DOI: 10.25728/mlsd.2023.1491

ИНТЕЛЛЕКТУАЛЬНЫЙ МОНИТОРИНГ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ ФРАГМЕНТНОГО АНАЛИЗА И ВЕСОВЫХ КОЭФФИЦИЕНТОВ

Русяева Е.Ю., Ахобадзе Г.Н., Полтавский А.В.
Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия
rusyaeva@ipu.ru, akhobadze@ipu.ru, avp57avp@yandex.ru

Аннотация. Предлагается интеллектуальный мониторинг текстов с использованием фрагментного анализа и весовых коэффициентов. Разработана методика расчета служебных слов с привязкой к количеству печатных знаков в тексте и возможностью вычисления авторства произведений на основе анализа весовых коэффициентов, характеризующих отношения служебных слов к объему текста.

Ключевые слова: интеллектуальный мониторинг, фрагментный анализ, определение авторства текста, весовые коэффициенты, служебные слова, вероятность, подсчет.

Введение

На фоне современных тенденций автоматизации по созданию текстов, мультипликативно растут разработки в области генеративного искусственного интеллекта (ГИИ), в частности, ChatGPT [1] и подобных ему ГИИ-инструментов. Но вместе с тем растут и риски [2], связанные как с тем, что люди могут слишком довериться автоматам в создании смыслов и принятии решений, так и с тем, что теперь генерировать и сомнительную информацию, фэйки и даже откровенную ложь становится еще проще, чем раньше [3], поскольку искусственный интеллект – ИИ не отличают правдивую информацию от ложной. Как известно, сейчас ГИИ-инструменты просто подбирают текстовые конструкции из общего доступного объема данных (интернета, мессенджеров и т.д.) и чем этого материала больше, тем и лучше для машинного обучения (самообучения) этих инструментов. На повестке дня остаются главные вопросы: *кто* и *как* будет «фильтровать» этот сгенерированный мультимедийный контент?

Авторы в предыдущих публикациях подчеркивали [4], что важен интегративный подход для решения задач анализа больших мультимедийный данных в управлении крупномасштабными системами. Мы предлагаем диалектический по своей сути подход [5], при котором комбинируются формализованные вычислительные параметры анализируемых объектов с их интеллектуальной, когнитивной обработкой, с анализом экспертных данных. Все полученные данные, обработанные формальными и когнитивными способами, преобразовываются в итоге в методы идентификации.

В данном исследовании авторы предлагают способы, помогающие решать задачу [6, 7], обратную разработкам генеративного ИИ. Если данные ресурсы анализируют текстовую информацию по неким параметрам в основном обезличено, авторство текста не важно, то мы ставим во главу угла задачу идентификации авторства текста и, также, анализируем то, как это скажется на качестве переводов текстов. Предлагаемый интегративный подход не просто соответствует духу времени, но и вписывается в актуальный тренд современности комплексную деятельность [8].

В исследовании представлены альтернативные подходы, позволяющие повысить уровень информативности данных для решения задач управления путем идентификации потоков мультимедийной информации. Это исследование продолжает поисковые исследования и разработки в русле интегративного подхода построения информационных конструкций и информационных систем [4, 6, 7].

Предлагаемые в данной работе подходы основываются на подсчитывании количества знаменательных частей речи и служебных слов в авторских текстах. Далее, путем фрагментного анализа, делим тексты и определяем весовые коэффициентов (интенсивностей) частоты использования служебных слов в них. Затем сравниваем их встречаемость в текстах разных авторов, определяя тем самым авторство, поскольку служебные части речи в определенной мере характеризуют лексико-синтаксические, отсюда и частично некоторые семантические аспекты авторских стилей.

Целью данной работы является установление авторства текста литературного произведения. Для этого используется фрагментный анализ и весовые коэффициенты.

1. Теоретическая часть

1.1. Фрагментный анализ

Суть данного анализа заключается в том, что анализируемый материал, например, два текста с близким сюжетом, разбиваем на фрагменты с определенным количеством слов, и вычисляем в каждом фрагменте количество появления служебных слов как минимум трех из общего количества служебных

частей речи, например, трех предлогов. Далее вычислением отношения количеств этих служебных слов к количеству слов по каждому фрагменту, определяем вероятность повторения всех трех служебных слов в каждом текстовом фрагменте. Затем суммируем полученные вероятности повторения служебных слов и делим эти суммы на количество фрагментов текстов. По результатам деления оцениваем средние арифметические значения вероятности повторения служебных слов и по сравнению средних арифметических значений вероятности повторения этих служебных слов с вероятностью повторения служебных слов можно с определенной долей вероятности судить об авторстве данного текста.

