Влияние саркастических признаков на анализ эмоциональной тональности русскоязычных текстов

И. С. Мухин, Н. А. Харковчук, Е. Ю. Авксентьева
Университет ИТМО
ilya.mukhinn@yandex.ru

Аннотация. **В статье представлен анализ роли саркастических признаков в задаче трёхклассовой классификации тональности на корпусе русскоязычных текстов. В работе сопоставляются линейные базовые модели и их модификации с включением признаков, извлечённых из модели детекции сарказма. Оценка проводится как при случайном разбиении данных, так и в условиях сдвига между датасетами. Полученные результаты не свидетельствуют об универсальном преимуществе таких признаков: при более строгом оценивании между наборами данных сарказм-модифицированные модели в целом не превосходят наиболее сильную базовую конфигурацию. Вместе с тем выявлен ограниченный, однако значимый эффект, который проявляется в чистом приросте исправленных ошибок при анализе тональности, улучшении на подмножестве некоторых спорных текстов и доменно-специфических преимуществах, наиболее выраженных в обзорах фильмов. В целом влияние саркастических признаков носит локальный, а не универсальный характер: в ряде случаев они действительно способствуют улучшению качества, однако текущий подход интеграции остаётся чрезмерно общим и чувствительным к домену, а также недостаточно эффективным в условиях доминирования нейтрального класса.**

Ключевые слова: анализ тональности с учётом сарказма; перенос между доменами; shortcut-признаки; детекция сарказма; русскоязычный текст

I. Введение

Связь между сарказмом и тональностью довольно тесна, однако недостаточно устойчива, чтобы её можно было свести к тривиальной задаче инверсии полярности. Сарказм действительно способен изменить буквальный тон высказывания, однако степень такого влияния зависит от контекста, жанра и действующих дискурсивных норм. Современные достижения в области автоматического распознавания сарказма связаны с применением графовых моделей, переноса обучения и участием в shared tasks, однако накопленные результаты показывают, что сарказм зачастую не может быть описан единым универсальным набором признаков.

Также стоит отметить, что проблема выходит за пределы только лишь распознавания сарказма. В более широком смысле высокие результаты на случайных разбиениях наборов данных не всегда свидетельствуют о действительно качественном решении задачи. Нередко они отражают наличие shortcut-признаков, перекосов распределения или доменно-специфических закономерностей, которые маскируют ограничения модели. Подходы, ориентированные на динамический бенчмаркинг, функциональное тестирование и оценку в условиях сдвига распределения, фактически формулируют то, что методы, успешные в более мягких режимах тестирования, могут заметно терять в качестве при изменении условий оценки. Анализы трансфера в кросс-доменных задачах оценки тональности обнаруживают схожую уязвимость. Переносимость в таких случаях зависит как от способа представления данных, так и от выбранного метода оценки.

Именно в этом контексте формулируется основной вопрос настоящей работы: способствует ли явное включение саркастических признаков улучшению классификации тональности при переходе от случайного разбиения данных к более жёсткой схеме тестирования. Далее будет показано, как такие признаки ведут себя в двух режимах оценки, насколько их вклад проявляется в агрегированных метриках и каким образом он отражается на структуре ошибок. Особое внимание будет уделено случаям, в которых внедренные признаки оказываются полезными, а также тем условиям, при которых их эффект остаётся ограниченным или неоднозначным. Представленный подход позволяет оценить не только общий потенциал введения признаков сарказма, но и границы их применимости с учётом класса и домена целевой задачи.

II. Литературный обзор

Современные работы в области детекции сарказма показывают, что использование подобных сигналов в задачах анализа тональности связано как с определёнными возможностями, так и с рядом существенных ограничений. Работы, сопоставляющие воспринятый и намеренный сарказм, подчёркивают значимость объекта аннотации и прагматического контекста [1]. Подходы, ориентированные на перенос моделей сарказма, в ряде случаев способствуют повышению устойчивости, однако сохраняют чувствительность к контекстуальному несоответствию [2]. Масштабное исследование iSarcasmEval, в свою очередь, демонстрирует, что неоднозначность интерпретации сарказма сохраняется в разных языках и экспериментальных условиях [3]. В совокупности эти результаты позволяют рассматривать сарказм как сложное и структурированное явление, которое едва ли может быть сведено к набору бинарных маркеров с универсальной переносимостью.

Исследования, посвящённые взаимодействию сарказма и тональности, показывают, что влияние добавочных признаков в таких задачах носит неоднородный характер. В частности, обнаружение

противоречивого тона может способствовать распознаванию сарказма [4], однако компоненты, учитывающие тональность, демонстрируют различную эффективность в зависимости от класса и условий применения [5]. В настоящей работе рассматривается смежная проблема: каким образом добавление саркастических признаков влияет на качество классификации тональности, а также в каких случаев их использование связано с исправлением ошибок, а в каких – с их появлением.

