Анализ русскоязычного корпуса сарказма: поверхностные признаки, прагматические подтипы и пограничные случаи

И. С. Мухин
Университет ИТМО
ilya.mukhinn@yandex.ru

Е. Ю. Авксентьева
Университет ИТМО
avksentievaelena@rambler.ru

Аннотация. **В работе анализируется внутреннее устройство сбалансированного русскоязычного корпуса сарказма, включающего 15146 текстов (7573 саркастических и 7573 несаркастических). Цель исследования состоит в том, чтобы отделить прагматически значимые признаки от влияния темы, жанра и других поверхностных особенностей текста. Для этого сопоставляются линейные лексические модели и модель, основанная на поверхностных признаках, проводится тематический анализ корпуса, рассматриваются прагматические подтипы сарказма и их взаимная переносимость, а также применяются определенные правки текста и сопоставление разных типов представлений. Продемонстрировано, что модели на основе TF-IDF достигают значения F1 около 0.93, однако тематические признаки сами по себе дают F1 = 0.833, что указывает на существенную роль поверхностных корреляций. Одновременно результаты выявляют внутреннюю неоднородность класса sarcasm = 1: диалогические формы сарказма распознаются устойчивее, тогда как более компактные и афористичные проявления оказываются менее переносимыми. В графическом представлении обнаруживается переходная зона между саркастическими и несаркастическими примерами, а некоторые правки текста изменяют предсказание лишь умеренно. Полученные результаты показывают, что высокий уровень итоговых метрик детекции саркастических выражений не исчерпывает структуры корпуса и что при разработке прикладных систем анализа необходимо учитывать как поверхностные корреляции, так и прагматическую неоднородность данных.**

Ключевые слова: сарказм; русский язык; детекция сарказма; анализ датасетов; shortcut learning; прагматические механизмы

I. Введение

Как правило, при автоматической обработке текста, в частности сарказма, качество анализа опирается на соответствующие метрики, однако для построения устойчивых систем детекции или классификации этого может быть недостаточно. При сопоставимых значениях accuracy и F1 алгоритм или модель может опираться на разные типы признаков: от прагматических до жанровых маркеров. Поэтому высокие результаты на отложенной выборке сами по себе еще не свидетельствуют о том, что модель использует содержательно значимые признаки сарказма.

Для рассматриваемого корпуса высокая эффективность линейных моделей была установлена заранее. В связи с этим внимание работы концентрируется не на дальнейшем повышении метрик, а на анализе факторов, за счёт которых достигается такой результат. В работе рассматриваются три аспекта структуры корпуса: доля поверхностных сигналов, связанных с доменом и формой текста; неоднородность класса sarcasm = 1; а также организация переходной зоны между саркастическими и несаркастическими высказываниями.

Анализ ограничен данным корпусом и не претендует на построение универсальной типологии сарказма. В этом смысле работа относится к исследованиям, ориентированным на диагностику данных: основное внимание уделяется внутренней структуре корпуса, а также тому, какие типы ошибок и ограничений она задаёт для моделей детекции сарказма.

II. Литературный обзор

В современной литературе, посвященной анализу сарказма и иронии, внимание постепенно смещается от простой бинарной классификации к более детальному анализу природы признаков явления. После iSarcasmEval [1], а также работ, посвящённых инконгруэнтности и механизмам рассуждения [2, 3], всё заметнее становится интерес к тому, за счёт каких именно признаков модель распознаёт саркастическое высказывание. Параллельно развивается мультимодальное направление [4-6], где расхождение между буквальным и подразумеваемым смыслом рассматривается через взаимодействие разных модальностей.

Наряду с этим в NLP активно изучается проблема ложных и побочных корреляций. Работы, посвящённые их диагностике [7–11], показывают, что устойчивость модели во многих случаях определяется не только архитектурой, но и особенностями самого обучающего корпуса.

