پردیس علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

گزارش پروژه پایانی درس بازیابی اطلاعات

نگارندگان

حسن اردشیر - زهرا خطیبی - مهیار محمدی متین

استاد درس: دکتر باباعلی

زمستان ۱۴۰۲

مقدمه

شناسایی نویسنده در ادبیات فارسی به دلیل عوامل مختلف با چالش های مهمی مواجه است. یکی از مسائل اصلی، وجود آثار بی‌نام یا با نام مستعار در ادبیات فارسی است. بسیاری از متون کلاسیک فارسی فاقد انتساب نویسندگی واضح هستند که تعیین نویسندگان واقعی را دشوار می کند. علاوه بر این، استفاده از نام‌های قلمی (تخلص) و انتساب آثار به شاعران یا دانشمندان مشهور نیز شناسایی نویسنده متون فارسی را پیچیده می‌کند. علاوه بر این، ادبیات فارسی طیف وسیعی از ژانرها، سبک‌ها و دوره‌ها را در بر می‌گیرد که این مسئله را پیچیده‌تر می‌کند. از شعر کلاسیک گرفته تا رمان‌های مدرن، هر دوره و هر ژانر ممکن است ویژگی‌های زبانی و قراردادهای سبکی متمایزی داشته باشند که می‌توان از آن‌ها برای انتساب استفاده کرد، اما نکته قابل توجه آن است که ویژگی‌های ذکر شده می‌تواند در مجموعه آثار یک نویسنده یکسان نیز بسیار متفاوت باشند.

همچنین در دسترس بودن متون دیجیتالی و ابزارهای محاسباتی برای تجزیه و تحلیل متن، منجر به تلاش برای به کارگیری روش‌های محاسباتی، مانند سبک‌سنجی و تحلیل زبان‌شناختی، برای حل مسئله شناسایی نویسنده در ادبیات فارسی شده است. با این حال، اثربخشی این روش ها را می توان با عواملی مانند در دسترس بودن و کیفیت داده های آموزشی، پیچیدگی زبان، و نیاز به تخصص در روش های محاسباتی و تحلیل ادبی محدود کرد. به طور کلی، شناسایی نویسنده در ادبیات فارسی همچنان یک مشکل چالش برانگیز میان رشته‌ای است که نیازمند تخصص در ادبیات، زبان‌شناسی و تحلیل محاسباتی است.

در این پروژه، برای ایجاد مجموعه داده، ابتدا یک مجموعه داده باید ساخته شود که شامل حداقل ۱۰ نویسنده در یک ژانر ادبی خاص در ادبیات فارسی باشد. برای هر نویسنده، حداقل ۳۰ سند

باید جمع‌آوری شود، که طول هر سند می بایست دقیقا ۵۰۰ کلمه باشد. این متون می‌توانند از تکنیک‌های وب اسکریپتینگ جمع‌آوری شوند یا از منابع دیگر مانند کتب و مقالات استفاده شود. تنوع در موضوعات و سبک‌های نوشتاری در این مجموعه داده حائز اهمیت است. همچنین، اطلاعات فراداده‌هایی مانند نام نویسنده و محتوای متن باید در هر سند ذخیره شوند.

به عنوان دومین بخش از پروژه، تمرکز بر روی مسئله شناسایی نویسنده خواهد بود. در این بخش، مدل‌های مختلفی از معماری BERT برای حل مسئله شناسایی نویسنده آموزش داده می‌شوند. از مدل‌های BERT موجود در Face Hugging استفاده می‌شود و عملکرد مدل‌های آموزش دیده با استفاده از معیارهایی مانند دقت، صحت، پوشش و F۱ Score ارزیابی می‌شود. در انتها، عملکرد مدل‌های مختلف BERT با یکدیگر مقایسه می‌شود تا بهترین مدل برای شناسایی نویسنده در ادبیات فارسی شناخته شود.

