Анализ тональности текста как метод моделирования русскоязычного нарратива

Е. В. Бирюкова, email: lenabir68@gmail.com
И. Е. Воронина, email: irina.voronina@gmail.com

Воронежский государственный университет

Аннотация. *Рассматриваются методы анализа тональности текста в контексте решения задачи моделирования русскоязычного нарратива. Приводятся результаты вычислительного эксперимента, демонстрирующего возможность использования готовых решений с целью дальнейшего использования результатов для моделирования русскоязычного нарратива.*

Ключевые слова: *нарратив, нарратология, моделирование нарратива, анализ тональности текста, график тональности, рекуррентные нейронные сети, трансформеры, метод опорных векторов.*

Введение

Нарратив — повествование о некотором множестве событий, связанных друг с другом, представленное читателю или слушателю в виде текста или образов.

Нарратология — дисциплина, изучающая повествовательные тексты (нарративы), их природу, формы и функционирование, общие черты, присущие всем возможным типам нарративов, равно как и критерии, позволяющие отличать последние между собой, систему правил, в соответствии с которыми нарративы создаются и развиваются [1].

Компьютерная нарратология получила свое развитие в рамках компьютерной лингвистики. Рост интереса к интерактивным развлечениям, необходимость производить большие объемы контента повысили интерес к исследованиям автоматической генерации нарративов.

Задача моделирования текстового нарратива — одна из задач современной компьютерной нарратологии. Часто при ее решении используется анализ тональности текста: с его помощью вычисляют эмоциональный окрас тех или иных фрагментов текста, используя эти данные моделирования динамики сюжета всего произведения, сюжетной линии конкретных персонажей, а также отношений между ними.

Анализ тональности текста — это класс методов компьютерной лингвистики, для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов по отношению к объектам, речь о которых идёт в тексте.

Задачи, которые решались при помощи анализа тональности текста по мере развития разных способов его реализации, исторически связаны с маркетингом и продажами. Как следствие большинство методов разрабатывалось для текстов определенного жанра: обзоры, отзывы, комментарии. Среди их отличительных черт можно выделить: небольшой размер, неформальную лексику, меньшее лексическое разнообразие. Современные методы при их использовании для анализа подобного рода текстов показывают высокую точность, а полученные результаты эффективно применяются [2].

Однако для решения задачи моделирования текстового нарратива необходимо анализировать тексты большего объема с более сложной лексикой. Качество результатов при подобных входных данных большинством существующих решений не гарантируется. Более того, при применении методов, основанных на машинном обучении, отдельно возникает вопрос о допустимости использования для обучения модели корпусов, отличающихся по жанру от тестовых данных. Несмотря на то, что существуют примеры использования готовых решений при моделировании англоязычного нарратива [3], точность результатов в этих исследованиях отдельно не измерялась. Примеров же применения анализа тональности текста для моделирования нарратива на русском языке не существует.

На конференции «Актуальные проблемы прикладной математики, информатики и механики» (Воронеж, 12-14 декабря, 2022) авторами были представлены потенциальные проблемы использования методов анализа тональности текста для задачи моделирования русскоязычного нарратива были рассмотрены более подробно, а также был представлен план исследования с необходимыми вычислительными экспериментами.

Для проведения одного такого эксперимента воспользуемся наиболее известными решениями для анализа тональности текста, поддерживающими русский язык: IBM Watson Natural Language Understanding, Microsoft Cognitive Services, Connexun Text Analysis API. Построим с их помощью графики тональности глав «Войны и мира» и сравним их. Проанализируем отдельно статистические значимые отклонения.

IBM Watson Natural Language Understanding

Используется метод опорных векторов (SVM, support vector machine) — набор алгоритмов обучения с учителем, использующихся

для задач классификации и регрессионного анализа. Для каждого из тонов модель обучается независимо с использованием парадигмы One-Vs-Rest. Во время прогнозирования определяются тона, которые были предсказаны с вероятностью не менее 0,5 в качестве итоговых. Для решения проблемы несбалансированности тонов вычисляется оптимальное значение веса функции стоимости для каждого из этих тонов во время обучения [4].

Microsoft Cognitive Services

Для анализа тональности используется машинное обучение, однако данные об архитектуре модели не публикуются.

Connexun Text Analysis API

Используется модель «xlm-roberta-large», дообученная для решения задачи Natural Language Inference [5]. Модель RoBERTa – это модель трансформер, предобученная на большом корпусе текстов без предварительной обработки. При обучении использовалась технология маскированной языковой модели, что выгодно отличается от часто используемых традиционных рекуррентных нейросетей.

