Построение алгоритма аннотирования русскоязычных текстовых данных социальных сетей с использованием переносимого обучения

Д.С. Баканов
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
dima.bakanov.1999@mail.ru

А.В. Куприянов
Самарский национальный исследовательский университет им. академика С.П. Королева
ИСОИ РАН
Самара, Россия
akupr@ssau.ru

***Аннотация*—В данной работе рассматриваются способы построения алгоритма аннотирования русскоязычных текстов из социальных сетей. В качестве аннотирования будем понимать оценку эмоционального окраса текста. Статья затрагивает как классические базовые методы статистического обучения, так и современные методы глубокого обучения, основанные на переносимом обучении и трансформерах. В заключении строится модель, которая совмещает модель трансформера и статистическую модель машинного обучения градиентного бустинга. Актуальность данной работы заключается в создании легковесной и независимой от тематики модели, которую можно использовать для анализа текстового содержимого постов в социальных сетях.**

Ключевые слова— обработка естественного языка, трансформер, переносимое обучение, анализ социальных сетей, TF-IDF, статистическое обучение, оценка эмоционального окраса, BERT

1. Введение

В наши дни все большую роль играют социальные сети, которые становятся местом притяжения все большего количества людей с разными интересами. Поэтому социальные сети могут служить хорошим местом при проведении социальных экспериментов.

Эмоции – это быстрые и короткие реакции человеческих чувств, их недопонимание при проведении рекламных компаний, маркетинга может повлечь за собой большие финансовые потери [1]. Уникальность социальной сети заключается в том, что сами пользователи создают контент, которого становится все больше. Но с прогрессом информационных технологий и развитием технологий машинного обучения и больших данных можно автоматически анализировать такую информацию [2].

На данный момент существует большое количество моделей, которые предсказывают эмоциональный окрас для постов с отзывами о продуктах и услугах [3, 4]. В данной работе описывается полный цикл построения алгоритма аннотирования данных, который не зависит от тематики постов в социальных сетях, что можно использовать при анализе любого поста в социальных сетях, с использованием трансформера и градиентного бустинга.

2. Исходные данные

В качестве обучающего набора данных были использованы следующие обучающие наборы: RuReviews [3], RuTweetCorp [4], Kaggle Russian News Dataset [5]. Всего набор данных насчитывает 358190 образцов. В данном наборе фигурирует три класса эмоционального окраса: негативный (-1), нейтральный (0) и положительный (1). Ниже представлено распределение образцов по классам (см. Рисунок 1).

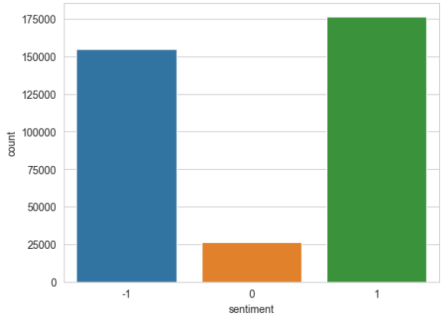

Рис. 1 Распределение образцов набора данных по классам

Из рисунка 1 можно видеть, что наблюдается сильный дисбаланс классов. Данную проблему стоит учесть при выборе метрики качества и моделей, которые устойчивы к дисбалансу классов.

3. Разведывательный анализ

Перед проведением разведывательного анализа данные были предобработаны: из текста была удалена HTML-разметка, текст приведен к общему регистру, удалены знаки препинания и стоп-слова.

На рисунке 2 приведена визуализация данных из обучающего набора, которая была сделана при помощи применения метода латентного семантического анализа (индексирования) [6].

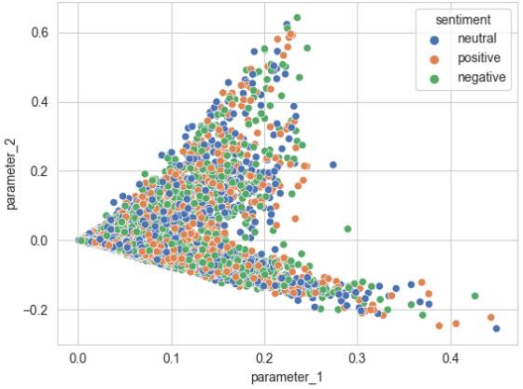

Рис. 2 Векторное представления набора данных на плоскости

Как можно видеть данные не являются линейно разделимыми и находятся вперемешку друг с другом.

4. Статистическое машинное обучение при анализе эмоционального окраса русскоязычных текстов

В качестве базового решения задачи классификации можно выбрать статистические методы машинного обучения, которые хорошо себя показывают при решении задач в области обработки естественного языка [6].

Перед обучением данные были приведены в векторное представление:

- удалены пустые значения;
- разбиты на N-граммы;
- созданы векторы на основе метрики TF-IDF.

