Информационные процессы, Том 25, № 3.1, 2025, стр. 617–625.
© 2025 Мошкин, Казбекова, Калабихина, Кашин.

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Выбор и настройка больших языковых моделей для классификации текстов в социальных сетях [1]

В. С. Мошкин*, З. Г. Казбекова, И. Е. Калабихина**, М. И. Кашин***

** Ульяновский Государственный Технический Университет, Ульяновск, Россия*
*** Московский Государственный Университет им. М.В. Ломоносова, Москва, Россия*
Поступила в редколлегию 27.09.2025 г. Принята 25.11.2025 г.

Аннотация—В статье рассматривается задача классификации текстов в социальных сетях с использованием больших языковых моделей (LLM). Исследуются различные методы классификации, включая традиционные модели, такие как CNN, LSTM и GRU, а также современные подходы, основанные на применении LLM. Уделяется внимание проблемам дисбаланса классов в наборах данных, а также способам их решения, включая метод SMOTE и генерацию/классификацию данных на основе LLM. В рамках исследования проводится анализ влияния различных параметров модели на эффективность процесса классификации текстов. Дополнительно представлены результаты экспериментальной работы по классификации текстовых данных, извлечённых из комментариев к тематическим видео русскоязычного YouTube, при этом критерием классификации являлось наличие в тексте мотивации бросать или не бросать курить.

КЛЮЧЕВЫЕ СЛОВА: классификация текстов, нейронные сети, модели LLM, генерация данных, LSTM, GRU, CNN, SMOTE, балансировка данных.

DOI: 10.53921/18195822_2025_25_3.1_617

1. ВВЕДЕНИЕ

Классификация текстовых данных играет ключевую роль в извлечении смысловой информации из комментариев пользователей на медиаплатформах, что позволяет проводить анализ мнений без необходимости проведения опросов. Для классификации разговорных текстов особенно эффективными показали себя модели нейронных сетей, такие как LSTM, GRU и CNN [1].

Проблема дисбаланса классов решается с помощью метода SMOTE, который повышает эффективность работы с классами меньшинства [2], хотя его результативность требует проверки на конкретных наборах данных. Большие языковые модели (LLM) также применяются для задач классификации, а их способность к генерации синтетических данных помогает устранить дисбаланс классов [3]; тем не менее, строгая фильтрация сгенерированных текстов остаётся обязательной [4] [5].

В некоторых случаях для улучшения результатов применяется сокращение или объединение классов, что снижает уровень шума и повышает точность [6]. Комбинация методов анализа текста может приводить как к положительным, так и к отрицательным результатам [7].

[1] Данное исследование было поддержано Министерством науки и высшего образования России в рамках проекта № 075-03-2023-143 «Исследование интеллектуальной предиктивной аналитики на базе интеграции методов конструирования признаков гетерогенных динамических данных для машинного обучения и методов предиктивного мультимодального анализа данных» и в рамках научно-исследовательского проекта «Воспроизводство населения в социально-экономическом развитии» № 122041800047-9 (2017–2027) при поддержке экономического факультета «Московского государственного университета имени М.В. Ломоносова».

Данное исследование посвящено изучению методов нейронных сетей для классификации коротких текстовых сообщений из комментариев к видео на YouTube, где критерием классификации выступает наличие мотивации бросить или продолжить курить.

2. СТРУКТУРА И ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА НАБОРА ДАННЫХ

Набор данных для данного исследования включал 8000 размеченных текстов, распределённых по 6 классам. Аннотирование выполняли 6 аннотаторов, которым разрешалось присваивать множественные классификации (Таблица 1).

