بهبود قدرت تعمیم مدل‌های تشخیص کلام نفرت‌انگیز مبتنی بر تطبیق دامنه

سیده فاطمه نوراللهی*، راضیه برادران و حسین امیرخانی

گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران

چکیده

امروزه با رشد فعالیت در شبکه‌های اجتماعی شاهد افزایش کلام نفرت‌انگیز به صورت برخط هستیم و به‌همین منظور مسئلهٔ تشخیص نفرت در فضای مجازی دارای اهمیت است. همچنین تطبیق دامنه نیز در این مسئله و به‌طورکلی در حوزهٔ پردازش زبان طبیعی، یکی از چالش‌های مهم است. در بسیاری از مسائل، ضمن تغییر دامنه با افت عملکرد مواجهیم که این موضوع در مسئلهٔ نفرت نیز صادق است. در این پژوهش با استفاده از روش‌های تطبیق دامنه سعی در افزایش قدرت تعمیم‌پذیری مدل‌های تشخیص نفرت خواهیم داشت. برای این منظور روش‌های مبتنی بر ترنسفورمر شامل آموزش خصمانهٔ دامنه و ترکیب متخصصان را به کار می‌گیریم و همچنین از آموزش چندمنبعی استفاده می‌کنیم. آزمایش‌ها با استفاده از چهار مجموعه‌داده در حوزهٔ نفرت انجام می‌شوند. در ابتدا مدل‌ها را به‌صورت درون‌دامنه‌ای و تک‌منبعی ارزیابی می‌کنیم. در مرحلهٔ بعد با اضافه‌کردن دامنه‌های دیگر به بخش آموزش، شاهد افت نتایج و انتقال منفی هستیم؛ سپس آزمایش‌های برون‌دامنه‌ای را ابتدا به‌صورت تک‌منبعی با مدل DistilBERT انجام می‌دهیم که با تغییر دامنه نتایج به‌طور قابل توجهی کاهش می‌یابند. به‌منظور افزایش قدرت تطبیق دامنهٔ مدل در بخش برون‌دامنه‌ای، روی چند منبع آموزش را انجام می‌دهیم که درحدود نیمی از موارد، سبب بهبود نتایج می‌شود که نتیجهٔ معناداری نیست. در ادامه با استفاده از روش‌های مبتنی بر ترنسفورمر شامل آموزش خصمانهٔ دامنه و ترکیب متخصصان سعی در افزایش قدرت تطبیق دامنهٔ مدل‌ها خواهیم داشت که در ۸۷٪ از آزمایش‌های برون‌دامنه‌ای چند منبعی شاهد افزایش عملکرد هستیم. البته این روش‌ها در عملکرد آزمایش‌های درون‌دامنه‌ای هم مؤثر هستند. مسئلهٔ مهمی که گاهی موجب افت‌وخیز چشمگیر نتایج می‌شود، مجموعه‌داده‌ها هستند. شباهت داده‌ها و تشابه توزیع بعضی دامنه‌ها باعث افزایش قدرت تطبیق دامنهٔ مدل می‌شوند.

واژگان کلیدی: کلام نفرت‌انگیز، تطبیق دامنه، تعمیم، طبقه‌بندی، ترنسفورمر

Domain adaptation-based method for improving generalization of hate speech detection models

Seyedeh Fatemeh Nourollahi*, Razieh Baradaran and Hossein Amirkhani

Department of Computer Engineering and Information Technology,
Technical and Engineering Faculty, Qom University, Qom, Iran

Abstract

Today, with the growth of activity in social media, we see an increase in hate speech online and for this reason, the issue of recognizing hate in cyberspace is important. Also, domain adaptation is one of the important challenges in this task and in general in the field of natural language processing. In many issues, while changing the domain, we face a drop in performance, which is also true in the task hate speech. In this research, we try to increase the generalizability of hate detection models by using domain adaptation methods. For this purpose, we use Transformer-based methods, including domain adversarial training and mixture of experts, and we also use multi-source training. Experiments are conducted using four datasets in the domain of hate. At first, we evaluate the models in an in-domain

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۴۰۱/۷/۲ • تاریخ پذیرش: ۱۴۰۲/۱۲/۱۲ • تاریخ انتشار: ۱۴۰۳/۵/۱۰ • نوع مطالعه: کاربردی

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]

and single-source manner. In the next step, by adding other domains to the education section, we see a drop in results and a negative transfer. Then we perform the out-of-domain tests first as a single source with the DistilBERT model, which significantly reduces the results by changing the domain. In order to increase the power of domain adaptation of the model in the out-of-domain part, we perform the training on several sources, leads to improve the results in about half of the cases, which is not significant. In the following, we try to increase the domain adaptation power of the models, using transformer-based methods including domain adversarial training and the mixture of experts, which leads to increase in performance in 87% of multi-source out-of-domain tests. Of course, these methods are also effective in the performance of in-domain tests. An important issue that sometimes causes a significant drop in results is datasets. The similarity of the data and the similarity of the distribution of some domains increase the power of domain adaptation of the model and on the contrary.

Keywords: hate speech, classification, transformer, domain adaptation, generalization

۱- مقدمه

هیچ تعریف دقیقی در مورد کلام نفرت‌انگیز[1] (به‌صورت برخط یا برون‌خط) وجود ندارد و این موضوع مورد بحث دانشگاهیان، کارشناسان حقوقی و سیاست‌گذاران قرار گرفته است. به‌طورمعمول گفته می‌شود که کلام نفرت‌انگیز دارای انگیزه‌های متعصبانه، خصمانه و بدخواهانه است که به شخص یا گروهی به‌خاطر ویژگی‌های ذاتی واقعی یا تصور شدهٔ آنها، اطلاق می‌شود [۷، ۱۲].

از یک طرف تعاریفی وجود دارد که طیف گسترده‌ای از کلام را بر اساس ویژگی‌ها علیه شخص یا گروه نشان می‌دهد [۲۹]. طرف دیگر تعاریفی قرار دارند که مستلزم آسیب هدفمند هستند. محدودترین تعاریف دلالت بر این دارد که کلام نفرت‌انگیز باید «کلام خطرناک» باشد؛ یعنی زبانی که به‌طورمستقیم با تحریک خشونت جمعی یا آسیب جسمی علیه گروه ارتباط دارد [۴]. کلام نفرت‌انگیز برخط می‌تواند شامل محرک‌ها، اهداف، انگیزه‌ها و شیوه‌های متفاوتی باشد [۳۷]. گاهی اوقات مجرمان کسانی را که به آنها حمله می‌کنند می‌شناسند، در حالی که دیگران ممکن است دنبال‌کنندگان ناشناس برخط را تحریک کنند تا افراد خاصی را هدف قرار دهند [۳۶]. برخی عناصر نیز به‌شدت با کلام نفرت‌انگیز مرتبط هستند؛ به‌عنوان مثال نژادپرستی، خشونت، تبعیض جنسیتی و ...[2] که از انواع مختلف کلام نفرت‌انگیز محسوب می‌شوند [۳۵].

تصمیم‌گیری در مورد اینکه آیا بخشی از متن حاوی کلام نفرت‌انگیز است، حتی برای انسان‌ها نیز ساده نیست. کلام نفرت‌انگیز یک پدیدهٔ پیچیده است که ذاتاً با روابط بین گروه‌ها مرتبط است و بر ظرافت‌های زبانی تکیه می‌کند [۳۵].

مردم به‌طور فزاینده‌ای از سکوهای شبکه‌های اجتماعی مانند توییتر، فیس‌بوک، یوتیوب و... برای برقراری ارتباط و به اشتراک‌گذاری اطلاعات استفاده می‌کنند. اگرچه تعامل بین کاربران در این پلتفرم‌ها می‌تواند منجر به مکالمات سازنده شود [۳، ۶]، اما به‌ویژه به‌دلیل دسترسی آسان و محیط ناشناس این پلتفرم‌های برخط [۲۷]، به‌طور چشم‌گیری برای انتشار زبان توهین‌آمیز و سازماندهی فعالیت‌های مبتنی بر نفرت مورد سوءاستفاده قرار گرفته‌اند [۳، ۶]. از سوی دیگر به دلایل بهره‌مندی از تنوع قوانین ملی کلام مشوق عداوت و نفرت، دشواری تعیین محدودیت برای فضای مجازی پیوسته در حال تحول، نیاز روزافزون افراد و کاربران شبکه‌های اجتماعی به بیان نظرات، ضدحمله‌های مخالفان و تأخیر در بررسی دستی توسط اپراتورهای اینترنتی، انتشار کلام نفرت‌انگیز برخط شتاب جدیدی پیدا کرده است که به‌طور مداوم هم سیاست‌گذاران و هم جامعهٔ پژوهشی را به چالش می‌کشد [۳۵].

همان‌طورکه در تعریف کلام نفرت‌انگیز اتفاق نظر روشنی وجود ندارد، در مورد مؤثرترین راه برای تشخیص آن در پلتفرم‌های مختلف هم اتفاق نظر وجود ندارد. اکثر رویکردهای خودکار برای شناسایی کلام مشوق عداوت و نفرت با یک کار طبقه‌بندی دوتایی شروع می‌شود که در آن پژوهش‌گران در حال کدگذاری یک سند به‌عنوان «کلام نفرت‌انگیز یا خیر» هستند [۳۷]، اگرچه رویکردهای چندکلاسه نیز استفاده شده است [۹].

با توسعهٔ فناوری پردازش زبان طبیعی (NLP)، پژوهش‌های زیادی در مورد تشخیص خودکار کلام نفرت‌انگیز متنی در سال‌های اخیر انجام شده است [۳۵]. چند مسابقهٔ مشهور (به‌عنوان مثال، SemEval-2019 [۴۶] و SemEval-2020 [۴۵]، GermEval-2018 [۴۲] رویدادهای مختلفی را برای یافتن راه‌حل بهتر برای تشخیص خودکار کلام نفرت‌انگیز برگزار کرده‌اند. در این

[1] Hate Speech

[2] Aggression برای مفهوم پرخاشگری (نیز همین طور است.)

