Автоматические методы текстовой атрибуции в диагностировании склонности личности к аутоагрессивному поведению

Анастасия Д. Комратова
НИУ «Высшая школа экономики», Нижний Новгород,
komratova.asia@yandex.ru

Аннотация. Работа посвящена проблеме автоматической идентификации предрасположенности личности к аутоагрессивному поведению. В ходе исследования был произведен сравнительный анализ различных подходов к диагностике личностных характеристик автора текста, основанных на машинном обучении. Исследование проводилось на материале русскоязычного корпуса с целью определения наиболее эффективного метода выявления аутоагрессии у автора письменного текста.

Проблема диагностики личностных особенностей автора письменного текста изучается исследователями на протяжении нескольких десятилетий, однако в последние годы в научной сфере возник особый интерес к выявлению поведения, отклоняющегося от социальных норм. Такой интерес вызван ростом преступности и смертности среди людей, страдающих различными формами девиантного поведения и как следствие, необходимостью более глубокого анализа данной проблемы. Отдельной междисциплинарной научной проблемой является проблема изучения склонности человека к аутоагрессивному поведению. Под аутоагрессивным поведением понимают действия, направленные на нанесение какого-либо ущерба своему соматическому или психическому здоровью. Вариант агрессивного поведения, при котором субъект и объект агрессии совпадают [1].

Целью данной работы является решение задачи из области диагностической атрибуции, а именно создание компьютерной модели для выявления склонности личности к аутоагрессивному поведению по речевому материалу.

Наиболее распространенным подходом к решению диагностической задачи текстовой атрибуции является подход, основанный на выделении из текста статистики по двум группам параметров: формальным (собственно языковым) и неформальным (несобственно языковым). В процессе диагностики, выявляются корреляции между текстом и выбранными параметрами. На основе выявленных корреляций создаются модели, которые по текстовым характеристикам делают заключение о личностных особенностях (поле, возрасте, психологической акцентуации) автора [4].

В ходе данного исследования был собран корпус текстов, который содержит речевой материал двух русскоязычных социальных групп: людей с аутоагрессивной акцентуацией и без неё. Материал был взят из открытых источников (речь людей с аутоагрессивной

акцентуацией: Presuicidal signals dataset Twitter (URL:https://data.mendeley.com/datasets/86v3z38dc7/1); речь людей без акцентуации: URL:https://vk.com/blog_27325, URL:https://mel.fm/) и из источников, предоставленных сотрудниками Приволжского исследовательского медицинского университета с последующей анонимизацией материала. Объем корпуса составил около 100 миллионов токенов.

Для выявления наиболее эффективного автоматического метода диагностики аутоагрессивных тенденций у автора текста были отобраны пять способов классификации: 2 традиционных алгоритма классификации (Logistic Regression, Random Forest) и 3 модели нейронных сетей (CNN, RNN (LSTM), Transformer (BERT)). Таким образом, на каждый способ классификации, кроме классификатора на основе BERT, делалось по две модели диагностики – с использованием вектора, составленного из статистических значений выбранных характеристик (частота встречаемости различных частей речи (местоимений, в частности местоимений я-группы; глаголов; прилагательных, предлогов); индекс удобочитаемости; лексическое разнообразие; тональность текста; средняя длина предложений), и с использованием вектора TF-IDF. Для реализации модели на основе BERT была использована предобученная модель rubert_base_cased_sentence, поэтому векторизация текстов производилось тем же способом, что и при предобучении (с помощью функций библиотеки PyTorch).

При оценке и сравнении эффективности работы представленных методов использовались метрики, которые традиционно применяются при оценке качества классификации: Precision, Recall и F-мера (Таблица 1,2).

	TF-IDF			**вектор, состоящий из значений лингвистических признаков**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
Random Forest	0.87	0.84	0.84	0.95	0.84	**0.89**
Logistic Regression	0.96	0.95	**0.95**	0.89	0.83	0.86
RNN (LSTM)	0.97	0.97	**0.97**	0.95	0.91	0.93
CNN	0.89	0.86	**0.85**	0.77	0.78	0.78

Таблица 1. Результаты классификации

	Precision	Recall	F-score
BERT	0.98	0.98	0.98

Таблица 2. Результаты классификации с использованием предобученной модели на основе BERT

Таким образом, в результате проведенного исследования были сделаны следующие выводы:

1. Лучше всего с задачей выявления аутоагрессивных тенденций справляется предобученная модель глубокого обучения на базе BERT
2. В основном модели, использовавшие в качестве способа векторизации числовые значения лингвистических параметров показывают точность классификации ниже, чем модели, построенные на тех же алгоритмах, но с использованием вектора TF-IDF
3. При решении задачи диагностики аутоагрессивных тенденций модели нейронных сетей в целом показывают результаты лучше, чем традиционные алгоритмы классификации.

Перспективами работы является создание универсального диагностического инструмента с удобным интерфейсом, который можно было бы использовать для выявления аутоагрессии на ранних стадиях.

Список литературы

1. Амбрумова А.Г., Трайнина Е.Г., Ратинова Н.А. Аутоагрессивное поведение подростков с различными формами социальных девиаций // VI Всероссийский съезд психиатров, Томск, 24-26 окт. 1990 г.: тез. докл. / Всерос. науч. мед. общество психиатров. – М., 1990. – Том 1. – С.105-106.
2. Литвинова Т. А., Литвинова О. А. Языковые особенности русскоязычных текстов лиц, совершивших суицид, и лиц с высоким риском аутоагрессивного поведения //Studia Humanitatis. – 2017. – №. 4. – С. 18.
3. Ji S. et al. Suicidal ideation detection: A review of machine learning methods and applications //IEEE Transactions on Computational Social Systems. – 2020. – Т. 8. – №. 1. – С. 214-226.
4. Litvinova T. et al. Profiling a set of personality traits of text author: what our words reveal about us //Research in Language. – 2016. – Т. 14. – №. 4. – С. 409-422.