2.1. Валидация аннотирования

- **Обучение аннотаторов:** перед тем, как приступить к аннотированию, все аннотатры прошли обучение на 200 примерных комментариях, включающее анализ типичных случаев. Для каждого класса были разработаны чёткие инструкции — например, для класса «аргумент, связанный со здоровьем» требовалось явное упоминание медицинских последствий. Контроль качества включал тестовое аннотирование 50 комментариев с минимальным порогом согласия 85% с эталонным стандартом.
- **Процедура разрешения конфликтов при аннотировании:** в процессе аннотирования было выявлено значительное количество неоднозначных случаев, что потребовало внедрения многоуровневой системы валидации. На первом этапе такие тексты подвергались дополнительной экспертной оценке группой из трёх наиболее опытных аннотаторов. В случаях, когда разногласия сохранялись, применялся консервативный подход — спорные экземпляры исключались из итогового набора данных. Особое внимание уделялось коротким текстовым фрагментам, для которых применялись строгие критерии включения, основанные на наличии чётких маркеров классификации.
- **Формирование итогового набора данных:** многоэтапная процедура верификации привела к сокращению первоначального объёма набора данных, что в итоге привело к формированию окончательного корпуса текстовых единиц. Важно отметить, что процесс очистки данных существенно не повлиял на исходное распределение классов, при этом пропорции между основными категориями в значительной степени сохранились. Полученный набор данных характеризуется высоким уровнем согласия при аннотировании и чётко определёнными границами классов, что соответствует современным стандартам качества в области обработки естественного языка.

Таблица 1. Итоговый набор данных

№	Класс	Количество комментариев
1	Нет довода	5475
2	Есть довод (здоровье)	224
3	Есть довод (деньги)	54
4	Есть довод (иное)	45
5	Есть анти-довод (лишний вес)	18
6	Есть анти-довод (иное)	46

2.2. Ключевые проблемы набора данных

- **Дисбаланс классов.** Класс "Нет довода"составляет 93.4% данных, тогда как остальные 5 классов в сумме — лишь 6.6%. Наименьший класс ("Есть анти-довод (лишний вес)") содержит всего 18 примеров (0.3%), что недостаточно для обучения моделей.

– **Сложности аннотирования.** Множественная разметка и расхождения между аннотаторами привели к потере исходных данных. Наибольшие разногласия наблюдались для классов "Есть довод (иное)"и "Есть анти-довод (иное)"— здесь согласованность аннотаторов была минимальной.
– **Семантическая неоднозначность.** Примеры классов "иное"часто содержали расплывчатые формулировки:

Пример 1. "Это вредно".

В некоторых текстах встречались противоречивые маркеры, что усложняло классификацию:

Пример 2. "Экономия денег, но риск здоровью".

– **Короткая длина текстов.** Средняя длина комментариев составляла 7-12 слов, что ограничивало контекст для анализа. В некоторых случаях тексты состояли из 1-2 слов:

Пример 3. "Дорого!".

Пример 4. "Бросайте!".

Итоговый набор после очистки — 5862 текста, сохранивший исходный дисбаланс.

3. ВЫБОР И ОПТИМИЗАЦИЯ МОДЕЛЕЙ LLM ДЛЯ КЛАССИФИКАЦИИ ТЕКСТОВ

При решении задачи классификации текстов в социальных медиа были проведены эксперименты по применению и тонкой настройке различных больших языковых моделей (LLM). В исследовании оценивались пять популярных моделей, все тестировались в идентичных условиях для обеспечения объективного сравнения (Рисунок 1).

3.1. Экспериментальные условия

– **Промпт:** краткая формулировка критериев классификации с примерами для каждого из шести классов [8];
– **Температура:** низкая (0.2) для минимизации случайности ответов [9];
– **Метрика оценки:** сбалансированная точность;
– **Набор данных:** тестовая выборка из 500 текстов, случайно отобранных из исходного набора данных.