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]

راستا، پژوهش‌گران مجموعه‌داده‌هایی با مقیاس بزرگ را از منابع متعدد جمع‌آوری کرده‌اند که به پژوهش در این زمینه کمک می‌کند. بسیاری از این مطالعات همچنین به کلام نفرت‌انگیز در چندین زبان غیرانگلیسی و جوامع برخط پرداخته‌اند. این موضوع منجر به بررسی و مقایسهٔ روند پردازش مختلف، از جمله انتخاب مجموعهٔ ویژگی و روش‌های یادگیری ماشین (مانند با نظارت، بدون نظارت و نیمه‌نظارتی) و الگوریتم‌های طبقه‌بندی (مانند بیز ساده[1]، رگرسیون خطی[2]، شبکهٔ عصبی کانولوشن[3]، حافظهٔ بلندمدت-کوتاه‌مدت[4]، معماری‌های یادگیری عمیق بازنمایی رمزگذار دو طرفه از ترنسفورمرها[5] و ...) شد. محدودیت رویکرد خودکار مبتنی بر متن برای تشخیص کارآمد، به‌طور گسترده مورد تأیید قرار گرفته که نیازمند پژوهش‌های آینده در این زمینه است [۳۵].

یک پیش‌فرض در بسیاری از الگوریتم‌های یادگیری ماشین این است که مجموعه‌های آموزشی و آزمایشی دارای توزیع مشابهی هستند. هنگامی که این توزیع‌ها با هم مطابقت ندارند، با تغییر مجموعه‌داده مواجه می‌شویم [۱۶] که در پردازش زبان طبیعی به‌عنوان تغییر دامنه[6] از آن یاد می‌شود. در این مسئله، دامنهٔ هدف و داده‌های آموزشی مبدأ متفاوت هستند و از یک توزیع مشابه نمونه‌برداری نشده‌اند. در نتیجه، عملکرد روی هدف کاهش می‌یابد که توانایی مدل‌ها برای تعمیم در دنیای واقعی را تضعیف می‌کند. تطبیق دامنه با یک موضوع باز و اساسی بزرگتر در یادگیری ماشین ارتباط دارد و آن هم تعمیم فراتر از توزیع آموزش است.

کار بر روی تطبیق دامنه اغلب بر تطبیق دامنهٔ با نظارت متمرکز شده است [۸، ۳۰]. در چنین تنظیم کلاسیک تطبیق دامنهٔ با نظارت، مقدار کمی از داده‌های دامنهٔ هدف دارای برچسب همراه با مقدار بیشتری از داده‌های دامنهٔ مبدأ دارای برچسب، در دسترس است. مسئله این است که با توجه به داده‌های محدود دامنهٔ هدف، از مبدأ به دامنهٔ هدف خاص منطبق شویم. تطبیق دامنهٔ بدون نظارت تنها با یادگیری از داده‌های هدف بدون برچسب، که به‌طورمعمول برای هر دو دامنهٔ مبدأ و هدف در دسترس است، باعث افزایش دقت در مسئلهٔ تطبیق دامنه می‌شود [۳۱].

[1] Naive Bayes
[2] Linear Regression
[3] Convolutional Neural Network (CNN)
[4] Long Short-Term Memory (LSTM)
[5] Bidirectional Encoder Representations from Transformers (BERT)
[6] Domain Shift

با ارزیابی مدل‌های متعدد روی دامنه‌های متفاوت و مطالعهٔ کارهای پیشین مشخص می‌شود که عملکرد مدل‌ها در تعمیم به سایر دامنه‌ها کاهش می‌یابد. در دنیای واقعی نیز بسیار دور از انتظار به نظر می‌رسد که دامنهٔ آموزش‌یافته با دامنهٔ هدف، دارای توزیع یکسان و داده‌های با شباهت بالا باشند. به همین دلیل نیاز است تا با استفاده از روش‌هایی سعی در افزایش قدرت تطبیق دامنهٔ مدل‌ها داشته باشیم.

در ادامه به مرور کارهای مرتبط خواهیم پرداخت. پس از آن تعدادی از روش‌های مبتنی بر ترنسفورمر[7] برای بهبود قابلیت تعمیم را بررسی خواهیم کرد؛ سپس در بخش بعدی نتایج آزمایش‌های درون‌دامنه‌ای و برون‌دامنه‌ای را ارائه خواهیم کرد. آزمایش‌های بخش درون‌دامنه‌ای را هم به‌صورت تک‌منبعی و هم به‌صورت چند منبعی (اضافه‌کردن بخشی از داده‌های دامنهٔ هدف به مجموعهٔ آموزشی) انجام می‌دهیم. در بخش برون‌دامنه‌ای نیز در نظر داریم با آموزش روی چند دامنه موجب افزایش قدرت تطبیق دامنهٔ مدل‌های تشخیص نفرت شویم؛ سپس با استفاده از روش‌های مختلف مبتنی بر ترنسفورمر شامل آموزش خصمانهٔ دامنه[8] و ترکیب متخصصان[9] که شامل میانگین‌گیری پیش‌بینی طبقه‌بندی‌کننده‌ها و توجه[10] است، سعی بهبود عملکرد مدل‌ها خواهیم داشت. در پایان نیز جمع‌بندی مطالب را ارائه می‌نماییم.

نشان خواهیم داد استفاده از چند دامنه برای آموزش، در بخش درون‌دامنه‌ای ممکن است باعث انتقال منفی[11] شود و در بخش برون‌دامنه‌ای در حدود نیمی از آزمایش‌ها سبب افزایش قدرت تطبیق دامنه می‌شود. همچنین به‌کارگیری روش‌های مبتنی بر ترنسفورمر در بهبود عملکرد تطبیق دامنه مؤثر است.

۲- کارهای مرتبط

رویکردهای تطبیق دامنه به‌طور کلی به سه دسته تقسیم می‌شوند: رویکردهای با نظارت [۸، ۱۳، ۲۳]، که برچسب هر دو دامنهٔ مبدأ و هدف در دسترس هستند، رویکردهای نیمه‌نظارتی [۱۱، ۴۴]، که در آن برچسب‌های مبدأ و مجموعهٔ کوچکی از برچسب‌ها برای دامنهٔ هدف ارائه می‌شود و در نهایت رویکردهای بدون نظارت [۵، ۱۵، ۲۵، ۳۸]، که در آن فقط برچسب‌هایی برای دامنهٔ مبدأ ذکر شده است.

[7] Transformer
[8] Domain Adversarial Training
[9] Mixture of Experts
[10] Attention
[11] Negative Transfer

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]
[DOI: 10.61186/jsdp.21.1.125]

یک رویکرد رایج برای تطبیق دامنهٔ بدون نظارت، ایجاد بازنمایی‌هایی است که در تغییر توزیع بین داده‌های مبدأ و هدف تغییرناپذیر هستند. برای شبکه‌های عمیق، این را می‌توان از طریق آموزش خصمانهٔ دامنه با استفاده از ترفند سادهٔ معکوس گرادیان[1] انجام داد [۱۵]. نشان داده شده است که این مورد در تنظیم تطبیق دامنهٔ چند مبدأ نیز کار می‌کند [۲۴]. سایر روش‌های یادگیری بازنمایی رایج شامل به حداقل رساندن کواریانس بین ویژگی‌های مبدأ و هدف [۳۸] و استفاده از حداکثر اختلاف میان توزیع حاشیه‌ای مبدأ و ویژگی‌های هدف به‌عنوان یک هدف خصمانه است [۱۹].

همچنین نشان داده شده که ترکیب متخصصان برای تطبیق دامنهٔ چند مبدأ مؤثر است. برای نمونه در [۲۲] از سازوکار توجه برای ترکیب پیش‌بینی‌های متخصصان[2] دامنه استفاده شده است و در [۱۹] نیز یادگیری ترکیب متخصصان که در دامنه‌های منحصربه‌فرد آموزش دیده‌اند، پیشنهاد شده است.

تعدادی از مطالعات به موضوع قابلیت تعمیم مدل‌ها در مسائل طبقه‌بندی زبان توهین‌آمیز در «مجموعه‌دادهٔ برون‌دامنه‌ای[3]» پرداخته‌اند. به‌عنوان مثال در [۴۱] بر روی سه مجموعه‌دادهٔ توییتر، W&H، Waseem و Davidson آزمایش شده است که Waseem توسعه‌یافته از W&H است. نویسندگان نشان می‌دهند که عملکرد طبقه‌بندی مجموعه‌دادهٔ برون‌دامنه‌ای پایین است. برای بهبود آن، داده‌های آموزشی از مجموعه‌دادهٔ دیگری مورد نیاز است به این صورت که یا مجموعه‌داده‌های مختلف ادغام شوند، یا مدل‌های آموزش داده‌شده بر روی یک مجموعه‌داده با استفاده از یادگیری انتقالی داده‌های مجموعه‌دادهٔ دیگر تنظیم شوند. همچنین در [۱۷] عملکرد ضعیف مجموعه‌دادهٔ برون‌دامنه‌ای را در مجموعه‌داده‌های بیشتر و با تنظیمات آزمایشی مختلف گزارش می‌شود. نویسندگان از رگرسیون خطی، پرسپترون چندلایه مبتنی بر کاراکتر[4]، واحد بازگشتی دروازه‌ای[5] به‌اضافهٔ شبکهٔ عصبی کانولوشن، LSTM و تنظیم دقیق‌یافتهٔ مدل زبان جهانی[6] برای تشخیص زبان توهین‌آمیز در مجموعه‌داده‌های W&H، Davidson، Wul2 و Zhang استفاده می‌کنند و نشان می‌دهند که عملکرد خوب تنها زمانی به دست می‌آید که روی یک مجموعه‌دادهٔ مشابه آزمایش شود. در [۲۱] از طیف وسیع‌تری از (نه مجموعه‌دادهٔ مختلف) استفاده می‌شود: W&H، Waseem، TRAC، Kol، Gao، Kaggle، Wul1، Wul2 و Wul3. قبل از آزمایش، برچسب همهٔ مجموعه‌داده‌ها به‌صورت «مثبت» (زبان توهین‌آمیز) و «منفی» (زبان غیرتوهین‌آمیز) دوتایی شده است. این بدان معناست که تمایز بین دسته‌های اصلی که مانع تجزیه‌وتحلیل دقیق ویژگی‌های هر یک از آنها و طبقه‌بندی با جزئیات زبان توهین‌آمیز می‌شود، از بین می‌رود. ماشین بردار پشتیبان با مدل‌های Unigram-Count ابتدا بر روی هر یک از مجموعه‌داده‌ها آموزش داده می‌شوند (با برچسب مجدد) و در هشت مجموعه‌دادهٔ دیگر آزمایش می‌شوند؛ سپس از یادگیری انتقالی [۴۱]، برای به‌دست‌آوردن تعمیم خاصی استفاده می‌شود. نویسندگان نتیجه می‌گیرند که برای عملکرد خوب در طبقه‌بندی مجموعه‌داده‌های هدف، داشتن داده‌های آموزشی، حداقل برخی داده‌های مجموعه‌دادهٔ مورد نظر، بسیار مهم است. بااین‌حال، باید توجه داشت که این نتیجه‌گیری با کار [۳۹] سازگار نیست.