Как известно, любому художественному тексту присуща авторская индивидуальность, стиль автора, т.е. авторы, отличаются своим складом речи. При этом особенности склада речи автора можно отчасти различать по незнаменательным, служебным частям речи, называемыми еще и распорядительными словами. К знаменательным частям речи относят имена существительные, прилагательные и глаголы, а к служебным часто встречающихся у многих авторов предлоги «в», «с» и «на». Практика показывает, что имена существительные, прилагательные и глаголы зависят от многих субъективных факторов, содержания текста и пр., так что частота их употребления в тексте практически мало что дает нам в метрическом плане для определения авторства текста. В предлагаемом подходе именно частота повторения выше указанных предлогов «в», «с» и «на» главным образом используется для установления авторства того или иного литературного текста, так как они служат определенного рода маркерами, определяющими индивидуальную авторскую стилистику.

Формально-математически этот расчет выглядит так. Пусть n_i – количество слов фрагментов исследуемого текста, $m_{iв}$ – число повторения служебной частицы «в» во фрагментах, m_{ic} – число повторения служебной частицы «с» во фрагментах $m_{iна}$ – число повторения служебной частицы «на» во фрагментах. Тогда по формуле $P_1 = m_{1в}/n_1$ можно вычислить частоту повторения частицы «в» в первом фрагменте. В силу этого по формулам $P_1 = m_{1c}/n_1$ и $P_1 = m_{1на}/n_1$ можно определить соответственно частоты повторения частиц «с» и «на» в первом фрагменте. Аналогичным образом, в зависимости от количества слов во фрагментах, можно оценить все возможные частоты повторения частиц «в», «с» и «на» и в других текстовых фрагментах. После этого, зная все частоты повторения предлогов в общему количеству слов во фрагментах P_1, P_2, P_3 … P_i, отношением суммы всех частот появления предлогов к общему числу фрагментов, можно вычислить среднее арифметическое значение $P_{ср}$. Оно, согласно известной теореме Я. Бернулли, принимается в качестве вероятности появления разыскиваемого события. Это формула:

$$P_{ср} = \sum_i \mathbf{Pi} /N, \quad (1)$$

где P_i - частота появления во фрагментах того или иного предлога, N - количество фрагментов исследуемого текста, показывающую факт повторяемости частот предлогов, можно принимать как закон устойчивости этих частот. Вычисления $P_{ср}$ по формуле (1) дает возможность использовать эти значения в качестве вероятностей появления указанных предлогов в рассматриваемом тексте.

1.2. Подход подсчета весовых коэффициентов

Данный подход определения авторства текста литературного произведения, предусматривает подсчет служебных слов, как минимум трех различных предлогов, в первом и втором текстах, вычисление отношений количества каждого предлога к количеству слов в соответствующих первом и втором текстах, равных по количеству слов, и вычисление весовых коэффициентов указанных служебных слов в этих текстах. При сравнении значений однотипных весовых коэффициентов служебных слов, совпадение значений как минимум по одному однотипному весовому коэффициенту служебных слов в анализируемых текстах, дает возможность судить об авторстве текстов литературного произведения. Для этого сначала, как уже было сказано выше, выбирают два текста (два произведения) одного того же автора с близкими сюжетами. Так как объем текстов по печатным знакам может отличаться, то в качестве первого текста принимается текст с меньшим количеством слов, и в этом анализируемом тексте (подсчет нужно вести от начала и до конца текста) подсчитывают количество слов и служебных слов, например, предлогов «в», «с» и «на». После этого отношениями $n_{1в}/N = K_{1в}$, $n_{1на}/N = K_{1на}$ и $n_{1c}/N = K_{1c}$, где $n_{1в}$ – количество предлогов «в» в первом тексте, $n_{1на}$ – количество предлогов «на» в первом тексте, n_{1c} – количество предлогов «с» в первом тексте, где N – количество слов в первом тексте, вычисляют соответственно весовые коэффициенты $K_{1в}$; $K_{1на}$; K_{1c} выше указанных предлогов.

В предлагаемом подходе для окончательного установления авторства двух рассматриваемых текстов производится сравнение вычисленных однотипных весовых коэффициентов предлогов. То есть, в данном случае сравнению подвергаются коэффициенты: $K_{1в}$ и $K_{2в}$; $K_{1на}$ и $K_{2на}$; $K_{1с}$ и $K_{2с}$.

