Оценка тональности новостных лент на основе методов машинного обучения

Чернышова Г.Ю.[1], Абакумова Н.В.[2]
[1]*cherny111@mail.ru*, [2]*vetsillia@mail.ru*
[1]*Саратовский государственный университет имени Н.Г. Чернышевского,*
[2]*ООО «СИБИНТЕК-СОФТ», Россия, Саратов*

Аннотация. Применение методов машинного обучения для анализа тональности связано с наличием обучающих наборов данных. В работе предлагается подход для формирования обучающей выборки с учетом специфики новостных сообщений. Разработанное приложение обеспечивает возможность использования различных классификационных алгоритмов, классификацию новостей в соответствии с тональностью на основе модифицированного датасета, визуализацию полученных результатов. В качестве апробации осуществлена оценка тональности экономических сообщений новостного агрегатора за 2019-2021 годы.
Ключевые слова: Sentiment Analysis, обучающая выборка для оценки тональности, оценка тональности новостных лент

Введение

Задача автоматического определения тональности текстов представляет практический интерес в условиях возрастающего объема информации[1]. Для автоматической классификации текста с точки зрения эмоциональной окраски активно применяется машинное обучение, обработка естественного языка и другие методы под управлением искусственного интеллекта, что позволяет более быстрым, экономичным и точным способом решать подобные задачи. Подход на основе машинного обучения используется для анализа тональности текстов (Sentiment Analysis) [2–4].

Анализ тональности можно применить для новостных сообщений, чтобы повысить объективность оценки эмоциональной окраски текстов из информационных источников. Однако можно столкнуться со следующей проблемой: новостные сообщения чаще всего подаются в нейтральной тональности, что затрудняет их классификацию. Для решения данной проблемы предлагается предобработка обучающего набора данных и применение соответствующих алгоритмов машинного обучения. Задачами исследования является сравнительный анализ существующих обучающих выборок для Sentiment Analysis; модификация обучающей выборки с учетом особенности новостных сообщений; разработка web-приложения для оценки тональности новостных лент; применение web-приложения на примере новостных текстовых сообщений.

Анализ обучающих выборок для Sentiment Analysis новостных сообщений. Существует ряд доступных наборов обучающих данных для анализа тональности, с помощью которых можно создать модель классификации тональности для русскоязычных текстов.

Новостная подборка на русском языке [5] включает тексты новостей и метки класса, к которому принадлежит конкретное сообщение. Выборка включает 8256 объектов. Набор данных с короткими сообщениями на русском языке в формате твитов [6] содержит 17000 записей. У данных наборов имеются

определенные недостатки. В частности, для каждого из трех классов в имеющихся выборках количество образцов относительно невелико, что повлияет на точность полученных классификационных моделей. Кроме того, несбалансированное количество записей для классов. Например, в датасете с новостями мало образцов с негативной окраской. Во второй выборке с твитами имеется большое количество пустых записей или записей, помеченных как пропуск.

Так как количество русскоязычных размеченных наборов данных для анализа тональности сообщений ограничено, для модификации обучающей выборки использованы также новости [7] и твиты [8] на английском языке. Была применена распространенная практика перевода иностранных обучающих выборок на русский язык. Путем перевода этих наборов данных на русский язык были получены 15000 записей.

В рамках вычислительного эксперимента был обучен целый ряд моделей на основе разных алгоритмов классификации (наивный байесовский классификатор, логистическая регрессия, *k*-NN, случайный лес). Различные алгоритмы были выбраны для того, чтобы объективнее исследовать точность моделей с различными обучающими выборками. Модели, обученные на обновленной выборке, показали большую точность.

В таблице 1 представлен результат оценки F1-score для каждой из моделей, полученных в результате применения различных алгоритмов машинного обучения для каждой из представленных обучающих выборок.

Таблица 1 – Оценка точности моделей для различных обучающих выборок

Алгоритм классификации	Новости на русском языке	Короткие сообщения на русском языке	Новости, переведенные на русский язык	Короткие сообщения, переведенные на русский язык
Наивный байесовский классификатор	0.70	0.53	0.84	0.65
Логистическая регрессия	0.72	0.52	0.80	0.73
k-NN	0.66	0.48	0.82	0.82
Случайный лес	0.66	0.52	0.85	0.84

Лучше всего себя показали иностранные выборки, переведенные на русский язык, это обусловлено большим и сбалансированным количеством записей в разных классах при отсутствии пропусков. Модели, обученные на российских твитах, имеют невысокую точность из-за небольшого количества записей и специфических особенностей подобных текстовых сообщений, связанных с их длиной и смысловой содержательностью. Чтобы повысить точность имеющихся моделей предложено расширить количество примеров в классах, где их недостаточно.

