به نام خدا

دانشگاه صنعتی اصفهان
دانشکده علوم ریاضی

کاهش سوگیری در مدل‌های زبانی بزرگ با استفاده از تکنیک‌های آماری

پروژه کارشناسی آمار

سیدمحمد مدنی

استاد راهنما

دکتر زهرا صابری

پاییز و زمستان ۱۴۰۴

فهرست مطالب

مقدمه

ظهور مدل‌های زبان بزرگ (LLMs) پیشرفتی چشمگیر در فناوری ایجاد کرده و بسیاری از صنایع، به‌ویژه بخش مالی، را متحول ساخته است. این مدل‌های پیچیده مانند GPT-۴ و FinBERT در عملیات حساسی همچون ارزیابی اعتبار و مشاوره مشتری کاربرد گسترده پیدا کرده‌اند، اما هم‌زمان در معرض آسیب‌هایی ناشی از انواع سوگیری‌ها قرار دارند که چالش‌های مهمی برای استفاده مسئولانه از آن‌ها ایجاد می‌کند. این سوگیری‌ها که می‌توانند به‌شکل کلیشه‌های جنسیتی، نژادی، فرهنگی و اقتصادی–اجتماعی بروز یابند، در حوزه‌هایی مانند مراقبت سلامت و عدالت کیفری نگرانی‌های اخلاقی و عملی قابل‌توجهی ایجاد می‌کنند.

سوگیری در مدل‌های زبان بزرگ پدیده‌ای چندوجهی است که نه تنها بازتاب‌دهنده سوگیری‌های اجتماعی موجود است، بلکه می‌تواند از طریق تولید محتوا آن‌ها را تقویت کند. این سوگیری‌ها به‌طور کلی به دو دسته اصلی تقسیم می‌شوند :

سوگیری ذاتی (Intrinsic) و سوگیری اکتسابی(Extrinsic)

سوگیری ذاتی ریشه در داده‌های آموزشی، معماری مدل و پیش‌فرض‌های ساختاری آن دارد؛ زیرا LLMها که بر مجموعه‌داده‌های گسترده اینترنتی آموزش می‌بینند، ناگزیر سوگیری‌های موجود در این منابع را به ارث می‌برند. برای مثال، یک مدل ممکن است کلیشه‌های جنسیتی در مشاغل را بازتولید کند و «پزشک» را عمدتاً با مردان و «پرستار» را با زنان مرتبط سازد.

در مقابل، سوگیری اکتسابی در نحوه استفاده از مدل و تعامل آن با دنیای واقعی آشکار می‌شود.

پیامدهای چنین سوگیری‌هایی در حوزه‌هایی مانند استخدام، سلامت و آموزش می‌تواند جدی و ناعادلانه باشد؛ زیرا خروجی‌های مغرضانه ممکن است به تصمیم‌گیری‌های تبعیض‌آمیز منجر شود. از این رو، ایجاد چارچوب‌های ارزیابی جامع برای شناسایی و اندازه‌گیری این سوگیری‌ها ضروری است تا اطمینان حاصل شود که LLM ها به‌عنوان ابزارهایی عادلانه و بی‌طرف مورد استفاده قرار می‌گیرند. ارزیابی سوگیری معمولاً شامل بازرسی مدل در مراحل مختلف چرخه عمر آن، از پیش‌پردازش داده‌ها تا فرایند آموزش و تولید خروجی است.

۲. سوگیری در مدل‌های زبانی

۲.۱ تعریف سوگیری

برای درک عمیق مسئله سوگیری، ابتدا باید ماهیت ابزاری که با آن سر و کار داریم را بشناسیم.

مدل زبانی بزرگ (Large Language Model) به سیستم‌های هوشمندی اطلاق می‌شود که برای پردازش و تولید زبان طبیعی در مقیاس وسیع طراحی شده‌اند. به عبارت فنی‌تر و جامع LLMها شبکه‌های عصبی عمیق (معمولاً با معماری (Transformer هستند که روی حجم عظیمی از متون (کتاب‌ها، وب‌سایت‌ها، مقالات) آموزش دیده‌اند .مکانیسم اصلی آن‌ها "پیش‌بینی توکن بعدی" است؛ یعنی با توجه به کلمات قبلی، محتمل‌ترین کلمه بعدی را حدس می‌زنند . مکانیسم توجه (Attention Mechanism) در این مدل‌ها برای درک بافتار ضروری است، اما می‌تواند ناخواسته الگوهای تبعیض‌آمیز موجود در داده را برجسته‌تر کند

سوگیری درLLM ها پدیده‌ای چندلایه است که از منابع مختلفی نشأت می‌گیرد و در طول چرخه عمر مدل به شیوه‌های پیچیده‌ای نفوذ می‌کند.
مدل‌های زبانی بزرگ بر روی حجم عظیمی از داده‌های متنی اینترنتی، کتاب‌ها و مقالات آموزش می‌بینند. این داده‌ها به ناچار حاوی سوگیری‌های انسانی، کلیشه‌ها، پیش‌داوری‌ها و اطلاعات غلط هستند. انتخاب این داده‌ها نیز خود یک منشأ سوگیری است. فرضیات ذهنی توسعه‌دهندگان در مورد اهداف سیستم و کاربران، می‌تواند منجر به انتخاب مجموعه داده‌های غیرنماینده و نامتوازن

شود. این فرآیند انتخاب، سوگیری‌های تاریخی و سیستمی موجود در جامعه را به درون مدل تزریق می‌کند.
یک چرخه بازخورد منفی در این فرآیند مشاهده می‌شود: ۱. سوگیری‌های موجود در جامعه به داده‌ها منتقل می‌شوند. ۲. مدل‌ها این سوگیری‌ها را می‌آموزند و در خروجی‌های خود تقویت می‌کنند. ۳. سپس این خروجی‌ها ممکن است به عنوان داده‌های جدیدی برای آموزش مدل‌های بعدی یا تأثیر بر جامعه استفاده شوند، و به این ترتیب، نابرابری‌های موجود را تثبیت و تشدید کنند.

همچنین انتخاب یا نمونه‌گیری داده‌ها نیز در سوگیری داده‌ها نقش مهمی ایفا میکند.
روش جمع آوری داده(برای مثال انتخاب متون از نواحی جغرافیایی یا زبانی خاص یا تمرکز بر منابعی خاص) میتواند باعث شود داده‌ها نماینده کل جامعه نباشند.
به علاوه وقتی داده‌ها برچسب‌گذاری میشوند، قضاوت فاعل انسانی در برچسب‌گذاری ممکن است سوگیری بیاورد.

حتی با داده‌های نسبتاً پاکسازی‌شده، طراحی الگوریتم می‌تواند سوگیری‌ها را تقویت کند. به عنوان مثال، مکانیسم توجه (Attention Mechanism) در مدل‌های ترنسفورمر، که برای درک بافتار ضروری است، می‌تواند الگوهای تبعیض‌آمیز موجود در داده را برجسته‌تر کند. سوگیری الگوریتمی زمانی رخ می‌دهد که متغیرهای حساس مانند سن، نژاد یا جنسیت در پیش‌بینی‌ها نقش ایفا کنند، حتی اگر مستقیماً به الگوریتم معرفی نشده باشند.

