УДК 004.891.3

СТИЛОМЕТРИЧЕСКИЙ АНАЛИЗ В ЗАДАЧЕ ПОИСКА ЗАИМСТВОВАНИЙ ТЕКСТОВ НА ТАТАРСКОМ ЯЗЫКЕ

И. З. Хаялеева[1] [0009-0007-5837-7010], М. М. Абрамский[2] [0000-0003-3063-8948]

[1, 2] *Казанский (Приволжский) федеральный университет, г. Казань, Россия*

[1]izidakh@yandex.ru, [2]mabramsk@kpfu.ru

Аннотация

Рассмотрена возможность применения методов стилометрического анализа для поиска заимствований в текстах на татарском языке. Разработаны соответствующие инструменты, в которых использованы алгоритмы машинного обучения, включая кластеризацию (метод *k*-средних), классификацию (метод случайного леса, метод опорных векторов, наивный байесовский классификатор) и гибридный подход (модель FastText + логистическая регрессия). Особое внимание уделено адаптации лингвистических метрик для татарского языка.

Ключевые слова*: поиск заимствований, обработка естественного языка, стилометрический анализ, татарский язык.*

ВВЕДЕНИЕ

В современном мире, где информация играет ключевую роль, анализ текстов и определение их авторства становятся все более актуальными задачами. Особенно это касается малоресурсных языков, чьи носители стремятся к сохранению и развитию своего культурного наследия. Одним из таких языков является татарский. В государственной программе «Сохранение, изучение и развитие государственных языков Республики Татарстан и других языков в Республике Татарстан на 2023–2030 годы», принятой Постановлением № 821 Кабинета Министров Республики Татарстан в 2020 г. [1], отмечено, что использование татарского языка в сфере науки, в том числе при написания квалификационных работ для присуждения академической или ученой степеней, сохраняет свою актуальность. А это, в свою очередь, требует современных и точных средств определения уникальности текста.

Целью настоящей работы являются исследование и разработка подходов поиска заимствований в текстах на татарском языке, анализирующих исходный документ с помощью стилометрических[1] методов. Для их достижения были поставлены следующие задачи:

– провести исследование стилометрических методов поиска заимствований;

– применить эти методы в задаче поиска заимствований на татарском языке;

– протестировать и оценить корректность применения стилометрических методов для поиска заимствований на татарском языке.

ОПРЕДЕЛЕНИЕ АВТОРСТВА ТЕКСТА

Одной из задач, решаемых стилометрическим анализом, является определение авторства текста. Для ее решения был реализован инструмент, основанный на алгоритме кластеризации *k*-средних.

Алгоритм *k*-средних – один из методов кластерного анализа, позволяющий разделить произвольный набор данных на заданное количество кластеров таким образом, чтобы объекты внутри одного кластера находились достаточно близко друг к другу, а объекты из разных кластеров не пересекались [2].

В настоящей работе алгоритм *k*-средних был использован для определения *k* различных центроидов в тексте, имеющем разные стили написания. Каждый центроид охватывает такие фрагменты, которые имеют одинаковый стиль написания. Следовательно, количество центроидов соответствует различному количеству стилей написания, присутствующих в документе. На основе предположения о том, что каждый отдельный стиль принадлежит каждому отдельному автору, можно получить оценку авторства каждой части текста. Схема работы созданного инструмента представлена на рис. 1.

[1] Стилометрия – система средств и приемов количественного измерения стилистических характеристик текста.

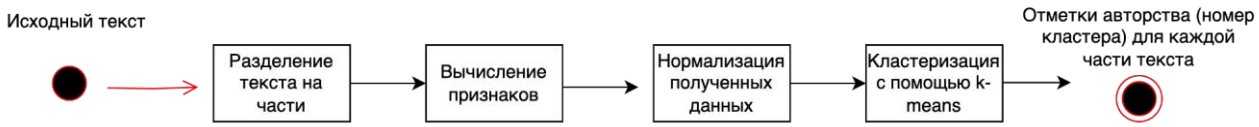

Рис. 1. Схема работы инструмента определения авторства

Для применения алгоритма кластеризации исходный текст был разделен на фрагменты определенной длины, каждый из которых был представлен в виде вектора следующих характеристик:

- сложность чтения текста;
- разнообразие используемых в тексте слов;
- лексические особенности текста.