Выбор схемы оценивания требует отдельного обоснования, поскольку именно он во многом определяет, насколько корректно можно судить об устойчивости модели. Работы, посвящённые ложным корреляциям и влиянию распределения данных, показывают, что особенности структуры датасета способны существенно смещать оценку качества модели [6-8]. Исследования в области бенчмаркинга, в свою очередь, подчёркивают важность таких режимов тестирования, которые позволяют выйти за пределы упрощающего предположения об IID-характере (independent and identically distributed) данных [9-11]. В задачах анализа тональности работы кросс-доменных переносов также демонстрируют существенное снижение качества и заметные различия между классами даже тогда, когда агрегированные показатели в более мягких условиях выглядят устойчивыми [12, 13]. Применяемая в данной работе групповая схема оценивания между различными датасетам позволяет рассматривать чувствительность к сдвигу распределения как одну из ключевых характеристик модели.

III. Материалы и методы

Эмпирическая база исследования включает два типа данных. Первый из них представляет собой русскоязычный корпус сарказма объёмом 15146 текстов, сбалансированный по наличию и отсутствию саркастических выражений, по 7573 примера в каждом классе. Проведённый предварительный анализ указывает на высокую лексическую различимость классов: гибридная модель, сочетающая поверхностные и лексические признаки, достигает F1 = 0.9617, тогда как модель, использующая только поверхностные признаки, показывает существенно более низкий результат (F1 = 0.7362). Вместе с тем корпус характеризуется выраженной тематической неоднородностью, а доля сарказма в отдельных тематических категориях варьирует в широких пределах, приблизительно от 0.0104 до 0.9325. Такое сочетание высокой разделимости и тематической дифференциации может приводить к завышенной оценке качества при случайном разбиении данных, оставляя открытым вопрос об обобщающей способности модели в кросс-доменных условиях.

Вторую часть материалов составляет датасет для классификации тональности, включающий 403478 примеров с метками негативного (110638), нейтрального (114964) и позитивного (177876) классов, охватывающих различные домены.

В экспериментальной части рассматривались три базовые линейные модели, различающиеся типом используемого признакового представления: word_logreg, основанная на словарных признаках; char_logreg, использующая символьные признаки; и hybrid_surface_logreg, сочетающая лексические и поверхностные признаки. На основе последней были построены две модификации с учётом сведений о сарказме. Модель sarcasm_prob дополняет исходное представление скалярной оценкой вероятности сарказма, тогда как sarcasm_bundle использует расширенный набор признаков, извлечённых из модели детекции сарказма. Таким образом, сравнение охватывает как различия между базовыми типами представления текста, так и два способа включения саркастических признаков в задачу классификации тональности.

Оценивание проводилось в двух режимах. Режим random_split соответствует стандартному случайному разбиению данных на обучающую и тестовую части и тем самым задаёт более близкие к IID условия проверки. Режим dataset_group_test_fold, напротив, предполагает разбиение с группировкой по отдельным датасетам, так что тестирование выполняется на данных, выделенных по источнику или домену и в большей степени отражающих сдвиг распределения. Такое сопоставление позволяет оценить, насколько результаты модели зависят от упрощённого сценария случайного разбиения и в какой мере сохраняется качество при переходе к более сложным условиям обобщения между наборами данных.

IV. Результаты

Сопоставление результатов приведено в табл. I, однако наибольший интерес представляет не столько различие между моделями в пределах одного режима оценивания, сколько изменение результатов при переходе от случайного разбиения данных к более требовательной схеме тестирования. В режиме random_split показатели всех моделей оказываются практически одинаковыми: модель hybrid_surface_logreg + sarcasm_prob достигает macro-F1 = 0.7095, тогда как гибридная базовая модель и word_logreg показывают 0.7092 и 0.7090 соответственно. При таком уровне расхождений случайное разбиение не даёт достаточных оснований для уверенного сопоставления моделей по качеству.

ТАБЛИЦА I. Результаты классификации тональности.

Протокол	Модель	Macro-F1	Accuracy
random_split	hybrid_surface_logreg + sarcasm_prob	0.7095	0.7351
	hybrid_surface_logreg + sarcasm_bundle	0.7093	0.7348
	hybrid_surface_logreg (baseline)	0.7092	0.7348
	word_logreg (baseline)	0.7090	0.7340
	char_logreg (baseline)	0.7078	0.7327
dataset_group_test_fold	word_logreg (baseline) hybrid_surface_logreg +	0.4418	0.4949
	hybrid_surface_logreg + sarcasm_prob	0.4296	0.4906
	hybrid_surface_logreg (baseline)	0.4293	0.4901
	sarcasm_bundle	0.4281	0.4921
	char_logreg (baseline)	0.4264	0.4831