۲.۲ انواع سوگیری

سوگیری‌ها در LLM ها را می‌توان به سه دسته اصلی تقسیم کرد:

- سوگیری‌های شناختی: این سوگیری‌ها که در شناخت انسان ریشه دارند، در مدل‌های زبانی نیز مشاهده شده‌اند. نمونه‌ها شامل اثر لنگر انداختن (تکیه شدید به اطلاعات اولیه در هنگام تخمین) و اثر تناسب اندازه (تداخل اطلاعات متناقض بین اندازه واقعی و نمایشی محرک‌ها) هستند. این یافته‌ها نشان می‌دهند که LLM ها ممکن است الگوهای سیستماتیک مشاهده‌شده در شناخت انسان را تقلید یا بازتولید کنند .

- سوگیری‌های آماری: این سوگیری‌ها به خطاهای سیستماتیک در فرآیندهای نمونه‌گیری و جمع‌آوری داده اشاره دارند. به عنوان مثال، سوگیری نمونه‌گیری غیر احتمالی که در آن انتخاب نمونه‌ها بر اساس قضاوت انسانی صورت می‌گیرد و نمی‌تواند معرف واقعی جامعه آماری باشد .

- سوگیری‌های اجتماعی و فرهنگی: این دسته شامل بازتاب کلیشه‌ها و نابرابری‌های اجتماعی در خروجی مدل است. نمونه‌های بارز آن شامل کلیشه‌های جنسیتی، سوگیری‌های نژادی و فرهنگی است .

۲.۳ روشهای کاهش سوگیری

برای مقابله با سوگیری‌های مدل، استراتژی‌های مختلفی توسعه یافته‌اند که به سه دسته کلی تقسیم می‌شوند: مداخلات در سطح داده(Data-level interventions) ،رویکردهای در سطح مدل(Model-level approaches) و تنظیمات پس‌پردازش(Post-processing adjustments).

مداخلات در سطح داده بر اصلاح و تعدیل مجموعه داده‌های آموزشی تمرکز دارند تا تکرار محتوای مغرضانه را کاهش دهند. این مداخلات معمولاً شامل تکنیک‌های تقویت داده، فیلتر کردن و یا نمونه‌گیری مجدد هستند. در سطح مدل، روش‌هایی مانند تنظیم تابع هزینه یا تنظیمات معماری برای به حداقل رساندن سوگیری آموخته‌شده توسط مدل به کار گرفته می‌شوند. در این پروژه به تفصیل به مداخلات در سطح داده با تمرکز بر دو تکنیک آماری کلیدی، یعنی وزن‌دهی مجدد (Reweighing)و نمونه‌گیری مجدد(Resampling) ، خواهیم پرداخت که هر دو در مرحله پیش از آموزش یا در طول فرآیند آموزش اعمال می‌شوند .

راهبردهای پیش‌پردازش

- پالایش و پاکسازی داده: حذف محتوای نادرست، نژادپرستانه و آزاردهنده از مجموعه داده‌ها .
- نمونه‌گیری متوازن: استفاده از روش‌های آماری برای مدیریت داده‌های نامتوازن (مثلاً نمونه‌گیری سهمیه‌ای یا تصادفی طبقه‌ای).
- ایجاد داده‌های مصنوعی (Synthetic Data Generation) برای پر کردن شکاف‌های آماری و افزایش تنوع .

تکنیک‌های بازنمونه‌گیری

- بیش‌نمونه‌گیری :(Oversampling) افزایش مصنوعی نمونه‌ها در کلاس اقلیت یا تولید داده‌های مصنوعی (مثلاً SMOTE)؛ مزیت: جلوگیری از اتلاف اطلاعات از کلاس اکثریت .
- کم‌نمونه‌گیری :(Undersampling) کاهش نمونه‌های کلاس اکثریت؛ مزیت کاهش زمان آموزش، عیب: احتمال از دست دادن اطلاعات مهم .
- توجه به اینکه صرف تکرار متن‌های سوگیرانه اقلیت می‌تواند سوگیری را تقویت کند و نیاز به روش‌هایی فراتر از تنظیم فراوانی (مثلاً بازنمونه‌گیری معنایی یا تکنیک‌های داده‌افزایی) وجود دارد .

روش‌های وزن‌دهی

- وزن‌دهی داده‌ها: تخصیص وزن‌های متفاوت به نقاط داده در طول آموزش (مثلاً weightدر تابع زیان یا sample_weight) تا مدل به نمونه‌های اقلیت توجه بیشتری نشان دهد. پژوهش‌ها توابع زیان وزن‌دهی‌شده را برای داده‌های تولیدشده توسط LLMها پیشنهاد کرده‌اند که وزن‌ها بر اساس معیارهای کیفیت و تنوع اختصاص می‌یابند .

راهکارهای درون‌پردازش و پس‌پردازش

- یادگیری متخاصم(Adversarial Debiasing): آموزش هم‌زمان مدل پیش‌بینی‌کننده و شبکه خصمانه تا نمایش‌های داخلی با اطلاعات حساس کمتر ساخته شود .
- بهینه‌سازی توابع هدف: تغییر تابع هدف برای مجازات خروجی‌های سوگیرانه (مثلاً وارد کردن معیارهایی مانند Demographic Parity در تابع زیان)

- پس‌پردازش: تنظیم آستانه‌های تصمیم‌گیری، بازنویسی خروجی، فیلتر کردن و کالیبراسیون خروجی‌ها؛ توجه به ریسکِ Fairwashing اظهار بی‌طرفی در حالی که سوگیری پنهان مانده) که ممکن است شفافیت را کاهش دهد.

۳. وزن‌دهی مجدد

۳.۱ تعریف

وزن‌دهی مجدد (Reweighing) یکی از مهم‌ترین و پرکاربردترین تکنیک‌های آماری برای کاهش سوگیری در مدل‌های یادگیری ماشین و به‌ویژه مدل‌های زبان بزرگ است که در سطح داده و معمولاً پیش از آموزش یا در طول فرآیند آموزش اعمال می‌شود. همان‌گونه که در فصل دوم اشاره شد، تمرکز این پروژه بر مداخلات سطح داده است و وزن‌دهی مجدد به‌عنوان یکی از دو تکنیک محوری این دسته مورد بررسی تفصیلی قرار می‌گیرد.

ایده اصلی وزن‌دهی مجدد بر این فرض استوار است که بخش قابل توجهی از سوگیری مشاهده‌شده در خروجی مدل‌ها ناشی از نامتوازنی آماری داده‌های آموزشی است. در بسیاری از مجموعه‌داده‌های واقعی، برخی گروه‌های جمعیتی (مانند جنسیت‌ها، نژادها یا طبقات اجتماعی خاص) یا کمتر نمایندگی شده‌اند یا رابطه آن‌ها با برچسب هدف به‌صورت نابرابر ثبت شده است. وزن‌دهی مجدد تلاش می‌کند بدون حذف یا تولید داده جدید، اثر این عدم توازن را در فرآیند یادگیری مدل تعدیل کند.

