RU EN ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ

УДК 004.89:004.93
10.25209/2079-3316-2026-17-1-57-84

Сравнительный анализ состязательных методов для нетематической классификации текстов

Михаил Николаевич **Лепехин**[1]✉, Сергей Александрович **Шаров**[2]

[1] Московский физико-технический институт, Москва, Россия
[2] Университет Лидса, Лидс, Великобритания
[1]✉*lepehin.mn@phystech.edu*

Аннотация. Нетематическая классификация текстов широко используется в современных приложениях. Одной из проблем, возникающих при решении этой задачи, является наличие смещений в распределении в тренировочных текстовых корпусах. Наиболее существенным видом смещений являются тематические смещения. Для решения этой проблемы в данной работе применяются состязательные методы - Adversarial Domain Adaptation, Energy-based ADA, BERT с контрастной функцией потерь и ADA с контрастной функцией потерь.

В работе впервые производится модификация контрастной функции потерь для снижения влияния тематических сдвигов и показывается, что использование состязательных методов повышает точность и надежность классификаторов для задачи определения пола автора текста. Также проводятся эксперименты с LLaMA-3B и показано, что большие языковые модели достигают в режиме few-shot более низкую точность чем дообученные модели с меньшим числом параметров, и требуют больше времени для предсказания.

Ключевые слова и фразы: Состязательные методы, контрастная функция потерь, классификация гендера, классификация текстов, нетематическая классификация, bert, доменная адаптация

Для цитирования: Лепехин М. Н., Шаров С. А. *Сравнительный анализ состязательных методов для нетематической классификации текстов* // Программные системы: теория и приложения. 2026. **Т. 17**. № 1(70). С. 57–84. https://psta.psiras.ru/read/psta2026_1_57-84.pdf

1. Введение

Нетематическая классификация текстов охватывает широкий спектр задач, направленных на выявление свойств текста, не связанных напрямую с его тематикой. К таким задачам относятся определение стиля, уровня сложности, эмоциональной окраски, вежливости, а также социолингвистических характеристик автора (возраст, пол, родной язык и др.). Решения подобных задач важны для информационного поиска, адаптивных образовательных систем, лингвистических исследований, рекомендательных систем и других областей.

В отличие от тематической классификации, где целевой признак часто коррелирует с набором ключевых слов, нетематические свойства являются более абстрактными и не определяются непосредственно по лексическому составу. Это создаёт дополнительные трудности для классификации. В частности, модели оказываются чувствительны к *сдвигам распределения*, особенно к *тематическим сдвигам* [1].

Тематический сдвиг возникает, когда в обучающих данных существует нежелательная корреляция между целевым признаком и тематикой текстов (например, спортивные тексты чаще написаны авторами-мужчинами). В таких условиях модель начинает опираться на тематические маркеры, что приводит к значительной деградации качества на данных с иной тематической структурой [2].

Для повышения устойчивости моделей к доменным и тематическим сдвигам используются методы, направленные на обучение доменно-инвариантных признаков. Одним из наиболее перспективных подходов является состязательная доменная адаптация (Adversarial Domain Adaptation, ADA) [3], позволяющая извлекать характеристики текста, не зависящие от источника данных. Её улучшенная версия – метод Energy-based ADA (EADA) [4] – использует энергетическую функцию и специальную функцию потерь для выравнивания распределений между доменами, снижая влияние доменных различий.

Другой класс подходов – методы *контрастного обучения* [5], формирующие более компактные и разделимые представления целевых классов. Несмотря на доказанную эффективность этих методов для повышения общей точности классификации [6, 7], их потенциал для уменьшения влияния тематических сдвигов в задачах нетематической классификации остаётся недостаточно изученным.

Параллельно с развитием специализированных архитектур стремительно совершенствуются большие языковые модели (LLM), такие как LLaMA [8], GPT-4 [9] и DeepSeek [10]. Они демонстрируют высокую

эффективность в решении широкого круга задач без явного дообучения [11, 12]. Однако их вычислительная дороговизна и задержки выполнения делают необходимым сопоставление их возможностей с более компактными моделями на основе BERT в контексте нетематической классификации.

Постановка задачи. Цель данной работы — разработать и экспериментально оценить методы, устойчивые к тематическим сдвигам, для задач нетематической классификации текстов. Особое внимание уделяется снижению зависимости качества моделей от домена (источника данных) и анализу того, в какой степени различные подходы позволяют уменьшить влияние тематических корреляций в обучающих данных.

Основной вклад работы заключается в следующем:

(1) Демонстрируется высокая чувствительность моделей на основе BERT к тематическим сдвигам и источнику данных (Раздел 6.1);

(2) предлагается модификация контрастной функции потерь, направленная на более эффективное подавление влияния тематических сдвигов;

(3) впервые для задачи нетематической классификации текстов исследуется совместное применение ADA и контрастной функции потерь;

(4) Проводится сравнение дообучаемых моделей меньшего размера с большой языковой моделью LLaMA-3B-Instruct в режиме few-shot; показывается, что LLM уступает дообучаемым моделям по качеству и времени выполнения в задаче классификации пола автора (Раздел 6.3).

Настоящее исследование существенно расширяет и углубляет результаты работы [13], включая эксперименты с контрастным обучением, его комбинацией с ADA, а также расширенный анализ с использованием современных больших языковых моделей.

2. Связанные исследования

Задача классификации текстов является одной из наиболее важных и часто возникающих в обработке естественного языка. Долгое время архитектура BERT [14] и её производные (например, XLM-RoBERTa [15]) служили основой для многих современных моделей классификации [16]. В данной работе BERT используется в качестве базовой архитектуры для всех сравниваемых методов.

Проблема доменной адаптации тесно связана с задачами классификации, особенно в условиях сдвигов распределения данных. Для нетематической классификации, где целевые признаки слабо выражены

на лексическом уровне, устойчивость к таким сдвигам приобретает критическое значение. В литературе можно выделить несколько основных подходов к решению этой проблемы.

Методы, основанные на модификации эмбеддингов. Ряд работ предлагает напрямую корректировать векторные представления слов. Например, в исследовании [17] предлагается выявлять и модифицировать эмбеддинги слов, специфичных для определённой предметной области (*weird words*). Несмотря на простоту и эффективность в рамках конкретного домена, этот подход требует предварительного знания ключевых слов и не решает общей задачи создания классификатора, устойчивого к любым тематическим сдвигам.

Состязательные методы доменной адаптации (ADA). Классический подход Adversarial Domain Adaptation [3] и его развитие – Energy-based ADA (EADA) [4] – используют состязательное обучение для извлечения признаков, инвариантных к домену. Однако в оригинальных работах эти методы применялись преимущественно для задачи переноса знаний из домена с большим количеством размеченных данных в домен с малым их количеством. Наша работа фокусируется на иной цели: минимизации влияния доменных (в частности, тематических) признаков на классификацию внутри одного набора данных, что позволяет снизить чувствительность модели к доминирующему источнику в обучающей выборке.

Методы с мета-обучением. Исследование [18] интегрирует мета-обучение с ADA, вводя дополнительный модуль – генератор метазнаний на основе BiLSTM. Целью является не только улучшение классификации, но и максимальное «запутывание» дискриминатора домена. Однако данная работа сфокусирована на задачах тематической классификации (анализ тональности, категоризация новостей), в то время как мы исследуем более сложный случай нетематической классификации.