فهرست مطالب

فصل ۱

تکنیک های ساخت مجموعه داده

در فرآیند ساخت مجموعه داده برای این پروژه، ژانر عاشقانه به عنوان یک ژانر محبوب که بخش زیادی از آثار ادبی در زبان فارسی را شامل می شود، انتخاب شده است. این انتخاب به علت تنوع و گستردگی موضوعات و سبک‌های مختلف نوشتاری در این ژانر صورت گرفت. مطابق با خواسته صورت مسئله، ۱۰ نویسنده مختلف از میان نویسندگان معروف در این ژانر ادبی انتخاب شدند. هر یک از این نویسندگان بر اساس شهرت و شناخته شدگی آثارشان و همچنین تنوع در سبک نگارش و موضوعات مطرح شده در آثار، انتخاب شده‌اند. این نویسندگان عبارتند از:

۱. فاطمه امیری
۲. نازنین موسوی
۳. زکیه اکبری
۴. مژگان زارع
۵. ساینا مقدسی
۶. نیلوفر شقاقی

۷. محرابه سادات قدیری

۸. خورشید روزبهی

۹. تورج هاشمی

۱۰. محمدعلی قجه

برای جمع‌آوری متون، از منابع مختلفی استفاده شد. این منابع شامل کتب و رمان‌های عاشقانه فارسی زبان و همچنین برخی ترجمه‌ها و تفسیرهای این متون به زبان فارسی بودند. لازم به ذکر است در این قسمت، داده نرمال‌سازی نیز شده‌اند که در قسمت‌های بعدی کار با این دادگان راحت‌تر باشد. در این فرآیند، چندین چالش مواجه شدیم که نیازمند راه‌حل‌های خاصی بودند. این چالش‌ها عبارتند از:

دسترسی به متون مناسب

برخی از متون در فرمت‌هایی مانند PDF بودند که امکان کپی کردن متن فراهم نبود. برای حل این چالش، از OCR استفاده شده است. این تکنولوژی به ما اجازه می‌دهد تا متن‌هایی که قابلیت کپی کردن ندارند را به متن قابل ویرایش و کپی تبدیل کنیم.

مشکلات نیم فاصله

برخی از کلمات فارسی از چند قسمت تشکیل شده‌اند. این کلمات معمولاً با یک فاصله نیم‌فاصله از هم جدا می‌شوند. اما در برخی از کتب و آثار، این نکته به درستی رعایت نمی‌شود. به عنوان مثال، کلمه «آن‌ها» به طور نادرست به شکل «آن ها» نوشته می‌شود. بنابراین این موارد نیاز به اصلاح داشتند.

تنوع فرمت‌های زبانی

یکی دیگر از عواملی که ممکن است منجر به مشکل در کدگذاری کاراکترها شود، تفاوت در کدگذاری حروف در استاندارد یونیکد است. به عنوان مثال، حرف «ی» را می‌توان به صورت‌های مختلف «ی» ، «ي» و «ئ» نوشت که دارای کدگذاری متفاوتی در استاندارد یونیکد هستند. برای حل این مشکل تمام حروف چند فرمته را به یک فرم یکسان می‌بریم. به این ترتیب، تفاوت در کدگذاری کاراکترها برطرف شده و مشکلات مربوط به نمایش کاراکترها در سیستم‌هایی که از استاندارد یونیکد پشتیبانی نمی‌کنند، رفع می‌شود.

حذف عناصر زائد

بعضی از متون شامل تیترها یا جملات تبلیغاتی و اضافی بودند که نیاز به حذف داشتند. در برخی موارد، به منظور تاکید بیشتر بر احساسات یا بیان شدت بیشتری در سخن، نویسندگان از تکرار حروف در کلمات استفاده می‌کنند. به عنوان مثال، به جای نوشتن «نه» ، از «نهههههه» استفاده می‌شود. با این حال، تکرار حروف در کلمات، تاثیری در نتیجه تحلیل احساسات ندارد و فقط برای تاکید بیشتر در بیان استفاده می‌شود. به همین دلیل، با استفاده از ابزار هضم، می‌توان حروف اضافه کلمات را حذف کرده و به شکل استاندارد آن‌ها را نمایش داد. این کار می‌تواند به بهبود دقت مدل کمک کند.

