

Построение кривых эмоциональной тональности текста с использованием словаря RuSentiLex

Цыпущатова М.А.¹, Крюкова К.Ю.², Ахмедов Э.Ф.³

1 - Ульяновский государственный педагогический университет имени И.Н. Ульянова, Факультет физико-математического и технологического образования, Ульяновская область, Россия, *E-mail: ts.marina.04@gmail.com*; 2 - Ульяновский государственный педагогический университет имени И.Н. Ульянова, Факультет физико-математического и технологического образования, Ульяновская область, Россия, *E-mail: kristina.kryukova.05@inbox.ru*; 3 - Ульяновский государственный педагогический университет имени И.Н. Ульянова, Факультет физико-математического и технологического образования, Ульяновская область, Россия, *E-mail: andy.ahmedov@gmail.com*

Анализ тональности (sentiment analysis) — это методы автоматической оценки эмоционального содержания текста, основанные на выявлении уровня позитивности/негативности с использованием специальных словарей, машинного обучения или комбинированных подходов. Как показывает анализ литературы, в настоящее время для русскоязычных текстов такие методы развиты недостаточно [4].

Одно из приложений анализа тональности — построение кривых эмоциональной тональности текста, представляющих собой временные ряды значений тональности. Для построения кривых применяются различные подходы. Методы, основанные на использовании словарей плохо справляются с короткими фрагментами [3], в то время как нейросетевые модели позволяют работать с отдельными предложениями, но кривые получаются сильно зашумленными [1].

В данной работе для построения кривых эмоциональной тональности предлагается использовать подход, основанный на использовании словаря оценочных слов и выражений RuSentiLex [2]. В каждой записи словаря приводятся лемматизированная (нормальная) форма слова или словосочетания и его оценка, например:

аварийная защита, NG, аварийный защита, neutral, fact

Построение кривой начинается с предварительной ручной подготовки текста (поиск и устранение опечаток, разметка). Дальнейший процесс автоматизирован и состоит из следующих этапов:

1. Поиск и удаление специальных символов (лишних пробелов, знаков табуляций, концов строк), разбиение текста на предложения.

2. Объединение предложений в группы заданного размера, вычисление тональности каждой группы с использованием словаря RuSentiLex и усреднение результата для отдельных предложений.

3. Обработка результатов: фильтрация, построение графиков.

На рисунке представлены две кривые. Первая построена в два прохода с размерами групп 25 и 50 с последующим усреднением, вторая получена из первой с помощью скользящего среднего.

Программная реализация выполнена на языке Python. При определении тональности предложений для лемматизации слов и словосочетаний использовались библиотеки Natasha и Rymorphy2.

Источники и литература

- 1) Пронин Д. Как провести анализ тональности текста // «Системный Блок»: <https://sysblok.ru/courses/kak-provesti-analiz-tonalnosti-teksta>

- 2) Loukachevitch N., Levchik A. Creating a General Russian Sentiment Lexicon // In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 2016, P. 1171–1176.
- 3) Reagan A. J., Mitchell L., Kiley D., Danforth C. M., Dodds P. S. The emotional arcs of stories are dominated by six basic shapes // EPJ Data Science. 2016. Vol. 5. Article number: 31.
- 4) Smetanin S. The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives // IEEE Access. 2020. Vol. 8. P. 110693–110719.

Иллюстрации

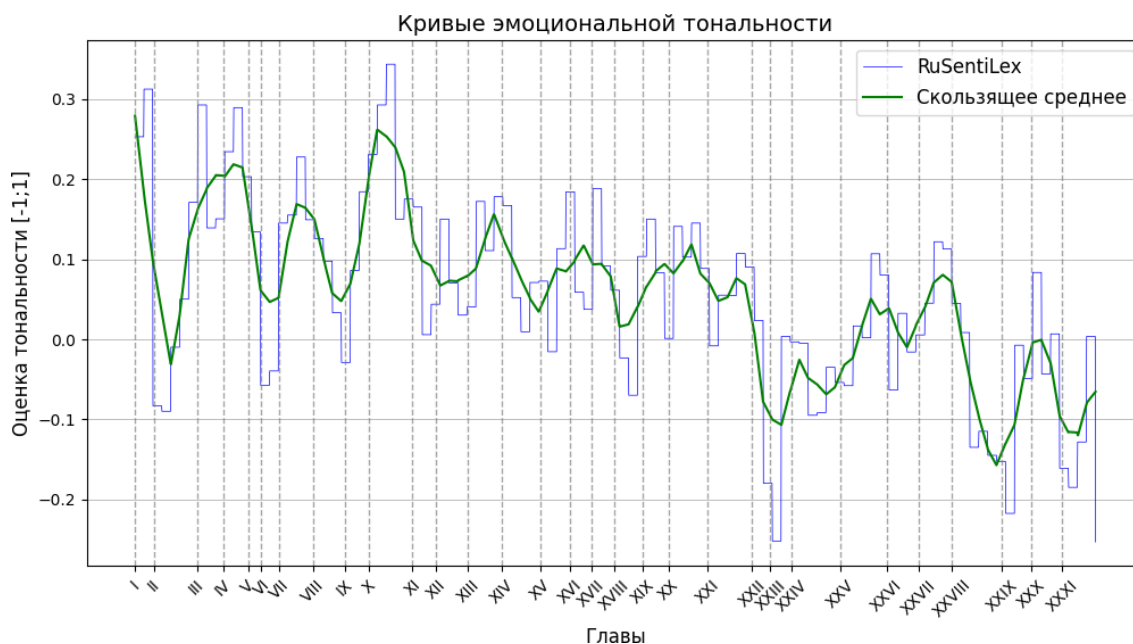


Рис. : Кривые эмоциональной тональности седьмой части «Анны Карениной» Л. Н. Толстого