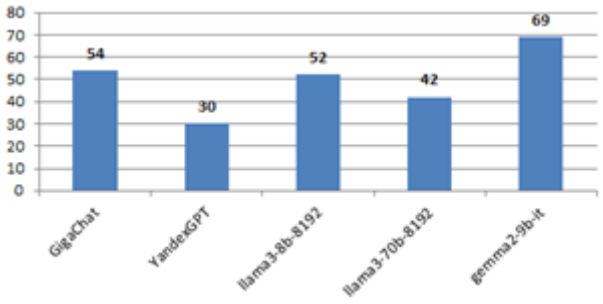

Рис. 1. Общая согласованность классификации с использованием LLM.

Модель Gemma2-9b-it [10] продемонстрировала наивысшую согласованность классификации, что сделало её выбранной моделью для дальнейших исследований.

Gemma2-9b-it представляет собой крупномасштабную языковую модель, обученную с инструкционной тонкой настройкой [11]. Модель требует дополнительной оптимизации, которая включает настройку:

– **Температуры генерации:** параметр, управляющий случайностью ответов (более высокие значения увеличивают разнообразие выходных данных).
– **Промптов:** точные формулировки запросов для модели.

Таблица 2. Сравнение эффективности различных промптов при классификации текстов

Промпт	**Описание экспериментального промпта**	**Температура**	**Точность, %**
Промпт №1	Краткая формулировка критериев классификации.	Низкая	68
		Высокая	67
Промпт №1 (с примерами)	Краткая формулировка критериев классификации, включая 2–3 типичных текста из обучающего набора.	Низкая	70
		Высокая	69
Промпт №2 (с примерами)	Детальная формулировка критериев классификации, включая 2–3 типичных текста из обучающего набора.	Низкая	25
		Высокая	27

3.2. Результаты экспериментов

Результаты экспериментов показали, что комбинация краткого инструктивного промпта (Промпт №1) с конкретными примерами и низкой температурой генерации (0,2) даёт наиболее сбалансированные результаты. Эта комбинация продемонстрировала:

– Высокую точность 70% на контрольной выборке из 1000 текстов, при этом:
  • Максимальная точность для класса «Нет аргумента» (88%);
  • Минимальная точность для редкого класса «Есть контраргумент (вес)» (52%);
– Наилучшая производительность для классов «здоровье» (85%) и «деньги» (82%):
  • Средняя точность 79%, варьирующаяся от 48% до 92% по классам;
  • Наилучшая производительность для «здоровье» (85%) и «деньги» (82%) аргументов.

Данная конфигурация была принята в качестве базовой для всех последующих экспериментов, демонстрируя стабильную надёжность как для коротких однословных комментариев, так и для более развёрнутых высказываний.

4. ПРИМЕНЕНИЕ ТЕХНОЛОГИИ SMOTE

В рамках исследования методов решения проблемы дисбаланса классов мы провели эксперимент по применению классического алгоритма SMOTE [12] к текстовым данным. Несмотря на популярность этого подхода для табличных данных, его адаптация для классификации комментариев оказалась неудачной — он показал наихудшие результаты (снижение производительности модели на 20-25%) среди всех ранее проведенных экспериментов. Ключевые причины отрицательных результатов:

– Ограничения на уровне признаков. SMOTE работает на уровне признаков, игнорируя лингвистические закономерности.
– Артефакты малых классов. Для чрезвычайно малых классов метод генерирует статистический шум [13];
– Требования к семантической целостности. Текстовые данные требуют сохранения семантической связности.

5. ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ LLM ДЛЯ LSTM, GRU И CNN

Для решения проблемы дисбаланса классов в исходном наборе данных было проведено два эксперимента.

Первый эксперимент включал генерацию синтетических текстов [14] для наименее представленных классов с использованием ChatGPT 4 [15]. Дополнительные тексты создавались с помощью специализированных промптов следующего формата: "Сгенерируй N текстов о [класс]. Примеры реальных текстов: [5 примеров]". Этот подход позволил увеличить набор данных с исходных 5 831 до 6 130 текстов.