Результаты вычислительных экспериментов

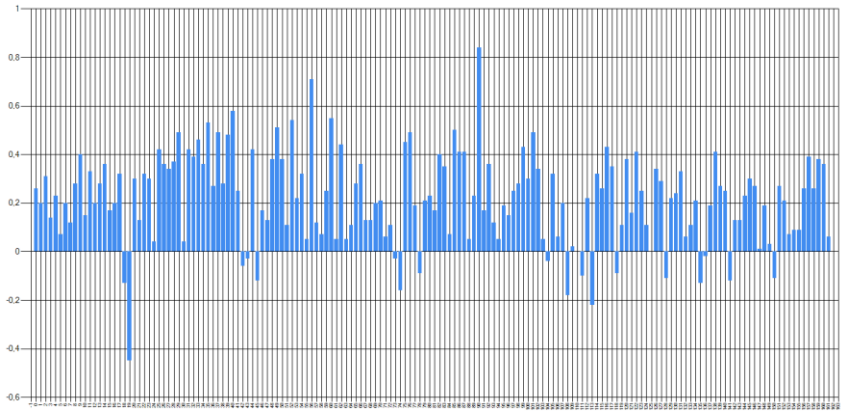

Рис. 1. IBM

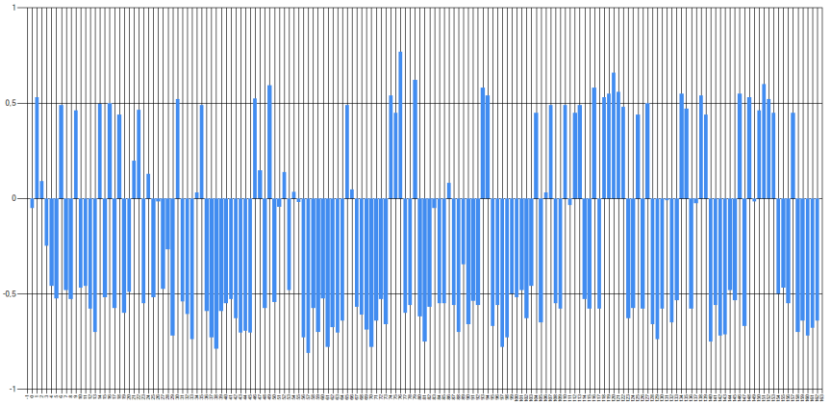

Рис. 2. Microsoft

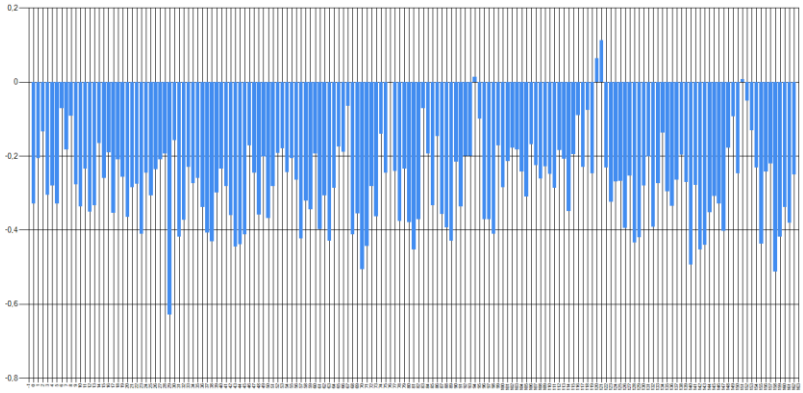

Рис. 3. Connexun

На графиках по вертикальной шкале представлена выраженность тональности от -1 до 1, по горизонтальной – номер главы. Очевидно, что полученные результаты во многом различны.

Разные решения предоставили противоположные ярлыки тональности даже для пиковых значений: из редких совпадений стоит отметить 121 главу, оцененную как положительную на всех 3 графиках; 76 главу, оцененную двумя решениями как положительную и одним как нейтральную; 19, на всех графиках оцененную как одну из самых негативных.

Графики, построенные при помощи решения от Microsoft и Connexun, имеют больше общего. Это можно объяснить тем, что они

используют более сложные модели, а также ограничивают анализируемые фрагменты текста по длине, что повышает точность.

При этом в некоторых главах (113, 89, 49, 158) при значениях близких к пиковым у Microsoft и Connexun противоположные ярлыки. Это во многом связано с тем, что почти все главы были оценены Connexun как негативные. Эта оценка, тем не менее, полностью справедлива для многих глав середины и начала и для всех глав в конце.

Заключение

Таким образом, наиболее архитектурно сложное решение дало наиболее правдивый согласно предварительной оценке результат. Однако относительно небольшой разброс выраженности сантимента необходимо исследовать более тщательно.

Так же улучшение результата с увеличением сложности модели актуализирует вопрос о построении собственного решения, поднятый еще в статье «Методы моделирования и анализа нарративов русскоязычных текстов» (материалы вышеупомянутой конференции, в печати) . Модель, используемую в Connexun, предполагается взять за основу как самую успешную в рамках эксперимента. Подход обучения без учителя при этом является принципиально важным в связи с отсутствием корпусов нужного жанра с нужной разметкой.

Список литературы

1. Шмид В. Нарратология // В. Шмид. – Языки славянской культуры. – 2003. – 312 с.
2. Gonçalves P, Araújo M, Benevenuto F, Cha M. Comparing and combining sentiment analysis methods. In: Proceedings of the first ACM conference on Online social networks. ACM; 2013.
3. Min S, Park J (2019) Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling. PLoS ONE 14(12): e0226025. https://doi.org/10.1371/journal.pone.0226025
4. Watson Natural Language Understanding // IBM. – URL: https://www.ibm.com/cloud/watson-natural-language-understanding (дата обращения: 13.01.2023).
5. Sentiment Analysis // Connexun. – URL: https://connexun.medium.com/sentiment-analysis-fe35fac4831e (дата обращения: 13.01.2023).