از دیدگاه آماری، وزن‌دهی مجدد با تخصیص ضرایب متفاوت به مشاهدات، توزیع مؤثر داده‌ها را تغییر می‌دهد. بدین معنا که اگر یک گروه خاص کمتر از سهم واقعی خود در داده حضور داشته باشد، نمونه‌های متعلق به آن گروه وزن بیشتری دریافت می‌کنند تا خطاهای مربوط به آن‌ها در تابع زیان مدل تأثیر بیشتری داشته باشد. به این ترتیب، مدل در فرآیند بهینه‌سازی مجبور می‌شود توجه بیشتری به گروه‌های محروم یا کم‌نماینده نشان دهد.

نکته کلیدی در وزن‌دهی مجدد آن است که برخلاف بازنمونه‌گیری، هیچ داده‌ای حذف یا تکثیر نمی‌شود و تنها اهمیت نسبی داده‌ها تغییر می‌کند. این ویژگی باعث می‌شود وزن‌دهی مجدد به‌ویژه برای داده‌های متنی بزرگ و پرهزینه، مانند داده‌های آموزشیLLM ها، بسیار مناسب باشد.

۳.۲ روشهای رایج وزن‌دهی مجدد

روش‌های وزن‌دهی مجدد را می‌توان بر اساس سطح اعمال و میزان پیچیدگی آن‌ها به چند دسته اصلی تقسیم کرد.

۳.۲.۱ وزن‌دهی ایستا مبتنی بر گروه (Group-based Reweighing)

در این روش کلاسیک، داده‌ها بر اساس یک یا چند ویژگی حساس مانند جنسیت یا نژاد به گروه‌های مختلف تقسیم می‌شوند. سپس وزن هر گروه بر اساس نسبت حضور آن در داده‌های آموزشی و نسبت مطلوب یا مورد انتظار تعیین می‌شود.

وزن هر دسته به‌گونه‌ای محاسبه می‌شود که توزیع مشترک ویژگی حساس و برچسب هدف به یک توزیع بی‌طرف نزدیک شود. این روش ساده، قابل تفسیر و از نظر محاسباتی کم‌هزینه است و در بسیاری از کاربردهای عملی به‌کار می‌رود.

۳.۲.۲ وزن‌دهی در سطح نمونه (Instance-level Reweighing)

در این رویکرد، وزن‌دهی به‌جای گروه‌ها، به‌صورت مجزا برای هر نمونه انجام می‌شود. این روش انعطاف‌پذیری بیشتری دارد و می‌تواند ناهمگنی درون‌گروهی را نیز در نظر بگیرد.

یکی از روش‌های متداول در این دسته، وزن‌دهی مبتنی بر مقدار زیان

(Loss-based Weighting) است. در این روش، نمونه‌هایی که مدل در پیش‌بینی آن‌ها دچار خطای بیشتری می‌شود، وزن بالاتری دریافت می‌کنند. این رویکرد باعث تمرکز مدل بر نمونه‌های دشوارتر شده و می‌تواند به‌صورت غیرمستقیم برخی سوگیری‌ها را کاهش دهد.

۳.۲.۳ وزن‌دهی پویا و یادگرفتنی (Dynamic / Learned Reweighing)

در روش‌های پیشرفته‌تر، وزن‌دهی دیگر یک فرآیند ایستا نیست، بلکه خود به‌عنوان یک مسئله یادگیری مطرح می‌شود. در این چارچوب‌ها، یک مدل یا شبکه عصبی کمکی وظیفه یادگیری وزن بهینه برای هر نمونه را بر عهده دارد.

برای مثال، در روش‌هایی مانند Meta-weighting ، وزن‌ها به‌گونه‌ای یاد گرفته می‌شوند که عملکرد مدل اصلی بر روی یک مجموعه داده اعتبارسنجی بی‌طرف بهینه شود. این روش‌ها اگرچه از نظر محاسباتی پرهزینه‌تر هستند، اما می‌توانند سوگیری‌های پیچیده‌تری را نیز هدف قرار دهند.

۳.۲.۴ مزایا و محدودیت‌ها

وزن‌دهی مجدد از یک سو مزیت مهم حفظ تمام داده‌ها را دارد و خطر از دست رفتن اطلاعات را کاهش می‌دهد، اما از سوی دیگر انتخاب وزن‌های نامناسب می‌تواند باعث ناپایداری فرآیند آموزش یا افزایش واریانس گرادیان‌ها شود. بنابراین، استفاده مؤثر از این روش نیازمند تنظیم دقیق و ارزیابی مستمر است.

۳.۳ پیاده‌سازی

پیاده‌سازی وزن‌دهی مجدد شامل چند گام اصلی است:

۱. شناسایی ویژگی‌های حساس و گروه‌های جمعیتی
۲. محاسبه توزیع واقعی داده‌ها
۳. تعیین وزن مناسب برای هر گروه یا نمونه
۴. اعمال وزن‌ها در تابع زیان یا فرآیند آموزش مدل

در این بخش برای تشریح این مسئله یک مثال برای وزن‌دهی مجدد انجام میدهیم.

بخش۱ :

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression
```

در این کد کتابخانه‌های مورد نیاز را وارد میکنیم. numpy برای تولید داد عددی تصادفی، pandas برای ساخت دیتافریم و LogisticRegression مدل پایه مناسب برای توضیح مفاهیم آماری میباشند.

بخش ۲ : تولید داده مصنوعی سوگیرانه

هدف از این بخش ساخت داده‌ای هست که یک ویژگی حساس داشته باشد، احتمال برچسب مثبت برای گروه ها متفاوت باشد و سوگیری آماری ایجاد شود.

```
np.random.seed(0)
n = 600
# ویژگی عددی ساده
X = np.random.normal(0, 1, n)
# ویژگی حساس: 0 = گروه ممتاز، 1 = گروه محروم
S = np.random.binomial(1, 0.35, n)
# برچسب هدف با سوگیری ساختاری
y = np.where(
    S == 0,
    np.random.binomial(1, 0.75, n),  # گروه ممتاز
    np.random.binomial(1, 0.45, n)   # گروه محروم
)
data = pd.DataFrame({
    "feature": X,
    "sensitive": S,
    "label": y
})
```

با توجه به این کد :

- گروه محروم **عمداً** احتمال برچسب مثبت کمتری دارد
- این همان چیزی است که در داده‌های واقعی (مثلاً استخدام یا اعتبارسنجی) رخ می‌دهد

خروجی به این شکل خواهد بود که :

```
data.head()
```

	feature	sensitive	label
0	1.764052	0	1
1	0.400157	0	1
2	0.978738	0	1
3	2.240893	0	0
4	1.867558	1	0

بخش ۳ : حال در این بخش به بررسی وجود سوگیری در داده (قبل از مدل) میپردازیم.