اصلاح کلمات محاوره‌ای

در تحلیل دادگان متنی به زبان فارسی، یکی از مشکلات اساسی و مهم، وجود کلمات محاوره‌ای و غیررسمی در متون است. تفاوت در گفتار و نوشتار بعضی از کلمات، باعث می‌شود برخی نویسندگان مطالب خود را به صورت محاوره‌ای بنویسند. این مشکل باعث می‌شود که تحلیل دادگان به صورت درست و دقیق صورت نگیرد و نتایج نادرستی به دست آید. برای حل این مشکل، از ابزار هضم استفاده شده است. ابزار هضم توانایی تبدیل کلمات محاوره‌ای به کلمات رسمی و استاندارد را

داراست. این ابزار با تحلیل کلمات و متون، قادر به شناسایی کلمات محاوره‌ای و غیررسمی است و با استفاده از لغت‌نامه‌های معتبر، آن‌ها را به کلمات رسمی و استاندارد تبدیل می‌کند. به این ترتیب، امکان تحلیل دقیق‌تر در دادگان متنی به زبان فارسی، فراهم می‌شود.

حذف کلمات توقف

کلمات توقف، کلماتی هستند که در فرآیند پردازش متن و تحلیل آنها معمولاً از آن‌ها چشم‌پوشی می‌شود زیرا معمولاً به تنهایی اطلاعات معنایی مهمی ارائه نمی‌دهند و درک موضوع یا مفهوم جمله را به طور معناداری تغییر نمی‌دهند. این کلمات معمولاً در زبان‌های طبیعی مانند انگلیسی، فارسی، و غیره وجود دارند و مثال‌هایی از آن‌ها عبارتند از "به"، "و"، "هم"، "از"، "باشد" و غیره.
حذف کلمات توقف یک مرحله مهم در پردازش متن است که به منظور بهبود عملکرد الگوریتم‌های پردازش متن و استخراج اطلاعات مفید از متن‌ها انجام می‌شود. این کار به کاهش ابعاد داده، افزایش سرعت پردازش و بهبود کیفیت نتایج کمک می‌کند.
برای حذف کلمات توقف از دو رویکرد استفاده شده است. در رویکرد اول تنها حروف ربط حذف شده‌اند که تنها ۹۳ کلمه توقف اصلی را شامل می‌شوند. در رویکرد دوم، از لیستی از کلمات توقف استفاده شده است. این لیست شامل کلماتی با بار معنایی خنثی است که باید از متن حذف شوند و ۱۴۷۲ کلمه را شامل می‌شود. با حذف کلمات توقف متن کاهش می‌یابد و تمرکز بر کلمات اصلی و مفید متن افزایش می‌یابد، که باعث بهبود کارایی و دقت الگوریتم‌های پردازش متن می‌شود.

فصل ۲

انتخاب مدل و تنظیم دقیق

۱.۲ انتخاب مدل

مدل HooshvareLab/bert-fa-base-uncased برای مسئله شناسایی نویسنده در آثار فارسی، به دلیل چندین عامل اساسی انتخاب شده است. اولین عامل این است که این مدل بر روی مجموعه‌ای گسترده از متون فارسی پیش‌آموزش دیده است. بنابراین با درک عمیقی از ساختار و مفاهیم زبان فارسی، مناسب برای مسائل پردازش زبان طبیعی در این زبان است. همچنین، این مدل از قابلیت تنظیم دقیق برای اجرای وظایف خاصی مانند شناسائی نویسنده، بهره‌مند است. مدل‌های از پیش آموزش دیده مانند ParsBERT با انجام یک فاز از یادگیری بدون نظارت بر روی متون بزرگ، قادر به درک الگوهای پیچیده و جزئیات زبانی زبان فارسی می‌شوند که این امکان را فراهم می‌کند تا در مسائلی مانند شناسایی نویسنده، عملکرد بسیار خوبی داشته باشند.

