تشخیص سوگیری جنسیتی در بردارهای کلمه‌ای مشاغل

نویسنده: علیرضا مهدی زاده

چکیده

این مطالعه به بررسی سوگیری‌های جنسیتی اجتماعی موجود در بردارهای کلمه‌ای از پیش آموزش‌دیده، با تمرکز بر اصطلاحات شغلی، می‌پردازد. با استفاده از مدل‌های Google News Word2Vec و GloVe 6B (300d) ، چندین روش تشخیص سوگیری شامل پروجکشن جهت جنسیتی، شباهت کسینوسی، آزمایش‌های قیاسی، و آزمون ارتباط بردار کلمه‌ای (WEAT) به کار گرفته شد. تحلیل‌ها نشان‌دهنده وجود ارتباطات جنسیتی قابل‌توجهی در بردارها هستند که با کلیشه‌های اجتماعی همخوانی دارند، مانند ارتباط قوی زنانه برای مشاغلی نظیر "پرستار" و "ماما" و ارتباط مردانه برای "مهندس" و "مدیر". این سوگیری‌ها به‌صورت متوسطی با توزیع جنسیتی واقعی از داده‌های اداره آمار کار ایالات متحده (BLS) همبستگی دارند، که نشان‌دهنده بازتاب هنجارهای اجتماعی در مدل‌های هوش مصنوعی است. همچنین یک روش ساده برای کاهش سوگیری آزمایش شد. یافته‌ها بر اهمیت رفع سوگیری‌های موجود در سیستم‌های هوش مصنوعی برای جلوگیری از تداوم کلیشه‌ها در کاربردهایی مانند استخدام یا توصیه شغلی تأکید دارند.

---

۱ .مقدمه

مدل‌های هوش مصنوعی، به‌ویژه آن‌هایی که مبتنی بر پردازش زبان طبیعی (NLP) هستند، به‌طور فزاینده‌ای در فرآیندهای تصمیم‌گیری مانند استخدام، توصیه محتوا، و راهنمایی شغلی استفاده می‌شوند. با این حال، این مدل‌ها ممکن است به‌طور ناخواسته سوگیری‌های اجتماعی موجود در داده‌های آموزشی خود را کدگذاری کنند، که منجر به نتایج تبعیض‌آمیز می‌شود. این پروژه بر تشخیص سوگیری‌های جنسیتی در زمینه‌های شغلی در بردارهای کلمه‌ای از پیش آموزش‌دیده تمرکز دارد—نمایش‌های برداری کلمات که روابط معنایی را ضبط می‌کنند. با تحلیل بردارهای مدل‌های Google News Word2Vec و GloVe ، هدف ما سنجش و درک میزان بازتاب کلیشه‌های اجتماعی درباره مشاغل جنسیتی در این مدل‌ها است.

درک و کاهش این سوگیری‌ها برای توسعه سیستم‌های هوش مصنوعی منصفانه و عادلانه حیاتی است. این گزارش تحلیلی جامع از سوگیری جنسیتی در بردارهای کلمه‌ای شغلی ارائه می‌دهد، این سوگیری‌ها را با داده‌های واقعی همبسته می‌کند، و یک رویکرد اولیه برای کاهش سوگیری را بررسی می‌کند.

---

۲ .روش‌شناسی

۲.۱ مدل‌های بردار کلمه‌ای

دو مدل بردار کلمه‌ای از پیش آموزش‌دیده به کار گرفته شدند:

- **Google News Word2Vec:** مدل ۳۰۰بعدی آموزش‌دیده بر مقالات خبری گوگل.
- **GloVe 6B (300d):** مدل ۳۰۰بعدی آموزش‌دیده بر ترکیبی از ویکی‌پدیا و داده‌های Gigaword.

این مدل‌ها به دلیل پوشش گسترده و استفاده رایج در تحقیقات تشخیص سوگیری انتخاب شدند.

۲.۲ تکنیک‌های تشخیص سوگیری

برای سنجش سوگیری جنسیتی در اصطلاحات شغلی، روش‌های زیر به کار رفتند:

۱ .پروجکشن جهت جنسیتی:

- یک بردار جهت جنسیتی با استفاده از جفت‌های کلمات جنسیتی (مانند "مرد - زن"، "او - او") تعریف شد. برای GloVe، چندین جفت میانگین گرفته شدند تا محور جنسیتی قوی‌تری ایجاد شود.
- بردارهای شغلی روی این جهت جنسیتی پروجکت شدند تا امتیازهای سوگیری محاسبه شوند، که در آن امتیازهای مثبت نشان‌دهنده ارتباط مردانه و امتیازهای منفی نشان‌دهنده ارتباط زنانه هستند.

۲ .شباهت کسینوسی:

- شباهت کسینوسی بین بردارهای شغلی و اصطلاحات جنسیتی (مانند "مرد"، "زن") برای سنجش ارتباطات مستقیم محاسبه شد.

۳ .آزمایش‌های قیاسی:

- از الگوهای قیاسی مانند "مرد به شغل مانند زن به چیست؟" برای شناسایی تکمیل‌های کلیشه‌ای کدگذاری‌شده در بردارها استفاده شد.

۴ .آزمون ارتباط بردار کلمه‌ای:(WEAT)

- WEAT برای سنجش ارتباط تفاضلی بین مشاغل به‌طور کلیشه‌ای مردانه و زنانه و ویژگی‌های جنسیتی (مانند "مرد"، "زن") اعمال شد.

۲.۳ داده‌ها

- **اصطلاحات شغلی :** فهرستی از ۳۸۲ شغل از اداره آمار کار ایالات متحده (BLS) استخراج شد که داده‌های توزیع جنسیتی واقعی (درصد زنان شاغل) را ارائه می‌داد.