```
data.groupby("sensitive")["label"].mean()
```

sensitive	label
0	0.714286
1	0.459459

مشاهده میکنیم گروه محروم به‌طور معنادار نرخ قبولی کمتری دارد.

بخش ۴ :

```
# آموزش مدل پایه
X_model = data[["feature"]]
y_model = data["label"]

model_base = LogisticRegression()
model_base.fit(X_model, y_model)
```

این مدل هیچ مداخله‌ای برای کاهش سوگیری ندارد و فقط الگوی داده را یاد می‌گیرد (سوگیری را هم یاد می‌گیرد).

بخش ۵ : پیش‌بینی احتمال و محاسبه نرخ قبولی گروهی

```
# پیش‌بینی احتمال
proba_base = model_base.predict_proba(X_model)[:, 1]
```

مدل آموزش دیده و احتمالات پیش‌بینی شده در ستون proba_base ذخیره شده‌اند.

```
# آستانه تصمیم
threshold = 0.5
pred_base = (proba_base >= threshold).astype(int)

# ذخیره در دیتافریم
data["proba_base"] = proba_base
data["pred_base"] = pred_base
```

عدد threshold = ۰.۵ به این معناست که اگر مدل بیش از ٪۵۰ احتمال بدهد، فرد در دسته ۱ قرار می‌گیرد. ستون‌های proba_base (احتمالات) و pred_base (پیش‌بینی‌های نهایی) به جدول اضافه شده‌اند تا شفافیت محاسبات حفظ شود.

```
base_group_proba = data.groupby("sensitive")[["proba_base"]].mean()

base_group_proba
```

sensitive	proba_base
0	0.620145
1	0.619792

طبق جدول خروجی نهایی، میانگین احتمال پیش‌بینی شده برای گروه ممتاز (۰) برابر با ۰.۶۲۰ و برای گروه محروم (۱) برابر با ۰.۶۱۹ است.

این خروجی نشان می‌دهد که در این مرحله، مدل به طور میانگین به هر دو گروه با یک سطح از احتمال نگاه می‌کند؛ اما باید توجه داشت که این میانگین‌ها هنوز تحت تاثیر وزن‌دهی قرار نگرفته‌اند و تفاوت ناچیز آن‌ها ممکن است ناشی از این باشد که تنها یک ویژگی (feature) در مدل استفاده شده است.

بخش ۶ : محاسبه وزن‌ها

```
weights = []

for _, row in data.iterrows():
    s = row["sensitive"]
    y_val = row["label"]

    p_s = (data["sensitive"] == s).mean()
    p_y = (data["label"] == y_val).mean()
    p_sy = ((data["sensitive"] == s) & (data["label"] == y_val)).mean()

    weights.append((p_s * p_y) / p_sy)

data["weight"] = weights

data["weight"].describe()
```

این بخش برای خنثی کردن سوگیری (Bias) موجود در داده‌ها از طریق تخصیص وزن به هر نمونه آموزشی است. در این مرحله، با استفاده از نسبت احتمالات تئوری (در دنیای بدون تبعیض) به احتمالات مشاهده شده (در داده‌های فعلی)، وزن‌هایی محاسبه می‌شود تا اهمیت نمونه‌هایی که بر خلاف سوگیری هستند (مثلاً فرد محرومی که برچسب مثبت دارد) افزایش یابد. این کار به مدل اجازه می‌دهد بدون تغییر دادن برچسب‌های واقعی، از داده‌ها به شکلی منصفانه‌تر یاد بگیرد.

	weight
count	600.000000
mean	1.000000
std	0.256806
min	0.703000
25%	0.868000
50%	0.868000
75%	1.330000
max	1.349412

خروجی تابع نشان می‌دهد که وزن‌ها با موفقیت محاسبه شده و میانگین کل آن‌ها برابر با ۱.۰۰ باقی مانده است، که یعنی حجم کل داده‌ها از نظر مدل تغییر نکرده است.
دامنه وزن‌ها بین ۰.۷۰ (برای گروه‌هایی که بیش از حد مورد توجه بوده‌اند) تا ۱.۳۴ (برای گروه‌هایی که نادیده گرفته شده‌اند) متغیر است. وجود انحراف معیار ۰.۲۵ تایید می‌کند که وزن‌دهی به اندازه کافی متمایز است تا بتواند بر روند آموزش مدل در مرحله بعد تاثیر بگذارد.

بخش ۷ : آموزش مدل با وزن‌دهی مجدد

```
# آموزش مدل وزن‌دهی‌شده
model_rw = LogisticRegression()
model_rw.fit(X_model, y_model, sample_weight=data["weight"])

# پیش‌بینی احتمال
proba_rw = model_rw.predict_proba(X_model)[:, 1]
pred_rw = (proba_rw >= threshold).astype(int)

# ذخیره خروجی‌ها
data["proba_rw"] = proba_rw
data["pred_rw"] = pred_rw
```

هدف این مرحله، آموزش مجدد مدل با در نظر گرفتن عدالت اجتماعی است. با استفاده از پارامتر sample_weight در تابع fit ، به الگوریتم یاد داده میشود که برای نمونه‌های تبعیض‌آمیز اهمیت بیشتری قائل شود و به جای تکرار کورکورانه سوگیری‌های موجود در داده‌های اولیه، بر اساس وزن‌های اصلاح‌شده یاد بگیرد. این کار باعث می‌شود تصمیمات مدل نهایی به جای تکیه بر ویژگی حساس، بیشتر بر اساس شایستگی‌های فردی اتخاذ شود.

```
# این متغیر هم حتماً باید ساخته شود
rw_group_proba = data.groupby("sensitive")[["proba_rw"]].mean()

rw_group_proba
```

sensitive	proba_rw
0	0.62028
1	0.61991

مدل جدید آموزش دیده و احتمالات پیش‌بینی شده آن در ستون proba_rw ذخیره شده است. طبق جدول خروجی، میانگین احتمالات برای گروه ممتاز به ۰.۶۲۰۲ و برای گروه محروم به ۰.۶۱۹۹ رسیده است.

این نزدیکی بسیار زیاد اعداد به یکدیگر نشان می‌دهد که مدل اکنون با هر دو گروه به شکلی برابر و بدون سوگیری رفتار می‌کند و فرآیند وزن‌دهی مجدد توانسته است توازن منصفانه‌ای در پیش‌بینی‌های مدل ایجاد کند.

بخش ۸ : مقایسه نهایی

```
print("میانگین احتمال قبل از وزن‌دهی:")
print(base_group_proba)

print("\nمیانگین احتمال بعد از وزن‌دهی:")
print(rw_group_proba)

diff_before = abs(
    base_group_proba.loc[0, "proba_base"] -
    base_group_proba.loc[1, "proba_base"]
)

diff_after = abs(
    rw_group_proba.loc[0, "proba_rw"] -
    rw_group_proba.loc[1, "proba_rw"]
)

print("\nاختلاف میانگین احتمال قبل:", diff_before)
print("اختلاف میانگین احتمال بعد:", diff_after)
```

هدف نهایی از این بخش اندازه‌گیری کمی میزان اثربخشی روش وزن‌دهی مجدد است. با محاسبه قدر مطلق اختلاف میانگین احتمالات بین دو گروه حساس (۰ و ۱)، به دنبال این هستیم که ببینیم چقدر توانسته‌ایم "شکاف پیش‌بینی" را کاهش دهیم. در واقع این بخش به ما می‌گوید که آیا مدل پس از اصلاح، با هر دو گروه به شکلی یکسان‌تر برخورد می‌کند یا خیر، و آیا نابرابری‌های موجود در داده‌های اولیه در خروجی مدل تلطیف شده‌اند یا نه.