علاوه بر این، مدل‌های از پیش آموزش دیده از نظر کارایی بهینه هستند. این مدل‌ها از طریق پیش‌آموزش، یک درک کلی از زبان فارسی را پیدا کرده‌اند که به طور قابل توجهی میزان داده و زمان مورد نیاز برای آموزش مخصوص وظیفه را کاهش می‌دهد. به همین دلیل، این مدل‌ها انتخاب مناسبی برای کاربردهای عملی می‌باشند که ممکن است دسترسی محدودی به منابع مانند داده‌های برچسب‌خورده و قدرت محاسباتی داشته باشند. علاوه بر این، امکان تنظیم دقیق به ما این اجازه

را می‌دهد که مدل‌های از پیش آموزش دیده را به طور موثر برای حل مسائل خاصی مانند شناسایی نویسنده تطبیق دهیم و با استفاده از داده‌های برچسب‌خورده، آن‌ها را آموزش دهیم. این فرآیند باعث بهبود دقت و عملکرد مدل در حل مسائل مشخص می‌شود و تأثیر مثبتی بر روی نتایج نهایی دارد.
به طور خلاصه، استفاده از مدل‌های از پیش آموزش دیده مانند ParsBERT برای شناسایی نویسنده در آثار فارسی، از پیش‌آموزش گسترده بر روی متون فارسی بهره می‌برد و آن‌ها را با درک عمیقی از زبان و ویژگی‌های آن مجهز می‌کند. همچنین، تنظیم دقیق به ما این امکان را می‌دهد که مدل‌ها را به طور موثر برای خواسته مسئله تطبیق دهیم و عملکرد آن‌ها را با داده‌های برچسب‌خورده بهبود بخشیم. این ترکیب از پیش‌آموزش و تنظیم دقیق، به عنوان یک رویکرد قوی برای ایجاد مدل‌های دقیق و کارآمد برای وظایف پردازش زبان طبیعی در فارسی و سایر زبان‌ها مورد استفاده قرار می‌گیرد.

۲.۲ تنظیم دقیق، اصلاحات ، معماری و پارامترهای مدل

با توجه به اینکه مسئله شامل طبقه بندی نویسندگان در مجموع ۱۰ کلاس است، یک لایه متراکم با ۱۰ گره به معماری مدل اضافه شد تا به عنوان لایه خروجی عمل کند. این موضوع آن را قادر می سازد تا احتمالات هر کلاس را پیشبینی کند. این اصلاح ساختار، خروجی مدل را با ماهیت طبقه‌بندی چند کلاسه کار هماهنگ می‌کند.
علاوه بر این، با توجه به الزامات ورودی مدل BERT که از پیش آموزش‌دیده است و به اسنادی با طول ثابت نیاز دارد، حداکثر طول سند را روی ۵۰۰ توکن تنظیم کردیم. برای گنجاندن اسناد با طول‌های مختلف، اسناد طولانی‌تر را کوتاه کردیم تا با این محدودیت طول مطابقت داشته باشد و در عین حال محتوای ضروری سند را بتوانیم حفظ کنیم. این موضوع، یکنواختی در ابعاد ورودی را در تمام اسناد تضمین می‌کند و پردازش کارآمد توسط مدل تسهیل می‌شود. شایان ذکر است که آزمایش با طول‌های مختلف سند برای بهینه سازی عملکرد انجام شد و طول ۵۰۰ توکن نتایج رضایت بخشی را به همراه داشت. از این رو طول ۵۰۰ توکن برای ورودی مدل انتخاب شد.
علاوه بر این، با توجه به ماهیت از پیش آموزش‌دیده‌شده مدل BERT که در آن پارامترها از قبل در طول دوره‌های پیش‌آموزشی آموخته شده‌اند و به‌طور پیش‌فرض قابل آموزش نیستند، فرآیند تنظیم