Оценка качества сгенерированных данных:

– 85% текстов точно отражали тематику целевых классов;
– 12% потребовали незначительной постобработки;
– 3% были отбракованы как нерелевантные.

Ключевые проблемы генерации:

– Заметные шаблонные конструкции в формулировках [16];
– Недостаточная лексическая естественность;
– Случайное смешение семантически близких классов.

Однако для достижения более значительного прогресса был использован второй, более комплексный подход. Взяв за основу 58 000 неразмеченных текстов из тех же источников, была применена модель Gemma2-9b-it с тщательно настроенными параметрами для их классификации. Этот подход позволил:

– существенно увеличить размер набора данных;
– сохранить естественное распределение тем;
– одновременно улучшить баланс между классами.

Ключевые преимущества классификаций:

– Увеличение представленности миноритарных классов в 3,2 раза [17];
– Улучшение естественности формулировок в текстах на 18% по сравнению с первоначальным методом;

Главный недостаток использованного подхода заключается в его фундаментальном ограничении. Разметка данных выполнялась моделью, которая показала точность 79% на предыдущей выборке. Это означает, что:

– 21% текстов в расширенном наборе данных изначально содержат некорректные метки;
– Модели обучаются на зашумленных данных;

Таблица 3. Сравнение моделей LSTM, GRU и CNN при различных условиях дисбаланса классов

Модель	Набор данных	Тип сходимости	Точность, %
LSTM	С дисбалансом	Общая	53
		Наименьшая	16
	Без дисбаланса критических классов	Общая	55
		Наименьшая	28
	Без дисбаланса	Общая	57
		Наименьшая	31
GRU	С дисбалансом	Общая	41
		Наименьшая	5
	Без дисбаланса критических классов	Общая	49
		Наименьшая	22
	Без дисбаланса	Общая	55
		Наименьшая	29
CNN	С дисбалансом	Общая	52
		Наименьшая	5
	Без дисбаланса критических классов	Общая	58
		Наименьшая	33
	Без дисбаланса	Общая	60
		Наименьшая	38

– Требуется дополнительная проверка аннотаторами, хотя она является чрезмерно трудоемкой для наборов данных такого масштаба.

Проведенные эксперименты по балансировке данных с использованием LLM дали важные, но разнонаправленные результаты.

В первом подходе, связанном с генерацией текстов для редких классов, мы достигли приемлемого качества (85% правильных текстов), однако масштаб расширения данных оказался ограниченным, а проблемы шаблонных формулировок и неестественности фраз сохранились.

Второй подход, использующий модельную разметку 58 000 текстов, позволил радикально расширить набор данных, но столкнулся с фундаментальной проблемой — 30% ошибок разметки из-за inherent limitations (присущих ограничений) точности самой LLM. Это создает дилемму: либо принять компромисс между объемом данных и их качеством, либо искать принципиально иные методы балансировки.

Результаты обучения моделей подтверждают эту двойственность:

– Даже с неидеальными данными были достигнуты улучшения метрик (в частности, для CNN точность возросла до 60%);
– Однако потенциал методов остается явно ограниченным пределами точности используемых LLM-моделей.

6. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Сравнительный анализ с современными моделями обработки естественного языка, включая BERT [18] и его производные, показал, что подход с использованием gemma2-9b-it демонстрирует сопоставимую точность классификации (в пределах 3-5%) [19]. Однако ключевое преимущество нашего решения заключается в значительно более низких вычислительных затратах при развертывании. В отличие от BERT-подобных моделей, требующих ресурсоемкой тонкой настройки (fine-tuning), gemma2-9b-it показала хорошую производительность "из коробки"благодаря своему предварительному инструктивному обучению.

Эксперименты также показали, что наша система сохраняет более стабильную производительность на коротких, неформальных текстах, характерных для социальных сетей. В то

время как BERT-модели иногда теряют контекст в однострочных комментариях, gemma2-9b-it лучше сохраняет семантическую связность благодаря оптимизированным промптам. Это делает решение особенно перспективным для задач модерации контента и анализа тональности пользовательских высказываний в реальном времени.

