0.1. *Тухтина Е.А., Пахомова К.И.* Библиотека VKsentiment как инструмент для анализа тональности текстовых данных социальных сетей

Анализ тональности текста — задача компьютерной лингвистики, заключающаяся в определении эмоциональной окраски текста. Относится к задаче классификации, где необходимо определить, является ли выраженное в тексте мнение позитивным, негативным или нейтральным. В данной работе авторы предлагают автоматизировать процесс анализа тональности публикаций социальных сетей посредством разработки библиотеки на языке Python.

Зачастую исследователи предпочитают выбирать уже готовые решения для интеллектуального анализа больших объемов текстовых данных. В настоящее время в свободном доступе существуют две Python библиотеки, позволяющие оценить тональность русскоязычного текста посредством моделей машинного обучения: DeepPavlov [1] и Dostoevsky. Однако для решения конкретных прикладных проблем функционала описанных библиотек недостаточно. Исследователю потребуется самостоятельно создавать отдельные функции для сбора данных и обобщения полученных моделью результатов. Библиотека VKsentiment позволяет оптимизировать эти процессы, предоставляя удобный функционал для оперативной обработки больших массивов текстовых данных и получения информативных результатов.

Библиотека VKsentiment представляет собой набор готовых функций для анализа тональности текстовых данных социальных сетей. Проект состоит из двух основных модулей:

- comments_scraping;
- comments_analysis.

Тематические сообщества социальной сети «ВКонтакте» являются источниками пользовательских мнений практически на любую тематику. Модуль comments_scrapping позволяет извлечь комментарии из сообществ «ВКонтакте», в которых открыт доступ к комментированию записей. Реализован с помощью пакета vk_api. Основная логика содержится в методе get_comments класса CommentsScrapper. В качестве аргументов метод принимает числовой идентификатор сообщества и число комментариев, которое необходимо получить. На выходе создается текстовый файл формата .tab с полученными комментариями. Если исследователю требуется проанализировать данные из другой социальной сети, то необходимо воспользоваться соответствующим API и представить данные в формате .tab.

Модуль comments_analysis позволяет провести анализ тональности текстовых данных и обобщение полученных результатов. Класс CommentsResearcher содержит метод get_sentiment, который в качестве аргумента принимает текстовый файл с данными формата .tab. В результате выполнения метода создается файл с отчетом в формате .txt. В нем содержится информация о распределении текстовых сообщений по трем классам: негативные, позитивные, нейтральные. В качестве классификатора используется модель FastTextSocialNetworkModel из библиотеки Dostoevsky, обученная на наборе данных RuSentiment [2].

Таким образом, библиотека VKsentiment позволяет значительно упростить процесс оценки тональности текстовых данных социальных сетей. Исследователь может провести анализ собственной выборки данных, или воспользоваться модулем для извлечения комментариев из сообществ «ВКонтакте», подобрав сообщества, тематика которых соответствует объекту исследования.

Список литературы

[1] Burtsev M., Seliverstov A., Airapetyan R. et al. DeepPavlov: Open-source library for dialogue systems // Proc. ACL 2018, System Demonstrations. 2018. P. 122–127.

[2] Rogers A., Romanov A., Rumshisky A. et al. RuSentiment: An enriched sentiment analysis dataset for social media in Russian // Proc. 27th Intern. Conf. on Computational Linguistics. 2018. P. 755–763.