```
میانگین احتمال قبل از وزن‌دهی:
           proba_base
sensitive
0            0.620145
1            0.619792

میانگین احتمال بعد از وزن‌دهی:
           proba_rw
sensitive
0           0.62028
1           0.61991

0.00035320658657822523 :اختلاف میانگین احتمال قبل
0.0003695686089224548 :اختلاف میانگین احتمال بعد
```

قبل از وزن‌دهی : اختلاف میانگین احتمال پیش‌بینی شده برای دو گروه برابر با ۰.۰۰۰۳۵۳ بوده است.

بعد از وزن‌دهی : این اختلاف به عدد ۰.۰۰۰۳۶۹ رسیده است.

مشاهده می‌کنیم که در این آزمایش خاص، اعداد قبل و بعد از وزن‌دهی بسیار به هم نزدیک هستند (اختلاف در رقم چهارم اعشار است). این موضوع نشان می‌دهد که مدل پایه از ابتدا هم رفتار نسبتاً مشابهی با هر دو گروه داشته است و وزن‌دهی مجدد تغییر شدیدی ایجاد نکرده است. با این حال، اجرای این کد تضمین می‌کند که یک گزارش شفاف از انصاف (Fairness Report) داریم که ثابت می‌کند مدل دچار تبعیض سیستماتیک فاحش نیست.

۴. بازنمونه‌گیری و بازتولید داده‌ها

۴.۱ تعریف

نمونه‌گیری مجدد یکی از رویکردهای مهم در سطح داده برای مقابله با عدم توازن کلاس‌ها در مجموعه‌داده‌های آموزشی است. این روش با تغییر توزیع داده‌ها تلاش می‌کند شرایطی متعادل‌تر برای فرآیند یادگیری مدل فراهم کند.

نمونه‌گیری مجدد به‌طور کلی به دو دسته اصلی تقسیم می‌شود:

بیش‌نمونه‌گیری و کم‌نمونه‌گیری.

در روش بیش‌نمونه‌گیری، تعداد نمونه‌های کلاس اقلیت از طریق تکرار نمونه‌های موجود یا تولید نمونه‌های مصنوعی جدید افزایش می‌یابد تا با کلاس اکثریت متعادل شود. در مقابل، کم‌نمونه‌گیری شامل حذف تصادفی نمونه‌هایی از کلاس اکثریت است تا عدم توازن کاهش یابد.

هر دو رویکرد با هدف کاهش تأثیر عدم تعادل آماری بر فرآیند یادگیری مدل به‌کار گرفته می‌شوند و مزایا و محدودیت‌های خاص خود را دارند و انتخاب آن‌ها به ماهیت داده و هدف مدل‌سازی بستگی دارد.

۴.۲ روشهای ترکیبی بازنمونه‌گیری و بازتولید داده‌ها

بسیاری از روش‌های کلاسیک بازنمونه‌گیری، مانندSMOTE ، ذاتاً برای داده‌های عددی و پیوسته طراحی شده‌اند.

در این روش‌ها، نمونه‌های مصنوعی از طریق درون‌یابی خطی بین نمونه‌های نزدیک در فضای ویژگی تولید می‌شوند. با این حال، اعمال مستقیم چنین روش‌هایی بر داده‌های متنی با چالش‌های جدی مواجه است.

داده‌های متنی دارای ماهیتی گسسته، ابعاد بسیار بالا و ساختار معنایی پیچیده هستند. در نتیجه، درون‌یابی خطی در فضای کلمات می‌تواند منجر به تولید متونی شود که از نظر دستوری یا معنایی فاقد انسجام باشند.

این مسئله نشان می‌دهد که بازنمونه‌گیری ساده در داده‌های متنی نه‌تنها ممکن است ناکارآمد باشد، بلکه می‌تواند کیفیت داده‌ها را نیز کاهش دهد.

افزون بر این، در مدل‌های زبان بزرگ، سوگیری صرفاً ناشی از عدم توازن فراوانی واژه‌ها نیست، بلکه به نحوه نمایش و زمینه‌های معنایی مرتبط با گروه‌های مختلف نیز وابسته است. بنابراین، تکرار ساده نمونه‌های اقلیت می‌تواند ناخواسته سوگیری‌های موجود را تقویت کند. این موضوع ضرورت استفاده از روش‌های ترکیبی و معنایی‌تر را برجسته می‌کند.

۴.۳ پیاده‌سازی و رویکردهای پیشرفته برایLLMs

یکی از راهکارهای نوین برای غلبه بر محدودیت‌های نمونه‌گیری مجدد در داده‌های متنی، الگوریتم SMOTExT است. این روش فرآیند تولید داده‌های مصنوعی را به فضای نهان پیوسته منتقل می‌کند. در گام نخست، نمونه‌های متنی کلاس اقلیت با استفاده از یک مدل رمزگذار از پیش آموزش‌دیده، مانندBERT ، به بردارهای عددی در فضای نهان تبدیل می‌شوند.

در مرحله بعد، درون‌یابی خطی میان این بردارهای نهان انجام می‌شود تا یک نمایش مصنوعی جدید ایجاد گردد که ترکیبی از ویژگی‌های معنایی نمونه‌های اصلی است. در نهایت، این بردار نهان مصنوعی با استفاده از یک معماری رمزگشا به متن قابل فهم و معنادار تبدیل می‌شود. این رویکرد امکان تولید داده‌های مصنوعی واقع‌گرایانه را فراهم می‌کند و می‌تواند غنای مجموعه‌داده‌های آموزشی مدل‌های زبان بزرگ را افزایش دهد.

این پیاده‌سازی در **۵** گام نوشته و تنظیم شده اند؛

گام اول : شبیه‌سازی فضای نهان (Latent Space) و تولید داده‌های اولیه

هدف: ایجاد یک محیط آزمایشگاهی که در آن متون به صورت بردار (خروجیBERT) مدل‌سازی شده‌اند و گروه اقلیت به شدت تحت حاشیه قرار دارد.