При равных значениях, по меньше мере, двух из трех указанных предлогов и их однотипных весовых коэффициентов, можно констатировать вероятность авторства этих двух текстов.

2. Практическая часть

Практическая реализация выше приведенных подходов предусматривает составление таблиц с учетом подсчета общего количества слов и служебных слов во фрагментах (по спектральному анализу) в исследуемых текстов и без фрагментов.

Для анализа выбраны произведения С. Лукьяненко «Геном» и «Линия грез». Результаты подсчитывания слов и служебных слов во фрагментах произведения «Линия грез» приведены в таблице 1.

Таблица 1. Частота появления предлогов «в», «на» и «с» во фрагментах произведении Лукьяненко С. «Линия грез» (фрагмент расчета на каждую 1000 слов (10 серий))

№ i серии	Количество n_i в серии	Абсолютные величины m_i появления предлогов			Частоты P_i появления предлогов		
		в	с	на	в	с	на
1	1000	21	10	17	0,021	0,01	0,017
2	1000	15	6	17	0,015	0,006	0,017
3	1000	24	6	19	0,024	0,006	0,019
4	1000	24	17	11	0,024	0,017	0,011
5	1000	25	13	16	0,025	0,013	0,016
6	1000	23	9	20	0,023	0,009	0,02
7	1000	17	15	13	0,017	0,015	0,013
8	1000	23	16	20	0,023	0,016	0,02
9	1000	19	13	11	0,019	0,013	0,011
10	1000	25	7	8	0,025	0,007	0,008
					Pср	Pср	Pср
					0,0216	0,0112	0,0152

Согласно приведенной таблице 1, характер изменения числовых значений частот появления предлогов «в», «с», «на» во фрагментах на 1000 слов при их изменении носит неоднозначный характер. Так, например, предлог «в» имеет минимальную величину во 2-м фрагменте и чуть выше в 7-м, а максимальной величины достигает в 5-м и 10-м фрагментах. Предлог «с» имеет минимальную величину во 2 и 3-м фрагментах, а максимума достигает сразу в 4-м и потом чуть меньше в 8-м фрагменте. Предлог «на» имеет минимальную величину в 10-м фрагменте, а максимальную в 3-м фрагменте и чуть ниже в 1-м и 2-м фрагментах, то есть, показывает поначалу плавный рост, а затем спад в 4-м фрагменте и скачок в 5-м фрагменте.

Результаты подсчитывания слов и служебных частиц во фрагментах произведения «Геном» приведены в таблице 2.

Таблица 2. Частота появления предлогов «в», «на» и «с» во фрагментах произведения С. Лукьяненко «Геном» (фрагмент расчета на каждую 1000 слов (10 серий))

№ i серии	Количество n_i в серии	Абсолютные величины m_i появления предлогов			Частоты P_i появления предлогов		
		в	на	с	в	на	с
1	1000	29	23	11	0,029	0,023	0,011
2	1000	16	24	10	0,016	0,024	0,01
3	1000	20	16	7	0,02	0,016	0,007
4	1000	23	18	12	0,023	0,018	0,012
5	1000	26	22	14	0,026	0,022	0,014
6	1000	25	26	11	0,025	0,026	0,011
7	1000	31	17	9	0,031	0,017	0,009
8	1000	29	15	13	0,029	0,015	0,013
9	1000	23	29	4	0,023	0,029	0,004
10	1000	28	20	28	0,028	0,02	0,028
					$P_{ср}$	$P_{ср}$	$P_{ср}$
					0,025	0,021	0,019

Согласно приведенной таблице 2, характер изменения числовых значений частот появления предлогов «в», «с», «на» во фрагментах на 1000 слов при их изменении также носит неоднозначный характер. Так, например, предлог «в» имеет минимальную величину во 2-м фрагменте и уже немного выше в 3-м, а максимальной величины достигает в 1-м и 8-м фрагментах, чуть ниже максимума в 10-м. Предлог «с» имеет минимальную величину в 9-м фрагменте и чуть выше в 3-м фрагменте, а максимума достигает в 10-м. Предлог «на» имеет минимальную величину в 8-м фрагменте и чуть выше в 3-м, а максимальную в 9-м фрагменте.

Согласно данному анализу для установления авторства выше рассмотренных двух произведений, необходимо произвести сравнение средних арифметических значений вероятности повторения однотипных по меньше мере одних служебных слов. Из табличных данных следует, что сходство цифровых значений искомых параметров наблюдается по предлогам «в» с цифровыми значениями, 02 и 0, 02. По остальным предлогам имеет место небольшое различие. Все это в итоге позволяет определить авторство текстов С. Лукьяненко у этих двух произведений.