Контрастное обучение. Методы, основанные на контрастных функциях потерь [5], направлены на формирование компактных кластеров в пространстве признаков для объектов одного класса. В нашей работе мы используем этот подход, но вносим ключевые модификации: добавляем механизм для сближения эмбеддингов текстов из *разных* предметных областей (доменов) при совпадении целевого класса, а также применяем нормирование контрастной функции потерь. Это позволяет напрямую бороться с тематическими сдвигами, а не только повышать общую разделимость классов.

Каузальные модели. Подходы, основанные на каузальном выводе, такие как CausaLM [19] и CausalNLP [20], направлены на выявление

и устранение влияния конфаундеров – признаков, создающих ложные зависимости между переменными. В отличие от оригинальных работ, где основной целью является оценка значимости признаков или анализ причинно-следственных связей, мы используем каузальную функцию потерь из [19] для прямой максимизации точности классификации, заставляя модель уделять меньше внимания конфаундерам (тематическим признакам).

Большие языковые модели (LLM). С появлением моделей типа GPT-4 [9], Gemini [21] и LLaMA [8] значительно возрос потенциал решения задач без явного дообучения [22]. Однако их использование сопряжено с высокими вычислительными затратами и ограничениями на доступ. В нашей работе мы исследуем более экономичный вариант – модель LLaMA-3.1-3B-Instruct в режиме few-shot, – чтобы оценить, насколько современные LLM способны решать задачу нетематической классификации без дополнительной тонкой настройки, и сравнить их с дообученными моделями меньшего размера.

Таким образом, существуют различные подходы к повышению устойчивости моделей классификации: модификация эмбеддингов, состязательная доменная адаптация, мета-обучение, контрастное и каузальное обучение. Однако многие из них либо не ориентированы специально на борьбу с тематическими сдвигами в нетематической классификации, либо требуют априорных знаний о доменах, либо фокусируются на других типах задач (тематическая классификация, анализ тональности). Недостаточно изученным остаётся совместное применение и адаптация этих методов для прямой минимизации влияния доминирующего источника данных на качество классификации.

В нашей работе мы систематически применяем ADA, EADA и контрастное обучение и исследуем их эффективность для решения именно этой задачи.

3. Методология

Состязательные и контрастные функции потерь показали свою эффективность в ряде задач, включая доменную адаптацию и классификацию текстов. Во всех методах, рассматриваемых далее, особое внимание уделяется тому, каким образом они способны снижать влияние *сдвигов в распределении* между обучающим и тестовым доменами. Под *тематическими сдвигами* подразумеваются изменения распределений тематик текстов между двумя источниками данных (например, различия между блогами Mail.Ru, содержащими тексты на самые разные темы, и

AWD с преобладанием текстов о путешествиях). Механизмы подавления доменных различий представлены в описании каждого метода.

3.1. Состязательная адаптация домена (ADA)

Метод ADA (Adversarial Domain Adaptation [3]) изначально был разработан для решения задачи доменной адаптации и относится к *задаче адаптации домена без учителя* (Unsupervised Domain Adaptation, UDA) [23]. В постановке задачи адаптации домена без учителя есть *исходный домен* с большим числом размеченных примеров и *целевой домен* с отсутствующими или крайне малочисленными метками и требуется обучить модель таким образом, чтобы она смогла показывать достаточную точность на целевом домене. Данный метод показывает высокую эффективность в задачах обработки естественного языка, включая случаи значительного различия тематик между доменами [3].

В доменной адаптации предполагается обучение модели на размеченных текстах исходного домена (X_s, Y_s) для применения на целевом домене (X_t, Y_t).

Модель ADA состоит из *извлекателя признаков* $f = G_f(x)$, *классификатора целевого класса* $y = G_y(x)$ и *доменного дискриминатора* $d = G_d(x)$. Дискриминатор пытается отличить тексты по происхождению, а извлекатель признаков — извлекать такие признаки, по которым это сделать сложно. Обучение формулируется как состязательный процесс:

$$\min_{G_f, G_y} L_y(X_s, Y_s) - \lambda L_f(X_s, X_t), \tag{1}$$

$$\min_{G_d} L_d(X_s, X_t). \tag{2}$$

Компонента L_f уменьшает различимость доменов в скрытом представлении. Тем самым ADA стремится сделать признаки *инвариантными к предметной области*.

В качестве извлекателя признаков используется BERT: классификатор и дискриминатор – полносвязные слои с активацией ReLU [24]. Гиперпараметр $\lambda_{ADA} > 0$ управляет степенью подавления доменной информации.

В нашей работе решается не задача доменной адаптации, а проблема снижения влияния тематических сдвигов. Поэтому важным отличием по сравнению с исходным методом ADA является то, что мы используем в обучении целевые метки обоих доменов.

Состязательный механизм делает скрытые представления нечувствительными к источнику данных. Это уменьшает зависимость классификатора от доменных маркеров (лексики, специфической для источника, стиля),

что важно при наличии сильного распределительного сдвига. Поскольку тема текста часто является сильным доменным маркером, дискриминатор вынуждает извлекатель признаков подавлять тематическую информацию. Таким образом, классификация становится менее зависимой от тематического контекста и более устойчивой к различиям в темах между Mail.Ru и AWD.

3.2. Состязательная адаптация домена на основе энергии (EADA)

Energy-based Adversarial Domain Adaptation (EADA) [4] является улучшенной версией метода ADA. EADA оптимизирует значение *энергетической функции*, характеризующей различие между доменами: *автоэнкодер* усиливает различия, а *извлекатель признаков* их подавляет, что приводит к более стабильному и гладкому распределению признаков. Это улучшает переносимость модели между доменами. В нашей постановке задачи это может сделать модели менее чувствительными к тематическим смещениям и повысить переносимость модели на другие источники.

Тематические различия между доменами проявляются в структуре текстовых эмбеддингов. Извлекатель признаков в EADA вынужден формировать такие эмбеддинги, которые *не содержат устойчивых тематических сигналов*, поскольку автоэнкодер их постоянно выделяет. Это уменьшает зависимость классификатора от тематики текстов и снижает проседание качества при смене тем.

Модель EADA включает классификатор целевого класса и автоэнкодер. Оптимизация записывается следующим образом:

$$\min_{G_f, G_y} L_{CE}(X_s, Y_s) + \lambda L_{AE}(X_t), \tag{3}$$

$$\min_{G_a} \big(L_{AE}(X_s) + \max(0, m - L_{AE}(X_t))\big). \tag{4}$$

Автоэнкодер G_a стремится усиливать различия между эмбеддингами исходного и целевого доменов, в то время как извлекатель признаков G_f пытается их уменьшить, обеспечивая сохранение информации, необходимой для классификации.

Гиперпараметр λ_{ADA} задаёт силу штрафа за доменно-специфические признаки, а параметр m определяет минимальную степень отделимости представлений доменов в пространстве автоэнкодера.

3.3. Контрастная функция потерь (Contrastive Loss)

Другим важным и потенциально эффективным методом является контрастная функция потерь [5]. Она устроена таким образом, что

стимулирует модель к сближению эмбеддингов текстов с одинаковыми целевыми метками.

Пусть $P(i)$ – индексы текстов той же метки, что и x_i. Функция потерь:

$$L_{cl} = \sum_{i \in I} \frac{-1}{|P(i)|} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}, \tag{5}$$

а итоговая функция потерь имеет вид:

$$L_{total} = (1 - \lambda_{CL}) L_{ce} + \lambda_{CL} L_{cl}. \tag{6}$$

В работе вводятся изменения:

- семплирование позитивных примеров только с противоположным значением конфаундера, чтобы стимулировать объединение по целевой метке вне зависимости от источника или темы;
- нормировка L_{cl} для стабилизации обучения.