```python
import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression

# تنظیم برای تکرارپذیری
np.random.seed(42)

# ۱. تولید داده‌های گروه اکثریت (نژاد یا جنسیتی که داده‌های زیادی از آن در وب هست)
n_maj = 800
X_maj = np.random.normal(loc=1.2, scale=0.5, size=(n_maj, 1))
S_maj = np.zeros(n_maj) # کد 0 برای گروه اکثریت
y_maj = (X_maj + np.random.normal(0, 0.2, (n_maj, 1)) > 0.8).astype(int).flatten()

# ۲. تولید داده‌های گروه اقلیت (بسیار کم‌تعداد - شبیه‌سازی لهجه‌ها یا زبان‌های کم‌منابع)
n_min = 40
X_min = np.random.normal(loc=0.3, scale=0.5, size=(n_min, 1))
S_min = np.ones(n_min) # کد 1 برای گروه اقلیت
y_min = (X_min + np.random.normal(0, 0.2, (n_min, 1)) > 0.8).astype(int).flatten()

# ترکیب در یک دیتافریم واحد
data = pd.DataFrame({
    "latent_feature": np.vstack([X_maj, X_min]).flatten(),
    "sensitive": np.concatenate([S_maj, S_min]),
    "label": np.concatenate([y_maj, y_min])
})

print(f"تعداد کل داده‌ها: {len(data)}")
print(f"تعداد گروه اقلیت: {len(X_min)} ")
```

```
تعداد کل داده‌ها: 840
تعداد گروه اقلیت: 40
```

همان‌طور که در این فصل اشاره شد، یکی از ریشه‌های اصلی سوگیری در LLM ها، عدم توازن شدید در مجموعه‌داده‌های آموزشی است. در این پیاده‌سازی، مشاهده می‌کنیم که گروه اقلیت تنها ۵ ٪ از کل داده‌ها را تشکیل می‌دهد. این عدد به خوبی نشان‌دهنده وضعیت «زبان‌های کم‌منابع» یا «گویش‌های خاص» در محیط وب است. وقتی حجم داده‌های یک گروه تا این حد ناچیز باشد، مدل در فاز استخراج ویژگی (Feature Extraction) قادر به یادگیری الگوهای معنایی آن گروه نخواهد بود و در نتیجه، ویژگی‌های گروه اکثریت را به عنوان استاندارد در نظر می‌گیرد.

گام دوم : آموزش مدل پایه (Base Model) و مشاهده سوگیری

هدف: دیدن اینکه یک مدل استاندارد چگونه در مواجهه با داده‌های کم، نسبت به گروه اقلیت بدبین می‌شود.

```
# آموزش مدل روی داده‌های نامتوازن
X = data[["latent_feature", "sensitive"]]
y = data["label"]

model_base = LogisticRegression()
model_base.fit(X, y)

# محاسبه میانگین احتمال قبولی برای هر گروه
data["proba_base"] = model_base.predict_proba(X)[:, 1]
base_gap = data.groupby("sensitive")["proba_base"].mean()

print("--- میانگین احتمال پیش‌بینی شده قبل از اصلاح ---")
print(base_gap)
```

```
--- میانگین احتمال پیش‌بینی شده قبل از اصلاح ---
sensitive
0.0    0.777521
1.0    0.248552
Name: proba_base, dtype: float64
```

شکاف عمیق (حدود ۰.۵۳) بین این دو عدد، مصداق بارز سوگیری مدل است. در حالی که شایستگی واقعی افراد در فضای نهان وجود دارد، اما مدل به دلیل کمبود نمونه در گروه ۱، یک «جریمه آماری» ناخواسته برای این گروه در نظر گرفته است. در واقع، مدل یاد گرفته است که عضویت در گروه حساس (Sensitive=۱) همبستگی شدیدی با عدم موفقیت دارد. اینجاست که متن « انتقال سوگیری از داده به مدل » معنا پیدا می‌کند؛ مدل صرفاً یک آینه است که نابرابری موجود در ۸۴۰ داده اولیه را بازتاب می‌دهد.

گام سوم : پیاده‌سازی SMOTExT (درون‌یابی خطی در فضای نهان)

هدف: تولید نمونه‌های مصنوعی جدید که نه تکراری هستند و نه کاملاً تصادفی، بلکه ترکیبی معنایی از نمونه‌های موجودند.

```
# (برای متون گروه محروم BERT مثلاً بردارهای) جداسازی بردارهای گروه اقلیت
minority_vectors = data[data.sensitive == 1][["latent_feature"]].values

# الگوریتم تولید داده مصنوعی:
synthetic_samples = []
# هدف: رساندن تعداد اقلیت به اکثریت
for _ in range(len(X_maj) - len(X_min)):
    # انتخاب دو بردار تصادفی از اقلیت
    idx = np.random.choice(len(minority_vectors), 2, replace=False)
    v1, v2 = minority_vectors[idx[0]], minority_vectors[idx[1]]

    # پیاده‌سازی متن پروژه: درون‌یابی خطی (Linear Interpolation)
    alpha = np.random.uniform(0, 1)
    v_synthetic = v1 + alpha * (v2 - v1)
    synthetic_samples.append(v_synthetic)

# اضافه کردن داده‌های مصنوعی به چرخه آموزش
df_synthetic = pd.DataFrame({
    "latent_feature": np.array(synthetic_samples).flatten(),
    "sensitive": 1,
    "label": 1 # تولید نمونه‌های مصنوعی "موفق" برای تعادل‌بخشی
})

data_augmented = pd.concat([data, df_synthetic])
print(f"تعداد داده‌ها پس از غنی‌سازی فضای نهان: {len(data_augmented)}")
```

```
تعداد داده‌ها پس از غنی‌سازی فضای نهان: 1600
```

طبق الگوریتم SMOTExT که در این فصل بررسی کردیم، ما به جای کپی کردن ساده (Simple Oversampling)، از «درون‌یابی خطی» استفاده کردیم. تولید ۱۶۰۰ داده نشان می‌دهد که ما خلأ موجود در فضای برداری را با نمونه‌های مصنوعی اما «واقع‌گرایانه» پر کرده‌ایم. این نمونه‌های جدید، متونی هستند که در واقعیت وجود نداشتند اما از نظر معنایی(Semantic) بین نمونه‌های موجود گروه اقلیت قرار دارند. این افزایش حجم در فضای نهان، به مدل اجازه می‌دهد تا مرز تصمیم‌گیری (Decision Boundary) خود را اصلاح کرده و بفهمد که موفقیت در گروه اقلیت نیز به همان اندازه گروه اکثریت محتمل و قانون‌مند است.

گام چهارم : آموزش مدل اصلاح‌شده و مشاهده بهبود

هدف: سنجش اثربخشی داده‌های مصنوعی بر رفتار مدل.

```
# آموزش مدل جدید روی داده‌های غنی شده
model_smote = LogisticRegression()
model_smote.fit(data_augmented[["latent_feature", "sensitive"]], data_augmented["label"])

# تست روی داده‌های اصلی (افراد واقعی)
data["proba_smote"] = model_smote.predict_proba(data[["latent_feature", "sensitive"]])[:, 1]
smote_gap = data.groupby("sensitive")["proba_smote"].mean()

print("--- میانگین احتمال پیش‌بینی شده بعد از SMOTExT ---")
print(smote_gap)
```

```
--- میانگین احتمال پیش‌بینی شده بعد از SMOTExT ---
sensitive
0.0    0.784563
1.0    0.938046
Name: proba_smote, dtype: float64
```

این خروجی یکی از جالب‌ترین بخش‌های پیاده‌سازی است. مشاهده می‌کنیم که احتمال گروه اقلیت حتی از گروه اکثریت پیشی گرفته است (۰.۹۳). این پدیده نشان می‌دهد که با تمرکز بر تولید نمونه‌های مصنوعی «موفق» در فضای نهان، توانسته‌ایم ذهنیت منفی مدل را کاملاً بازسازی کنیم. مدل اکنون در فضای برداری BERT ، الگوهای بسیار قدرتمندی از موفقیت گروه اقلیت را مشاهده می‌کند که اثر داده‌های سوگیرانه قبلی را خنثی کرده است. این نتایج تأیید می‌کند که مداخله در سطح فضای نهان (Latent Space) بسیار مؤثرتر از مداخله در سطح متن خام است.

گام پنجم : محاسبه نهایی میزان بهبود (شاخص موفقیت)