Для определения метрик измерения перечисленных выше характеристик в тексте на татарском языке были проанализированы работы [3–5]. Стоит отметить, что некоторые подходы к оценке должны рассчитываться с учетом возраста, образования или уровня развития читателя. Соответствующие метрики не были включены в рассмотрение из-за отсутствия данных о пользователях.

Выявленные стилометрические метрики, используемые для векторизации текста, можно отнести к трем группам: лексические, вычисляющие разнообразие используемой лексики и вычисляющие сложность чтения. К лексическим метрикам относятся:

- средняя длина слова;
- среднее количество символов в предложении;
- среднее количество слов в предложении;
- среднее количество слогов в слове;
- количество пунктуационных символов;
- частота специальных символов;
- частота служебных частей речи.

Описанные выше метрики опираются на устоявшиеся понятия в области лингвистики и языкознания. Гораздо больший интерес представляют две другие группы метрик. К метрикам, вычисляющим разнообразие используемой лексики, относятся:

- количество слов *hapax legomenon* – слов, которые встречаются в тексте только один раз. Этот термин часто используют для изучения уникальных слов, которые могут содержать важную информацию о тексте или культуре, в которой был написан текст;
- количество слов *dis legomenon* – таких слов, которые встречаются в тексте только два раза [5];
- мера Оноре – мера, зависящая от количества hapax legomenon и вычисляемая по формуле $H = 100 \log N/(1 - l/d)$, где N – количество слов в тексте, l – количество hapax legomena, d – количество уникальных слов в тексте [6];
- мера Сичела – мера, зависящая от количества dis legomenon и вычисляемая по формуле

$$S = \frac{\mathrm{dis}}{d},$$

где dis – количество dis legomenon, d – количество уникальных слов в тексте [5].

- мера Брюнета – мера, опирающаяся на количество hapax legomenon и вычисляемая по формуле

$$W = N^{d^{-0.17}},$$

где N – количество слов в тексте, d – количество уникальных слов в тексте [6];

- соотношение количества уникальных слов к общему количеству;
- энтропия Шеннона – мера количества информации, которую несет текст, вычисляемая по формуле

$$E = \sum_{i=0}^{N-1} P_i \log P_i,$$

где P_i – вероятность того, что слово под номером i встретится в тексте, а *N* – количество слов в тексте [7].

В качестве метрики сложности текста был отобран индекс удобочитаемости Флеша, оценивающий сложность текста по следующей формуле:

$$\text{УФ} = 206.835 - (1.015\ a) - (84.6\ b),$$

где *a* – средняя длина предложения в словах, *b* – среднее число слогов в слове [3].

Для валидации предложенного подхода был проведен эксперимент на синтетических данных, где в один текст искусственно объединялись фрагменты

от двух разных авторов. Алгоритм показал точность сегментации (accuracy) – 0.78, precision – 0.81 и recall – 0.75 при обнаружении границ стилей.

ОПРЕДЕЛЕНИЕ ЖАНРА ТЕКСТА

Еще одной задачей, решаемой с применением стилометрического анализа, является определение жанра текста. В настоящей работе для ее решения была разработана модель классификации, основанная на векторном представлении текста в виде словаря известных слов.

Для обучения модели были использованы 3450 текстов из разных татароязычных источников, имеющих научный, новостной или художественный жанры. Распределение жанров текстов приведено на рис. 2.