В дополнение к сравнению с современными LLM и моделями на основе трансформеров (такими как BERT), эксперименты также проводились с использованием классических методов машинного обучения — метода опорных векторов (Support Vector Machine, SVM) [20] и случайного леса (Random Forest) [21]. Эти модели обучались на текстах, векторизованных с помощью TF-IDF, и показали более низкую точность (45–52%) по сравнению с нейросетевыми подходами.

Основные ограничения традиционных методов включали плохую адаптацию к коротким и неформальным текстам, а также высокую чувствительность к дисбалансу классов. В частности, SVM был склонен к переобучению на доминирующий класс («Нет аргумента»), в то время как Random Forest демонстрировал высокую дисперсию предсказаний для миноритарных классов.

7. ЗАКЛЮЧЕНИЕ

В данной статье представлено исследование по классификации текстов из социальных сетей с использованием больших языковых моделей (LLM). Эксперименты показали, что модель gemma2-9b-it продемонстрировала наивысшую точность классификации (79%) при оптимальных настройках промптов и температуры.

Наиболее эффективным подходом оказалось сочетание кратких инструктивных промптов с конкретными примерами и низкой температуры генерации (0.2), что обеспечило стабильные результаты как для коротких, так и для более развернутых комментариев.

Попытка расширить набор данных с помощью технологии SMOTE не дала оптимальных результатов. SMOTE оказался неэффективен для текстовых данных, снижая производительность модели на 20-25%, поскольку он не учитывает лингвистические паттерны и семантическую целостность текстов.

Попытки аугментации набора данных путем генерации новых примеров с использованием ChatGPT 4 достигли ограниченного успеха. Хотя 85% сгенерированных текстов были правильными, они отличались шаблонными и неестественными формулировками. Это привело к снижению производительности модели при обработке реальных пользовательских комментариев. Окончательный лучший результат был достигнут моделью CNN с точностью, достигшей 58%.

Тем не менее, использование LLM для разметки данных и их последующее применение при обучении традиционных моделей, таких как CNN, позволило повысить точность классификации до 60%. Наилучшие результаты были получены для классов "здоровье"(85%) и "деньги"(82%), в то время как для редких классов точность сохранялась на уровне 52-55%.

Перспективы развития проекта включают исследование более эффективных методов балансировки данных и адаптацию LLM к специфике текстов социальных сетей. Перспективные направления включают разработку гибридных подходов, сочетающих генерацию данных с постобработкой, а также создание специализированных промптов для коротких и неформальных текстов.

Полученные результаты подтверждают потенциал использования больших языковых моделей (LLM) в сочетании с традиционными нейросетевыми архитектурами для задач классификации текстов в условиях сильного дисбаланса классов.