Мотивация заключается в том, чтобы сделать эмбеддинги текстов с разными значениями конфаундера максимально близкими друг к другу при условии совпадения значений целевого класса. Чем более похожими становятся эмбеддинги текстов с разными значениями конфаундера, тем сложнее восстановить значение конфаундера из признакового пространства, выученного нейронной сетью. Таким образом, данная модификация контрастной функции потерь должна способствовать снижению влияния тематических сдвигов на предсказание целевого класса.

3.4. ADA + контрастная функция потерь

ADA и контрастная функция потерь по-своему полезны для уменьшения влияния тематических сдвигов. При этом актуальным является вопрос о том, насколько эти методы эффективны при совместном применении. Для этого проводятся эксперименты с комбинированной функцией потерь:

$$L_{total} = \lambda_{cl} L_{cl} + (1 - \lambda_{cl}) L_{ada}. \tag{7}$$

Контрастная часть сближает элементы одного класса, а ADA делает признаки доменно-инвариантными, обеспечивая двойную регуляризацию. ADA подавляет тематические сигналы как доменные особенности, а контрастная функция потерь делает похожими эмбеддинги текстов с одинаковым значением целевой метки, даже если темы различаются. В сочетании это может быть особенно эффективно против тематических сдвигов.

3.5. Эксперименты с LLaMA 3.2 3B Instruct

Также проводятся эксперименты с LLaMA 3.2 3B Instruct [8]. Эта модель обладает значительно большей обобщающей способностью за счёт

приблизительно 3 миллиардов параметров.

Крупные языковые модели обладают высокой степенью доменной обобщаемости. Эксперименты позволяют проверить, насколько одного лишь проектирования промптов (инструкций) достаточно для подавления тематических эффектов без доменной адаптации.

Основная цель экспериментов – выяснить, может ли LLaMA обеспечивать устойчивость к тематическим сдвигам просто за счёт составления промптов без дополнительного обучения. Пример промпта (инструкции) на английском языке приведён в подразделе 6.3.

4. Эксперименты

Основной метрикой для сравнения моделей в данной работе является точность (accuracy).

Целью данного исследования является обучение надежного классификатора для пола автора текста, для которого изменение предметной области по сравнению с обучающим набором данных как можно меньше ухудшало бы точность модели.

Для этого наборы данных Блоги Mail.Ru и AWD табл. 1 разбиваются в соотношении 4:1 на тренировочную и тестовую часть. Обучение проводилось на наборе данных, полученном сэмплированием $\alpha\%$ текстов из тренировочной части Блогов Mail.Ru и $(100 - \alpha)\%$ текстов из AWD. Тестовая часть из каждого источника данных используется для оценки качества моделей. Для того, чтобы рассмотреть поведение моделей и в случае преобладания Mail.Ru и в случае преобладания AWD в тренировочном множестве текстов, используются $\alpha = 25, 75, 90$.

Для оценки моделей вычисляется:

- точность текстов из источника, преобладающего в тренировочном наборе данных;
- точность текстов из недостаточно представленного источника данных;
- Определим разницу в точности текстов из источника данных, преобладающего в обучающем наборе данных, и текстов из источника данных, недостаточно представленных в тренировочном наборе данных, как *разница в точности*. Обозначим её δ.

Параметр δ показывает, насколько снижается точность классификации при тестировании на текстах из источника, мало представленного в обучении. По своей сути, это значение является метрикой для оценки эффекта тематических сдвигов в источнике текста, преобладающем в обучающем наборе данных.

В рамках экспериментов, обучаются модели на основе BERT с применением алгоритмов ADA, EADA, Contrastive Loss, Contrastive Loss + ADA и CausaLM. Используется многоязычный BERT с базовой конфигурацией (12 слоев, 768 скрытых элементов, 12 голов, 125 миллионов параметров, google-bert/bert-base-многоязычный интерфейс в оболочке HuggingFace) в качестве основы для всех экспериментов. Во всех экспериментах используется learning rate=10^{-5}, поскольку это значение предложено в [16] и [4].

Для экспериментов с контрастной функцией потерь и её комбинацией с ADA используются те же самые параметры архитектуры.

Но для уменьшения числа экспериментов, при подборе оптимальных значений λ_{CL} и λ_{ADA} для контрастной функции потерь и ADA+CL фиксируются значения следующих гиперпараметров:

- $\tau = 0.5$;
- $K = 5$;
- стратегия отбора проб – взятие K случайных положительных примеров.

После нахождения оптимальных значений λ_{CL} и λ_{ADA} производится перебор $\tau \in [0.25, 0.5, 0.75]$ и $K \in [3, 5, 7, 9]$.

5. Данные

Классификация пола автора текста это нетематическая классификация, поскольку целевой классификационной переменной является не тема или тематический признак, а более сложное понятие, которое невозможно описать с помощью определенных ключевых слов.

Для экспериментов использовались два набора данных. Первый содержит тексты из Блогов Mail.Ru, второй – из AWD. Каждый датасет содержит около 10000 текстов.

В таблице 1 показано распределение по полу и длине текста. Здесь и далее $L_{10\%}$ – 10-й перцентиль длин, $L_{\text{ср}}$ – среднее арифметическое длин, L_{median} – медиана длин.

Таблица 1. Наборы данных для обучения и тестирования классификаторов с указанием распределения длин

Источник	#M	#W	$L_{\text{ср}}$	$L_{10\%}$	$L_{25\%}$	L_{median}	$L_{75\%}$	$L_{90\%}$
MAIL	3236	6764	217	71	83	115	188	370
AWD	5984	4016	84	13	21	39	76	144

В наборе данных Блоги Mail.Ru около 32% текстов написаны мужчинами, а остальные – женщинами. В AWD распределение полов распределение

иное, поскольку около 60% текстов написаны мужчинами. Более того, тексты из этих датасетов имеют разное распределение по длине.

Во всех экспериментах берутся 8000 примеров для обучения и 2000 текстов для тестирования. И Блоги Mail.Ru и AWD являются российскими платформами социальных сетей. Однако их контент и целевая аудитория различаются. Mail.Ru Blogs – это универсальная платформа, которая включает в себя широкий спектр тем, включая спорт, политику, технологии, здравоохранение, науку, туризм и так далее. В то же время, AWD – это платформа о туризме. Это приводит к значительным тематическим сдвигам в наборе данных AWD.

Эти наборы данных имеют эталонные метки для половой принадлежности авторов текстов, поскольку пол указывается самими пользователями платформ.

6. Результаты

6.1. Предсказание источника данных для текста

Сопоставим источникам данных mail.ru и awd бинарную метку: $awd = 0, mail = 1$. Номер источника данных ($mail$ или awd) используется в качестве искажающего фактора как для состязательных методов, так и для каузальных моделей. Чтобы убедиться, что распределение текстов в этих двух источниках данных существенно отличается, проведен ряд экспериментов с классификаторами источника текста.

Результаты экспериментов показывают, как на точность исходных классификаторов влияет доля источников данных в тренировочном наборе данных. Многоязычные классификаторы BERT обучены на наборах данных, содержащих $\alpha = 10, 25, 50$ и 75 процентов текстов Mail.Ru. Таблица 2 показывает, что вне зависимости от доли текстов из блогов Mail.Ru в обучении точность классификатора источника текста превышает 90%.

Таблица 2. Зависимость точности классификатора от состава тренировочного набора данных; α - процент текстов из Блогов Mail.Ru в тренировочном наборе данных

α	точность
25	0.901
50	0.931
75	0.920
90	0.905

Это означает, что тексты в Mail.Ru и AWD сильно отличаются, и даже базовый BERT без применения каких-либо дополнительных методов и алгоритмов способен заметить разницу между ними. Предполагается, что это обусловлено тематическими смещениями в AWD, вызванными специализацией этого веб-сайта. Различия показывают уязвимость базовой модели BERT к такого рода тематическим сдвигам.

