

УДК 004

Шамало В.В.

магистрант кафедры информатики и вычислительной техники

Кубанский государственный технологический университет

Россия, г. Краснодар

**ГИБРИДНАЯ НЕЙРО-СИМВОЛИЧЕСКАЯ СИСТЕМА ДЕТЕКЦИИ
ТЕКСТА,
СГЕНЕРИРОВАННОГО ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ,
В СРЕДЕ СОЦИАЛЬНЫХ МЕДИА**

Аннотация: развитие больших языковых моделей сделало машинную генерацию текста неотличимой от человеческой речи, что породило «нейрокомментинг» и астротурфинг в социальных сетях. Цель работы — детектор, сохраняющий точность на коротких комментариях длиной 5–15 слов, где трансформеры теряют семантическую опору. Предложена гибридная нейро-символическая система «Святогор», объединяющая признаки RuModernBERT со стилометрическими инвариантами через мета-классификатор CatBoost. На корпусе из 116 000 образцов достигнута F1-мера 0,960 против 0,890 у чистого трансформера.

Ключевые слова: искусственный интеллект, обработка естественного языка, детекция генерации, нейро-символический подход, стилометрия, информационная безопасность.

Shamalo V.V.

Master's Student

Kuban State Technological University

Russia, Krasnodar

A HYBRID NEURO-SYMBOLIC SYSTEM FOR DETECTION OF TEXT GENERATED BY ARTIFICIAL INTELLIGENCE IN SOCIAL MEDIA

Abstract: the development of large language models has made machine-generated text indistinguishable from human speech, which has given rise to mass automated commenting and astroturfing in social networks. The aim is a detector that retains accuracy on short comments of 5–15 words, where transformers lose semantic support. The proposed hybrid neuro-symbolic system “Svyatogor” combines RuModernBERT features with stylometric invariants through a CatBoost meta-classifier. On a corpus of 116,000 samples it achieved an F1-score of 0.960 against 0.890 for a pure transformer.

Keywords: artificial intelligence, natural language processing, generation detection, neuro-symbolic approach, stylometry, information security.

Введение

Генеративные модели нового поколения сделали автоматическую генерацию связного текста практически неотличимой от человеческой речи. Следствием стали астротурфинг, «нейрокомментинг» и спам-атаки ботов, а методы детекции, эффективные в 2022 году, сегодня несостоятельны: модели дообучаются быстрее, чем обновляются детекторы [1].

Наиболее сложный случай — короткие комментарии длиной 5–15 слов: трансформерные модели, построенные на механизме самовнимания [2], теряют на них точность из-за недостатка семантической информации, а статистические методы дают высокий уровень ложных срабатываний. Цель работы — детектор, сохраняющий точность в этом диапазоне длин. Предложена гибридная нейро-символическая система «Святогор», где семантический анализ дополнен анализом структурных инвариантов текста.

Методы и исследования

Система реализует конвейерную архитектуру из четырёх этапов.

Первый этап — противодействие состязательным атакам. Критичный вектор атаки — гомоглифы, подмена кириллических букв визуально идентичными латинскими [3]. Если доля кириллицы превышает 70 %, активируется итеративная замена латинских гомоглифов кириллическими, что восстанавливает целостность токенов.

Второй этап — эвристическая фильтрация. На микротекстах короче 5–7 слов самовнимание не успевает сформировать устойчивый вектор, и классификация реплик вроде «согласен» вырождается в угадывание. Тексты короче трёх слов отбраковываются со статусом «Неопределено», что защищает пользователей от ошибочной блокировки.

Третий этап — двухпоточное извлечение признаков. Нейросетевой поток передаёт текст в трансформер RuModernBERT [4]; вектор скрытого состояния классификационного токена сжимается методом главных компонент. Символический поток работает со структурными инвариантами: модели после RLHF-сглаживания пишут безупречно, но сохраняют «математически усреднённую» частотность функциональных слов. Алгоритм строит L1-нормализованный вектор их частот и вычисляет косинусное расстояние до центроидов классов «ИИ» и «Человек» [5].

Четвёртый этап — мета-классификация. Эмбединги, первичные вероятности и стилометрические дистанции объединяются в мета-вектор и подаются на вход бустинга CatBoostClassifier [6]. Ансамблирование даёт калибровку итоговой вероятности и устойчивость к смене генерирующей модели.

Открытые датасеты непригодны: они ориентированы на англоязычный сегмент или длинные тексты, а русскоязычные корпуса часто оказываются машинным переводом. Собран нативный корпус

коротких текстов NRS-H объёмом 116 000 образцов, сбалансированных поровну между классами «Человек» и «ИИ». Естественная часть получена парсингом комментариев публичных сообществ [7]; синтетическая — ансамблем из пяти языковых моделей, что снижает риск переобучения под особенности одной из них.

Результаты

Оценка выполнена по метрикам Precision, Recall и F1-Score на отложенной выборке. Изолированные подходы уступают на коротких текстах: TF-IDF дал F1-Score = 0,74, принимая сленг за аномалию, а чистый RuModernBERT при 0,89 допускал много ложноположительных срабатываний.

Объединение семантических эмбеддингов и стилометрических инвариантов мета-классификатором дало синергетический эффект: Precision — 0,962, Recall — 0,958, F1-Score — 0,960. Прирост в 0,07 относительно чистого трансформера получен без увеличения нейросетевой части и обеспечен символическим потоком: стилометрические признаки информативны там, где семантических данных недостаточно.

Заключение

Спроектирована и программно реализована гибридная нейро-символическая система детекции машинно-сгенерированного текста для коротких сообщений социальных сетей. Достигнутая F1-мера 0,960 превышает результат чистого трансформера на той же выборке, что подтверждает исходную гипотезу: при недостатке семантического контекста структурные инварианты текста несут самостоятельную различающую информацию. Система применима в контурах автоматизированной модерации для противодействия бот-сетям и астротурфингу; дальнейшие исследования — расширение корпуса за счёт моделей новых поколений и оценка устойчивости детектора к перефразированию.

Использованные источники:

1. Данилов А.А., Смирнов И.В. Архитектуры больших языковых моделей и их уязвимости в задачах классификации текстов // Искусственный интеллект и принятие решений. 2023. № 4. С. 45–58.
2. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. 2019. P. 4171–4186.
3. Boucher N., Shumailov I., Anderson R., Papernot N. Bad Characters: Imperceptible NLP Attacks // Proceedings of the 2022 IEEE Symposium on Security and Privacy. 2022. P. 1985–2002.
4. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // Computational Linguistics and Intellectual Technologies. 2019. P. 333–340.
5. Романов А.С. Методология стилометрического анализа авторства русскоязычных текстов в условиях малого контекста // Компьютерная лингвистика и вычислительные онтологии. 2022. Т. 3. С. 112–120.
6. Документация библиотеки CatBoost: градиентный бустинг на деревьях решений [Электронный ресурс]. URL: <https://catboost.ai/docs/> (дата обращения: 12.04.2026).
7. VK API. Официальная документация для разработчиков [Электронный ресурс]. URL: <https://dev.vk.com/ru/docs> (дата обращения: 14.04.2026).