Чтобы учесть дисбаланс классов при оценке моделей для задачи многоклассовой классификации, была использована метрика F1-мера.

Была исследована зависимость метрики точности модели от выбора N-грамм (см. Таблицу I).

Таблица I. Зависимость F1-меры для каждой модели в зависимости от выбора N-граммы

Модель машинного обучения	**N-грамма**		
	Униграмма	***Биграмма***	***Триграмма***
Наивный байесовский классификатор	0,67	0,59	0,48
Метод опорных векторов (SVM)	0,45	0,28	0,25
Линейный метод опорных векторов	0,67	0,57	0,43
Логистическая регрессия	0,67	0,55	0,4
Деревья решений	0,62	0,47	0,38
Градиентный бустинг (CatBoost)	0,63	0,48	0,37

Точность модели не превосходит 0,67 и с увеличением N-граммы уменьшается, что показывает значимость каждого слова при анализе эмоционального окраса.

5. Трансформеры и переносимое обучение

Трансформеры – класс моделей глубокого обучения, которые с использованием механизма внутреннего внимания решают задачу преобразования последовательности в другую последовательность. Рекомендуемым методом обучения трансформеров является переносимое обучение (transfer learning), которое заключается в настройке уже обученной модели на своем наборе данных [7].

A. Выбор BERT-модели

В качестве модели для дообучения была выбрана легковесная модель DeepPavlov/rubert-base-cased. Данная модель отличается своей легковесностью и обучена на русскоязычных статьях Википедии [8].

Б. Обучение и оценка точности

Для точной настройки трансформера были использованы следующие параметры:

- оптимизатор: Adam;
- размер пакета: 32;
- learning rate: 2e-5;
- размер вектора: 128.

На рисунке 3 показан график потерь на обучающем и проверочном наборе для 10 эпох обучения.

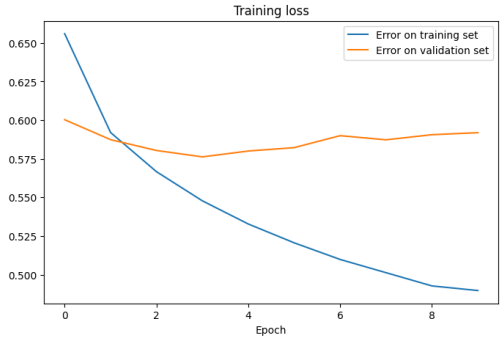

Рис. 3 Потери при обучении и проверки трансформера

На тестовых данных F1-мера трансформера оказалась равной 0,7.

B. Сочетание с моделями статистисческого обучения

После получения векторов трансформера к ним были применены статистические модели машинного обучения. Самую большую точность показал градиентный бустинг (CatBoost). Точность по F1-мере составила 0,76. Данная точность превышает точности предсказаний статистических моделей машинного обучения, а также модели трансформера.

6. Заключение

В данной работе был рассмотрен метод построения алгоритма аннотирования текстовых данных социальных сетей. Полученный алгоритм обладает рядом особенностей:

- независимость оценки эмоционального окраса от тематики поста;
- использование легковесных моделей;
- точность составила по F1-мере 0,76, что превышает показатели точности тяжеловесных моделей при анализе эмоционального окраса русскоязычных текстов [3].

Литература

[1] Канарская Л. Как работает эмоциональный контент в SMM (на примере популярных групп «Вконтакте») [Электронный ресурс]. – Режим доступа: https://texterra.ru/blog/kak-rabotaet-emotsionalnyy-kontent-v-smm-na-primere-populyarnykh-grupp-vkontakte.html (дата обращения: 06.06.2022).

[2] Рыцарев, И.А. Кластеризация медиаконтента из социальных сетей с использлованием технологий Big Data / И.А. Рыцарев, Д.В. Кирш, А.В. Куприянов // Компьютерная оптика. – 2018. – Т. 42, № 5. – С. 921-927. .

[3] Smetanin, S. "Sentiment Analysis of Product Reviews in Russian using Convolutional Neural Networks" / S. Smetanin, M. Komarov // 2019 IEEE 21st Conference on Business Informatics (CBI). – 2019. – P. 482-486.

[4] Рубцова, Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора / Ю. Рубцова // Инженерия знаний и технологии семантического веба. – 2012. – Т. 1. – С. 109-116.

[5] Sentiment Analysis in Russian [Электронный ресурс]. — Режим доступа: https://www.kaggle.com/c/sentiment-analysis-in-russian (01.06.2022).

[6] Маннинг, Д. Введение в информационный поиск / Д. Маннинг, Р. Прабхакар, Х. Шютце. – СПб.: ООО «Диалектика», 2020. – 528 с.

[7] Vaswani, A. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin // Computer Science. – 2017.

[8] Kuratov, Y. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language / Y. Kuratov, M. Arkhipov // Computer Science, Linguistics. –2019.