6.2. Дообучение состязательных моделей

Разница δ между точностью на текстах из источника данных, преобладающего в обучении, и точностью на источнике данных с недостаточным представлением используется в качестве второго ключевого показателя для оценки уязвимости модели к тематическим сдвигам в тестовых данных.

Наши эксперименты показали, что ADA помогает уменьшить разницу в точности между тестированием на текстовом источнике, преобладающем в тренировочном наборе данных, и тестированием на текстах из недостаточно представленного источника. Увеличение значения гиперпараметра λ_{ADA} соответствует уменьшению значения δ. При этом точность на текстах из тестового набора данных для источника, преобладающего в тренировочном наборе данных, снижается, в то время как точность на текстах из недопредставленного источника увеличивается, что делает модели более устойчивыми к изменениям предметной области.

Зависимость эффективности метода CausaLM от λ_{ADA} оказалась аналогична зависимости для состязательных методов: чем выше значение λ_{ADA}, тем ниже значение δ и тем выше точность на текстах из источника, недостаточно представленного в тренировочных данных. Хотя при этом CausaLM обеспечил меньшую разницу δ, чем состязательные методы, но снижение точности на текстах из чрезмерно представленного источника оказалось более значительным, чем для метода ADA.

Оказалось, что точность на текстах из Mail.Ru выше, чем на текстах из AWD, даже когда в обучении преобладает AWD ($\alpha = 25$). Это вызвано большей средней длиной текстов из Mail.Ru, поскольку длина текста влияет на точность моделей, основанных на архитектуре Трансформер. В этом случае значение δ становится отрицательным и менее информативным, чем прямая оценка точности на текстах из источника, недостаточно представленного в обучающей выборке.

Предложенная нами комбинация методов ADA и Contrastive Loss позволила достичь наивысшей точности на недопредставленном источнике AWD при сбалансированной ($\alpha = 75$) и высокой ($\alpha = 90$) доле Mail.Ru в обучающих данных. Результаты экспериментов представлены в таблице 3.

Таблица 3. Результаты экспериментов (accuracy);
наилучшие значения выделены **жирным**, вторые по величине – *курсивом*

Метод	λ_{ADA}	точность, mail		точность, awd		δ	
		$\alpha = 75$	$\alpha = 90$	$\alpha = 75$	$\alpha = 90$	$\alpha = 75$	$\alpha = 90$
BERT	0	0.838		0.716		0.122	
CausaLM	0.05	0.785		0.725		*0.060*	
CausaLM	0.2	0.781		0.728		**0.054**	
ADA	0.05	0.830		0.725		0.105	
ADA	0.2	0.825		0.731		0.094	
ADA	0.5	0.815		0.719		0.096	
EADA, m=2	0.05	0.819		0.688		0.131	
EADA, m=2	0.2	0.819		0.673		0.146	
EADA, m=4	0.05	0.819		0.692		0.127	
EADA, m=4	0.2	0.815		0.685		0.130	
EADA, m=8	0.05	0.832		0.694		0.138	
EADA, m=8	0.2	0.823		0.685		0.138	
CL, $\lambda_{CL} = 0.05$		0.810		0.722		0.088	
CL, $\lambda_{CL} = 0.2$		0.800		0.722		0.078	
CL, $\lambda_{CL} = 0.2$		0.800		0.726		0.074	
ADA + CL, $\lambda_{CL} = 0.05$	0.05	0.810	0.800	**0.734**	0.709	0.076	0.091
ADA + CL, $\lambda_{CL} = 0.1$	0.05	0.810	0.810	0.716	**0.715**	0.094	*0.095*
ADA + CL, $\lambda_{CL} = 0.2$	0.05	0.804	0.790	0.704	0.683	0.1	0.107
ADA + CL, $\lambda_{CL} = 0.05$	0.1	0.820		0.725		0.095	
ADA + CL, $\lambda_{CL} = 0.1$	0.1	0.820		*0.732*		0.088	
ADA + CL, $\lambda_{CL} = 0.05$	0.2	0.820	0.800	0.721	*0.713*	0.099	**0.087**
ADA + CL, $\lambda_{CL} = 0.1$	0.2	0.810	0.800	0.719	0.673	0.091	0.127
ADA + CL, $\lambda_{CL} = 0.2$	0.2	0.760	0.760	0.618	0.594	0.142	0.166

Наилучшая точность достигается при значениях λ_{CL} и λ_{ADA}, близким тем, которые были подобраны при отдельном применении ADA и контрастной функции потерь. Такое наблюдение позволило существенно сократить время, требуемое на подбор оптимальных гиперпараметров, выбирая λ_{CL} и λ_{ADA} по отдельности для контрастной функции потерь и ADA соответственно.

Наилучшие результаты достигнуты при умеренных значениях обоих гиперпараметров: $\lambda_{ADA} = 0.05, \lambda_{CL} = 0.05$ для $\alpha = 75$ и $\lambda_{ADA} = 0.2, \lambda_{CL} = 0.05$ для $\alpha = 90$. Это свидетельствует о сбалансированном вкладе обоих компонентов и синергетическом эффекте их комбинации.

Контрастная функция потерь как отдельно, так и при совместном использовании с ADA повышает точность классификации при преобладании Mail.Ru в обучающем наборе данных ($\alpha = 75$ и $\alpha = 90$).

При $\alpha = 25$ наивысшая и вторая по величине точность на Mail.Ru получилась с помощью ADA, в то время как EADA сработал хуже, чем ADA, но лучше, чем базовый BERT.

После фиксирования подобранных оптимальных значений $\lambda_{CL} = 0.05$ и $\lambda_{ADA} = 0.05$ производится перебор гиперпараметров $\tau \in [0.25, 0.5, 0.75]$ и $K \in [3, 5, 7, 9]$ аналогично экспериментам в статье [5]. По результатам экспериментов, представленных в таблице 4, оптимальными значениями оказались $\tau = 0.5, K = 7$.

Таблица 4. Зависимость от k и τ значений точности (*accuracy*) для *mail* и *awd* и их разности δ при $alpha = 75$ и $\lambda_{CL} = 0.05$

Метод	k	τ	*mail*	*awd*	δ
CL	5	0.5	0.800	0.726	*0.074*
CL	3	0.5	0.810	0.726	0.084
CL	7	0.25	0.813	0.708	0.095
CL	7	0.5	0.817	0.729	0.088
CL	7	0.75	0.807	0.715	0.092
CL	9	0.25	0.813	0.708	0.095
CL	9	0.5	0.817	0.729	0.088
CL	9	0.75	0.807	0.715	0.092
ADA+CL	5	0.5	0.810	*0.734*	0.076
ADA+CL	7	0.5	0.818	0.721	0.097
ADA+CL	7	0.75	0.811	0.723	*0.074*
ADA+CL	9	0.5	0.810	0.720	0.090
ADA+CL	9	0.75	0.805	**0.736**	**0.069**

Для каждого метода и каждого α в таблице 5 выбрана конфигурация с наивысшей точностью на недостаточно представленном в обучении источнике данных (выделено **жирным**) и/или наименьшим значением δ (выделено **жирным** или *курсивом*).