Наконец, всё более значимым становится направление, связанное с интерпретацией моделей и диагностикой данных. Для анализа сложных текстовых корпусов используются поведенческие тесты [12], картография датасетов [13], а также обзорные исследования по интерпретируемому NLP [14].

III. Материалы и методы

Исходные данные включают 15146 русскоязычных текстов, при этом датасет сбалансирован: 7573 саркастических и 7573 несаркастических выражений. Средняя длина текстов несколько различается: порядка 26 слов для sarcasm = 1 и около 20 для sarcasm = 0. Уже просматривается различие в том, что сарказм в среднем кодируется большим количество признаков.

Далее применялась многошаговая схема анализа. Сначала сопоставлялись модели TF-IDF + Linear SVM, TF-IDF + Logistic Regression и модель на вручную заданных признаках, включавших пунктуацию, регистр, длину текста и другие поверхностные характеристики. При этом различия оценивались по показателю (d) Коэна. Затем проводился анализ – в какой степени тематические и формальные особенности корпуса позволяют предсказывать сарказм без прямого обращения к прагматике.

На следующем этапе выполнялись кластеризация в пространстве эмбеддингов, полученных с помощью модели sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, построение слабой таксономии прагматических механизмов и анализ переносимости между различными подтипами саркастических высказываний. Затем в пайплайне контролируемых правок текста рассматривались три типа преобразований – desarcastify (снятие саркастического эффекта), flip_literal_polarity (смена буквальной полярности) и surface_strip (удаление поверхностных маркеров) – после чего оценивались сдвиг предсказанной вероятности и изменение локального положения примера в пространстве эмбеддингов. Дополнительно выполнялись графовый анализ, поиск типичных примеров и сопоставление трёх типов представлений – разреженного, плотного и реляционного – на подвыборке из 5000 примеров обоих классов.

IV. Результаты

Для рассматриваемого корпуса линейные лексические модели показывают высокое качество: модель TF-IDF + SVM достигает accuracy = 0.931 и F1 = 0.932, а результаты TF-IDF + Logistic Regression практически совпадают с этими значениями (accuracy = 0.930, F1 = 0.932). Хотя модель на вручную заданных признаках существенно уступает им по качеству (F1 = 0.707), и она демонстрирует, что часть предсказательного сигнала связана не с содержанием высказывания как таковым, а с его формальными характеристиками. Это подтверждается и анализом отдельных признаков: наиболее выраженные различия между классами наблюдаются для тире (d = 0.821), вопросительных конструкций (d = 0.666) и доли верхнего регистра (d = 0.488). Аналогичный вывод следует и из тематического анализа: в тематической группе с наибольшей долей сарказма доля саркастических примеров составляет 0.785, тогда как в группе с наименьшей долей сарказма –лишь 0.021. Кроме того, тематические признаки сами по себе дают F1 = 0.833. Сопоставление источников сигнала и эффектов различных упомянутых действий над текстами представлено на рис. 1.

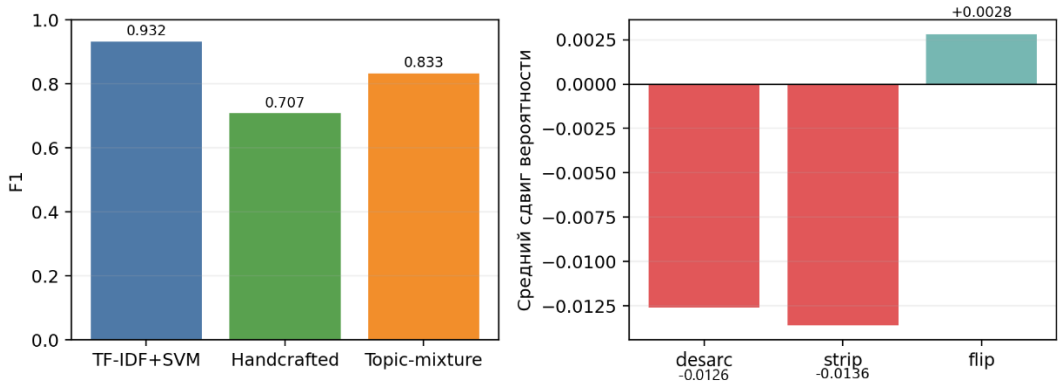

Рис. 1. Поверхностные и тематические источники сигнала в задаче детекции сарказма

