УДК 004.855.5

СРАВНЕНИЕ БИБЛИОТЕК PYTHON ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ

Белоусова А.Л. (Университет ИТМО), **Железникова П.А.** (Университет ИТМО)
Научный руководитель – доцент, кандидат педагогических наук, Авксентьева Е.Ю.
(Университет ИТМО)

Введение. В последнее время в научном сообществе популярна проблема классификации текстов по эмоциональной окраске, которая также известна как анализ тональности. Анализ тональности является частью компьютерной лингвистики, которая изучает мнения и эмоции в текстовых данных и представляет собой набор методов, предназначенных для автоматического определения реакции эмоций или отношения (настроения), выраженных в тексте. Язык программирования Python является одним из наиболее популярных языков, применяемых для решения задач анализа тональности текста. Множество инструментов и библиотек языка предоставляют возможность обработки как русскоязычных, так и англоязычных текстов. Сравнение библиотек Python для анализа тональности текста на русском и английском языках позволит проанализировать имеющиеся решения, выявить их преимущества и недостатки, выбрать инструмент, осуществляющий наиболее точную классификацию.

Основная часть. Анализ тональности – это метод, который идентифицирует текст из различных изображений и наборов данных. Его целью является выделение в тексте тональных компонентов. Единица, из которой выделяется одно мнение, называется уровнем анализа тональности. Существуют следующие уровни [1]: уровень документа, уровень предложения, уровень сущности и аспекта. В пилотном и основном экспериментах анализ тональности осуществлялся на уровне документа. Проводилось сравнение следующих библиотек: VADER, TextBlob и Dostoevsky. Для пилотного и основного экспериментов был применен бинарный метод классификации текстов. Бинарный метод является наиболее простым и распространенным и достигает высокой отметки точности.

Существует три основных инструмента, доступных для проверки качества модели классификации [2]: матрицы ошибок (confusion matrices), графики роста (lift charts), и кривые ошибок (ROC, receiver operator characteristic). Эффективность классификации лучше всего описывается матрицей ошибок, которая была использована в пилотном эксперименте. Базовым показателем матрицы ошибок для измерения производительности является метрика точности (Accuracy).

В рамках пилотного эксперимента была создана выборка из 50 отзывов пользователей об университете ИТМО с сервиса Яндекс.Карты. Разметка полярности собранных отзывов осуществлялась вручную. Негативным отзывам присваивалось значение -1, нейтральным – значение 0, а положительным – значение 1. Затем значения, полученные после применения выбранных библиотек для анализа тональности, сравнивались с оценками тональности, полученными вручную. Для оценки точности полученных результатов составлялись матрицы ошибок и рассчитывались метрики точности, полноты и F-меры. По результатам пилотного эксперимента показатели точности у библиотек TextBlob, VADER и Dostoevsky составили 82%, 70% и 70% соответственно.

В рамках основного эксперимента использовались две выборки из сети Интернет – выборки русскоязычных (размер выборки – 226834) и англоязычных (размер выборки – 498) “твитов”. Элементы выборки уже были классифицированы по бинарному методу. Данные выборки были предобработаны перед проведением сентимент-анализа: были убраны имена пользователей, хештеги и пунктуационные символы. Были получены следующие результаты в основном эксперименте:

– Для анализа русскоязычных отзывов по результатам эксперимента лучшие показатели точности были у библиотеки для анализа тональности текста Dostoevsky – 41,99%.

– Для анализа англоязычных отзывов по результатам эксперимента рекомендуется применять VADER, так как он точнее остальных рассмотренных библиотек делал вывод о тональности текстов. Точность библиотеки VADER составила 60,64%.

Выводы. Проведены пилотный и основной эксперимент для сравнения библиотек для анализа тональности текстов VADER, TextBlob и Dostoevsky. Пилотный эксперимент показал, что библиотека TextBlob точнее всего определила тональность для русскоязычной выборки. Результаты основного эксперимента для текстов на русском языке показали, что самой высокой точностью обладает библиотека Dostoevsky, а для англоязычных – библиотека VADER. Проведенный сравнительный анализ позволит минимизировать появление проблем при классификации текстовых данных и качественно ранжировать информацию с применением наиболее эффективных инструментов.

Список использованных источников:

1. Bhatt R., Gupta P. Sentiment Analysis // Indian Journal of Science and Technology. – 2019. – Vol. 12. – P. 1-6.
2. Kotu V., Deshapande B. Model Evaluation // Predictive Analytics and Data Mining. – 2015. – P. 257-274.