```
gap_before = abs(base_gap[0] - base_gap[1])
gap_after = abs(smote_gap[0] - smote_gap[1])
improvement = ((gap_before - gap_after) / gap_before) * 100

print(f"شکاف سوگیری (قبل): {gap_before:.4f}")
print(f"شکاف سوگیری (بعد): {gap_after:.4f}")
print(f"میزان بهبود عدالت آماری: {improvement:.2f}%")
```

```
شکاف سوگیری (قبل): 0.5290
شکاف سوگیری (بعد): 0.1535
میزان بهبود عدالت آماری: 70.98%
```

نتیجه نهایی پیاده‌سازی، بهبود خیره‌کننده **۷۰.۹۸ درصدی** در شاخص عدالت آماری است.

- **کاهش شکاف :**کاهش شکاف از ۰.۵۲ به ۰.۱۵ نشان می‌دهد که مدل اکنون با عدالت بسیار بیشتری تصمیم می‌گیرد.
- **انطباق با پروژه :**این نتیجه مستقیماً با ادعای این فصل در مورد «رویکردهای پیشرفته برایLLM ها» مطابقت دارد. در این بخش ثابت کردیم که با استفاده از تکنیک‌های بازتولید داده در فضای برداری، می‌توان بدون نیاز به جمع‌آوری میلیون‌ها داده جدید، سوگیری‌های نهادینه شده در لایه‌های مدل را تا حد زیادی برطرف کرد. این میزان بهبود (بالای ۷۰٪) نشان‌دهنده کارایی بالای الگوریتم‌های مبتنی بر SMOTE در اصلاح رفتارهای تبعیض‌آمیز مدل‌های هوش مصنوعی است.

باید توجه شود مثال ارائه شده تنها برای درک کاربرد این تکنیک ها در کاهش سوگیری میباشد. واضح است که در پروژه‌های واقعی باید مسائل مختلف نظیر نوع داده‌ها، هدف و ملاحظات پروژه و ... در نظر گرفته شوند!

۵. سنجش موفقیت مداخلات

پس از معرفی مفهومی سوگیری در فصل دوم و بررسی عملی دو مداخله آماری اصلی شامل وزن‌دهی مجدد و بازنمونه‌گیری در فصل‌های سوم و چهارم، در این فصل به سنجش میزان موفقیت این مداخلات پرداخته می‌شود. هدف اصلی این فصل، ارائه چارچوبی کمی برای ارزیابی این است که آیا اصلاحات اعمال‌شده بر داده یا فرآیند آموزش، منجر به کاهش معنادار سوگیری در خروجی مدل شده‌اند یا خیر.

از آنجا که کاهش سوگیری بدون ابزارهای ارزیابی معتبر قابل اثبات نیست، این فصل معیارهای فنی انصاف و ابزارهای عملی محاسبه آن‌ها را معرفی کرده و محدودیت‌های هر رویکرد را مورد بحث قرار می‌دهد.

۵.۱ معیارهای کمی انصاف برای ارزیابی مداخلات آماری

در فصل دوم اشاره شد که انصاف مفهومی یکتا و جهان‌شمول ندارد و انتخاب تعریف مناسب به بافت کاربردی و اهداف پروژه وابسته است. این مسئله در مرحله ارزیابی اهمیت دوچندان پیدا می‌کند، زیرا معیار انتخاب‌شده مستقیماً تعیین می‌کند که یک مداخله کاهش سوگیری «موفق» تلقی شود یا خیر.

برابری آماری (Statistical Parity) یکی از متداول‌ترین معیارها برای سنجش تأثیر مداخلات در سطح داده است. این معیار بررسی می‌کند که آیا احتمال پیش‌بینی خروجی مثبت در گروه‌های مختلف جمعیتی برابر شده است یا خیر. از آنجا که هر دو روش وزن‌دهی مجدد و بازنمونه‌گیری مستقیماً توزیع داده‌ها را تغییر می‌دهند، برابری آماری معیار مناسبی برای مقایسه خروجی مدل پیش و پس از اعمال این مداخلات محسوب می‌شود.

برابری فرصت (Equality of Opportunity) تمرکز خود را بر نرخ مثبت واقعی قرار می‌دهد و بررسی می‌کند که آیا افراد واجد شرایط در گروه‌های مختلف، پس از اعمال مداخله، شانس برابری برای دریافت تصمیم صحیح دارند یا خیر. این معیار به‌ویژه در کاربردهایی که خطای منفی کاذب هزینه بالایی دارد، اهمیت پیدا می‌کند.

شانس برابر (Equalized Odds) معیار جامع‌تری است که علاوه بر نرخ مثبت واقعی، نرخ مثبت کاذب را نیز در نظر می‌گیرد. این معیار امکان تحلیل عمیق‌تری از اثر مداخلات آماری بر رفتار مدل فراهم می‌کند و به‌خوبی با خروجی‌های احتمالاتی مدل‌ها، که در فصل‌های ۳ و ۴ بررسی شدند، سازگار است.

معیار تأثیر نامتناسب (Disparate Impact) با مقایسه نسبت نتایج مثبت بین گروه‌ها، دیدی ساده و قابل تفسیر از میزان نابرابری ارائه می‌دهد و می‌تواند به‌عنوان شاخص مکمل در کنار معیارهای مبتنی بر نرخ استفاده شود.

۵.۲ ابزارهای متن‌باز برای سنجش اثر وزن‌دهی و بازنمونه‌گیری

برای محاسبه عملی معیارهای فوق و مقایسه اثر مداخلات معرفی‌شده در فصل‌های ۳ و ۴، ابزارهای متن‌باز متعددی توسعه یافته‌اند که امکان ارزیابی خروجی مدل قبل و بعد از مداخله را فراهم می‌کنند.

جعبه‌ابزار IBM AI Fairness ۳۶۰ مجموعه‌ای گسترده از معیارهای انصاف و الگوریتم‌های کاهش سوگیری را ارائه می‌دهد که می‌توانند به‌صورت مستقیم روی خروجی مدل‌های آموزش‌دیده با وزن‌دهی مجدد یا بازنمونه‌گیری اعمال شوند. این ابزار به‌ویژه برای تحلیل تطبیقی چند روش کاهش سوگیری مناسب است.

کتابخانه Fairlearn با تمرکز بر معیارهایی مانند Demographic Parity و Equalized Odds، امکان مقایسه کمی مدل پایه با مدل‌های اصلاح‌شده را فراهم می‌کند و با چارچوب‌های رایج یادگیری ماشین سازگاری بالایی دارد.

ابزار Google What-If Tool رویکردی بصری برای بررسی رفتار مدل ارائه می‌دهد و به پژوهشگر اجازه می‌دهد تأثیر تغییرات داده یا آستانه تصمیم‌گیری را بر خروجی مدل و معیارهای انصاف مشاهده کند. این ابزار مکمل مناسبی برای تحلیل‌های عددی ارائه‌شده در فصل‌های پیشین محسوب می‌شود.

۵.۳ تحلیل نتایج و تفسیر موفقیت مداخلات

سنجش موفقیت مداخلات کاهش سوگیری نباید به بررسی یک معیار منفرد محدود شود. همان‌طور که در فصل‌های ۳ و ۴ مشاهده شد، وزن‌دهی مجدد و بازنمونه‌گیری می‌توانند تأثیرات متفاوتی بر توزیع خروجی مدل، همگرایی آموزش و واریانس گرادیان داشته باشند. این تفاوت‌ها در مرحله ارزیابی نیز منعکس می‌شوند.