هر سه مطالعهٔ بالا نیز تأثیر خصوصیات مجموعه‌داده‌ها را در طبقه‌بندی مجموعه‌دادهٔ برون‌دامنه‌ای ارزیابی می‌کنند؛ بنابراین، Waseem و همکاران عنوان می‌کنند که مجموعه‌دادهٔ Davidson راحت‌تر از مجموعه‌دادهٔ W&H طبقه‌بندی می‌شود، زیرا واژگان موجود در مجموعه‌دادهٔ Davidson حاوی درصد بالایی از زبان انگلیسی بومی آمریکایی آفریقایی است و بنابراین همگن‌تر است. فرض پژوهش [۲۱] این است که تفاوت در عملکرد طبقه‌بندی مجموعه‌دادهٔ برون‌دامنه‌ای از مجموعه‌داده‌ای به مجموعه‌دادهٔ دیگر، به‌دلیل تفاوت بین دسته‌های مجموعه‌داده و اندازهٔ مجموعه‌داده است. همچنین در [۱۷] استدلال می‌شود که نوع داده‌ها و معیارهای برچسب زدن نسبت به مدل ارتباط بیشتری دارند. بااین‌حال، در [۳۹] نشان داده می‌شود که با مدل‌های پیشرفته‌ای مانند BERT [۱۰] می‌توان یک مدل زبانی به دست آورد که تا حدی به تعمیم برسد، که بستگی زیادی به داده‌های آموزشی دارد. همان‌طورکه [۲۱، ۳۹] دسته‌های مجموعه‌داده‌های در نظر گرفته‌شده را در دو دستهٔ کلی «مثبت» (توهین‌آمیز) و «منفی» (غیرتوهین‌آمیز یا ملایم) ادغام می‌کنند؛ اگرچه همهٔ مجموعه‌داده‌های مورد استفاده یک نوع زبان توهین‌آمیز را به کار نمی‌برند.

[1] Gradient Reversal
[2] Experts
[3] Cross-Dataset
[4] Character-Based Multilayer Perceptron
[5] Gated Recurrent Unit (GRU)
[6] Universal Language Model Fine-Tuning (ULMFiT)

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]
[DOI: 10.61186/jsdp.21.1.125]

مدل‌های BERT در چهار مجموعه‌دادهٔ توییتر (W&H، Davidson، Offenseval و Founta) اعمال می‌شوند. نویسندگان بیان می‌کنند که اگر یک مدل در داده‌هایی استفاده شود که بیشتر شبیه داده‌های مورد استفاده برای آموزش است، بهتر تعمیم می‌یابد؛ بنابراین، مدلی که بر روی مجموعه‌دادهٔ Founta آموزش دیده است، هنگامی که روی مجموعه‌دادهٔ مشابه Offenseval آزمایش می‌شود و بالعکس، عملکرد خوبی دارد. در آزمایشی جداگانه، [۳۹] مدل‌هایی با تمام دسته‌های موجود در مجموعه‌دادهٔ Offenseval ساخته شد و آنها را در همهٔ دسته‌های سه مجموعه‌دادهٔ دیگر نیز آزمایش کردند. این امر تشخیص برخی از همپوشانی‌های بین مجموعه‌داده‌های در نظر گرفته‌شده را تسهیل می‌کند.

همچنین در [۳۹] هنگام رفتن از یک مجموعه‌دادهٔ آموزشی بزرگ به یک مجموعهٔ آزمایشی کوچک و بالعکس، افت عملکرد مشاهده شد. این مطابق با نتیجه‌گیری دیگری است که توسط [۲۱] انجام شده است که مجموعه‌داده‌هایی با درصد بیشتری از نمونه‌های مثبت نسبت به مجموعه‌داده‌هایی با نمونه‌های مثبت کمتر تمایل به تعمیم بهتری دارند، به‌ویژه هنگامی که در برابر مجموعه‌داده‌های متفاوت آزمایش می‌شوند. برای مثال، مدل‌های آموزش داده‌شده بر روی مجموعه‌دادهٔ Davidson، که اکثراً شامل موارد توهین‌آمیز است، هنگام آزمایش بر روی مجموعه‌دادهٔ Founta، که بیش‌تر شامل موارد غیرتوهین‌آمیز است، به‌خوبی عمل می‌کند.

در مطالعهٔ دیگری [۲۸]، نویسندگان تأیید می‌کنند که مدلی که بر روی مجموعه‌داده‌های با پوشش گسترده‌تری از پدیده‌ها آموزش دیده است، می‌تواند انواع دیگر زبان توهین‌آمیز را نیز نسبت به مواردی که در آن آموزش دیده است تشخیص دهد. نویسندگان از مجموعه‌داده‌های W&H، Hateval، Offenseval و Golbeck، با ماشین بردار پشتیبان خطی با کیف کلمات[1] و LSTM به‌عنوان مدل استفاده می‌کنند. بااین‌حال، باید توجه داشت که کیفیت تعمیم در این آزمایش (با بیشینه امتیاز F1 ۰/۵۵) به‌نسبه متوسط است.

همچنین در [۲] یک ویژگی اضافی را ارائه می‌شود که باید در زمینهٔ طبقه‌بندی مجموعه‌دادهٔ برون‌دامنه‌ای کلام نفرت‌انگیز در نظر گرفته شود که آن تعداد نویسندگان مطالب گرفته شده در یک مجموعه‌داده است. آنها نشان می‌دهند که پتانسیل عمومیت مجموعه‌داده‌های نفرت Waseem، که پیام‌هایشان به‌عنوان نفرت‌انگیز (تبعیض جنسیتی یا نژادپرستی) از چند حساب سرچشمه می‌گیرد، زمانی افزایش می‌یابد که مجموعه‌داده‌ها با مثال‌های کلام نفرت‌انگیز از حساب‌های دیگر برگرفته از مجموعه‌داده‌های Davidson غنی شود [۹]. اما حتی در این مورد هم امتیاز F1 به دست آمده فقط ۰/۵۴ برای گروه کلام نفرت‌انگیز است، به این معنا که مجموعه‌داده‌های متنوع‌تر مفید هستند، اما مشکل طبقه‌بندی ضعیف مجموعه‌دادهٔ برون‌دامنه‌ای را حل نمی‌کنند.

مطالعهٔ اخیر دیگری که هدف آن غلبه بر قابلیت تعمیم محدود مدل‌ها در دامنه‌ها و در نتیجه مجموعه‌داده‌ها بود، ارائه شد [۳۳]. نویسندگان استدلال می‌کنند که مدل‌های آموزش داده شده بر روی مجموعه‌داده‌های جمع‌آوری شده از یک بستر برخط را می‌توان فرض کرد که به‌طور پیش‌فرض به آن بستر محدود شده‌اند و یک مجموعه‌داده از بستر برون‌دامنه‌ای تعمیم‌پذیری بیشتر یک مدل را تضمین می‌کند. آنها از چهار مجموعه‌دادهٔ جمع‌آوری‌شده از یوتیوب [۳۲]، ویکی‌پدیا [۲۰]، Reddit [۱] و توییتر [۹] استفاده می‌کنند. به‌طور مشابه در [۲۱، ۳۹] نمونه‌ها در دو دستهٔ کلی، مثبت (نفرت‌انگیز) و منفی (غیرنفرت‌انگیز) ادغام می‌شود؛ سپس چندین الگوریتم طبقه‌بندی (رگرسیون لجستیک، بیز ساده، ماشین بردار پشتیبان، XGBoost[2] و شبکه‌های عصبی)، بازنمایی ویژگی‌ها (کیف کلمات، TF-IDF[3]، Word2Vec، BERT و ترکیب آنها) بر روی دسته‌های کلی اعمال می‌شوند. با XGBoost و همهٔ ویژگی‌ها، بهترین عملکرد امتیاز F1 ۰/۹۲ گزارش شده‌است.

در مورد استفاده از مدل‌ها، مطالعات قبلی در مورد تعمیم مدل بر طیف وسیعی از مدل‌های طبقه‌بندی با نظارت مختلف استوار است. برخی از ماشین‌های بردار پشتیبان بیشتر به‌عنوان مبنای اولیه استفاده می‌کنند [۲۱، ۲۸]، برخی دیگر از یادگیری عمیق استفاده می‌کنند [۱۷]. درهمین اواخر، نویسندگان از مدل‌های مبتنی بر ترنسفورمر مانند BERT استفاده می‌کنند که عملکرد بهتری را ارائه می‌دهند [۳۳، ۳۹].

۳- روش‌ها

مدل‌ها بر اساس روش‌های ارائه‌شده در [۴۲] در حوزهٔ تحلیل احساسات به مسئلهٔ کلام نفرت‌انگیز تطبیق

[1] Bag-of-Words

[2] eXtreme Gradient Boosting

[3] Term Frequency–Inverse Document Frequency

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]

یافته‌اند. در روش‌های آموزش خصمانهٔ دامنه و ترکیب متخصصان، تعداد دامنه‌ها K، دامنهٔ مبدأ S و دامنهٔ هدف T هستند. دامنه‌های مبدأ شامل مجموعه‌داده‌های دارای برچسب D_s, $s \in \{1, \ldots, K\}$ هستند و دامنهٔ هدف فقط از داده‌های بدون برچسب U_t تشکیل شده است. هدف این است که طبقه‌بندی‌کنندهٔ f آموزش یابد، تا فقط با استفاده از داده‌های برچسب زده‌شده از S و داده‌های اختیاری بدون برچسب از T به‌خوبی به T تعمیم یابد. در اینجا شبکهٔ اصلی f_z, $z \in S \cup \{g\}$ در نظر گرفته می‌شود که مربوط به یک شبکهٔ خاص دامنه یا یک شبکه مشترک جهانی است. این شبکه‌های f_z با استفاده از مدل‌های ترنسفورمر بزرگ از پیش آموزش‌دیده[1]، به‌ویژه DistilBERT، راه‌اندازی می‌شوند [۳۴].