Для обновления словаря терминов и расширения примеров в классах с негативными и позитивными записями был использован парсинг информационного источника Лента.ру. В результате был получен массив новостей, которые в дальнейшем были размечены и добавлены в модифицированную обучающую выборку. При этом для экспертной тестовой

выборки точность моделей, оцененная с помощью F1-score, увеличилась в среднем на 5%.

Применение приложения для анализа тональности текстов на примере экономических новостей. Было разработано приложение на языке Python 3.8 с использованием полученных моделей анализа тональности. Для создания интерфейса приложения использовалась библиотека PySimpleGUI.

Разработанное приложение для анализа тональности новостной ленты имеет следующий функционал:

- загрузка массива данных в формате .csv, в котором хранится текстовое сообщение и дата создания сообщения;
- классификация текстовых сообщений по тональности, при этом используется три класса (нейтральное, позитивное и негативное сообщение);
- выбор классификационной модели (наивный байесовский классификатор, логистическая регрессия, *k*-NN, случайный лес);
- вывод диаграммы, показывающей количество сообщений принадлежащих разным классам в разные даты; вывод диаграммы, показывающей процентное соотношение количества записей в разных классах. На рисунке 1 представлен интерфейс приложения.

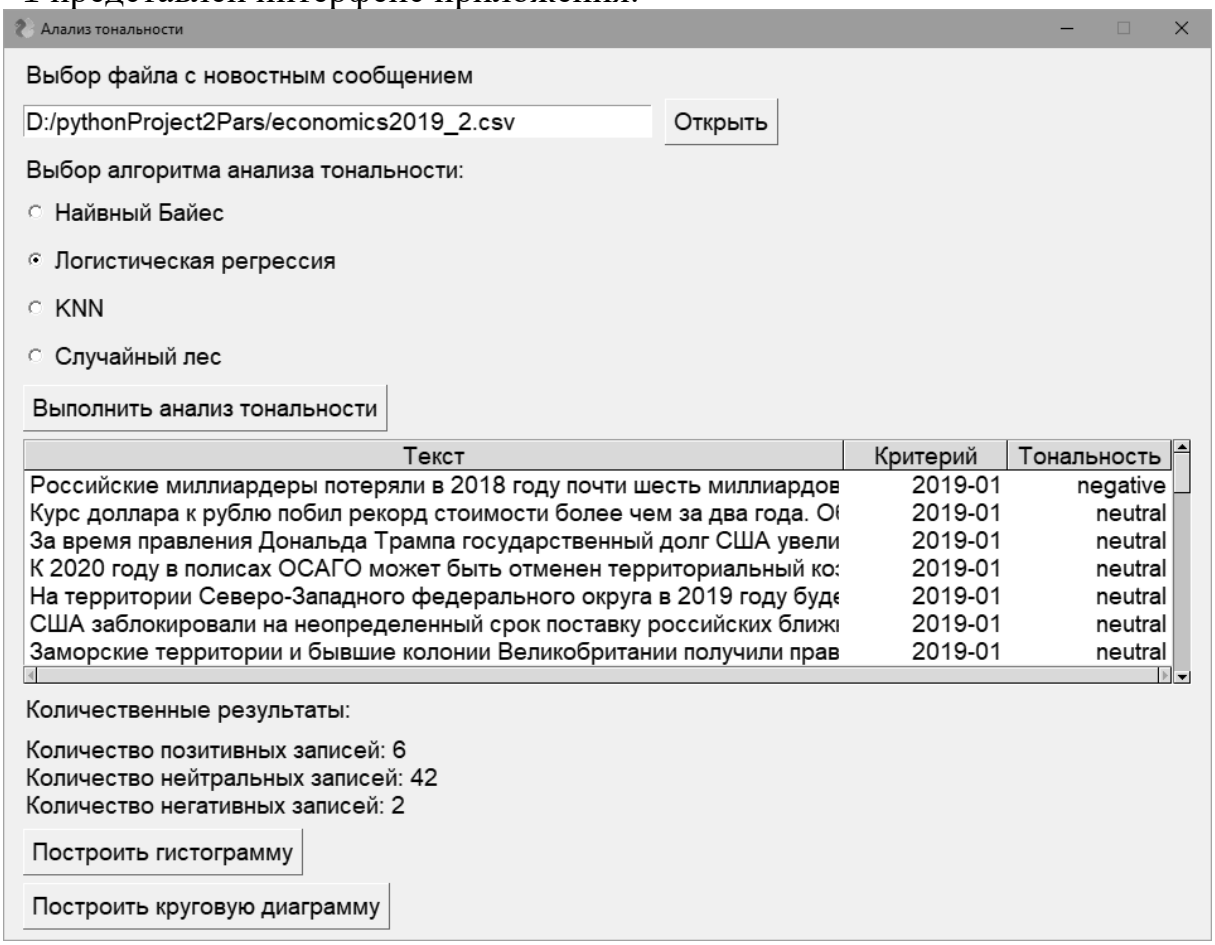

Рис. 1. Приложение для оценки тональности

Для апробации приложения с помощью парсера новостей были собраны новости из рубрики «Экономика» в период с 2019 по 2021 г. Пример