Класс sarcasm = 1 не является однородным. По упрощённой разметке прагматических механизмов сарказма (weak-mechanism) в нём особенно выделяются диалогические и адресные паттерны. Для сарказма, оформленного как вопрос или диалогическая реплика (question_dialogue), значение odds ratio составляет 23.27 при частотах 0.448 и 0.034, а при более детализированной разметке для краткого разговорного обращения к адресату (colloquial_short_address) оно достигает 25.97. Неоднородность заметна и при анализе подтипов сарказма: диалогическое вопрошание (dialogic_questioning) распознаётся наиболее устойчиво (recall = 1.000, n = 72), тогда как минимальное значение recall наблюдается у афористичных резких высказываний (aphoristic_blunt) (recall = 0.948, n = 630), что может быть связано с их большей близостью к прямому, несаркастическому высказыванию.

Классификатор, использующий таксономическую разметку как опорное представление, сохраняет высокое качество (accuracy = 0.912, F1 = 0.914), однако переносимость между различными подтипами сарказма остаётся асимметричной. При поочерёдном исключении отдельных категорий значение F1 варьирует от 0.823 до 0.896. Анализ матрицы переноса показывает наличие пар, для которых качество резко снижается — до 0.518; как правило, это связано с переносом кратких форм сарказма на подтипы, организованные по иному прагматическому принципу (рис. 2).

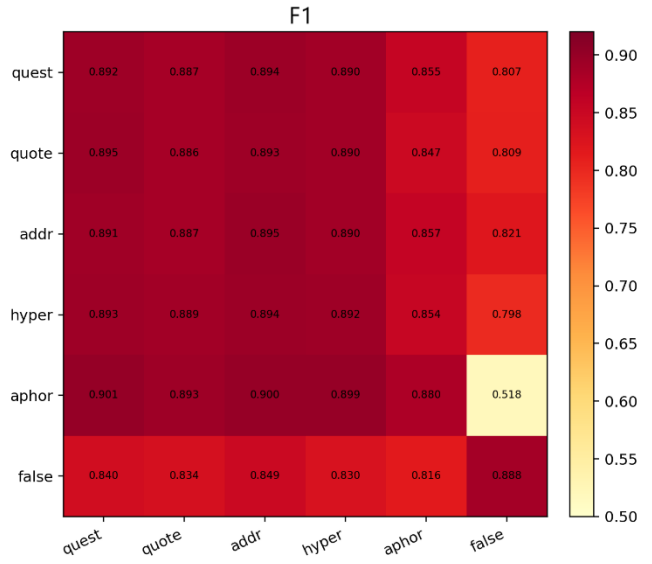

Рис. 2. Взаимная переносимость прагматических подтипов сарказма

Графовый анализ подтверждает наличие структурированной переходной области между классами. Центральная саркастическая область (n = 2835) и центральная несаркастическая область (n = 1869) почти не содержат примеров противоположного класса и характеризуются низкой энтропией (0.129 и 0.098). Напротив, пограничная область с высокой энтропией (n = 1536) имеет sarcasm_rate = 0.433, а также максимальные средние значения энтропии и посреднической центральности (mean_bridge = 0.493, entropy = 0.967). При этом крупнейшая пограничная компонента связности охватывает 96.8 % всех пограничных узлов, что позволяет говорить не о наборе разрозненных исключений, а о связной переходной зоне между саркастическими и несаркастическими примерами (рис. 3).