۳-۱- DistilBERT

با فراگیرترشدن یادگیری انتقالی از مدل‌های از پیش آموزش‌دیده در مقیاس بزرگ در NLP، محدودیت‌های ذخیره‌سازی و محاسبات نیز رایج‌تر می‌شوند. ایجاد DistiIBERT سعی در کاهش برخی از این مسائل دارد [۴۷]. این مدل زبانی همه‌منظوره کوچک‌تر از پیش آموزش‌دیده، روندی با عملکرد خوب ارائه می‌کند که سپس می‌تواند برای انجام سایر وظایف (طبقه‌بندی احساسات، SQuAD و ...) به‌خوبی تنظیم شود. پژوهش‌گران در حین پیش‌آموزش بر روی همان ساختار BERT، از روش‌های مختلفی برای متراکم‌کردن اندازهٔ مدل و افزایش سرعت استنتاج استفاده و در عین حال بسیاری از عملکرد را حفظ می‌کنند. این روش‌ها شامل تقطیر دانش، یک تابع Loss سه‌گانهٔ جدید و انتخاب‌های مختلف معماری، مقداردهی اولیه و هایپرپارامتر است [۵۱].

۳-۲- آموزش خصمانهٔ دامنه

روش تطبیق خصمانهٔ دامنه، روشی است که مطالعات خوبی روی آن صورت گرفته که در [۱۵] شرح داده شده و نشان داده شده است که هم در شبکه‌های پیچشی[2] و هم در شبکه‌های بازگشتی[3] در حل مسائل پردازش زبان‌های طبیعی مؤثر بوده‌اند [۱۸، ۲۴]؛ بنابراین یک نامزد اصلی برای مطالعه در زمینهٔ مدل‌های LPX است. علاوه‌براین، برخی شواهد اولیه نشان می‌دهد که آموزش خصمانه ممکن است تعمیم LPX را برای تطبیق دامنهٔ تک‌مبدأ بهبود بخشد [۲۶].

برای یادگیری بازنمایی‌های متغیر دامنه، مدلی آموزش داده می‌شود که بازنمایی‌های آموزش داده‌شده حداکثر طبقه‌بندی‌کنندهٔ دامنهٔ f_d را با هم اشتباه می‌گیرند. این امر از طریق هدف min-max بین پارامترهای طبقه‌بندی دامنهٔ θ_D و پارامترهای θ_G یک کدگذار f_g انجام می‌شود؛ سپس می‌توان هدف را به‌شرح زیر توصیف کرد [۴۳]:

$$L_D = \max_{\theta_D} \min_{\theta_G} - d \log f_d(f_g(x)) \qquad (۱)$$

که در آن d دامنهٔ نمونهٔ ورودی x است. تأثیر این امر در بهبود توانایی طبقه‌بندی‌کننده در تعیین دامنهٔ یک نمونه است، درحالی‌که مدل را تشویق می‌کند تا از طریق به حداقل رساندن خطای منفی، حداکثر بازنمایی‌های درهم‌ریخته را ایجاد کند. در عمل، این امر با آموزش مدل با استفاده از تابع هزینهٔ آنتروپی متقاطع استاندارد[4] انجام می‌شود، اما گرادیان‌های خطا را با توجه به پارامترهای مدل θ_G معکوس می‌کند [۴۲].

۳-۳- روش‌های ترکیب متخصصان

ترکیب متخصصان نوع خاصی از شبکهٔ عصبی است که نورون‌ها در بسیاری از خوشه‌های کوچک به هم متصل هستند و هر خوشه فقط تحت شرایط خاص فعال است. لایه‌های پایین‌تر شبکه ویژگی‌ها را استخراج می‌کنند و متخصصان برای ارزیابی آن ویژگی‌ها فراخوانی می‌شوند (برای هر مورد، فقط برخی از متخصصان فراخوانی می‌شوند) [۵۳].

مدل‌های ترکیب متخصصان [۵۴] فضای مسئله را به چند فضای فرعی تقسیم می‌کنند و به متخصصان اجازه می‌دهند در هر زیرفضا تخصصی شوند. در سال‌های اخیر این مفهوم با موفقیت در NLP اعمال شده است [۵۵] و مدل‌هایی با مقیاس پارامتری میلیاردی یا حتی تریلیون را ممکن می‌سازد [۵۲، ۵۶، ۵۷]. با این حال، این برنامه‌ها عمدتاً بر جنبه‌های مقیاس‌بندی تمرکز دارند؛ علاوه‌براین، بیش‌تر آنها متخصصان را بر اساس هر نمونه انتخاب می‌کنند [۵۸].

ترکیب متخصصان دارای مزایایی متمایز است: آنها می‌توانند به شرایط خاص با تخصص بیشتر پاسخ دهند و به شبکه اجازه می‌دهند تا رفتارهای متنوع‌تری را نشان دهد. متخصصان می‌توانند ترکیبی از محرک‌ها را دریافت

[1] Large Pretrained Transformer (LPX)
[2] Convolutional Networks
[3] Recurrent Networks
[4] Standard Cross Entropy Loss

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]
[DOI: 10.61186/jsdp.21.1.125]

کنند و داده‌ها را با حس‌گرهای مختلف یک‌پارچه کنند. هنگامی که شبکه در حال کار است، تنها چند متخصص فعال هستند؛ حتی یک شبکهٔ بزرگ تنها به مقدار کمی از قدرت پردازش نیاز دارد. همان‌طورکه شبکه‌های عصبی پیچیده‌تر می‌شوند، جریان‌های زیادی از داده‌ها را یک‌پارچه و پاسخ‌های متنوع‌تری را ارائه می‌کنند، مدل‌های ترکیب متخصصان غالب خواهند شد؛ بنابراین، به درک چگونگی تکامل ترکیب متخصصان کمک می‌کند [۵۳].

دو ترکیب متخصص مختلف در ادامه بررسی می‌شوند: میانگین‌گیری ساده و سازوکار توجه جدید مبتنی بر توجه بر مبنای ترنسفورمر [۴۰]. گزارش‌های قبلی نشان می‌دهد که استفاده از ترکیب متخصصان دامنه و طبقه‌بندی‌کننده‌های مشترک منجر به بهبود عملکرد هنگام دسترسی به چندین دامنهٔ مبدأ می‌شود [۱۹، ۲۴]. با توجه به این موارد، در ادامه بررسی می‌شود که آیا ترکیب متخصصان هنگام استفاده از مدل‌های LPX فایده‌ای دارند یا خیر.

برای تنظیمات با دامنهٔ K، K مدل مختلف LPX $f_k, k \in \{1, \ldots, K\}$ متناظر با هر دامنه تنظیم شده است. همچنین یک مدل LPX اضافی f_g متناظر به یک مدل مشترک جهانی وجود دارد. پیش‌بینی خروجی این مدل‌ها به‌ترتیب D_s, $s \in \{1, \ldots, K\}$ و p_g است. از آنجا که مسائل موردنظر طبقه‌بندی دودویی هستند، این مقادیر در محدودهٔ (0, 1) هستند. احتمال خروجی نهایی به‌عنوان ترکیب وزنی از مجموعه‌ای از احتمال‌های متخصص دامنهٔ $\overline{K} \subseteq S$ و احتمال مدل مشترک جهانی محاسبه می‌شود. از چهار روش برای محاسبهٔ وزن استفاده می‌شود [۴۲].

- میانگین‌گیری

روش نخست یک میانگین‌گیری ساده از پیش‌بینی طبقه‌بندی‌کننده‌های خاص و مشترک است. خروجی نهایی مدل [۴۳]:

$$p_A(x, \overline{K}) = \frac{1}{|\overline{K}|+1} \sum_{k \in \overline{K}} p_k(x) + p_g(x) \quad (۲)$$

- مدل توجه

در نهایت، یک مدل توجه پارامتری شدهٔ جدید آموخته می‌شود که بر اساس نمونهٔ ورودی به دامنه‌های مختلف توجه می‌کند. روش توجه مبتنی بر مقیاس توجه تولید شده به مقیاس کوچک است که در مدل‌های ترنسفورمر اعمال می‌شود [۴۰] که در آن یک مدل مشترک جهانی به‌عنوان یک شبکهٔ پرس‌وجو از هر یک از مدل‌های متخصص و مشترک استفاده می‌کند. به‌این‌ترتیب، یک مدل مشترک f_g یک بردار $r_k \in R^d$ و هر متخصص دامنه یک بردار $r_g \in R^d$ تولید می‌کند. نخست، برای یک نمونهٔ ورودی x، یک احتمال برای کار نهایی از طبقه‌بندی‌کنندهٔ هر مدل به دست می‌آید که احتمال‌های p_g و p_k, $k \in \{0, \ldots, K\text{-}1\}$ را به‌همراه دارد؛ سپس بردار توجه α_X از طریق تبدیل‌های زیر به‌دست می‌آید:

$$q = gQ^T \quad (۳)$$

$$k = \begin{bmatrix} r_1 \\ \cdot \\ \cdot \\ r_k \\ r_g \end{bmatrix} \quad (۴)$$

$$\alpha_X = \mathrm{softmax}(qk^T) \quad (۵)$$

که در آن $K \in R^{d \times d}$ و $Q \in R^{d \times K}$. سپس بردار توجه α_X به پیش‌بینی‌های منحصربه‌فرد هر متخصص دامنه و مدل مشترک جهانی توجه می‌کند [۴۳].

$$p_X(x, \overline{K}) = \sum_{k \in \overline{K}} p_k(x) * \alpha_X^{(k)} + p_g(x) * \alpha_X^{(g)}(x) \quad (۶)$$

برای اطمینان از اینکه هر مدل به‌عنوان یک متخصص در دامنهٔ خاص آموزش دیده است، یک روش آموزشی مشابه با [۱۹] استفاده می‌شود.

۴- آزمایش‌ها و بحث

۴-۱- تنظیمات

ما برای همهٔ آزمایش‌ها از چهار مجموعه‌داده به نام‌های Davidson [۹]، Stormfront [۴۸]، TRAC [۴۹] و W&H [۵۰] استفاده می‌کنیم که دسته‌بندی‌های مختلف نفرت‌پراکنی را پوشش می‌دهند و از چند سکو جمع‌آوری شده‌اند. مشخصات مجموعه‌داده‌ها در جدول (۱) به‌اختصار نشان داده شده است.

در بخش درون‌دامنه، شیوهٔ تقسیم‌بندی مجموعه‌داده‌ها به بخش‌های آموزشی و آزمایشی در مجموعه‌داده‌های Davidson، Stormfront و W&H به‌صورت ۷۰٪-۳۰٪ و در مجموعه‌دادهٔ TRAC که تعداد داده‌های کمتری دارد، به‌صورت ۸۰٪-۲۰٪ است.