визуализации результатов анализа тональности по экономической тематике за 2021 г. представлен на рисунке 1.

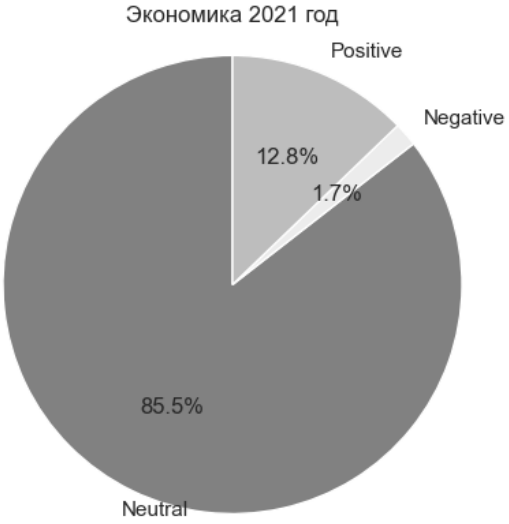

Рис. 2.Оценка тональности новостей за 2021 год

Результаты применения предложенного подхода для оценки тональности новостей за 2019-2021 гг. приведены в таблице 2.

Таблица 2 – Результаты анализа тональности новостной ленты за период с 2019 по 2021 год

Тональность	Позитивная		Нейтральная		Негативная	
Год	Кол-во	%	Кол-во	%	Кол-во	%
2019	874	14.4	4979	82.3	196	3.2
2020	888	11.9	6357	85.2	213	2.9
2021	1057	12.8	7068	85.5	138	1.7

При общем увеличении количества новостей в процентном соотношении количество нейтральных составляет 82-85%. Позитивных новостей стало меньше в 2020 г., но через год результат вырос на 1%, и составляет в 2021 г. 14.4%. Негативно окрашенных сообщений стало меньше (1.7% в 2021 г.).

В каждой из рассмотренных обучающих выборок большая часть записей относится к нейтральной тональности, что ухудшает качество обучения классификационных моделей. Для повышения точности следует расширить словарь терминов, применяемых в новостных сообщениях, так как ситуация постоянно меняется, и в начальном наборе данных отсутствовали термины, которые активно применяются последние несколько лет. Кроме того, следует увеличить количество записей относящихся к позитивной и негативной тональностям.

В качестве апробации были использованы экономические новости за период с 2019 по 2021 год из новостного источника Лента.ру. Модели анализа тональности показали снижение негативных новостей на фоне общего увеличения числа новостей за год

Список литературы

[1] Machine Learning for Text/ Aggarwal C. – New York: Springer, 2018. – 493 с.

[2] Reis, J., Olmo, P., Benevenuto, F., Kwak, H., Prates, R., An, J. Breaking the news: First impressions matter on online news // Proceedings of the International Conference on Web and Social Media, ICWSM. Oxford, UK, 2015. – Pp. 357-366.

[3] Godbole N, Srinivasaiah M, Skiena S. Large-Scale Sentiment Analysis for News and Blogs. // Proceedings of the International Conference on Web and Social Media, ICWSM, Oxford, UK. 7(21), 2007. – Pp. 219-222.

[4] Smetanin S., Khomarov M. Deep transfer learning baselines for sentiment analysis // Russian Information Processing & Management, 58(3), 2021. 102484.

[5] Sentiment Analysis in Russian [Электронный ресурс] URL: https://www.kaggle.com/competitions/sentiment-analysis-in-russian/data (дата обращения:05.08.2023).

[6] RuSentiTweet: A Sentiment Analysis Dataset of General Domain Tweets in Russian [Электронный ресурс] URL: https://github.com/sismetanin/rusentitweet (дата обращения:05.01.2023).

[7] AG News (News articles) [Электронный ресурс] URL: https://www.kaggle.com/datasets/thedevastator/new-dataset-for-text- classification-ag-news (дата обращения:05.08.2023).

[8] Twitter Sentiment Analysis [Электронный ресурс] URL: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment- analysis?resource=download (дата обращения:05.08.2023).