دقیق در درجه اول بر روی آموزش لایه متراکم اضافه‌شده و حفظ لایه‌های از پیش‌آموزش‌دیده متمرکز بود. این موضوع به طور قابل توجهی تعداد پارامترهای قابل آموزش در مدل را کاهش داد و به سرعت و کارایی آن در طول آموزش کمک کرد. با استفاده از نمایش‌های از پیش آموزش‌دیده‌شده توسط BERT و تنظیم دقیق تنها لایه خروجی، مدل می‌تواند به طور موثر اطلاعات زمینه‌ای خاص برای مسئله طبقه‌بندی نویسنده را گرفته و از آن استفاده کند و منجر به پیش‌بینی‌های دقیق شود.

به طور خلاصه، فرآیند تنظیم دقیق شامل افزودن یک لایه خروجی متراکم، تنظیم طول سند ثابت و آموزش تنها لایه جدید اضافه شده در حالی که لایه‌های BERT از پیش‌آموزش دیده را ثابت نگه می‌دارد. این اصلاحات، مدل را برای مسئله شناسایی نویسنده بهینه‌سازی کرده و از سرعت و دقت در طبقه‌بندی اطمینان حاصل شده و در عین حال از درک متنی غنی ارائه شده توسط مدل BERT از پیش آموزش‌دیده استفاده کرده است.

در فرآیند تنظیم دقیق، چندین جنبه در مورد بهینه‌سازی و پارامترهای آموزشی برای بهینه‌سازی عملکرد مدل مورد بررسی قرار گرفت. بهینه ساز Adam یک انتخاب رایج و محبوب برای حل مسائل یادگیری عمیق به دلیل متدولوژی نرخ یادگیری تطبیقی، برای به روزرسانی پارامترهای مدل در طول آموزش استفاده شد. Adam به صورت پویا نرخ‌های یادگیری را برای هر پارامتر بر اساس گرادیان های گذشته تنظیم می‌کند و امکان همگرایی کارآمد و تعمیم بهتر را فراهم می‌کند.

علاوه بر این، نرخ‌های یادگیری مختلف برای یافتن نرخ بهینه برای کار خاص و معماری مدل مورد آزمایش قرار گرفتند. آزمایش با نرخ‌های مختلف یادگیری به یافتن تعادل بین سرعت یادگیری و همگرایی به راه حل بهینه کمک می‌کند. با تنظیم دقیق نرخ یادگیری، هدف ما افزایش کارایی و ثبات فرآیند آموزش و در نهایت بهبود دقت مدل بود.

همچنین، انتخاب تابع ضرر در هدایت فرآیند یادگیری مدل بسیار مهم است. برای حل مسئله شناسایی نویسنده، Sparse Categorical Crossentropy به عنوان تابع ضرر استفاده شد.

Sparse Categorical Crossentropy برای کارهای طبقه‌بندی چند طبقه‌ای مناسب است، هنگامی که هر نمونه دقیقاً به یک کلاس تعلق دارد. این اختلافات آنتروپی متقابل بین برچسب‌های واقعی و احتمالات پیش‌بینی‌شده را محاسبه می‌کند و مدل را بر اساس اختلاف بین توزیع‌های کلاس پیش‌بینی‌شده و واقعی جریمه می‌کند.

در این راستا اندازه دسته، که تعداد نمونه‌های پردازش شده در هر تکرار آموزشی را تعیین می‌کند، برای ارزیابی تأثیر آن بر پویایی آموزش و عملکرد مدل، متفاوت است. اندازه‌های دسته‌ای مختلف برای ایجاد تعادل بین کارایی محاسباتی و هم‌گرایی مدل مورد آزمایش قرار گرفتند. تنظیم اندازه دسته به بهینه‌سازی استفاده از حافظه و سرعت حل مسئله کمک می‌کند و در عین حال از آموزش مدل پایدار و موثر اطمینان می‌دهد.