Как уже было сказано выше, для иллюстрации данного способа были исследованы на авторство два произведения С. Лукьяненко «Линия грез» и «Геном» (см. таблицы 1 и 2). Количество слов N составило в обоих произведениях N =10 000. По подсчетам в случае произведения (первый текст) «Линия грез» $n_{1в}$ =226; $n_{1на}$ = 113; $n_{1с}$ = 154. В результате весовые коэффициенты для этого случая составили $K_{1в}$ =0,03; $K_{1на}$ = 0,02; $K_{1с}$ = 0,02. В случае рассказа (второй текст) «Геном» $n_{2в}$ = 250; $n_{2на}$ = 210; $n_{2с}$ =118. Итого весовые коэффициенты для данного случая $K_{2в}$ = 0,03; $K_{2на}$ = 0,03; $K_{2с}$ = 0,02.

Отсюда при сравнении однотипных весовых коэффициентов можно заключить равенство $K_{1в}$ =0,03; $K_{2в}$ = 0,03, так же коэффициентов $K_{1с}$ = 0,02; $K_{2с}$ = 0,02. Эти равенства дают возможность зафиксировать авторство С. Лукьяненко, исследуемых выше двух произведений по весовым коэффициентам.

В ряде случае для удобства, размерности весовых коэффициентов (как относительные величины), могут быть выражены в процентах. Для этого числовые значения весовых коэффициентов следуют умножить на 100 %. В соответствии с этим можно заключить, что так, например, в произведении «Линия грез» наличие предлога «в» составляет 3%, предлога «на» - 2% и предлога «с» - 2%. В произведении «Геном» наличие предлога «в» составляет 3%, предлога «на» - 3% и предлога «с» - 2%.

3. Заключение

По результатам проведенных исследований можно сделать следующие выводы:

- показана возможность использования весовых коэффициентов, характеризующих отношения служебных слов в текстовом материале к его печатному знаку, для установления авторства рассматриваемых как минимум двух произведений;
- предложенные подходы – фрагментный анализ и весовые коэффициенты – для установления авторства неизвестного произведения, гарантируют полный охват текстового материала и направлены на повышение точности оценивания авторства произведений.

Литература

1. Обучение в эпоху ChatGPT: как преподавателям принять неизбежное. [Электронный ресурс]. URL: https://trends.rbc.ru/trends/education/6440cd219a7947834e9e39d0 (дата обращения: 29.04.2022).
2. *Новиков Д.А.* Вокруг искусственного интеллекта складывается очень тревожная структура знаний и компетенций, – академик Новиков. [Электронный ресурс] URL:https://new.ras.ru/mir-nauky/news/vokrug-iskusstvennogo-intellekta-skladyvaetsya-ochen-trevozhnaya-struktura-znaniy-i-kompetentsiy-aka/.
3. *Соколов Е*. ИИ несет в себе опасности, но совершенно не те, о которых все говорят. Коммерсант. 23.05.2023 [Электронный ресурс]. URL:https://www.kommersant.ru/doc/6000421 (дата обращения: 29.05.2023).
4. *Полтавский А.В., Русяева Е.Ю.* Интегративный подход к построению информационной системы мониторинга для комплексов беспилотных летательных аппаратов / Труды 14-й Международной конференции «Управление развитием крупномасштабных систем» (MLSD-2021). – М.: ИПУ РАН, 2021. С. 1670-1677.
5. *Rusyaeva E., Kravets A*. Creative Knowledge Representation for Knowledge Management: The Dialectical Approach / Communications in Computer and Information Science (book series CCIS, volume 1448). Volgograd: Springer, 2021. С. 97-109. DOI: 10.1007/978-3-030-87034-8_8
6. *Полтавский А.В., Русяева Е.Ю., Бурба А.А.* Устройство для содержательного анализа текстовой информации: Патент на изобретение № 2568272 РФ; Дата публикации 27.10.2015. Бюл. №30
7. *Русяева Е.Ю., Полтавский А.В., Ахобадзе Г.Н*. Integrative Approach to Creation of Information Systems and Entropy Analysis of Linguistic Information / Proceedings of the 2023 International Russian Smart Industry Conference (SmartIndustryCon). Sochi: IEEE, 2023. С. 105-110 https://ieeexplore.ieee.org/document/10110775.
8. *Белов М.В., Новиков Д.А.* Методология комплексной деятельности. М.: ЛЕНАНД, 2018. – 320 с.