При переходе к групповому тестированию показатели всех моделей заметно снижаются. В этих условиях наилучший результат демонстрирует базовая

модель word_logreg с macro-F1 = 0.4418 и accuracy = 0.4949. Тогда как сарказм-модифицированные версии гибридной модели показывают близкие, но несколько более низкие значения macro-F1. Наиболее показательным здесь является различие между результатами, полученными при случайном разбиении данных и в условиях групповой схемы оценивания. Так, значение macro-F1 для hybrid_surface_logreg уменьшается с 0.7092 до 0.4293 (на 0.2799), для word_logreg – с 0.7090 до 0.4418 (на 0.2672), а для char_logreg – с 0.7078 до 0.4264 (на 0.2814).

Хотя универсального преимущества сарказм-признаков не наблюдается, их влияние вряд ли можно считать случайным. Среди всех 403478 примеров группового теста в 8075 случаях предсказания базовой и сарказм-осведомлённой гибридной моделей расходятся (2.00%). Из них 3527 изменений оказываются полезными, тогда как 2767 приводят к ухудшению, что даёт чистый положительный баланс +760. Иными словами, вмешательство затрагивает относительно небольшую долю примеров, но не сводится к простому добавлению шума.

Несмотря на отсутствие общего выигрыша от включения саркастических признаков, их вклад едва ли следует считать случайным. Среди всех 403478 примеров группового тестирования предсказания базовой и сарказм-модифицированной гибридной моделей различаются в 8075 случаях (2.00%). В 3527 из этих случаев изменение предсказания оказывается корректным, тогда как в 2767 случаях оно связано с ошибочным изменением метки, что соответствует чистому приросту в 760 корректировках тональности.

Анализ подмножества расхождений позволяет уточнить, каким образом включение саркастических признаков влияет на предсказания модели. На тех примерах, где предсказания базовой и модифицированной моделей различаются (табл. II), accuracy увеличивается с 0.3427 до 0.4368, а macro-F1 — с 0.3400 до 0.3705. Вместе с тем это улучшение распределяется по классам неравномерно. Наиболее заметный прирост наблюдается для позитивного класса, где F1 возрастает с 0.2536 до 0.6009, тогда как для негативного и нейтрального классов значения снижаются.

ТАБЛИЦА II. МЕТРИКИ ОЦЕНКИ РАСХОЖДЕНИЙ АНАЛИЗА ТОНАЛЬНОСТИ.

Подход	Acc	Macro-F1	F1neg	F1neu	F1pos
Baseline	0.3427	0.3400	0.3686	0.3978	0.2536
Sarcasm-Aware	0.4368	0.3705	0.3313	0.1793	0.6009

Влияние добавочных признаков сарказма различается в зависимости от домена. Как показано в табл. III, в Kinopoisk обзорах наблюдается наиболее выраженный положительный сдвиг (+1114), тогда как обзорах одежды и иных доменах оценки тональности их включение несет с отрицательный эффект. На уровне отдельных доменных выборок улучшение accuracy фиксируется в сегменте фильмов (movie): увеличение с 0.6816 до 0.6900, хотя значение macro-F1 в данном случае остаётся практически неизменным при переходе из 0.5563 в 0.5558.

ТАБЛИЦА III. РАСПРЕДЕЛЕНИЕ ИСПРАВЛЕННЫХ И ОШИБОЧНО ИЗМЕНЁННЫХ ОЦЕНОК ПО ДОМЕНАМ

Домен	N	Исправлено	Ошибочно изменено	delta	Доли изменений
Kinopoisk reviews	131573	2842	1728	+1114	2.16% / 1.31%
Different sentiment samples	213330	532	831	-299	0.25% / 0.39%
Clothing reviews	58575	153	208	-55	0.26% / 0.36%

Отдельно следует отметить трудности, связанные с оценкой нейтрального класса, который в рассматриваемой задаче остаётся наименее устойчивым с точки зрения качества распознавания. В ряде доменов recall нейтрального класса остаётся крайне низким у обеих моделей, в частности в geo (0.0421), perekrestok (0.0434) или news (0.0158). Поскольку включение саркастических признаков связано со смещением предсказаний в сторону позитивного класса, в доменах с преобладанием нейтральных или слабо поляризованных данных это может приводить к дополнительному снижению качества классификации.

V. ОБСУЖДЕНИЕ

По полученным результатам возможно сделать следующие выводы. Отсутствие общего улучшения при групповом тестировании не обязательно означает, что сарказм-ориентированные признаки неинформативны для классификации тональности. Напротив, положительный чистый прирост исправленных ошибок и улучшение на подмножестве спорных выражений указывают на то, что такие признаки действительно вносят дополнительную информацию. Однако их вклад остаётся ограниченным, зависит от домена и неоднозначно проявляется для разных классов.

