УДК 004.89

СРЕДСТВА АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА НА РУССКОМ ЯЗЫКЕ

Тирский Б.Д. (Университет ИТМО)
Научный руководитель – к. п. н., доцент Авксентьева Е. Ю.
(Университет ИТМО)

Анализ тональности — это подзадача обработки естественного языка, ключевой целью которой является классификация текста в соответствии с содержащимся в нем настроением. Основные подходы обычно направлены на бинарную классификацию текстов как положительных, так и отрицательных. В некоторых случаях шкала классификации включает еще один класс нейтральных текстов. Более продвинутые подходы пытаются идентифицировать эмоциональные состояния, связанные с написанным текстом, например, страх, гнев, печаль или счастье. Также существуют методы, которые сводятся к тому, чтобы выставлять оценку для текстов, например, от 0 для негативных текстов до 5 для позитивных текстов, тем самым сводя анализ настроений к задаче регрессии. Аспектный анализ настроений – это подраздел анализа настроений, где основной задачей является выявление настроений по отношению к конкретному аспекту, связанному с заданной целью. Подходы к анализу настроений можно в общих чертах разделить на три типа.

Первый — это подходы, основанные на правилах, которые определяют вручную правила классификации, плюс на словарях настроений. В этих правилах обычно используются эмоциональные ключевые слова и их совместное использование в текстах с другими ключевыми словами [1] [2]. Несмотря на отличную производительность в пределах узкой предметной области, методы, основанные на правилах, страдают слабой способностью к обобщению и масштабированию. Кроме того, их создание, как правило, чрезвычайно трудоемко, особенно в тех случаях, когда недостаточно качественных словарей настроений.

Второй тип - подходы, основанные на машинном обучении, которые работают на автоматическом извлечении признаков из текста и последующем применении алгоритмов классификации машинного обучения. Наивная байесовская классификация, дерево решений, логистическая регрессия и метод опорных векторов могут быть определены в качестве базовых алгоритмов классификации полярностей. В последние годы методы глубокого обучения привлекли внимание исследователей благодаря их способности значительно превосходить традиционные методы в задаче анализа настроений [3]. Это также подтверждается успешным использованием сверточных нейронных сетей (CNNS) и рекуррентных нейронных сетей (RNNs). Одна из ключевых характеристик системы машинного обучения — это автоматическое извлечение признаков из текстов. Простые подходы обычно используют модель набора слов для представления текстов в векторном виде. Более сложные системы используют семантические модели для генерации связей между словами, например, Word2Vec – это модель позволяет представить слова в виде векторов с фиксированными размерностями на основе их семантической близости, GloVe и FastText. Существуют также методы, которые генерируют связи на уровне предложения или абзаца и специально нацелены на обучение в различных задачах обработки естественного языка, например, ELMo, Universal sentence encoder, а также Bidirectional Encoder Representations from Transformers (BERT) – это модель глубокого обучения, использующая механизм трансформера для представления текста, BERT способен учитывать контекст и зависимость между словами в тексте. Одним из основных недостатков использования этих систем для генерации встраиваний является тот факт, что для обучения им требуются большие массивы текстов. В целом, эта проблема актуальна для всех методов машинного обучения, поскольку все алгоритмы классификации требуют аннотированных наборов данных для обучения и большое количество времени для обработки датасетов.

Третий - гибридные подходы, которые сочетают в себе как подходы, основанные на правилах, так и подходы, основанные на машинном обучении. Например, Кумар и его коллеги разработали persian sentiment analysis framework, которая интегрировала лингвистические правила и модули CNN и LSTM для классификации настроений [4]. Мескеле и Фрасинкар предложили гибридную модель для анализа настроений на основе аспектов, ALDONAr, которая объединяет онтологию домена настроений для сбора информации, BERT для

получения вложений слов и два уровня CNN для улучшения классификации настроений [5]. Модель достигла показателей точности 83,8% и 87,1%. Подобно подходам, основанным на правилах, языковые модели, как правило, широко используются в гибридных подходах. С одной стороны, сочетание преимуществ подходов, основанных на правилах, и подходов, основанных на машинном обучении, обычно позволяет получать более точные результаты. С другой стороны, комбинированные и гибридные подходы также сталкиваются с проблемами и ограничениями, которые встречаются и в машинном обучении.

Выводы

Проведен анализ различных методов по выявлению тональности, а также было проведено сравнение различных методов и алгоритмов.

Литература

1. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai and A. Kappas, "Sentiment strength detection in short informal text", J. Amer. Soc. Inf. Sci. Technol., vol. 61, no. 12, pp. 2544-2558, Dec. 2010.

2. C. Gómez-Rodríguez, I. Alonso-Alonso and D. Vilares, "How important is syntactic parsing accuracy? An empirical evaluation on rule-based sentiment analysis", Artif. Intell. Rev., vol. 52, no. 3, pp. 2081-2097, Oct. 2019.

3. D. Tang, B. Qin and T. Liu, "Deep learning for sentiment analysis: Successful approaches and future challenges", Wiley Interdiscipl. Rev. Data Mining Knowl. Discovery, vol. 5, no. 6, pp. 292-303, Nov. 2015.

4. A. Kumar, K. Srinivasan, W.-H. Cheng and A. Y. Zomaya, "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data", Inf. Process. Manage., vol. 57, no. 1, Jan. 2020, [online] Available: http://www.sciencedirect.com/science/article/pii/S0306457319306934.

5. D. Meškelė and F. Frasincar, "ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model", Inf. Process. Manage., vol. 57, no. 3, May 2020, [online] Available: http://www.sciencedirect.com/science/article/pii/S0306457319310222.