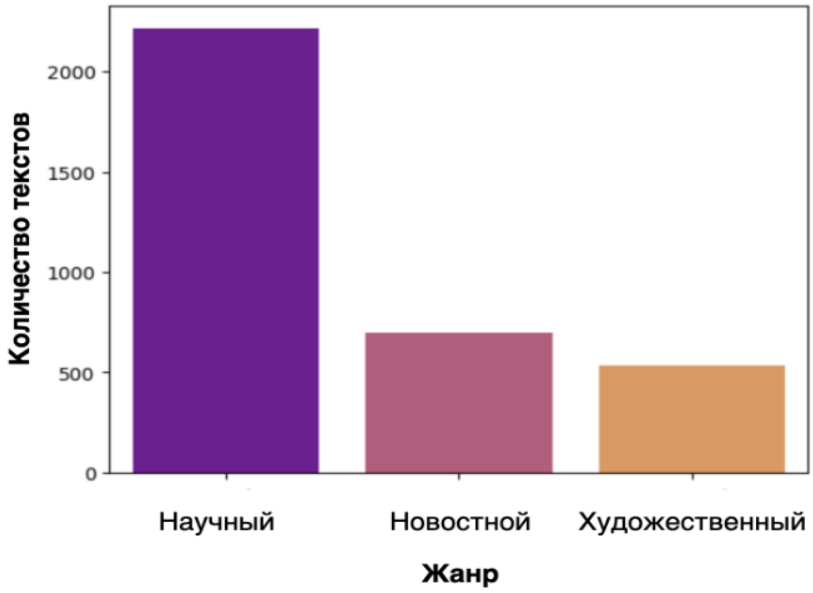

Рис. 2. Распределение текстов по жанрам в обучающем наборе данных (количество текстов – 3450)

Опираясь на работу [9], для данной задачи были реализованы и протестированы три алгоритма классификации: метод случайного леса [10], метод опорных векторов [11] и мультиномиальный наивный байесовский классификатор [12]. Наибольшую точность показал метод случайного леса. Результаты тестирования алгоритмов представлены в табл. 1.

Табл. 1. Сравнительные результаты оценки методов классификации текстов по жанрам

	Метод случайного леса	Метод опорных векторов	Мультиномиальный байесовский классификатор
Доля правильных ответов алгоритма	0.982	0.976	0.852
Точность	0.983	0.977	0.854
Полнота	0.982	0.976	0.852
F1-мера	0.982	0.976	0.844

ОПРЕДЕЛЕНИЕ ЭМОЦИОНАЛЬНОГО ТОНА ТЕКСТА

Определение эмоционального тона текста представляет собой еще одну важную задачу, решаемую в рамках стилометрического анализа. Эта задача заключается в автоматическом определении общего настроения текста: положительного, отрицательного или нейтрального. В более детализированных постановках задачи можно также говорить о классификации по типу эмоций (радость, гнев, печаль и др.). В рамках настоящей работы была рассмотрена базовая модель с трехклассовой классификацией.

Эта задача особенно актуальна для татароязычных текстов, представленных в социальных сетях, комментариях, форумах и пользовательских отзывах. В условиях отсутствия готовых корпусов и моделей на татарском языке разработка инструментов анализа тональности позволяет расширить возможности автоматической обработки текстов и способствует применению языка в современных цифровых сервисах.

Для решения этой задачи был использован гибридный подход, сочетающий методы векторизации текста при помощи предобученной модели FastText и классического машинного обучения. FastText включает в себя модель, обученную на татарском языке (cc.tt.300.vec) [13]. Она позволяет представить любой текст в виде вектора фиксированной размерности, основанного на усреднении векторов слов, входящих в текст.

Алгоритм определения эмоционального тона включал следующие этапы:

- предобработка текста (приведение к нижнему регистру, удаление пунктуации, токенизация);
- получение векторного представления текста;
- обучение классификатора на размеченном корпусе.

Для обучения модели был собран корпус текстов, размеченных вручную по трем категориям: положительный, отрицательный, нейтральный. Каждый текст представлял собой короткое высказывание (1–3 предложения), имитирующее типичные фрагменты отзывов, комментариев или пользовательских мнений. Примеры таких текстов приведены в табл. 2.

Табл. 2. Примеры размеченных текстов

Текст на татарском	Перевод на русский язык	Метка
Бу фильм бик күңелле иде	Этот фильм был очень интересным	положительный
Мин бу китапны яратмадым	Мне эта книга не понравилась	отрицательный
Кичә яңгыр яуды, һава салкын	Вчера дождь шел, было холодно	нейтральный

Для классификации была использована логистическая регрессия, реализованная с помощью библиотеки Scikit-learn [14] и обученная на векторах, полученных с помощью FastText. Модель показала удовлетворительные результаты на тестовом множестве (точность – 0.89, F1-мера – 0.88). Несмотря на небольшой размер корпуса, уже на этой стадии система способна различать базовые эмоциональные категории в татарских текстах.

Таким образом, определение эмоционального тона представляет собой перспективное направление в рамках стилометрического анализа текстов на татарском языке и может быть полезным как в задаче оценки субъективной окраски

текста, так и в качестве дополнительной информации при обнаружении заимствований и определении авторства.