Таблица 5. Сводка лучших результатов для каждого метода в зависимости от α

Метод	m	α	λ_{ADA}	λ_{CL}	$mail$	awd	δ
ADA	-	25	0.2	0	**0.818**	0.759	*-0.059*
EADA	4	25	0.2	0	*0.809*	0.746	**-0.063**
CL	-	25	0	0.05	0.800	0.730	-0.055
CausaLM	-	75	0.2	0	0.781	0.728	**0.054**
ADA	-	75	0.2	0	0.825	*0.731*	0.094
EADA	8	75	0.05	0	0.832	0.694	0.138
CL	-	75	0	0.2	0.817	0.729	0.088
ADA + CL	-	75	0.05	0.05	0.805	**0.736**	0.069
ADA	-	90	0.2	0	0.825	*0.712*	*0.113*
EADA	4	90	0.05	0	0.832	0.699	0.133
ADA + CL	-	90	0.05	0.1	0.800	**0.715**	**0.095**

Напрашивается вывод, что при преобладании Блогов Mail.Ru в тренировочном наборе данных совместное использование ADA и контрастной функции потерь приводит к наилучшим результатам среди всех рассмотренных методов как по точности на AWD, так и по значению δ. ADA при этом существенно улучшает результат стандартного BERTa, уступая только комбинации ADA и контрастной функции потерь.

Кроме того, для моделей на основе BERT, ADA, контрастной функции потерь и ADA+CL с лучшими значениями гиперпараметров была проведена 5-fold кросс-валидация с использованием 5 различных случайных разбиений на обучение и тест для проверки стабильности получаемых результатов. В итоге среднее значение точности на валидационном множестве по всем кросс-валидационным разбиениям оказалось равным 0.790, 0.787, 0.788 и 0.782 для BERT, ADA, CL, ADA+CL соответственно. При этом, среднее стандартное отклонение получилось равным 0.011, 0.006, 0.007 и 0.009 соответственно. Это показывает, что модели, обученные с использованием состязательных методов, более устойчивы к изменению разбиения на тренировочный и тестовый датасет.

Для оценки статистической значимости улучшений, достигнутых с помощью состязательных методов, применен односторонний t-критерий. Для каждой экспериментальной настройки ($\alpha = 25, 75$ и 90) проводилось

многократное случайное разбиение данных с последующим расчетом точности модели. Полученные p-значения оказались меньше 0.05 для ADA при $\alpha = 75$ и $\alpha = 90$, что подтверждает статистическую значимость улучшений. Для разбиения $\alpha = 25$ результаты не достигли статистической значимости (p-value = 0.18), что, предположительно, связано с особенностями распределения длин текстов в наборе данных AWD, который преобладает в данном тренировочном наборе.

В большинстве случаев повышение точности на тестовом наборе данных является статистически значимым, что подтверждает эффективность методов состязательной адаптации предметной области для снижения последствий тематических сдвигов.

В таблице 6 можно увидеть пример правильного и неправильного

Таблица 6. Пример корректного и некорректного предсказания BERT

Обучение: awd, тест: awd	Обучение: awd, тест: mail
Летали в сентябре 2009 года в Барселону на 10 дней (с пересадкой в Амстере), потом по этой же визе *я ездила* в Таллин на 3 дня, в феврале *летала* в Мадрид на 10 дней;	Зачем же вы так принижаете своих собеседников?? - Вы хотите, чтобы мне стало стыдно? Чтобы *стала* оправдываться? А я не буду оправдываться.
Пол автора: W Предсказано: W	Пол автора: W Предсказано: M

предсказания классификатора на основе BERT. Можно увидеть, что в обоих текстах присутствуют глаголы женского рода в прошедшем времени, что явно показывает пол автора текста. При этом предсказания получились разными.

6.3. Большие языковые модели в режиме zero-shot. LLaMA 3B Instruct

Поскольку большая часть данных для обучения модели LLaMA [8] на английском языке, было решено использовать инструкцию для предсказания половой принадлежности автора текста с помощью LLaMA тоже на английском языке, см. таблицу 7.

В каждой инструкции использовались 5 текстов. При этом, рассматривались 2 вида инструкций:

- 4 случайных текста из Mail.Ru + 1 случайный текст из AWD;
- 4 случайных текста из AWD + 1 случайный текст из Mail.Ru.

Это сделано, чтобы максимально приблизить эксперименты с LLaMA 3.2 3B Instruct к экспериментам с дообучением моделей меньшего размера.

Таблица 7. Формат инструкции для LLaMA 3.2 3B Instruct. Текст с описанием задачи выделен жирным шрифтом.

These are examples of correct responses:
Text: «Текст 1>". Response: «Пол автора текста 1>".
Text: «Текст 2>". Response: «Пол автора текста 2>".
Text: «Текст 3>". Response: «Пол автора текста 3>".
Text: «Текст 4>". Response: «Пол автора текста 4>".
Text: «Текст 5>". Response: «Пол автора текста 5>".
Given this text, you should understand the gender of its author.
Return only one character: M (for male) or W (for female)
Text:
"Всем доброго времени суток. у меня 4 перелета Львов-Стамбул-Белград и Сараево-Стамбул-Харьков"

По результатам таблицы 8 видно, что точность больших языковых

Таблица 8. Точность LLaMA 3.2 3B Instruct

Обучение	тест	точность	время/текст, сек.
mail	mail	0.74	4.8
awd	mail	0.73	4.6
mail	awd	0.42	3.9
awd	awd	0.54	3.8

моделей без дообучения существенно уступает точности дообученных моделей меньшего размера. При этом, время предсказания на один текст существенно превышает время для BERT и состязательных методов на основе BERT. Это показывает, что даже несмотря на быстрое развитие больших языковых моделей, всё ещё имеет смысл дообучать архитектуры меньшего размера и исследовать их возможные улучшения.

По результатам экспериментов можно сделать следующие выводы:

(*1*) Тематические смещения оказывают значительное влияние на классификатор пола автора текста;

(*2*) использование состязательных и каузальных методов немного повышает точность модели при тестировании на текстах из другой предметной области;

(*3*) хотя при использовании CausaLM разница уменьшается в большей степени, методы ADA и Contrastive Loss позволяют добиться этого без существенного снижения точности на текстах из преобладающего источника. Время, затрачиваемое на обучение причинно-следственных моделей с использованием ADA, значительно меньше, чем при использовании CausaLM. Это делает использование ADA более раци-

ональным с точки зрения требуемых временных и вычислительных ресурсов;

(4) несмотря на то, что это более сложный алгоритм, EADA показывает улучшение только при семплирования тренировочного набора данных из 75% mail.ru и 25% данных awd. Следовательно, метод оказывается менее эффективным, чем ADA и контрастная функция потерь;

(5) контрастная функция потерь позволяет повысить точность классификации по сравнению со стандартным BERT при преобладании текстов из mail.ru в тренировочном наборе данных;

(6) при преобладании текстов из mail.ru в тренировочном наборе данных, совместное использование ADA и Contrastive Loss оказывается более эффективным, чем использование просто ADA;

(7) LLaMA 3.2 3B Instruct без дообучения показывает более низкую точность на тесте, но при этом требует больше времени на предсказание одного текста, что делает её применение к задаче классификации пола автора текста не оправданной.