برای به‌دست‌آوردن تصویری عینی از پتانسیل تعمیم مدل‌ها در مجموعه‌داده‌های مختلف، از روشی برای استانداردسازی دسته‌بندی‌ها استفاده شده است [۱۴]. در

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]

مجموعه‌دادهٔ Davidson علاوه‌بر کلاس‌های Hate Speech، Offense و Normal، یک کلاس جدید دیگر به نام Toxicity اضافه شده است که حاوی همهٔ داده‌های دو کلاس Hate Speech و Offense است. در مجموعه‌دادهٔ TRAC هم علاوه بر کلاس‌های Convert Aggression، Overt Aggression و Normal، کلاسی به نام Aggression اضافه شده است که شامل تمام داده‌های دو کلاس Convert Aggression و Overt Aggression است. مجموعه‌دادهٔ W&H نیز سه کلاس به نام‌های Sexism، Racism و Normal، دارد و یک کلاس جدید دیگر به نام Hate Speech به آن اضافه شده است که مجموع داده‌های دو کلاس Sexism و Racism را شامل می‌شود. مجموعه‌دادهٔ Stormfront شامل دو کلاس Hate Speech و Normal است و کلاس دیگری به آن اضافه نمی‌کنیم.

در هر آزمایش، یک کلاس از هر مجموعه‌داده به‌عنوان کلاس مثبت انتخاب می‌شود و باقی کلاس‌ها به‌عنوان کلاس منفی در نظر گرفته می‌شوند. درواقع مسئله دوکلاسه است. معیار ارزیابی نیز Macro F1 است. برای اطمینان بیشتر هر آزمایش را سه بار تکرار کرده‌ایم و میانگین نتایج را به همراه درصد خطا (واریانس) ارائه می‌کنیم.

در همهٔ آزمایش‌ها از اندازهٔ دستهٔ ۴ و نرخ یادگیری ۲e-۵ استفاده شده است و کلیهٔ آزمایش‌ها در سه دوره آموزش دیده‌اند. با توجه به اینکه این تنظیمات بر اساس محدودیت‌های منابع انتخاب شدند، اما نتایج خوبی را در پی داشتند.

(جدول-۱): مشخصات مجموعه‌داده‌ها

(Table-1): Specifications of the datasets

Name	Category	Size (Post Based)	Source
Davidson ('David')	Offense ('Offe')	19190	Twitter
	Hate Speech ('HS')	1430	
	Toxicity (Offense or Hate Speech) ('Toxic')	20620	
	Normal	4163	
Stormfront ('Storm')	Hate Speech ('HS')	1197	Stormfront
	Normal	9720	
TRAC	Convert Aggression ('Cag')	5297	Facebook
	Overt Aggression ('Oag')	3418	
	Aggression (Convert or Overt Aggression) ('Aggr')	8715	
	Normal	6284	
W&H	Sexism ('Sex')	3430	Twitter
	Racism ('Race')	1976	
	Hate Speech (Sexism or Racism) ('HS')	5406	
	Normal	11501	

۴-۲- آزمایش‌ها

انواع روش‌های به‌کاررفته در آزمایش‌ها که در بخش ۳ به تفصیل شرح داده شدند، در اینجا به‌اختصار بیان می‌شوند.

Basic: DistilBert پایه با یک لایه طبقه‌بندی واحد در خروجی

Adv-X: DistilBert با نظارت خصمانهٔ دامنهٔ اعمال‌شده در لایهٔ X

MoE-Avg: DistilBert ترکیب متخصصان با استفاده از میانگین‌گیری

MoE-Att: DistilBert ترکیب متخصصان با استفاده از مدل توجه

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]

۴-۲-۱- آزمایش‌های درون‌دامنه‌ای

آزمایش‌های درون‌دامنه‌ای را در مرحلهٔ نخست به‌صورت تک‌دامنه‌ای و در مرحلهٔ بعد به‌صورت چندمنبعی انجام می‌دهیم. آزمایش‌های تک منبعی به‌دلیل تشابه و توزیع یکسان داده‌های بخش آموزشی و آزمایشی، از عملکرد بالایی برخوردارند؛ سپس شش ترکیب تصادفی چهارتایی از دامنه‌های مختلف در نظر می‌گیریم و در آزمایش‌های چندمنبعی، سه دامنهٔ دیگر را به بخش آموزشی اضافه می‌کنیم. همان طور که در جدول (۲) مشخص است، عملکرد آزمایش‌های تک‌منبعی مقداری بهتر است. در واقع در آزمایش‌های چند منبعی انتقال منفی رخ داده است و دامنه‌های اضافه‌شده باعث کمی کاهش در نتایج تطبیق دامنه شده‌اند.

(جدول-۲): میانگین نتایج آزمایش‌های درون‌دامنه‌ای همراه واریانس با معیار Macro F1 (بخش In نتایج مربوط به آزمایش‌های درون دامنهٔ تک منبعی، بخش Mix نتایج مربوط به آزمایش‌های چند منبعی به همراه آموزش روی بخشی از دامنهٔ هدف)

(Table 2): Average results of in-domain tests along with variance with Macro F1 measure (In section results related to single-domain in-domain tests, Mix section results related to multi-domain tests along with training on a part of the target domain)

Method	Davidson-tox		Stormfront-hs		TRAC-aggr		W&H-hs	
	In	Mix	In	Mix	In	Mix	In	Mix
Basic	.935 ±.001	.923 ±.004	.785 ±.000	.711 ±.007	.748 ±.004	.725 ±.000	.844 ±.007	.828 ±.001
Adv-6	.935 ±.001	.924 ±.002	.774 ±.006	.701 ±.005	.754 ±.003	.719 ±.002	.847 ±.003	.828 ±.002
Adv-3	.937 ±.002	.922 ±.002	.724 ±.045	.702 ±.004	.748 ±.003	.720 ±.001	.846 ±.001	.830 ±.001
MoE-Avg	.938 ±.001	.927 ±.002	.784 ±.007	.700 ±.006	.752 ±.001	.734 ±.004	.852 ±.001	.827 ±.001
MoE-Att	.935 ±.001	.922 ±.004	.781 ±.010	.703 ±.006	.751 ±.003	.716 ±.004	.852 ±.003	.829 ±.000
Method	**Davidson-tox**		**Stormfront-hs**		**TRAC-cag**		**W&H-sex**	
	In	**Mix**	**In**	**Mix**	**In**	**Mix**	**In**	**Mix**
Basic	.935 ±.001	.925 ±.001	.785 ±.000	.687 ±.024	.621 ±.001	.593 ±.002	.878 ±.001	.858 ±.001
Adv-6	.935 ±.001	.928 ±.003	.774 ±.006	.696 ±.010	.626 ±.004	.576 ±.005	.874 ±.003	.857 ±.005
Adv-3	.937 ±.002	.924 ±.004	.724 ±.045	.704 ±.003	.625 ±.019	.585 ±.003	.873 ±.001	.850 ±.017
MoE-Avg	.938 ±.001	.925 ±.001	.784 ±.007	.682 ±.002	.602 ±.020	.584 ±.004	.882 ±.001	.845 ±.004
MoE-Att	.935 ±.001	.923 ±.005	.781 ±.010	.698 ±.012	.613 ±.004	.573 ±.003	.874 ±.012	.856 ±.002
Method	**Davidson-offe**		**Stormfront-hs**		**TRAC-oag**		**W&H-sex**	
	In	**Mix**	**In**	**Mix**	**In**	**Mix**	**In**	**Mix**
Basic	.887 ±.000	.874 ±.001	.785 ±.000	.656 ±.052	.700 ±.004	.649 ±.024	.878 ±.001	.859 ±.009
Adv-6	.881 ±.003	.872 ±.003	.774 ±.006	.676 ±.032	.699 ±.006	.691 ±.010	.874 ±.003	.851 ±.018
Adv-3	.885 ±.002	.875 ±.000	.724 ±.045	.679 ±.046	.707 ±.009	.645 ±.025	.873 ±.001	.853 ±.016
MoE-Avg	.887 ±.003	.879 ±.002	.784 ±.007	.695 ±.007	.715 ±.003	.678 ±.002	.882 ±.001	.853 ±.005
MoE-Att	.886 ±.003	.878 ±.002	.781 ±.010	.678 ±.035	.706 ±.010	.693 ±.006	.874 ±.012	.857 ±.010

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]

(ادامهٔ جدول-۲): میانگین نتایج آزمایش‌های درون‌دامنه‌ای همراه واریانس با معیار Macro F1 (بخش In نتایج مربوط به آزمایش‌های درون دامنهٔ تک منبعی، بخش Mix نتایج مربوط به آزمایش‌های چند منبعی به همراه آموزش روی بخشی از دامنهٔ هدف)

(Table 2): Average results of in-domain tests along with variance with Macro F1 measure (In section results related to single-domain in-domain tests, Mix section results related to multi-domain tests along with training on a part of the target domain)

Method	Davidson-offe		Stormfront-hs		TRAC-cag		W&H-race	
	In	Mix	In	Mix	In	Mix	In	Mix
Basic	.887 ±.000	.875 ±.000	.785 ±.000	.675 ±.021	.621 ±.001	.584 ±.006	.872 ±.001	.843 ±.002
Adv-6	.881 ±.003	.874 ±.001	.774 ±.006	.657 ±.048	.626 ±.004	.526 ±.029	.874 ±.005	.840 ±.002
Adv-3	.885 ±.002	.876 ±.004	.724 ±.045	.658 ±.050	.625 ±.019	.575 ±.011	.876 ±.001	.828 ±.017
MoE-Avg	.887 ±.003	.880 ±.000	.784 ±.007	.662 ±.039	.602 ±.020	.590 ±.003	.873 ±.003	.848 ±.004
MoE-Att	.886 ±.003	.871 ±.002	.781 ±.010	.682 ±.027	.613 ±.004	.545 ±.032	.876 ±.005	.817 ±.026
Method	**Davidson-hs**		**Stormfront-hs**		**TRAC-aggr**		**W&H-race**	
	In	Mix	In	Mix	In	Mix	In	Mix
Basic	.699 ±.015	.670 ±.015	.785 ±.000	.710 ±.008	.748 ±.004	.715 ±.009	.872 ±.001	.849 ±.003
Adv-6	.675 ±.017	.688 ±.007	.774 ±.006	.708 ±.009	.754 ±.003	.721 ±.002	.874 ±.005	.848 ±.002
Adv-3	.669 ±.026	.669 ±.013	.724 ±.045	.711 ±.010	.748 ±.003	.714 ±.014	.876 ±.001	.847 ±.001
MoE-Avg	.704 ±.011	.649 ±.011	.784 ±.007	.694 ±.004	.752 ±.001	.730 ±.002	.873 ±.003	.862 ±.004
MoE-Att	.669 ±.028	.680 ±.025	.781 ±.010	.715 ±.005	.751 ±.003	.718 ±.002	.876 ±.005	.852 ±.004
Method	**Davidson-hs**		**Stormfront-hs**		**TRAC-oag**		**W&H-hs**	
	In	Mix	In	Mix	In	Mix	In	Mix
Basic	.699 ±.015	.652 ±.010	.785 ±.000	.699 ±.011	.700 ±.004	.676 ±.014	.844 ±.007	.823 ±.010
Adv-6	.675 ±.017	.680 ±.001	.774 ±.006	.707 ±.030	.699 ±.006	.688 ±.009	.847 ±.003	.823 ±.005
Adv-3	.669 ±.026	.667 ±.004	.724 ±.045	.702 ±.036	.707 ±.009	.692 ±.004	.846 ±.001	.824 ±.010
MoE-Avg	.704 ±.011	.663 ±.014	.784 ±.007	.697 ±.014	.715 ±.003	.683 ±.006	.852 ±.001	.828 ±.001
MoE-Att	.669 ±.028	.670 ±.016	.781 ±.010	.690 ±.012	.706 ±.010	.663 ±.024	.852 ±.003	.824 ±.006