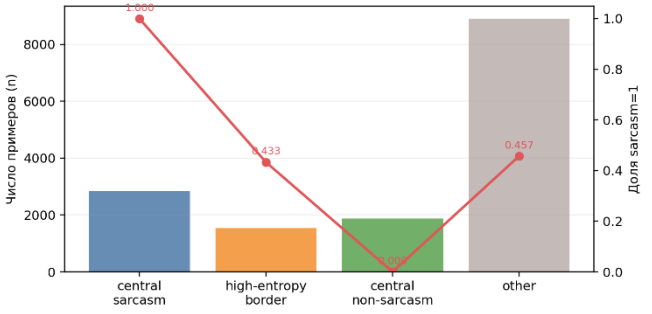

Рис. 3. Структура графового пространства корпуса: ядро и пограничная область

Правки текста изменяют предсказание лишь умеренно: средний сдвиг вероятности сарказма составляет –0.0126 для снятия саркастического эффекта, –0.0136 для удаления поверхностных маркеров и +0.0028 для изменения буквальной полярности. После таких преобразований выше порога 0.5 остаются 96–97 % примеров. Это позволяет предположить, что значимая часть предсказательного сигнала закреплена глубже, чем на уровне отдельных поверхностных признаков.

Сопоставление разных типов представлений подтверждает неоднородность устойчивости. Для реляционного представления доля примеров, сохраняющих p > 0.5 после правок, ниже (0.778–0.805), чем для разреженного и плотного представлений (до 0.98). Для саркастического класса разрыв по показателю мера неожиданности (surprisal) оказывается положительным (0.904), а для несаркастического — отрицательным (-1.038); вместе с тем корреляции изменений между типами представлений далеки от единицы (например, 0.137 для пары dense–relation при снятии саркастического эффекта). Это указывает на существование стабильного набора (ядра) примеров, однако его объём невелик по отношению ко всему корпусу.

V. Обсуждение

Полученные результаты позволяют описать структуру корпуса через сочетание трёх взаимосвязанных факторов. Во-первых, значимая часть предсказательного сигнала связана с темой, жанром и поверхностными особенностями текста. Во-вторых, внутри саркастического класса выделяется прагматическое ядро, которое относительно устойчиво распознаётся в разных типах представлений. В-третьих, между саркастическими и несаркастическими примерами обнаруживается широкая переходная зона, в которой буквальный и подразумеваемый смыслы соотносятся неоднозначно и реализуются через несколько частично пересекающихся подтипов сарказма.

С точки зрения развития систем детекции сарказма эти результаты показывают, что повышение качества не следует сводить только к росту итоговой метрики. Высокая предсказуемость оценки по тематическим признакам при стабильных лексических базовых моделях указывает на необходимость таких анализов, которые контролируют влияние тематических, жанровых и иных побочных корреляций. В этой связи особенно важны разделенные по доменам и жанрам механизмы валидации, а также дополнительные тесты, выходящие за рамки случайного деления на обучающую и тестовую выборки.

Важным результатом является и внутренняя неоднородность класса sarcasm = 1. Различия между подтипами сарказма и асимметрия их взаимной переносимости показывают, что сарказм в корпусе не образует единого однородного явления с общим набором признаков. В прикладном смысле это открывает возможность для применения иерархических или многозадачных моделей, где наряду с основной меткой учитываются и более слабые признаки прагматической организации текста.

Не менее существенным представляется наличие связной пограничной области между классами. Полученные графовые характеристики показывают, что речь идёт не о небольшом наборе случайных исключений, а о переходной зоне высокой плотности, где граница между саркастическим и несаркастическим выражением оказывается размытой.

Дополнительный интерес представляет различие между разреженными, плотными и реляционными представлениями по степени устойчивости к правкам текста. Оно показывает, что разные типы представлений фиксируют не один и тот же сигнал, а частично несовпадающие его аспекты. Поэтому их сопоставление может быть полезно не только для повышения качества,

но и для оценки надёжности решения: согласованность или расхождение между представлениями может служить индикатором того, насколько устойчиво распознаётся конкретный пример.