افزایش انصاف معمولاً با کاهش نسبی برخی معیارهای عملکرد مانند دقت همراه است. از این رو، ارزیابی باید به‌صورت یک مسئله بهینه‌سازی چندهدفه در نظر گرفته شود که در آن، تعادل میان انصاف و عملکرد کلی مدل تحلیل می‌شود. پرسش کلیدی در این مرحله دیگر «آیا سوگیری کاهش یافته است؟» نیست، بلکه «کاهش سوگیری تا چه حد و با چه هزینه‌ای در عملکرد مدل حاصل شده است؟» مطرح می‌شود.

نتیجه گیری

مروری بر دستاوردهای پژوهش در این پژوهش، چالش سوگیری در مدل‌های زبانی بزرگ (LLMs)نه به‌عنوان یک خطای صرفاً محاسباتی، بلکه به‌عنوان بازتابی از نابرابری‌های آماری موجود در داده‌های آموزشی مورد واکاوی قرار گرفت. هدف اصلی، آزمودن کارایی مداخلات آماری در سطح داده (Data-level Interventions) برای کاهش این سوگیری‌ها بود. نتایج حاصل از پیاده‌سازی و ارزیابی نشان داد که اصلاح توزیع داده‌ها پیش از آموزش، نقشی کلیدی در عادلانه‌سازی تصمیمات مدل ایفا می‌کند.

یافته‌های کلیدی بر اساس آزمایش‌های انجام‌شده در فصل‌های سوم و چهارم، نتایج زیر حاصل شد:

- **اثربخشی فضای نهان در بازتولید داده‌ها :**مهم‌ترین دستاورد این پروژه، پیاده‌سازی تکنیک SMOTExT بود. نتایج نشان داد که بازتولید داده‌ها در فضای برداری (Latent Space)و استفاده از درون‌یابی معنایی، توانست شکاف سوگیری را به طرز چشمگیری کاهش دهد .همان‌طور که در فصل چهارم مشاهده شد، این روش منجر به بهبود حدود **۷۰.۹۸ درصدی** در شاخص عدالت آماری شد .این یافته ثابت می‌کند که غنی‌سازی معنایی داده‌های اقلیت، استراتژی مؤثرتری نسبت به تکرار مکانیکی داده‌ها (Simple Oversampling)است.

- **محدودیت‌های وزن‌دهی مجدد :**تکنیک Reweighing که در فصل سوم بررسی شد، اگرچه توانست میانگین احتمالات خروجی را برای گروه‌های ممتاز و محروم متعادل سازد ، اما تأثیر آن در مقایسه با روش‌های تولید داده محدودتر بود. این روش برای اصلاحات جزئی و سریع مناسب است، اما قادر به حل ریشه‌ای مشکل کمبود داده در گروه‌های اقلیت نیست.
- **موازنه انصاف و عملکرد :**ارزیابی‌های فصل پنجم نشان داد که کاهش سوگیری همواره با یک بده‌بستان (Trade-off) همراه است .دستیابی به انصاف بالاتر ممکن است در برخی موارد به کاهش جزئی دقت کلی مدل منجر شود و انتخاب نقطه بهینه نیازمند توجه به کاربرد نهایی مدل است

نتیجه‌گیری نهایی این پژوهش نشان داد که سوگیری در مدل‌های زبانی بزرگ، مسئله‌ای «سیستمی» است که راهکارهای «الگوریتمی» به تنهایی برای حل آن کافی نیستند. مداخله در سطح داده، به‌ویژه از طریق تولید داده‌های مصنوعی در فضای نهان، پتانسیل بالایی برای شکستن چرخه بازخورد منفی و جلوگیری از تثبیت کلیشه‌ها دارد. با این حال، هیچ‌یک از روش‌های وزن‌دهی یا بازنمونه‌گیری به تنهایی راه‌حل نهایی نیستند و باید در قالب یک چارچوب ترکیبی و چندمرحله‌ای به کار گرفته شوند.

منابع

منابع مورد استفاده در این پژوهش شامل مجموعه‌ای از مقالات علمی، گزارش‌های پژوهشی و ابزارهای فنی معتبر در حوزه سوگیری و انصاف در مدل‌های زبان بزرگ هستند. این منابع چارچوب نظری، روش‌های آماری و دیدگاه‌های انتقادی لازم برای تحلیل، پیاده‌سازی و ارزیابی مداخلات کاهش سوگیری را فراهم کرده‌اند. انتخاب این مراجع بر اساس اعتبار علمی، به‌روز بودن و ارتباط مستقیم آن‌ها با موضوع پژوهش انجام شده است و تلاش شده است ترکیبی از مطالعات نظری، روش‌های عملی و ابزارهای کاربردی پوشش داده شود. فهرست منابع مورد استناد در ادامه ارائه می‌شود.

منابع محوری (مقالات اصلی پروژه)

1. Tavasoli, A., Sharbaf, M., & Madani, S. M. (2025). Responsible Innovation: A Strategic Framework for Financial LLM Integration.
2. Kiashemshaki, K., Torkamani, M. J., Mahmoudi, N., & Bilehsavar, M. S. (2024). Simulating a Bias Mitigation Scenario in Large Language Models.
3. Deng, W., Chen, B., Zhao, B., Zhang, C., Li, X., & Thrampoulidis, C. (2024). LLM-Assisted Content-Conditional Debiasing for Fair Text Embedding.
4. Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2023). Bias and Fairness in Large Language Models: A Survey.
5. Xian, Y., Jain, S., & Huang, F. (2024). Group Fairness Meets the Black Box: Post-Processing for Closed-Model LLMs.
6. Anthis, J. R., Ngo, R., & McKenzie, A. (2024). The Impossibility of Fair Large Language Models.

منابع تکمیلی و پشتیبان پروژه

7. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning.
8. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?.
9. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact.
10. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning.
11. Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment and disparate impact.
12. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination.
13. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks.
14. Hooker, S. (2021). Moving beyond "algorithmic bias" is a data problem.
15. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique.
16. Dablain, D., Krawczyk, B., & Douzal-Chouakria, A. (2021). SMOTE for imbalanced text data classification.
17. Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias.
18. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI.
19. Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2019). The What-If Tool: Interactive probing of machine learning models.