- **مدیریت مشاغل چندکلمه‌ای :**برای مشاغلی که مستقیماً در واژگان وجود نداشتند (مانند "مددکار اجتماعی")، از انواع یا کلمات تکی (مانند "اجتماعی"، "کارگر") استفاده شد.

---

۳ .نتایج

۳.۱ امتیازهای ارتباط جنسیتی

با استفاده از روش پروجکشن جهت جنسیتی، امتیازهای سوگیری برای اصطلاحات شغلی محاسبه شدند. یافته‌های کلیدی عبارتند از:

- **ارتباط‌های قوی زنانه:**
  - ماما -)۱.۲۵۹۹(، پرستار -)۱.۰۲۰۶(، مهماندار هواپیما -)۱.۱۷۱۱(
- **ارتباط‌های قوی مردانه:**
  - ژنرال)۱.۲۶۳۷(، رئیس)۱.۰۳۶۱(، بسته‌بندان)۱.۰۳۴۲(

این امتیازها کلیشه‌های اجتماعی را بازتاب می‌دهند، با نقش‌های مراقبتی که سوگیری زنانه و نقش‌های رهبری یا کارهای دستی که سوگیری مردانه نشان می‌دهند.

۳.۲ تحلیل شباهت کسینوسی

نتایج شباهت کسینوسی یافته‌های پروجکشن را تأیید کردند:

- پرستار :شباهت به "زن) "۰.۴۴۱۴" < (مرد) "۰.۲۵۴۷(
- مهندس :شباهت به "مرد) "۰.۱۵۱۳" < (زن) "۰.۰۹۴۴(

۳.۳ همبستگی با داده‌های واقعی

امتیازهای سوگیری با داده‌های BLS درباره درصد مردان در هر شغل همبسته شدند:

- **همبستگی پیرسون :**۰.۳۵۴۶ > p) ۰.۰۰۰۱(، که نشان‌دهنده رابطه‌ای متوسط اما از نظر آماری معنادار بین سوگیری‌های بردار و توزیع‌های جنسیتی واقعی است.

۳.۴ تحلیل WEAT

تحلیل WEAT نتایج زیر را به دست داد:

- **اندازه اثر :**۰.۲۱۶۷

- **مقدار :p** ۰.۶۴۹۰ (غیرمعنادار)، که نشان‌دهنده ارتباط تفاضلی ضعیف است، احتمالاً به دلیل وجود اصطلاحات پر سر و صدا.

۳.۵ ارزیابی کاهش سوگیری

یک روش کاهش سوگیری نرم (کاهش ۵۰٪ مؤلفه جنسیتی) روی بردارهای شغلی اعمال شد:

- **کاهش سوگیری :**کاهش مداوم ۵۰٪ در امتیازهای سوگیری در تمام مشاغل.
- **محدودیت :**عدم ارزیابی سودمندی معنایی پس از کاهش سوگیری.

---

۴ .بحث

۴.۱ پیامدهای یافته‌ها

نتایج نشان می‌دهند که بردارهای کلمه‌ای کلیشه‌های جنسیتی اجتماعی را کدگذاری می‌کنند، با مشاغلی مانند "پرستار" و "مهندس" که به‌ترتیب ارتباطات قوی زنانه و مردانه را نشان می‌دهند. همبستگی متوسط با داده‌های BLS نشان می‌دهد که بردارها گرایش‌های واقعی جهان را ضبط می‌کنند، اما همچنین کلیشه‌ها را تقویت می‌کنند، همان‌طور که در مورد "پزشک" دیده می‌شود که سوگیری زنانه نشان می‌دهد، در حالی که در واقعیت ۴۱.۴٪ زن هستند.

این سوگیری‌ها می‌توانند پیامدهای مهمی برای کاربردهای هوش مصنوعی داشته باشند، مانند:

- **سیستم‌های استخدام :**بردارهای دارای سوگیری ممکن است جنسیت‌های خاصی را برای نقش‌های خاص ترجیح دهند.
- **توصیه‌های شغلی :**بردارها ممکن است مسیرهای شغلی کلیشه‌ای را بر اساس جنسیت پیشنهاد دهند.

۴.۲ محدودیت‌ها

- **مشاغل چندکلمه‌ای :**مدیریت اصطلاحات چندکلمه‌ای (مانند "مددکار اجتماعی") ناکافی بود، و اغلب به کلمات تکی مانند "اجتماعی" وابسته بود که ممکن است معنای شغل را به‌طور کامل ضبط نکند.
- **جهت جنسیتی :**استفاده از یک جفت جنسیتی در Word2Vec ممکن است محور جنسیتی را بیش از حد ساده کند.
- **نتایج :WEAT** گنجاندن اصطلاحات مبهم (مانند "و"، "دیگر") تحلیل WEAT را تضعیف کرد.
- **کاهش سوگیری :**روش کاهش سوگیری ساده بود و حفظ معنای معنایی را ارزیابی نکرد.

---

۵ .نتیجه‌گیری

این پروژه با موفقیت سوگیری‌های جنسیتی در بردارهای کلمه‌ای شغلی را سنجش کرد و نشان داد که مدل‌های هوش مصنوعی کلیشه‌های اجتماعی را بازتاب داده و احتمالاً تقویت می‌کنند. همبستگی بین سوگیری‌های بردار و داده‌های واقعی بر نیاز به توجه دقیق به سوگیری در کاربردهای هوش مصنوعی تأکید دارد. در حالی که رویکرد کاهش سوگیری ما سوگیری را کاهش داد، کارهای آینده باید روش‌های پیچیده‌تر را بررسی کرده و تأثیر آن‌ها را در وظایف پایین‌دستی ارزیابی کنند.