Также следует отметить, что наблюдаемый эффект не совпадает с наиболее очевидным ожиданием. Можно было предположить, что учёт сарказма будет прежде всего способствовать лучшему распознаванию негативных высказываний, поскольку саркастические конструкции, как правило, выражают скрытую отрицательную оценку при внешне положительной форме. Однако результаты демонстрируют иное поведение моделей: в рассматриваемом алгоритме оценки саркастические признаки выступают не как средство прямой инверсии полярности, а как более общий индикатор субъективности и оценочной маркированности текста. Вследствие чего их включение ведёт не столько к усилению негативных предсказаний, а к смещению части наблюдений в сторону позитивного класса (как негативных, так и нейтральных).

Такая интерпретация согласуется и со свойствами исходного корпуса сарказма, который при высокой лексической разделимости остаётся тематически неоднородным. В этих условиях извлекаемые признаки отражают не только скрытую негативность, но и более широкий набор характеристик экспрессивной речи, что также помогает объяснить доменную неоднородность эффекта: наиболее отчётливое улучшение наблюдается в обзорах фильмов, где ирония и оценочная инверсия, вероятно, играют более заметную роль, тогда как в других доменах тот же сигнал оказывается менее выраженным.

Также результаты на случайном разбиении создают значительно более благоприятную картину, чем оценивание в условиях сдвига между наборами данных. Групповая схема оценивания между различными датасетами позволяет выявить снижение качества, чувствительность к нейтральному классу и доменные различия, которые остаются малозаметными при случайном разбиении.

При этом исследование имеет ряд ограничений. Так, остаётся не до конца изученным, какую именно информацию сарказм-ориентированные признаки переносят в задачу классификации тональности: скрытую негативность, общую субъективность или доменно-специфические маркеры экспрессивной речи. Также наблюдаемый эффект носит выраженно доменно-зависимый характер и не воспроизводится равномерно на всех подвыборках.

VI. Заключение

В работе представлен анализ роли саркастических признаков в задаче анализа тональности на русскоязычном корпусе. Полученные результаты не подтверждают их универсального преимущества: при оценивании в условиях сдвига между наборами данных сарказм-модифицированные модели в целом не превосходят наиболее сильную базовую модель. Вместе с тем выявлен ограниченный, но явный положительный эффект, проявляющийся в абсолютном приросте исправленных ошибок в оценке тональности. Результаты позволяют рассматривать саркастические признаки как потенциально полезные, но контекстно-зависимые данные, эффективное использование которых требует более точной интеграции с учётом класса, домена и особенностей нейтрального класса.

Список литературы

[1] Plepi J., Flek L. Perceived and Intended Sarcasm Detection with Graph Attention Networks // Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT). 2021. С. 97–105.

[2] Guo Z., Zhou G., Zhang X. Latent-Optimized Adversarial Neural Transfer for Sarcasm Detection // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021. С. 3193–3203.

[3] Abu Farha I., Oprea S., Wilson S., Magdy W. SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic // Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). 2022. С. 802–814.

[4] Liu L., Zhang D., Song W. A Dual-Channel Framework for Sarcasm Recognition by Detecting Sentiment Conflict // Findings of the Association for Computational Linguistics: NAACL 2022. 2022. С. 1720–1732.

[5] Hantsch A., Chkroun I. Connotation_clashers at SemEval-2022 Task 6: The Effect of Sentiment Analysis on Sarcasm Detection // Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). 2022. С. 836–842.

[6] Wang R., Vosoughi S., Danilevsky M. Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022. С. 5922–5937.

[7] Schwartz R., Stanovsky G. On the Limitations of Dataset Balancing: The Lost Battle against Spurious Correlations // Findings of the Association for Computational Linguistics: NAACL 2022. 2022. С. 2002–2014.

[8] Joshi A. R., Chatterjee S., Agrawal A. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022. С. 9800–9817.

[9] Kiela D. et al. Dynabench: Rethinking Benchmarking in NLP // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021. С. 4110–4124.

[10] Röttger P. et al. HateCheck: Functional Tests for Hate Speech Detection Models // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021. С. 41–58.

[11] Koh P. W. et al. WILDS: A Benchmark of in-the-Wild Distribution Shifts // Proceedings of the 38th International Conference on Machine Learning (ICML 2021). 2021. PMLR 139. С. 5637–5664.

[12] Wu Y., Shi B. Adversarial Soft Prompt Tuning for Cross-Domain Sentiment Analysis // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022. С. 2438–2448.

[13] Li Z., Wang X., Kuo C.-C. J. Cross-Domain Sentiment Classification using Semantic Representation // Findings of the Association for Computational Linguistics: EMNLP 2022. 2022. С. 3072–3085.