6.4. Анализ ошибки

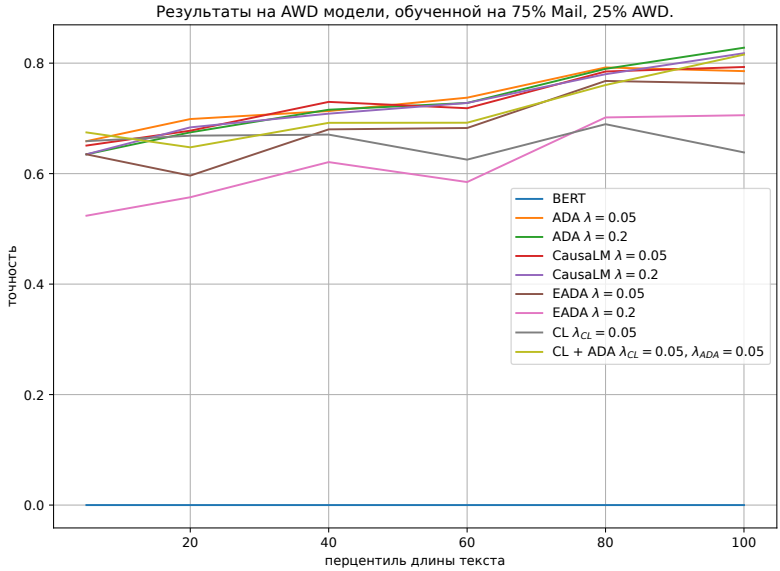

Рисунок 1. Влияние длины текста на эффективность состязательных методов при семплировании, $\alpha = 75$

На графике рисунка 1 показано, как длина текстов влияет на точность

модели. Тексты разбиваются на группы одинакового размера по количеству слов в них. Видно, что ADA с $\lambda = 0,05$ при обучении на наборе данных $\alpha = 75$ превзошёл стандартный BERT при любой длине текста. Однако при семплировании в тренировочную выборку $\alpha = 25$ состязательные методы менее стабильны на самых коротких и самых длинных текстах.

В отличие от ADA, в EADA есть дополнительный гиперпараметр $m-$ разница между представлениями из исходного домена и целевого домена. Значение по умолчанию, рекомендованное в [4], равно 4. Также были проверены $m = 2$ и $m = 8$ для тренировочного набора данных $\alpha = 75$. Поскольку все они показали худший результат, чем $m = 4$, мы оставили $m = 4$ для всех остальных экспериментов.

График для обучения на $\alpha = 75$ показывает чёткую закономерность, согласно которой точность модели имеет сильную положительную корреляцию с длиной текста в тестовом наборе данных. Это видно и для всех моделей, обученных на этом распределении mail/awd, что подтверждает выводы из [25].

Заметно, что использование контрастной функции потерь повышает точность классификации пола автора текста на awd в случае, когда в обучении преобладают тексты из mail.ru. Это соответствует ожиданиям об эффективности метода.

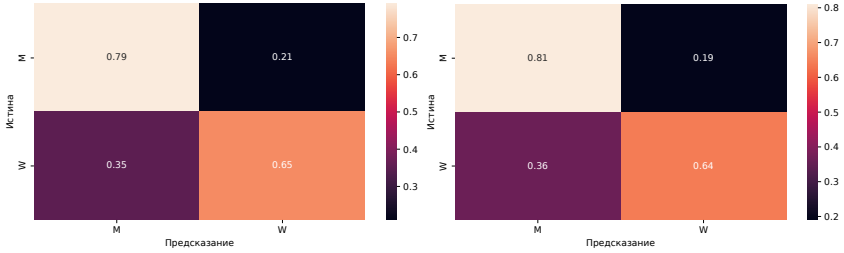

Рисунок 2. Матрица ошибок при обучении на $\alpha = 75$ и тестировании на awd для BERT и ADA с $\lambda_{ADA} = 0.05$ соответственно.

Рисунки 2 и 3 показывают как распределены ошибки при классификации с использованием стандартного BERT и классификации при использовании Adversarial Domain Adaptation с $\lambda_{ADA} = 0.05$. Видно, что при обучении на наборе данных с преобладанием текстов из Блогов Mail.Ru и тестировании на AWD, точность на текстах женского мужского выше. При обучении на наборе данных с преобладанием текстов из AWD и тестировании на Блогах Mail.Ru результат оказывается противоположным.

При преобладании Блогов Mail.Ru в обучении доля текстов женских авторов становится выше 50% в связи с распределением полов в наборах данных табл. 1 и ADA снижает долю ошибок при предсказании для текстов авторов как мужского, так и женского пола. При преобладании AWD в обучении ADA показывает существенное повышение точности для текстов авторов женского пола, но при этом доля ошибок для текстов авторов мужского пола увеличивается.

Вне зависимости от распределения источников данных в обучении, модель, обученная с использованием ADA, снижает долю ошибок на текстах того пола, который менее представлен в обучении.

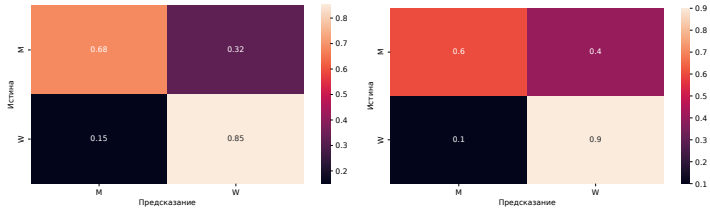

Рисунок 3. Матрица ошибок при обучении на 25% mail.ru + 75% awd и тестировании на Блогах mail.ru для BERT и ADA с $\lambda_{ADA} = 0.05$ соответственно

7. Время обучения

Рассчитано время, необходимое для обучения модели каждой архитектуры, упомянутой в экспериментах. Все модели обучены на одном графическом процессоре на базе Nvidia TITAN RTX. Доступный объем графического процессора: 24 ГБ. Результаты представлены в таблице 9.

Таблица 9. Время обучения всех моделей на наборе данных Блоги Mail.Ru. Модель, которая обучается быстрее всех, выделена жирным шрифтом. Курсивом выделена вторая по скорости модель.

Модель	Источник	Время на эпоху, сек.
BERT	mail	**117**
ADA	mail	*128*
EADA	mail	132
CausaLM	mail	234
CL	mail	162
CL + ADA	mail	174

Мы видим, что добавление состязательной функции потерь при использовании методов доменной адаптации (как ADA, так и EADA) к моделям BERT не приводит к существенному увеличению времени обучения. При этом заметное увеличение времени обучения происходит при использовании контрастной функции потерь, поскольку для неё требуется предподсчёт всех эмбеддингов текстов, полученных на предыдущей эпохе. Самым долгим по обучению методом является CausaLM.

Заключение

В работе рассмотрена проблема изменений в распределении обучающих данных, в частности, проблема тематических сдвигов. Предложена модификация семплирования для контрастной функции потерь, нацеленная на снижение эффектов тематических сдвигов. Впервые проведено исследование комбинированного метода ADA с контрастной функцией потерь.

Результаты экспериментов показывают, что состязательные методы полезны для улучшения нетематических классификаторов текстов при наличии тематических сдвигов и изменении предметной области. В целом, использование контрастной функции потерь может быть рекомендовано для случаев, когда в тестовых данных происходит значительный сдвиг предметной области. При этом, ADA остаётся наиболее эффективным методом в случае преобладания одной темы в обучающем наборе данных. Это важный практический результат, учитывая распространённость и актуальность задач нетематической классификации текстов в современном мире.

Кроме того, в работе показывается, что несмотря на быстрый прогресс в развитии больших языковых моделей, для ряда задач они все ещё остаются менее эффективными, чем дообученные модели меньшего размера.

Дальнейшие исследования будут направлены на адаптацию состязательных и контрастных методов для задач регрессии на текстах.