ЗАКЛЮЧЕНИЕ

Успешно применены методы стилометрического анализа для решения задач определения авторства, жанра и эмоционального тона текстов. Результаты тестирования созданных инструментов показали хорошие результаты, однако исследование имеет ряд ограничений, таких, например, как размер корпуса для анализа тональности и отсутствие валидации на реальных данных для задачи определения авторства. Дальнейшие направления исследований включают:

- увеличение корпуса с привлечением пользовательских данных из открытых источников;
- дообучение предобученных многоязычных трансформеров (например, XLM-RoBERTa) [15] на татароязычных текстах;
- внедрение более тонкой классификации (радость, грусть, тревога и др.);
- обнаружение иронии, сарказма и других сложных эмоциональных проявлений.

СПИСОК ЛИТЕРАТУРЫ

1. Постановление кабинета министров Республики Татарстан «Об утверждении государственной программы Сохранение, изучение и развитие государственных языков Республики Татарстан и других языков в Республике Татарстан на 2023 – 2030 годы» // Официальный портал правовой информации Республики Татарстан. Казань, 2020.
URL: https://pravo.tatarstan.ru/npa_kabmin/post/?npa_id=625356 (Дата обращения: 19.08.2025)

2. *Каримов К.Х., Василий Е.А.* Теоретические основы кластеризации данных // Актуальные вопросы фундаментальных и прикладных научных исследований. 2023. С. 242–247.

3. *Балясова И.И.* Параметры сложности текста в татарском языке // Вызовы и тренды мировой лингвистики. 2020. Т. 16. С. 302.

4. *Солнышкина М.И., Макнамара Д.С., Замалетдинов Р.Р.* Обработка естественного языка и изучение сложности дискурса // Russian Journal of Linguistics.

2022. Т. 26. № 2. С. 317–341.

5. *Scott M., Tribble C.* Textual patterns: Key words and corpus analysis in language education. Амстердам: John Benjamins Publishing, 2006. 203 с.

6. *Honoré A. et al.* Some simple measures of richness of vocabulary // Association for literary and linguistic computing bulletin. 1979. Vol. 7, No. 2. P. 172–177.

7. *Flesch R.* A new readability yardstick // Journal of applied psychology. 1948. Vol. 32, No. 3. P. 221–233.

8. *Kincaid J.P., Fishburne Jr R.P., Rogers R.L., Chissom B.S.* Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel // Institute for Simulation and Training. 1975. 49 p.

9. *Kuzman T., Ljubešić N.* Automatic genre identification: a survey // Language Resources and Evaluation. 2025. Vol. 59, No. 1. P. 537–570.

10. *Salman H.A., Kalakech A., Steiti A.* Random forest algorithm overview // Babylonian Journal of Machine Learning. 2024. Vol. 2024. P. 69–79.

11. *Bansal M., Goyal A., Choudhary A.* A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short-term memory algorithms in machine learning // Decision Analytics Journal. 2022. Vol. 3. P. 100071.

12. *Rastogi S., Sambyal R., Tyagi P., Kushwaha R.* Multinomial Naive Bayes Classification Algorithm Based Robust Spam Detection System // 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0. IEEE, 2024. P. 1–5.

13. *Khusainova A., Khan A., Rivera A.R.* Sart-similarity, analogies, and relatedness for tatar language: New benchmark datasets for word embeddings evaluation // International Conference on Computational Linguistics and Intelligent Text Processing. Cham: Springer Nature Switzerland, 2019. P. 380–390.

14. *Pedregosa F. et al.* Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2020. Vol. 12. P. 2825–2830.

15. *Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Stoyanov V.* Unsupervised Cross-lingual Representation Learning at Scale // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 8440–8451.

STYLOMETRIC ANALYSIS IN THE TASK OF SEARCHING FOR BORROWINGS OF TEXTS IN THE TATAR LANGUAGE

I. Z. Khayaleeva[1] [0009-0007-5837-7010], M. M. Abramskiy[2] [0000-0003-3063-8948]

[1, 2]*Kazan (Volga Region) Federal University Kazan, Russia*

[1]izidakh@yandex.ru, [2]mabramsk@kpfu.ru

Abstract

This article discusses the use of stylometric analysis in searching for borrowings of text in the Tatar language. Relevant tools have been developed, utilizing machine learn-ing algorithms, including clustering (k-means method), classification (random forest method, support vector machine method, naive Bayes classifier), and a hybrid approach (FastText model + logistic regression). Special attention is paid to the adaptation of lin-guistic metrics for the Tatar language.