۴-۲-۲- آزمایش‌های برون‌دامنه‌ای

در آزمایش‌های برون‌دامنه‌ای نیز شش ترکیب چهارتایی تصادفی از دامنه‌ها را در نظر می‌گیریم. ابتدا به‌صورت تک‌منبعی با مدل پایه آزمایش می‌کنیم. در هر یک از این آزمایش‌ها روی یک دامنه آموزش را انجام می‌دهیم و روی سه دامنه‌ی دیگر که هر یک دامنهٔ هدف هستند، آزمایش می‌کنیم.

همان‌طورکه نتایج تغییر دامنهٔ آموزشی در جدول (۳) قابل مشاهده است، ضمن تغییر دامنه همهٔ نتایج کاهش پیدا کرده‌اند؛ اما میزان کاهش در دامنه‌های مختلف معنادار است. در بخش‌هایی که داده‌های دامنه‌ها شباهت بیشتری به هم دارند یا از توزیع مشابهی برخوردارند، شاهد عملکرد بهتر تطبیق دامنه هستیم. برای مثال کلاس Hate Speech مجموعه‌دادهٔ Davidson نسبت به کلاس Offense و Toxicity همین مجموعه‌داده، تشابه بیشتری با کلاس Hate Speech مجموعه‌دادهٔ Stormfront دارد. به همین دلیل وقتی روی کلاس Hate Speech مجموعه‌دادهٔ Stormfront آموزش انجام می‌شود، نتیجهٔ آزمایش روی کلاس Hate Speech مجموعه‌دادهٔ Davidson حدود ۳۰٪ بهتر از آزمایش روی باقی کلاس‌های Davidson است.

در ادامه برای بهبود عملکرد آزمایش‌های برون‌دامنه‌ای، از آموزش چند منبعی (در اینجا سه تا) استفاده می‌کنیم. برای این منظور، همان شش ترکیب چهارتایی تصادفی دامنه‌ها در آزمایش‌های قبل را در نظر می‌گیریم. در هر آزمایش روی سه دامنهٔ مختلف آموزش را

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]
[DOI: 10.61186/jsdp.21.1.125]

انجام می‌دهیم و میزان تطبیق دامنه را در دامنهٔ هدف می‌سنجیم. در این آزمایش‌ها بررسی می‌شود که تجمیع دامنه‌ها از مجموعه‌داده‌های مختلف چه میزان می‌توانند در تطبیق دامنه مؤثر واقع شوند و میزان تأثیرگذاری به چه عواملی بستگی دارد.

همان‌طورکه پیش‌تر گفته شد، میزان تشابه داده‌ها و شباهت توزیع آنها در دامنه‌های مختلف در نتایج مؤثر هستند. نکتهٔ قابل‌توجه در رابطه با همین موضوع این است که دامنه‌هایی که میزان نزدیکی آنها به یکدیگر کمتر است، اغلب در آزمایش‌های برون‌دامنه‌ای چندمنبعی عملکرد ضعیف‌تری نسبت به آزمایش‌های تک‌منبعی دارند و معنی آن این است که در این موارد افزایش دامنه‌های دیگر برای آموزش کمکی به افزایش قدرت تطبیق دامنه نکرده است. البته مطابق جدول (۳) در ۴۰٪ آزمایش‌های برون‌دامنه‌ای با مدل پایه، اضافه کردن دامنه‌های دیگر سبب افزایش قدرت تطبیق دامنه شده است؛ سپس برای بهبود عملکرد آزمایش‌های برون‌دامنه‌ای چند منبعی، روش‌های آموزش خصمانهٔ دامنه و ترکیب متخصصان را به کار می‌گیریم. در آموزش خصمانهٔ دامنه دو روش به نام‌های روش آموزش خصمانهٔ دامنهٔ ۶ و روش آموزش خصمانهٔ دامنهٔ ۳ به کار می‌گیریم و در هر دو روش به ارزیابی عملکرد تطبیق دامنه می‌پردازیم. در ترکیب متخصصان نیز از دو روش استفاده می‌کنیم که شامل میانگین‌گیری و سازوکار توجه است و در اینجا نیز میزان تطبیق دامنه را در آزمایش‌های مختلف بررسی می‌کنیم.

همان طور که در جدول (۳) مشخص است، در آزمایش‌های چند منبعی، استفاده از مجموع روش‌های آموزش خصمانهٔ دامنه و ترکیب مخصصان نسبت به مدل پایه، در ۸۷٪ از موارد شاهد بهبود عملکرد هستیم که نشان‌دهندهٔ عملکرد مثبت این روش‌ها است. در ادامه به بررسی میزان برتری هر یک از روش‌ها به طور مجزا می‌پردازیم.

(جدول-۳): میانگین نتایج آزمایش‌های برون‌دامنه‌ای به‌صورت تک‌منبعی، چند منبعی و با استفاده از روش‌های آموزش دامنهٔ خصمانه و ترکیب متخصصان همراه واریانس با معیار Macro F1

(Table-3): The average results of out-of-domain tests in single-domain, multi-domain and using adversarial domain training methods and combination of experts with variance with Macro F1 measure

Method	Train Set	David-tox	Storm-hs	TRAC-aggr	W&H-hs
Basic	David-tox	-	.608 ±.011	.372 ±.021	.577 ±.009
	Storm-hs	.237 ±.042	-	.385 ±.013	.503 ±.020
	TRAC-aggr	.674 ±.012	.502 ±.036	-	.569 ±.013
	W&H-hs	.547 ±.026	.539 ±.002	.387 ±.007	-
	Multi Source	.619 ±.025	.614 ±.013	.404 ±.021	.621 ±.015
Adv-6		.620 ±.054	.589 ±.010	.448 ±.012	.622 ±.010
Adv-3		.618 ±.019	.594 ±.003	.422 ±.010	.617 ±.004
MoE-Avg		.620 ±.005	.614 ±.005	.401 ±.005	.626 ±.006
MoE-Att		.610 ±.037	.602 ±.008	.402 ±.003	.614 ±.014
Method	**Train Set**	**David-tox**	**Storm-hs**	**TRAC-cag**	**W&H-sex**
Basic	David-tox	-	.608 ±.011	.424 ±.014	.579 ±.020
	Storm-hs	.237 ±.042	-	.425 ±.004	.443 ±.001
	TRAC-cag	.167 ±.006	.488 ±.004	-	.472 ±.009

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]

	W&H-sex	.630 ±.007	.481 ±.003	.415 ±.005	-
	Multi Source	.589 ±.047	.567 ±.017	.421 ±.001	.556 ±.002
Adv-6		.538 ±.044	.560 ±.005	.419 ±.002	.550 ±.009
Adv-3		.539 ±.020	.565 ±.017	.427 ±.010	.554 ±.009
MoE-Avg		.606 ±.019	.564 ±.004	.405 ±.002	.550 ±.004
MoE-Att		.596 ±.018	.560 ±.012	.411 ±.002	.558 ±.008

(ادامهٔ جدول-۳): میانگین نتایج آزمایش‌های برون‌دامنه‌ای به‌صورت تک‌منبعی، چند منبعی و با استفاده از روش‌های آموزش دامنهٔ خصمانه و ترکیب متخصصان همراه واریانس با معیار Macro F1

(Table-3): The average results of out-of-domain tests in single-domain, multi-domain and using adversarial domain training methods and combination of experts with variance with Macro F1 measure

Method	Train Set	David-offe	Storm-hs	TRAC-oag	W&H-sex
Basic	David-offe	-	.524 ±.011	.475 ±.008	.561 ±.031
	Storm-hs	.261 ±.036	-	.540 ±.011	.443 ±.001
	TRAC-oag	.672 ±.024	.634 ±.002	-	.514 ±.002
	W&H-sex	.669 ±.012	.481 ±.003	.456 ±.003	-
	Multi Source	.678 ±.015	.547 ±.006	.481 ±.005	.531 ±.005
Adv-6		.582 ±.098	.563 ±.018	.512 ±.003	.528 ±.002
Adv-3		.645 ±.055	.588 ±.008	.511 ±.007	.530 ±.003
MoE-Avg		.672 ±.011	.563 ±.013	.472 ±.006	.531 ±.005
MoE-Att		.701 ±.010	.569 ±.010	.486 ±.008	.545 ±.009

Method	Train Set	David-offe	Storm-hs	TRAC-cag	W&H-race
Basic	David-offe	-	.524 ±.011	.403 ±.005	.457 ±.005
	Storm-hs	.261 ±.036	-	.425 ±.004	.667 ±.015
	TRAC-cag	.207 ±.005	.488 ±.004	-	.480 ±.002
	W&H-race	.183 ±.000	.608 ±.011	.596 ±.033	-
	Multi Source	.201 ±.017	.554 ±.004	.422 ±.006	.490 ±.020
Adv-6		.235 ±.004	.549 ±.006	.436 ±.002	.474 ±.008
Adv-3		.232 ±.020	.555 ±.013	.428 ±.007	.460 ±.006
MoE-Avg		.199 ±.006	.546 ±.011	.429 ±.003	.478 ±.011
MoE-Att		.210 ±.023	.549 ±.008	.424 ±.003	.495 ±.005

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]