Список использованных источников

[1] Sharoff S., Wu Z., Markert K. *The web library of babel: evaluating genre collections // Proceedings of the Seventh Conference on International Language Resources and*

Evaluation (Valletta, Malta, 17–23 May 2010).– European Language Resources Association.– 2010.– ISBN 2-9517408-6-7.– Pp. 3063–3070. ↑58

[2] Petrenz P., Webber B. *Stable classification of text genres* // Computational Linguistics.– 2011.– Vol. **37**.– No. 2.– Pp. 385–393. ↑58

[3] Tzeng E., Hoffman J., Saenko K., Darrell T. *Adversarial discriminative domain adaptation* // *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, Hawaii, USA, 21–26 July 2017).– IEEE.– 2017.– ISBN 9781538604588.– Pp. 2962–2971. ↑58, 60, 62

[4] Zou H., Yang J., Wu X. *Unsupervised energy-based adversarial domain adaptation for cross-domain text classification* // *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP 2021 (Online, 1–6 August 2021).– ACL.– 2021.– ISBN 9781713833116.– Pp. 1208–1218. ↑58, 60, 63, 66, 75

[5] Jiang T. *Learn from failure: Causality-guided contrastive learning for generalizable implicit hate speech detection* // *Proceedings of the 31st International Conference on Computational Linguistics*, COLING 2025 (Industry Track, Abu Dhabi, UAE, 19–24 January 2025).– ACL.– 2025.– ISBN 979-8-3313-1356-2.– Pp. 8858–8867. ↑58, 60, 63, 70

[6] Suresh V., Ong D. C. *Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification.*– 2021.– 14 pp. arXiv 2109.05427 ↑58

[7] Lin N., Qin G., Wang G., Zhou D.-p., Yang A. *An effective deployment of contrastive learning in multi-label text classification* // *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, ACL 2023 (Toronto, Canada, 9–14 July 2023).– 2023.– ISBN 978-1-959429-72-2.– Pp. 8730–8744.– 13 pp. arXiv 2212.00552 ↑58

[8] Llama Team, AI@Meta *The Llama 3 herd of models.*– 2024.– 92 pp. arXiv 2407.21783 ↑58, 61, 64, 72

[9] OpenAI *GPT-4 technical report.*– 2023.– 100 pp. arXiv 2303.08774 ↑58, 61

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao et al. *DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning* // Nature.– 2025.– Vol. **645**.– No. 8081.– Pp. 633-–638. ↑58

[11] Gao T., Jin J., Ke Z. T., Moryoussef G. *A comparison of DeepSeek and other LLMs.*– 2025.– 30 pp. arXiv 2502.03688 ↑59

[12] Rahman A., Mahir S. H., An Tashrif Md. T., Aishi A. A., Karim Md. A., Kundu D., Debnath T., Ala Moududi Md. A., Eidmum Md. Z. A. *Comparative analysis based on DeepSeek, ChatGPT, and Google Gemini: Features, techniques, performance, future prospects.*– 2025.– 20 pp. arXiv 2503.04783 ↑59

[13] Lepekhin M., Sharoff S. *Causal Models and Adversarial Training: Selecting the right properties for robust non-topical text classification*, Proceedings of the International Conference "Dialogue 2025" (April 23-–25, 2025), Computational Linguistics and Intellectual Technologies.– vol. **23**.– 2025.– Pp. 224–233.– 10 pp. doi URL ↑59

[14] Devlin J., Chang M.-W., Lee K., Toutanova K. *BERT: Pre-training of deep bidirectional transformers for language understanding.*– 2018.– 16 pp. arXiv 1810.04805 ↑59

[15] Conneau A., Khandelwal K., Goyal N., Vishrav C., Wenzek G., Guzman F. J., Grave E., Ott M., Zettlemoyer L., Stoyanov V. *Unsupervised cross-lingual representation learning at scale // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, 5–10 July 2020).– ACL.– 2019.– Pp. 8440–8451. doi arXiv 1911.02116 ↑59

[16] Sun C., Qiu X., Xu Y., Huang X. *How to Fine-Tune BERT for Text Classification?*– 2019.– 10 pp. arXiv 1905.05583 ↑59, 66

[17] Basile V. *Domain adaptation for text classification with weird embeddings // Proceedings of the Seventh Italian Conference on Computational Linguistics* (Bologna, Italy, 1–3 March 2021), CEUR Workshop Proceedings.– vol. **2769**.– 2020.– id. 34.– 7 pp. URL ↑60

[18] Han C., Fan Z., Zhang D., Qui M., Gao M, Zhou A. *Meta-learning adversarial domain adaptation network for few-shot text classification // 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP 2021 (Online, 1–6 August 2021).– ACL.– 2021.– ISBN 9781713833116.– Pp. 1664-1673. doi arXiv 2107.12262 ↑60

[19] Feder A., Oved N., Shalit U., Reichart R. *CausaLM: Causal model explanation through counterfactual language models.*– 2020.– 54 pp. arXiv 2005.13407 ↑60, 61

[20] Maiya A. S. *CausalNLP: A practical toolkit for causal inference with text.*– 2021.– 9 pp. arXiv 2106.08043 ↑60

[21] Gemini Team, Google *Gemini: A Family of highly capable multimodal models.*– 2023.– 90 pp. arXiv 2312.11805 ↑61

[22] Zhao W. X., Zhou K., Li J., Tang T., Wang X., Hou Y., Min Y., Zhang B., Zhang J., Dong Z., Du Y., Yang Ch., Chen Y., Chen Zh., Jiang J., Ren R., Li Y., Tang X., Liu Z., Liu P., Nie J.-Y., Wen J.-R. *A Survey of large language models.*– 2023.– 144 pp. arXiv 2303.18223 ↑61

[23] Ramponi A., Plank B. *Neural unsupervised domain adaptation in NLP — A survey // Proceedings of the 28th International Conference on Computational Linguistics*, COLING 2020 (Barcelona, Spain, Online, 8–13 December 2020).– ACL.– 2020.– ISBN 9781713825210.– Pp. 6838–6855. doi ↑62

[24] Agarap A. F. *Deep learning using rectified linear units (ReLU).*– 2018.– 7 pp. arXiv 1803.08375 ↑62

[25] Baillargeon J.-T., Lamontagne L. *Assessing the impact of sequence length learning on classification tasks for transformer encoder models* // The International FLAIRS Conference Proceedings.– 2024.– Vol. **37**.– 7 pp. arXiv 2212.08399 ↑75

Поступила в редакцию 11.12.2025;
одобрена после рецензирования 26.01.2026;
принята к публикации 12.02.2026;
опубликована онлайн 04.03.2026.

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Информация об авторах:

Михаил Николаевич Лепехин

Аспирант Московского физико-технического института (национального исследовательского университета) (МФТИ). Область научных интересов: классификация текстов, большие языковые модели, машинный перевод.

0009-0004-1114-8403
e-mail: lepehin.mn@phystech.edu

Сергей Александрович Шаров

Профессор языковых технологий университета Лидса, кандидат физико-математических наук, автор множества работ в области обработки естественного языка. Научные интересы связаны с тремя областями: лингвистикой (в первую очередь компьютерной лингвистикой и корпусной лингвистикой), когнитивными науками и коммуникационными исследованиями.

0000-0002-4877-0210
e-mail: s.sharoff@leeds.ac.uk

Вклад авторов: *М. Н. Лепехин* – 70% (методология, программное обеспечение, расследование, написание черновой версии); *С. А. Шаров* – 30% (идея, курирование данных, доработка и редактирование).