Ограничения исследования связаны прежде всего с тем, что все выводы получены на материале одного корпуса и зависят от выбранных подходов слабой разметки и анализа представлений. При переносе на другой домен или иной тип данных соотношение между подтипами сарказма, роль поверхностных признаков и устройство пограничной зоны могут измениться. При этом для задач диагностики корпуса подобный локальный анализ остаётся целесообразным, поскольку позволяет заранее выявить области риска и тем самым точнее определить направления дальнейшей разработки моделей или алгоритмов.

VI. Заключение

Проведённый анализ демонстрирует то, что высокая величина F1 не исчерпывает интерпретацию высокого качества корпуса или модели или алгоритма, которые его могли анализировать. В структуре данных сочетаются явно распознаваемое ядро сарказма, доменно- и жанрово-обусловленные поверхностные сигналы, а также связная пограничная область между классами. В связи с этим представление о сарказме как об однородном классе применительно к данному датасету оказывается чрезмерно упрощённым.

Полученная диагностика позволяет наметить несколько направлений дальнейшего исследования. К ним относятся оценка, учитывающая влияние темы и жанра, моделирование с учётом внутренней неоднородности класса sarcasm=1, механизмы, ориентированные на работу с пограничными и неопределёнными случаями, а также гибридные архитектуры, объединяющие лексические, семантические и реляционные представления.

Список литературы

[1] Abu Farha I., Oprea S. V., 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.

[2] Qiu Z., Yu J., Zhang Y., Lai H., Rao Y., Su Q., Yin J. Detecting Emotional Incongruity of Sarcasm by Commonsense Reasoning // Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025). 2025.

[3] Yang Q., Zeng J., Yang L., Ma K., Lin H. Sarcasm-R1: Enhancing Sarcasm Detection through Focused Reasoning // Findings of the Association for Computational Linguistics: EMNLP 2025. 2025.

[4] Tian Y., Xu N., Zhang R., Mao W. Dynamic Routing Transformer Network for Multimodal Sarcasm Detection // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023). 2023.

[5] Farabi S., Ranasinghe T., Kanojia D., Kong Y., Zampieri M. A Survey of Multimodal Sarcasm Detection // Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024). 2024.

[6] Yue T., Shi X., Mao R., Hu Z., Cambria E. SarcNet: A Multilingual Multimodal Sarcasm Detection Dataset // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024.

[7] Wang T., Sridhar R., Yang D., Wang X. Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models // Findings of the North American Chapter of the Association for Computational Linguistics (NAACL 2022). 2022. С. 1719–1729.

[8] Joshi N., Pan X., He H. 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 (EMNLP 2022). 2022.

[9] Sun Z., Xiao Y., Li J., Ji Y., Chen W., Zhang M. Exploring and Mitigating Shortcut Learning for Generative Large Language Models // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024.

[10] Zhou Y., Tang R., Yao Z., Zhu Z. Navigating the Shortcut Maze: A Comprehensive Analysis of Shortcut Learning in Text Classification by Language Models // Findings of the Association for Computational Linguistics: EMNLP 2024. 2024.

[11] Yuan Y., Zhao L., Zhang K., Zheng G., Liu Q. Do LLMs Overcome Shortcut Learning? An Evaluation of Shortcut Challenges in Large Language Models // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024). 2024.

[12] Ribeiro M. T., Wu T., Guestrin C., Singh S. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). 2020. С. 4902–4912.

[13] Swayamdipta S., Schwartz R., Lourie N., Wang Y., Hajishirzi H., Smith N. A., Choi Y. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020). 2020. С. 9275–9293.

[14] Gurrapu S., Kulkarni A., Huang L., Lourentzou I., Batarseh F. A. Rationalization for Explainable NLP: A Survey // Frontiers in Artificial Intelligence. 2023. Vol. 6.