(ادامهٔ جدول-۳): میانگین نتایج آزمایش‌های برون‌دامنه‌ای به‌صورت تک‌منبعی، چندمنبعی و با استفاده از روش‌های آموزش دامنهٔ خصمانه و ترکیب متخصصان همراه واریانس با معیار Macro F1

(Table-3): The average results of out-of-domain tests in single-domain, multi-domain and using adversarial domain training methods and combination of experts with variance with Macro F1 measure

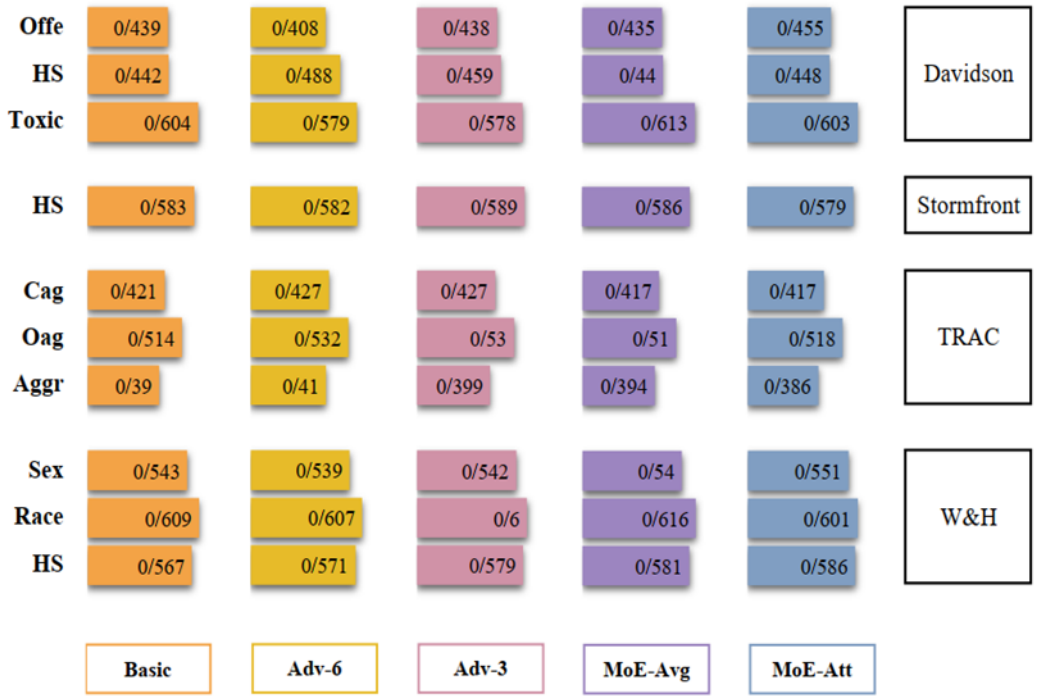
Method	Train Set	David-hs	Storm-hs	TRAC-aggr	W&H-race
Basic	David-hs	-	.598 ±.008	.354 ±.010	.710 ±.045
	Storm-hs	.549 ±.001	-	.385 ±.013	.667 ±.015
	TRAC-aggr	.202 ±.017	.502 ±.036	-	.481 ±.033
	W&H-race	.510 ±.011	.596 ±.033	.378 ±.018	-
	Multi Source	.542 ±.002	.596 ±.002	.377 ±.012	.729 ±.007
Adv-6		.538 ±.006	.608 ±.010	.373 ±.011	.740 ±.038
Adv-3		.532 ±.003	.609 ±.008	.377 ±.007	.740 ±.005
MoE-Avg		.539 ±.009	.610 ±.003	.387 ±.007	.754 ±.004
MoE-Att		.535 ±.003	.592 ±.010	.371 ±.021	.708 ±.021
Method	**Train Set**	**David-hs**	**Storm-hs**	**TRAC-oag**	**W&H-hs**
Basic	David-hs	-	.598 ±.008	.473 ±.043	.486 ±.046
	Storm-hs	.549 ±.001	-	.540 ±.011	.503 ±.020
	TRAC-oag	.321 ±.022	.634 ±.002	-	.609 ±.007
	W&H-hs	.385 ±.011	.539 ±.002	.544 ±.003	-
	Multi Source	.342 ±.008	.623 ±.003	.547 ±.012	.513 ±.004
Adv-6		.439 ±.027	.628 ±.004	.552 ±.007	.521 ±.014
Adv-3		.386 ±.019	.625 ±.011	.550 ±.009	.542 ±.022
MoE-Avg		.341 ±.009	.623 ±.009	.548 ±.007	.536 ±.007
MoE-Att		.362 ±.017	.603 ±.006	.550 ±.016	.559 ±.012

به‌منظور این‌که در بخش چند منبعی بین روش‌های مختلف مقایسه‌ای انجام دهیم، میانگین عملکرد هر دامنه در آزمایش‌های متفاوت را محاسبه کردیم. نتایج در شکل (۱) قابل مشاهده است.

هر یک از روش‌های آموزش خصمانهٔ دامنهٔ ۶، ترکیب متخصصان با روش میانگین‌گیری و ترکیب متخصصان با مدل توجه، در مقایسه با مدل پایه در نیمی از موارد باعث بهبود عملکرد شده‌اند. روش آموزش خصمانهٔ ۳ در بیش از نیمی از موارد مقداری عملکرد بهتری داشته است. همان‌طورکه در شکل (۱) نیز مشخص است، در همهٔ دامنه‌ها روشی از بین روش‌های آموزش خصمانهٔ دامنه و ترکیب متخصصان وجود داشته است که سبب افزایش قدرت تطبیق دامنه شده است.

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]



(شکل-۱): میانگین نتایج دامنه‌ها در آزمایش‌های مختلف برون‌دامنه‌ای با معیار Macro F1 در چند روش

(Figure-1): The average results of domains in different out-of-domain tests with the Macro F1 criterion in several methods

۵- نتیجه‌گیری

در این پژوهش ما به موضوع تطبیق دامنه در مسئلهٔ تشخیص کلام نفرت‌انگیز پرداخته‌ایم. در ادامه آزمایش‌های درون‌دامنه‌ای کلام نفرت‌انگیز را هم به‌صورت تک‌منبعی و هم به‌صورت چندمنبعی انجام دادیم. آزمایش‌های درون‌دامنه‌ای تک‌دامنه نتایج بهتری داشتند. در بخش چندمنبعی علاوه‌بر بخش آموزشی دامنهٔ هدف، از دامنه‌های دیگر نیز استفاده کردیم که به‌دلیل انتقال منفی، کمی نتایج کاهش یافتند؛ سپس آزمایش‌های برون‌دامنه‌ای را ابتدا به‌صورت تکمنبعی انجام دادیم تا برای هر دامنهٔ هدف، تأثیر تغییر دامنه‌های آموزشی را بررسی کنیم که نتایج ضمن کاهش، بسیار نوسان داشتند؛ سپس سعی کردیم تا با استفاده از آموزش چندمنبعی و روش‌های آموزش خصمانهٔ دامنه و ترکیب متخصصان، عملکرد تطبیق دامنه را بهبود بخشیم. مشاهده کردیم که در تعداد کمتری از آزمایش‌های مدل پایه، استفاده از چند دامنه موجب ارتقای عملکرد شد و با به‌کارگیری روش‌های آموزش خصمانهٔ دامنه و ترکیب متخصصان در ۸۷٪ از آزمایش‌ها شاهد بهبود نتایج بودیم.

با ارزیابی عملکرد مدل‌ها دریافتیم که نقش مجموعه‌داده‌ها دارای اهمیت است. بعضی دامنه‌ها دارای مجموعه واژگان خاص و داده‌هایی با مفاهیم آشکار هستند که تشخیص آن برای ماشین آسان‌تر است. برعکس این موضوع نیز صادق است و دامنه‌هایی که دارای مفاهیم ضمنی هستند، کار تشخیص را برای ماشین سخت‌تر می‌کنند. از طرف دیگر بعضی دامنه‌ها به‌دلیل ویژگی‌هایشان می‌توانند با دامنه‌های دیگر ترکیب شوند و به تطبیق دامنه کمک کنند. یکی از این ویژگی‌ها می‌تواند مشابه‌بودن داده‌ها یا توزیع‌شان باشد یا این‌که ممکن است چند دامنه با ترکیب با یکدیگر بتواند در پیش‌بینی دامنهٔ هدف مؤثرتر باشند.

توزیع داده‌های هر دامنه نیز به‌شکل ویژه در عمکرد تطبیق دامنه مؤثرند. در برخی دامنه‌ها کلاس مثبت دارای داده‌های بیشتری بود که باعث می‌شد وقتی ماشین بخواهد ازطریق دامنه‌ای که روی آن آموزش دیده است و توزیعی معکوس دارد، پیش‌بینی کند، دچار خطا شود. از طرف دیگر دامنه‌ای که نسبت تعداد داده‌های کلاس مثبت آن به کل داده‌های آن دامنه، بیشتر است، به‌الزام عامل تعیین‌کننده‌ای در نتیجه نیست.

6- References ۶- مراجع

[1] H. Almerekhi, H. Kwak, B. J. Jansen and J. Salminen, "Detecting Toxicity Triggers in Online Discussions," in *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, Hof, Germany, pp. 291-292, September 2019.

[2] A. Arango, J. Pérez and B. Poblete, "Hate Speech Detection Is Not as Easy as You May Think: A Closer Look at Model Validation," in, *Proceedings of the 42nd International*

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]

[14] P. Fortuna, J. Soler-Company and L. Wanner, "How Well Do Hate Hpeech, Toxicity, Abusive and Offensive Language Classification Models Generalize Across Datasets?," *Information Processing & Management*, vol. 58, pp. 102524, May 2021.

[15] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France, pp. 1180-1189, June 2015.

[16] A. Gretton, K. Borgwardt, M. Rasch, B. Scholkopf and A. J. Smola, "A Kernel Method for the Two-Sample-Problem," *Advances in neural information processing systems*, vol. 8, pp. 513–520, December 2007.

[17] T. Gröndahl, L. Pajola, M. Juuti, M. Conti and N. Asokan, "All You Need Is "Love": Evading Hate Speech Detection," in *AISec '18, Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, Toronto, Canada, pp. 2-12, January 2018.

[18] T. Gui, Q. Zhang, H. Huang, M. Peng and X. Huang, "Part-of-Speech Tagging for Twitter with Adversarial Neural Networks," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2411-2420, September 2017.

[19] J. Guo, D. Shah and R. Barzilay,"Multi-Source Domain Adaptation with Mixture of Experts," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 4694-4703, 2018.