Декларация об отсутствии личной заинтересованности: *благополучие авторов не зависит от результатов исследования.*

EN RU PROGRAM SYSTEMS: THEORY AND APPLICATIONS ISSN 2079-3316
Research Article ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

UDC 004.89:004.93
10.25209/2079-3316-2026-17-1-57-84

Comparative Analysis of the Adversarial Methods For Non-Topical Classification of Texts

Mikhail Nikolaevich **Lepekhin**[1], Sergey Aleksandrovich **Sharoff**[2]

[1] Moscow Institute of Physics and Technology, Moscow, Russia
[2] University of Leeds, Leeds, UK
[1]✉*lepekhin.mn@phystech.edu*

Abstract. Non-topical text classification is widely used in modern applications. One of the issues related to this problem is the presence of biases and shifts in the distribution in the training text datasets. The most significant type of shift is the topical shift. To handle this issue we apply competitive methods such as Adversarial Domain Adaptation, Energy-based ADA, BERT with contrast loss function, ADA with contrast loss function.

In this paper, we first modify the contrast loss function to reduce the influence of thematic shifts and show that the use of adversarial methods improves the accuracy and reliability of classifiers for the task of determining the gender of the author of a text. We also apply LLaMA-3B and show that the large language models attain lower accuracy in the few-shot mode and require more time for prediction than the pre-trained models based on smaller architectures. (*In Russian*).

Key words and phrases: adversarial methods, contrastive loss, gender classification, text classification, non-topical classification, bert, domain adaptation

2020 *Mathematics Subject Classification*: 68T50; 68T07, 68T20

For citation: Mikhail N. Lepekhin, Sergey A. Sharoff. *Comparative Analysis of the Adversarial Methods For Non-Topical Classification of Texts*. Program Systems: Theory and Applications, 2026, **17**:1(70), pp. 57–84. (*In Russ.*). https://psta.psiras.ru/read/psta2026_1_57-84.pdf

References

[1] S. Sharoff, Z. Wu, K. Markert. "The web library of babel: evaluating genre collections", *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (Valletta, Malta, 17–23 May 2010), European Language Resources Association, 2010, ISBN 2-9517408-6-7, pp. 3063–3070.

[2] P. Petrenz, B. Webber. "Stable classification of text genres", *Computational Linguistics*, **37**:2 (2011), pp. 385–393.

[3] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell. "Adversarial discriminative domain adaptation", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, Hawaii, USA, 21–26 July 2017), IEEE, 2017, ISBN 9781538604588, pp. 2962–2971.

[4] H. Zou, J. Yang, X. Wu. "Unsupervised energy-based adversarial domain adaptation for cross-domain text classification", *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP 2021 (Online, 1–6 August 2021), ACL, 2021, ISBN 9781713833116, pp. 1208–1218.

[5] T. Jiang. "Learn from failure: Causality-guided contrastive learning for generalizable implicit hate speech detection", *Proceedings of the 31st International Conference on Computational Linguistics*, COLING 2025 (Industry Track, Abu Dhabi, UAE, 19–24 January 2025), ACL, 2025, ISBN 979-8-3313-1356-2, pp. 8858–8867.

[6] V. Suresh, D. C. Ong. "Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification", 2021, 14 pp. arXiv 2109.05427

[7] N. Lin, G. Qin, G. Wang, D.-p. Zhou, A. Yang. "An effective deployment of contrastive learning in multi-label text classification", *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, ACL 2023 (Toronto, Canada, 9–14 July 2023), 2023, ISBN 978-1-959429-72-2, pp. 8730–8744, 13 pp. arXiv 2212.00552

[8] Llama Team, AI@Meta. *The Llama 3 herd of models*, 2024, 92 pp. arXiv 2407.21783

[9] OpenAI. *GPT-4 technical report*, 2023, 100 pp. arXiv 2303.08774

[10] Guo Daya, Yang Dejian, Zhang Haowei, Song Junxiao, Wang Peiyi, Zhu Qihao, Xu Runxin, Zhang Ruoyu, Ma Shirong, Bi Xiao, Zhang Xiaokang, Yu Xingkai, Wu Yu, F. Wu Z., Gou Zhibin, Shao Zhihong, Li Zhuoshu, Gao et al. Ziyi. "DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning", *Nature*, **645**:8081 (2025), pp. 633-–638.

[11] T. Gao, J. Jin, Z. T. Ke, G. Moryoussef. *A comparison of DeepSeek and other LLMs*, 2025, 30 pp. arXiv 2502.03688

[12] A. Rahman, S. H. Mahir, Md. T. An Tashrif, A. A. Aishi, Md. A., Karim Kundu D., Debnath T., Ala Moududi Md. A., Eidmum Md. Z. A.. *Comparative analysis based on DeepSeek, ChatGPT, and Google Gemini: Features, techniques, performance, future prospects*, 2025, 20 pp. arXiv 2503.04783

[13] M. Lepekhin, S. Sharoff. "Causal Models and Adversarial Training: Selecting the right properties for robust non-topical text classification", Proceedings of the International Conference "Dialogue 2025" (April 23-–25, 2025), Computational Linguistics and Intellectual Technologies, vol. **23**, 2025, pp. 224–233, 10 pp.

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. *BERT: Pre-training of deep bidirectional transformers for language understanding*, 2018, 16 pp. arXiv 1810.04805

[15] A. Conneau, K. Khandelwal, N. Goyal, C. Vishrav, G. Wenzek, F. J. Guzman, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov. "Unsupervised cross-lingual representation learning at scale", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, 5–10 July 2020), ACL, 2019, pp. 8440–8451. arXiv 1911.02116

[16] C. Sun, X. Qiu, Y. Xu, X. Huang. *How to Fine-Tune BERT for Text Classification?* 2019, 10 pp. arXiv 1905.05583

[17] V. Basile. "Domain adaptation for text classification with weird embeddings", *Proceedings of the Seventh Italian Conference on Computational Linguistics* (Bologna, Italy, 1–3 March 2021), CEUR Workshop Proceedings, vol. **2769**, 2020, id. 34, 7 pp.

[18] C. Han, Z. Fan, D. Zhang, M. Qui, M Gao, A. Zhou. "Meta-learning adversarial domain adaptation network for few-shot text classification", *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP 2021 (Online, 1–6 August 2021), ACL, 2021, ISBN 9781713833116, pp. 1664-1673, 10 pp. arXiv 2107.12262

[19] A. Feder, N. Oved, U. Shalit, R. Reichart. *CausaLM: Causal model explanation through counterfactual language models*, 2020, 54 pp. arXiv 2005.13407

[20] A. S. Maiya. *CausalNLP: A practical toolkit for causal inference with text*, 2021, 9 pp. arXiv 2106.08043

[21] Gemini Team, Google. *Gemini: A Family of highly capable multimodal models*, 2023, 90 pp. arXiv 2312.11805

[22] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J., Zhang Dong Z., Du Y., Yang Ch., Chen Y., Chen Zh., Jiang J., Ren R., Li Y., Tang X., Liu Z., Liu P., Nie J.-Y., Wen J.-R.. *A Survey of large language models*, 2023, 144 pp. arXiv 2303.18223

[23] A. Ramponi, B. Plank. "Neural unsupervised domain adaptation in NLP — A survey", *Proceedings of the 28th International Conference on Computational Linguistics*, COLING 2020 (Barcelona, Spain, Online, 8–13 December 2020), ACL, 2020, ISBN 9781713825210, pp. 6838–6855.

[24] A. F. Agarap. *Deep learning using rectified linear units (ReLU)*, 2018, 7 pp. arXiv 1803.08375

[25] J.-T. Baillargeon, L. Lamontagne. "Assessing the impact of sequence length learning on classification tasks for transformer encoder models", *The International FLAIRS Conference Proceedings*, **37** (2024), 7 pp. URL arXiv 2212.08399 doi