DOI: 10.37791/2687-0649-2022-17-3-5-15

Задача анализа эмоциональной окраски текстов в банковской деятельности

С. П. Строев[1*], *А. В. Захаров*[2], *Ж. В. Мекшенева*[2], *В. В. Шоколов*[3], *А. М. Нечаев*[2], *Н. Н. Люблинская*[2]

[1] *Орловский государственный университет имени И. С. Тургенева, Орел, Россия*
[2] *Университет «Синергия», Москва, Россия*
[3] *АО «Райффайзенбанк», Москва, Россия*
** stroewsp@mail.ru*

Аннотация. В работе излагается авторский подход к решению задачи анализа тональности русскоязычных сообщений в сети Интернет о деятельности банков. Материалами исследования выступают отзывы клиентов о банках в целом, о продуктах, сервисах и качестве обслуживания, размещенные на портале Банки.ру. В работе задача анализа тональности текстов рассматривается как задача бинарной классификации на множестве позитивных и негативных отзывов. Для представления собранных и предварительно обработанных текстов использовалась векторная модель со схемой взвешивания tf-idf. Поиск решения задачи бинарной классификации осуществлялся следующими алгоритмами с подбором оптимальных параметров на сетке: наивный байесовский классификатор, метод опорных векторов, логистическая регрессия, случайный лес и градиентный бустинг. Для оценки качества решения задачи классификации применялись стандартные статистические метрики – точность, полнота и F-мера. По указанным метрикам наилучшие результаты получены на классификационной модели, построенной с помощью метода опорных векторов. С целью выделения наиболее характерных тем сообщений клиентов рассматривалась также задача тематического моделирования текстов. Для ее решения применялся метод латентного размещения Дирихле. В результате установлено, что наиболее популярными темами сообщений являются «Карты» и «Качество облуживания». Полученные результаты работы могут использоваться в деятельности банка для автоматизации мониторинга его репутации в медиапространстве и при маршрутизации клиентских запросов по решению различных проблем. При решении задач активно применялись возможности языка программирования Python, а именно библиотеки для веб-скрейпинга, машинного обучения, обработки естественного языка.

Ключевые слова: анализ тональности текста, методы обработки текстов, алгоритмы машинного обучения, векторная модель представления текста, банковская деятельность

Для цитирования: *Строев С. П., Захаров А. В., Мекшенева Ж. В., Шоколов В. В., Нечаев А. М., Люблинская Н. Н.* Задача анализа эмоциональной окраски текстов в банковской деятельности // Прикладная информатика. 2022. Т. 17. № 3. С. 5–15. DOI: 10.37791/2687-0649-2022-17-3-5-15

Text sentiment analysis in banking

S. Stroev[1*], *A. Zakharov*[2], *Zh. Meksheneva*[2], *V. Shokolov*[3], *A. Nechaev*[2], *N. Lyublinskaya*[2]
[1] *Orel State University named after I. S. Turgenev, Orel, Russia*
[2] *Synergy University, Moscow, Russia*
[3] *Raiffeisenbank JSC, Moscow, Russia*
** stroewsp@mail.ru*

Abstract. The paper presents the author's approach to solving the problem of sentiment analysis of online Russian-language messages about the activities of banks. The study data are customer reviews about banks in general and their products, services and quality of service posted on the Banki.ru portal. In this paper, the problem of text sentiment analysis is considered as a binary classification task based on a set of positive and negative reviews. A vector model with a tf-idf weighting scheme was used to represent the collected and preprocessed texts. The following algorithms with the selection of optimal parameters on the grid were used for binary classification task: naive Bayesian classifier, support vector machine, logistic regression, random forest and gradient boosting. Standard statistical metrics, such as accuracy, completeness, and F-measure, were used to evaluate the quality of solving the classification problem. For the indicated metrics, the best results were obtained on the classification model developed with the use of Support Vector Machine. Thematic text modeling was also carried out using the Dirichlet latent placement method to define the most typical topics of customer messages. As a result, it was concluded that the most popular message topics are "cards" and "quality of service". The obtained results can be used in the activities of banks to automate its reputation monitoring in the media and when routing client requests to solve various problems. When solving problems, the features of the Python programming language were actively used, namely, libraries for web scraping, machine learning, and natural language processing.

Keywords: text sentiment analysis, text processing methods, machine learning algorithms, vector text representation model, banking

For citation: Stroev S., Zakharov A., Meksheneva Zh., Shokolov V., Nechaev A., Lyublinskaya N. Text sentiment analysis in banking. *Prikladnaya informatika*=Journal of Applied Informatics, 2022, vol.17, no.3, pp.5-15 (in Russian). DOI: 10.37791/2687-0649-2022-17-3-5-15

Введение

Взрывное развитие социальных сетей, рекомендательных сервисов и тематических форумов привело к появлению огромных массивов неструктурированных данных, содержащих в том числе мнения потребителей о товарах или услугах [16]. У компаний, предоставляющих такие товары или услуги, таким образом, появляется необходимость ориентироваться в этих потоках данных и анализировать не только их содержание, но и эмоциональную окраску. Для этих целей активно применяются методы машинной обработки естественного языка.

В общем случае под анализом эмоциональной окраски принято понимать выявление положительного или отрицательного отношения автора текста к обсуждаемому предмету [1, 4–6]. Следует заметить, что в области обработки естественного языка оформилось отдельное направление по анализу эмоциональной окраски текстов, которое также называют анализом тональности или сентимент-анализом. В дальнейшем в данной работе эти термины будут использоваться взаимозаменяемо.

Методы анализа эмоциональной окраски активно применяются для исследования сообщений пользователей в социальных сетях [11], рекомендательных сервисах [2]. В корпо-