Keywords*: plagiarism detection, natural language processing, stylometric analysis, Tatar language.*

REFERENCES

1. Postanovlenie Kabineta Ministrov Respubliki Tatarstan "Ob Utverzhdenii Gosudarstvennoy Programmy Sokhranenie, Izucheniye i Razvitie Gosudarstvennykh Yazykov Respubliki Tatarstan i Drugikh Yazykov v Respublike Tatarstan na 2023–2030 Gody" // Official Portal of Juridical Information of Republic of Tatarstan. Kazan, 2020. URL: https://pravo.tatarstan.ru/npa_kabmin/post/?npa_id=625356 (access date: 19.08.2025).

2. *Karimov K.Kh., Vasily E.A.* Teoreticheskie osnovy klasterizacii dannyh // Aktual'nye voprosy fundamental'nyh i prikladnyh nauchnyh issledovanij. 2023. P. 242–247.

3. *Balyasova I.I.* Parametry Slozhnosti Teksta v Tatarskom Yazyke // Vyzovy i Trendy Mirovoy Lingvistiki. 2020. Vol. 16. P. 302.

4. *Solnyshkina M.I. McNamara D.S., Zamaletdinov R.R.* Obrabotka Yestestvennogo Yazyka i Izucheniye Slozhnosti Diskursa // Russian Journal of Linguistics. 2022. Vol. 26, No. 2. P. 317–341.

5. *Scott M., Tribble C.* Textual patterns: Key words and corpus analysis in language education. Amsterdam: John Benjamins Publishing, 2006. 203 c.

6. *Honoré A. et al.* Some simple measures of richness of vocabulary // Association for literary and linguistic computing bulletin. 1979. Vol. 7, No. 2. P. 172–177.

7. *Flesch R.* A new readability yardstick // Journal of applied psychology. 1948. Vol. 32, No. 3. P. 221–233.

8. *Kincaid J.P., Fishburne Jr R.P., Rogers R.L., Chissom B.S.* Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel // Institute for Simulation and Training. 1975. 49 p.

9. *Kuzman T., Ljubešić N.* Automatic genre identification: a survey // Language Resources and Evaluation. 2025. Vol. 59, No. 1. P. 537–570.

10. *Salman H.A., Kalakech A., Steiti A.* Random forest algorithm overview // Babylonian Journal of Machine Learning. 2024. Vol. 2024. P. 69–79.

11. *Bansal M., Goyal A., Choudhary A.* A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short-term memory algorithms in machine learning // Decision Analytics Journal. 2022. Vol. 3. P. 100071.

12. *Rastogi S., Sambyal R., Tyagi P., Kushwaha R.* Multinomial Naive Bayes Classification Algorithm Based Robust Spam Detection System // 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0. IEEE, 2024. P. 1–5.

13. *Khusainova A., Khan A., Rivera A.R.* Sart-similarity, analogies, and relatedness for tatar language: New benchmark datasets for word embeddings evaluation // International Conference on Computational Linguistics and Intelligent Text Processing. Cham: Springer Nature Switzerland, 2019. P. 380–390.

14. *Pedregosa F. et al.* Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2020. Vol. 12. P. 2825–2830.

15. *Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Stoyanov V.* Unsupervised Cross-lingual Representation Learning at Scale // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 8440–8451.

СВЕДЕНИЯ ОБ АВТОРАХ

ХАЯЛЕЕВА Изида Зуфаровна – аспирант Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

Izida Zufarovna KHAYALEEVA – PhD student at the Institute of Information Technology and Intelligent Systems, Kazan Federal University.

email: izidakh@yandex.ru

ORCID: 0009-0007-5837-7010

АБРАМСКИЙ Михаил Михайлович – директор Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета, кандидат технических наук.

Mikhail Mikhailovich ABRAMSKIY – director of the Institute of Information Technology and Intelligent Systems, Kazan Federal University, PhD (Cand Sci. – Tech.)

email: mabramsk@kpfu.ru

ORCID: 0000-0003-3063-8948

Материал поступил в редакцию 15 октября 2025 года