به طور خلاصه، فرآیند تنظیم دقیق شامل آزمایش با بهینه‌ساز Adam، تنظیم نرخ‌های یادگیری، استفاده از Sparse Categorical Crossentropy به‌عنوان تابع ضرر و آزمایش اندازه‌های مختلف دسته بود. هدف این کاوش‌ها بهینه‌سازی پویایی آموزش، افزایش همگرایی و بهبود دقت کلی مدل برای مسئله شناسایی و طبقه‌بندی نویسنده بود.

فصل ۳

نتایج مدل

۱.۳ نتایج مدل با استفاده از 5-fold cross-validation

در این پروژه، برای شناسایی نویسنده از مدل‌های مختلف با پارامترهای متفاوت استفاده کرده‌ایم. برای ارزیابی عملکرد این مدل‌ها، از روش cross-validation با ۵ فولد استفاده کرده‌ایم. در این روش، داده‌ها به ۵ بخش مساوی تقسیم شده و هر بار یکی از این بخش‌ها به عنوان مجموعه آزمون و بقیه به عنوان مجموعه آموزش استفاده می‌شوند. مدل با استفاده از مجموعه آموزش، آموزش داده می‌شود و سپس بر روی مجموعه آزمون ارزیابی می‌شود. این فرآیند به طور متوالی برای ۵ بار انجام می‌شود و نتایج به دست آمده از هر بار ارزیابی متوسط‌گیری می‌شود تا نتیجه نهایی به دست آید. نتایج به‌دست آمده در جدول زیر آورده شده است. حال به شرح جزییات هر مدل می‌پردازیم:

جدول ۱.۳: نتایج cross-validation با ۵ فولد

مدل	دقت تست	دقت فولد ۲	دقت فولد ۳	دقت فولد ۴	دقت فولد ۵	دقت فولد ۱	میانگین دقت
مدل ۱	۸۳.۷	۸۷.۵	۸۸.۵	۸۵.۴	۹۲.۷	۸۴.۳	۸۷.۷
مدل ۲	۹۱.۲	۸۸.۵	۸۱.۲	۸۸.۵	۷۱.۸	۸۶.۴	۸۳.۳
مدل ۳	۸۹.۹	۶۹.۷	۵۴.۱	۵۴.۱	۷۵	۵۵.۲	۶۱.۶
مدل ۴	۸۳.۷	۸۲.۲	۸۵.۴	۸۵.۴	۸۴.۳	۸۵.۴	۸۴.۵
مدل ۵	۸۰	۸۸.۷	۸۱.۴	۸۸.۷	۷۰.۱	۸۶.۱	۸۳

جدول ۲.۳: نتایج cross-validation با ۵ فولد

مدل	اسناد	اسم مدل	نرخ یادگیری آغازین	نرخ یادگیری نهایی
مدل ۱	no-stop-word۲	bert-fa-base-uncased	۰.۱	۰.۰۰۱
مدل ۲	no-stop-word۱	bert-fa-base-uncased	۰.۰۱	۰.۰۰۰۱
مدل ۳	no-stop-word۲	bert-fa-base-uncased	۰.۰۰۱	-
مدل ۴	no-stop-word۲	bert-fa-base-uncased	۰.۱	۰.۰۰۰۱
مدل ۵	no-stop-word۲	bert-fa-base-uncased	۰.۱	۰.۰۰۱

در مدل شماره ۲، از اسنادی استفاده شده است که تنها ۹۳ کلمه توقف به صورت دستی از دادگان را حذف کرده بودیم. در مابقی مدل‌ها، از اسنادی استفاده شده است که به کمک لیستی ۱۴۷۲ـتایی از کلمات توقف، دادگان را بهینه‌تر کرده‌ایم.

در مدل سوم، need-schedualer را برابر false و در مدل پنجم، use-dropout را برابر true قرار داده‌ایم.

با توجه به نتایج، مدل اول بهترین نتیجه را برای ما به همراه داشته است.