[20] J. Sorensen, J. Elliott, L. Dixon, M. McDonald and W. Cukierski, "Toxic Comment Classification Challenge," *kaggle.com*, Mar. 21, 2018. [Online]. Available: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview/citation.

[21] M. Karan and J. Šnajder, "Cross-Domain Detection of Abusive Language Online," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, pp. 132-137, October 2018.

[22] Y. Kim, K. Stratos and D. Kim, "Domain Attention With an Ensemble of Experts," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 643-653, July 2017.

[23] B. Kulis, K. Saenko and T. Darrell, "What You Saw is Not What You Get: Domain Adaptation Using Asymmetric Kernel Transforms," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, Los Alamitos, USA, pp. 1785-1792, June 2011.

[24] Y. Li, T. Baldwin and T. Cohn, "What's in a Domain? Learning Domain-Robust Text Representations Using Adversarial Training," in *Proceedings of the 2018 Conference of the ACM SIGIR Conference on Research and Development in Information Retrieval*, Paris, France, pp. 45-54, July 2019.

[3] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*, Perth Australia, pp. 759–760, April 2017.

[4] S. Benesch, "Dangerous Speech: A Proposal to Prevent Group Violence," in *Voices That Poison: Dangerous Speech Project Proposal Paper*, 2013.

[5] J. Blitzer, R. McDonald and F. Pereira, "Domain Adaptation with Structural Correspondence Learning," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, pp. 120-128, July 2006.

[6] P. Burnap and M. Williams, "Us and Them: Identifying Cyber Hate on Twitter Across Multiple Protected Characteristics," *EPJ Data Science*, vol. 5, pp. 1-15, March 2016.

[7] R. Cohen-Almagor, "Fighting Hate and Bigotry on the Internet," *Policy & Internet*, vol. 3, pp. 1-26, August 2011.

[8] H. Daume III, "Frustratingly Easy Domain Adaptation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 256-263, June 2007.

[9] T. Davidson, D. Warmsley, M. Macy and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in *International AAAI Conference on Web and Social Media*, Montreal, Canada, vol. 11, pp. 512-515, May 2017.

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, USA, pp. 4171-4186, June 2019.

[11] J. Donahue, J. Hoffman, E. Rodner, K. Saenko and T. Darrell, "Semi-supervised Domain Adaptation with Instance Constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, USA, pp. 668-675, June 2013.

[12] R. Faris, A. Ashar, U. Gasser and D. Joo, "Understanding Harmful Speech Online," *Berkman Klein Center Research Publication*, vol. 21, December 2016.

[13] J. R. Finkel and C. D. Manning, "Hierarchical Bayesian Domain Adaptation," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, USA, pp. 602-610, June 2009.

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]

Smaller, Faster, Cheaper and Lighter," in *CoRR*, 2019.
[35] M. S. Jahan and M. Oussalah, "A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing," in *arXiv*, May 2021.
[36] A. Sellars, "Defining Hate Speech," *Berkman Klein Center Research Publication*, Boston Univ. School of Law, Public Law Research Paper, 2016.
[37] A. A. Siegel, "Online Hate Speech," in *Social Media and Democracy: The State of the Field, Prospects for Reform*, N. Persily and J. A. Tucker, Cambridge, UK: Cambridge University Press, pp. 56-58, 2020.
[38] B. Sun, J. Feng and K. Saenko, "Return of Frustratingly Easy Domain Adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Phoenix, USA, vol. 30, pp. 2058–2065, March 2016.
[39] S. D. Swamy, A. Jamatia and B. Gambäck, "Studying Generalisability Across Abusive Language Detection Datasets," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Hong Kong, China, pp. 940-950, November 2019.
[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in Neural Information Processing Systems*, vol. 11, pp. 5998-6008, December 2017.
[41] Z. Waseem, J. Thorne and J. Bingel, "Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection," in *Online Harassment*, pp. 29-55, July 2018.
[42] M. Wiegand, M. Siegel and J. Ruppenhofer, "Overview of the germeval 2018 shared task on the identification of offensive language," in *Proceedings of the GermEval 2018 Workshop*, September 2018.
[43] D. Wright and I. Augenstein, "Transformer Based Multi-Source Domain Adaptation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7963-7974, November 2020.
[44] T. Yao, Y. Pan, C. Ngo, H. Li and T. Mei, "Semi-supervised Domain Adaptation with Subspace Learning for Visual Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 2142–2150, June 2015.
[45] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis and Ç. Çöltekin, "SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona, Spain, pp. 1425-1447, December 2020.
North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, USA, pp. 474-479, June 2018.
[25] Z. C. Lipton, Y. Wang and A. J. Smola, "Detecting and Correcting for Label Shift with Black Box Predictors," in *Proceedings of the 35th International Conference on Machine Learning*, pp. 3128-3136, July 2018.
[26] X. Ma, P. Xu, Z. Wang, R. Nallapati. and B. Xiang, "Domain Adaptation with BERT-based Domain Classification and Data Selection," in *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP*, Hong Kong, China, pp. 76-83, November 2019.
[27] M. Mozafari, R. Farahbakhsh and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media," in *Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019*, Cham, Germany, pp. 928-940, December 2019.
[28] E. W. Pamungkas and V. Patti, "Cross-Domain and Cross-Lingual Abusive Language Detection: A Hybrid Approach with Deep Learning and a Multilingual Lexicon," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Florence, Italy, pp. 363-370, July 2019.
[29] B. Parekh, "Is There a Case for Banning Hate Speech?," in *The Content and Context of Hate Speech: Rethinking Regulation and Responses*, M. Herz and P. Molnar, Cambridge: Cambridge University Press, pp. 37–56, 2012.
[30] B. Plank, "Domain Adaptation for Parsing", Ph.D. Thesis, University of Groningen, 2011.
[31] A. Ramponi and B. Plank, "Neural Unsupervised Domain Adaptation in NLP—A Survey," in *The 28th International Conference on Computational Linguistics*, Barcelona, Spain, pp. 6838-6855, December 2020.
[32] J. Salminen, H. Almerekhi, M. Milenkovic, S. Jung, J. An, H. Kwak and B. Jansen, "Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media," in *Proceedings of the twelfth International Conference on Web and Social Media, ICWSM 2018*, Palo Alto, USA, vol. 12, pp. 330-339, June 2018.
[33] J. Salminen, M. Hopf, S. A. Chowdhury, S. g. Jung, H. Almerekhi and B. J. Jansen, "Developing an Online Hate Classifier for Multiple Social Media Platforms," *Human-Centric Computing and Information Sciences*, vol. 10, pp. 1-34, January 2020.
[34] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a Distilled Version of BERT:

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]
[DOI: 10.61186/jsdp.21.1.125]

[46] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra and R. Kumar, "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, USA, pp. 75–86, June 2019.

[47] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383-2392, October 2016,

[48] O. de Gibert, N. Pérez, A. García-Pablos and M. Cuadros, "Hate Speech Dataset from a White Supremacy Forum," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pp. 11-20, September 2018.

[49] R. Kumar, A. K. Ojha, S. Malmasi and M. Zampieri, "Benchmarking Aggression Identification in Social Media," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 1-11, August 2018.

[50] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in *Proceedings of the NAACL Student Research Workshop*, pp. 88-93, June 2016.

[51] T. Consigny, "Sesame Street Ensemble: A Mixture of DistiIBERT Experts," 2021.

[52] M. Artetxe, S. Bhosale, N. Goyal, T. Mihaylov, M. Ott, S. Shleifer, X. V. Lin, J. Du, S. Lyer, R. Pasunuru, G. Anantharaman, X. Li, S. Chen, H. Akin, M. Baines, L. Martin, X. Zhou, P. Singh Koura, B. O'Horo, J. Wang, L. Zettlemoyer, M. Diab, Z. Kozareva and V. Stoyanov, "Efficient Large Scale Language Modeling with Mixtures of Experts," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11699-11732, December 2022.

[53] A. Repetto, "Neural Networks: A Mixture of Experts with Attention," *towardsdatascience.com*, Jul. 23, 2017. [Online]. Available: https://towardsdatascience.com/neural-networks-a-mixture-of-experts-with-attention-30e196657065.

[54] R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton, "Adaptive Mixtures of Local Experts," in *Neural Computation*, vol. 3, pp. 79-87, March 1991.

[55] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," in *International Conference on Learning Representations*, November 2016.

[56] W. Fedus, B. Zoph, and N. Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," in *Journal of Machine Learning Research*, vol. 23, pp. 1-39, 2022.

[57] Z. Du, J. Li, H. Su, L. Zhu, and K. Lu, "Cross-Domain Gradient Discrepancy Minimization for Unsupervised Domain Adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3937-3946, June 2021.

[58] Q. Ye, J. Zha, and X. Ren, "Eliciting Transferability in Multi-Task Learning with Task-Level Mixture-of-Experts," in *arXiv*, 2022.

سیده فاطمه نوراللهی در سال ۱۳۹۶ مدرک کارشناسی را در رشته مهندسی کامپیوتر از دانشگاه قم دریافت کرد و در سال ۱۴۰۱ کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات در همان دانشگاه به اتمام رسانید. در حال حاضر نیز دانشجوی دکترای مهندسی فناوری اطلاعات در دانشگاه قم است. از زمینه‌های پژوهشی مورد علاقهٔ او یادگیری ماشینی است.

نشانی رایانامهٔ ایشان عبارت است از:

sfn1373@gmail.com

راضیه برادران مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه قم در سال ۱۳۸۹ دریافت کرد. مدرک کارشناسی ارشد و دکتری را نیز در رشته مهندسی فناوری اطلاعات به‌ترتیب در سال‌های ۱۳۹۱ و ۱۴۰۰ از همان دانشگاه دریافت کرد. زمینه‌های پژوهشی مورد علاقهٔ وی پردازش زبان طبیعی، یادگیری ماشینی و داده‌کاوی است.

نشانی رایانامهٔ ایشان عبارت است از:

r.baradaran@stu.qom.ac.ir

حسین امیرخانی در سال ۱۳۹۴ مدرک دکترای هوش مصنوعی را از گروه مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران) دریافت کرد. وی در حال حاضر استادیار گروه مهندسی کامپیوتر و فناوری اطلاعات دانشگاه قم است. علایق پژوهشی او شامل یادگیری ماشینی و پردازش زبان طبیعی است.

نشانی رایانامهٔ ایشان عبارت است از:

amirkhani@qom.ac.ir

[Downloaded from jsdp.rcisp.ac.ir on 2025-07-04]

[DOI: 10.61186/jsdp.21.1.125]