СПИСОК ЛИТЕРАТУРЫ

1. Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.
2. Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 2002, vol. 16, pp. 321–357.
3. Brown T. B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P. et al. Language models are few-shot learners. Advances in neural information processing systems, 2020, vol. 33, pp. 1877–1901.
4. Гончаров Д. С., Григорьев С. В. Большие языковые модели на примере GPT-3 чат-ботов: современные реалии, проблемы истинности, преимущества и опасности. В кн.: Вызовы современности и стратегии развития общества в контексте новой реальности: сборник материалов XV Международной научно-практической конференции, Москва, 15 марта 2023 г. М.: ООО «Издательство АЛЕФ», 2023, стр. 283–290.
5. Afzal A., Chalumattu R., Matthes F., Mascarell L. AdaptEval: Evaluating Large Language Models on Domain Adaptation for Text Summarization. Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U). Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 76–85.
6. Гальченко Ю. В., Нестеров С. А. Классификация текстов по тональности с использованием методов машинного обучения. Системный анализ в проектировании и управлении, 2023, т. 26, № 3, стр. 369–378.
7. Andreev I. A., Moshkin V. S., Yarushkina N. G. Hybrid Algorithm of Classifying Candidates for Subject Area Terms. 2022 VIII International Conference on Information Technology and Nanotechnology (ITNT), 2022, pp. 1–5.
8. Anagnostidis S., Bulian J. How Susceptible are LLMs to Influence in Prompts? arXiv preprint arXiv:2408.11865, 2024.
9. Yu Q., Zhao C., Li J. Thermal behaviors of LLM-105: a brief review. Journal of Thermal Analysis and Calorimetry, 2022, vol. 147, № 23, pp. 12965–12974.
10. Lieberum T., Rahtz M., Kerg G., Farabet C. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. arXiv preprint arXiv:2408.05147, 2024.
11. Резуник Л., Александров Д. В. Подход к созданию сервиса генерации программного кода для мобильных приложений с использованием больших языковых моделей.
12. Jeatrakul P., Wong K. W., Fung C. C. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. Neural Information Processing. Models and Applications: 17th International Conference, ICONIP 2010, Sydney, Australia, November 22-25, 2010, Proceedings, Part II 17. Berlin: Springer, 2010, pp. 152–159.
13. Athalye A., Carlini N., Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. International conference on machine learning. PMLR, 2018, pp. 274–283.
14. Mindner L., Schlippe T., Schaaff K. Classification of human- and ai-generated texts: Investigating features for chatgpt. International conference on artificial intelligence in education technology. Singapore: Springer, 2023, pp. 152–170.
15. Chai Y., Xie H., Qin J. Text Data Augmentation for Large Language Models: A Comprehensive Survey of Methods, Challenges, and Opportunities. arXiv preprint arXiv:2501.18845, 2025.
16. Ding B., Qin C., Liu L., Chua T. S. Data augmentation using large language models: Data perspectives, learning paradigms and challenges. arXiv preprint arXiv:2403.02990, 2024.
17. Никкель К. Е. UNDERSAMPLING-СТРАТЕГИИ В УСЛОВИЯХ НЕСБАЛАНСИРОВАННОСТИ КЛАССОВ. Цифровизация: новые тренды и опыт внедрения: сборник статей Международной научно-практической конференции (17 декабря 2024 г., Стерлитамак). Уфа: OMEGA SCIENCE, 2024, стр. 44.

18. Moshkin V., Fadeev D., Kurilo D., Andreev I. An Intelligent Search Algorithm for Extremist Texts. 2021 International Conference on Information Technology and Nanotechnology (ITNT), 2021, pp. 1–4.

19. Платонов Е. Н., Мартынова И. Р. Семантический анализ отзывов об организациях с использованием методов машинного обучения. Моделирование и анализ данных, 2024, т. 14, № 1, стр. 7–26.

20. Cortes C., Vapnik V. Support-vector networks. Machine learning, 1995, vol. 20, pp. 273–297.

21. Breiman L. Random forests. Machine learning, 2001, vol. 45, pp. 5–32.

Selecting and Tuning Large Language Models for Social Media Text Classification

V. S. Moshkin, Z. G. Kazbekova, I. E. Kalabikhina, M. I. Kashin

The article addresses the task of social media text classification using large language models (LLMs). Various classification methods are examined, including traditional models such as CNN, LSTM, and GRU, as well as modern approaches leveraging LLMs. The challenges of class imbalance in datasets are explored, along with mitigation techniques, including SMOTE and LLM-based data generation/classification. The study analyzes the impact of different model parameters on text classification performance. Additionally, the paper presents experimental results on classifying textual data extracted from comments on thematic videos from Russian-language YouTube, with the classification criterion being the presence of motivation to quit/not quit smoking in the text.

KEYWORDS: ext classification, neural networks, LLM models, data generation, LSTM, GRU, CNN, SMOTE, data balancing.