DOI: 10.37791/2687-0649-2020-15-6-58-77

Разработка тонально-тематического словаря EcSentiThemeLex для анализа экономических текстов на русском языке

Е. А. Федорова[1], Д. О. Афанасьев[2], И. С. Дёмин[1], И. В. Пыльцин[3], Р. Г. Нерсесян[4], А. М. Лазарев[5]*
[1] *Финансовый университет при Правительстве РФ, Москва, Россия*
[2] *АО «Гринатом»**, Москва, Россия,* [3] *НИУ ВШЭ, Москва, Россия ,* [4] *ООО «Цифра», Москва, Россия*
[5] *МГУ имени М. В. Ломоносова, Москва, Россия*
** ecolena@mail.ru*

Аннотация. Цель исследования – разработка общедоступного тонально-тематического словаря на русском языке, позволяющего выявлять смысловую направленность по группам экономических текстов, а также определять их сентиментные (тональные) характеристики. В статье описаны основные этапы составления словаря с применением методов машинного обучения (кластеризация, выделения частотности слов, построение коррелограмм) и экспертной оценки определения тональности и расширение словаря за счет включения терминов из аналогичных зарубежных словарей. Эмпирическая база исследования включала в себя: годовые отчеты компаний, новости министерств и ЦБ РФ, финансовые твиты компаний и новостные статьи РБК по направлению «Экономика, финансы, деньги и бизнес». Составленный словарь отличается от предыдущих по следующим направлениям: 1) является одним из первых словарей, позволяющих оценивать тональность экономических и финансовых текстов на русском языке по пяти степеням тональности; 2) позволяет оценить тональность и смысловую направленность текста по 12 экономическим темам (например, макроэкономика, монетарная политика, фондовые и товарные рынки и т. д.); 3) итоговый словарь EcSentiThemeLex включен в программный пакет (библиотеку) rulexicon для среды программирования R[1] и Python[2]. Приведены пошаговые примеры использования разработанной библиотеки в среде R, позволяющие оценить тональность и тематическую направленность экономического или финансового текста на основе лаконичного кода. Структура библиотеки позволяет использовать оригинальные тексты для их оценки без предварительной лемматизации (приведения к начальным формам). Составленный в данной работе тонально-тематический словарь EcSentiThemeLex со всеми словоформами позволит упростить решение прикладных задач текстового анализа в финансово-экономической сфере, а также потенциально сможет послужить базисом для наращивания числа соответствующих исследований в российской литературе.

Ключевые слова: тематический словарь, экономические тексты, новости, машинное обучение, текстовый анализ, база данных, программные средства

Для цитирования: *Федорова Е. А., Афанасьев Д. О., Дёмин И. С., Пыльцин И. В., Нерсесян Р. Г., Лазарев А. М.* Разработка тонально-тематического словаря EcSentiThemeLex для анализа экономических текстов на русском языке // Прикладная информатика. 2020. Т. 15. № 6. С. 58–77. DOI: 10.37791/2687-0649-2020-15-6-58-77

[1] https://dmafanasyev.github.io/rulexicon/index.html

[2] https://pypi.org/project/ecsentithemelex

** Позиция Афанасьева Д. О., отраженная в данном исследовании, не является официальной позицией АО «Гринатом» и может не совпадать с ней.