық ережелердің үлесін растады (F1 көрсеткішінің 5–7% өсуі). Статистикалық маңыздылық тесттері (McNemar, $p<0.05$) нәтижелердің тұрақтылығын растады. Ұсынылған тәсіл дәйексөздерді автоматты талдау және ғылыми әсерді бағалау үшін тиімді.

Тірек сөздер: дәйексөз тоналдылығын талдау, машиналық оқыту, лингвистикалық ережелер, гибридті модельдер, SVM, библиометрия, macro-F1, кросс-валидация.

[1]Akhmediyarova A.T.,
PhD, Professor, ORCID ID: 0000-0003-4439-7313,
e-mail: a.akhmediyarova@satbayev.university
[1]*Alibiyeva Zh.M.,
PhD, Associate Professor, ORCID ID: 0000-0001-9565-5621,
*e-mail: zh.alibiyeva@satbayev.university
[2]Oralbekova D.O.,
PhD, Junior Researcher, ORCID ID: 0000-0003-4975-6493,
e-mail: dinaoral@mail.ru
[1]Nauryzbayeva A.I.,
Senior Lecturer, ORCID ID: 0000-0001-6004-103X,
e-mail: a.nauryzbaeva@satbayev.university
[3]Kassymova D.T.,
PhD, Head of the Department, Assistant Professor, ORCID ID: 0000-0001-6004-103X,
e-mail: d.kassymova@alt.edu.kz

[1]Satbayev University, Kazakh National Research Technical University
named after K.I. Satpayev, Almaty, Kazakhstan
[2]Institute of Information and Computational Technologies,
Almaty, Kazakhstan
[3]ALT University named after M. Tynyshpayev,
Almaty, Kazakhstan

A HYBRID APPROACH TO THE ANALYSIS OF CITATION TONALITY BASED ON LINGUISTIC FEATURES AND MACHINE LEARNING

Abstract

The analysis of tonality in scientific texts, including citations, is actively advancing, enabling the identification of emotional coloring in references and their impact on scientific discourse. This study focuses on developing and evaluating a hybrid approach that integrates linguistic rules (analysis of parts of speech, syntactic dependencies, and negations) with machine learning algorithms (SVM, RF, NB, J48) to classify citation tonality. Experiments were conducted on the ACL Anthology (8700 sentences) and Clinical Trials (6500 additional sentences) corpora using stratified splitting (70/15/15 for train/val/test) and 5-fold cross-validation. The proposed method achieved 90% macro-F1 and 95% F1-score on the Athar dataset, and 85% macro-F1 on Clinical Trials, showing a 10–15% improvement over baseline models (BERT, LSTM). Ablation studies confirmed the contribution of linguistic rules (F1 increase of 5–7% when excluded). Statistical significance tests (McNemar, $p<0.05$) validated the robustness of the results. The approach proves effective for automated citation analysis and scientific impact assessment.

Keywords: citation sentiment analysis, machine learning, linguistic rules, hybrid models, SVM, bibliometrics, macro-F1, cross-validation.

Received: May, 19, 2025; revised: October 10, 2025; accepted: January 14, 2026.