۲.۳ معیارهای ارزیابی عملکرد

	precision	recall	f1-score	support
0	0.78	1.00	0.88	7
1	0.83	1.00	0.91	5
2	1.00	0.88	0.93	8
3	1.00	0.64	0.78	11
4	0.80	0.73	0.76	11
5	1.00	0.33	0.50	6
6	0.89	1.00	0.94	8
7	1.00	0.80	0.89	5
8	0.73	1.00	0.85	11
9	0.73	1.00	0.84	8
accuracy			0.84	80
macro avg	0.88	0.84	0.83	80
weighted avg	0.87	0.84	0.83	80

شکل ۱.۳: نتایج مدل اول

	precision	recall	f1-score	support
0	1.00	1.00	1.00	9
1	1.00	1.00	1.00	6
2	0.89	0.80	0.84	10
3	1.00	1.00	1.00	7
4	1.00	0.86	0.92	7
5	0.92	0.92	0.92	12
6	1.00	0.67	0.80	6
7	1.00	0.88	0.93	8
8	1.00	1.00	1.00	6
9	0.64	1.00	0.78	9
accuracy			0.91	80
macro avg	0.94	0.91	0.92	80
weighted avg	0.93	0.91	0.91	80

شکل ۲.۳: نتایج مدل دوم

	precision	recall	f1-score	support
0	0.86	0.86	0.86	7
1	1.00	1.00	1.00	6
2	1.00	0.70	0.82	10
3	0.91	1.00	0.95	10
4	0.82	1.00	0.90	9
5	0.80	1.00	0.89	8
6	0.88	0.78	0.82	9
7	1.00	0.83	0.91	6
8	1.00	1.00	1.00	6
9	0.89	0.89	0.89	9
accuracy			0.90	80
macro avg	0.91	0.91	0.90	80
weighted avg	0.91	0.90	0.90	80

شکل ۳.۳: نتایج مدل سوم

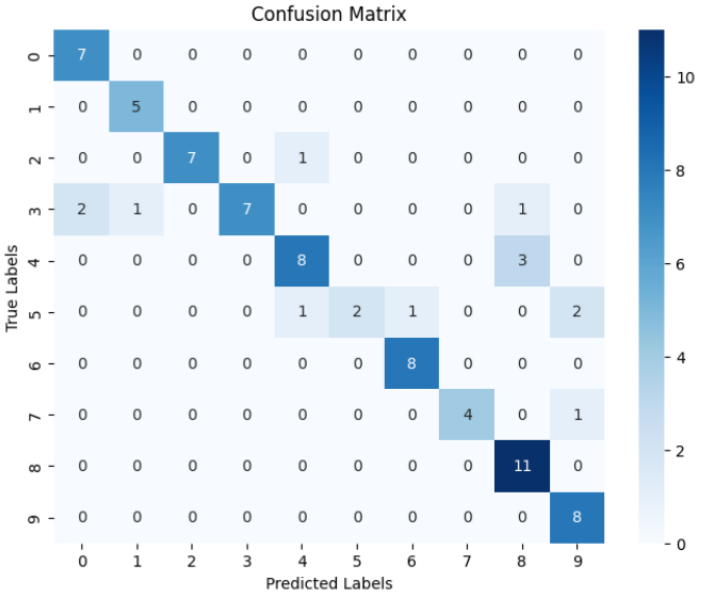
	precision	recall	f1-score	support
0	0.75	1.00	0.86	6
1	0.64	1.00	0.78	9
2	0.54	1.00	0.70	7
3	1.00	0.10	0.18	10
4	1.00	0.82	0.90	11
5	1.00	1.00	1.00	11
6	1.00	1.00	1.00	5
7	1.00	0.67	0.80	3
8	1.00	1.00	1.00	9
9	1.00	0.89	0.94	9
accuracy			0.84	80
macro avg	0.89	0.85	0.82	80
weighted avg	0.90	0.84	0.81	80

شکل ۴.۳: نتایج مدل چهارم

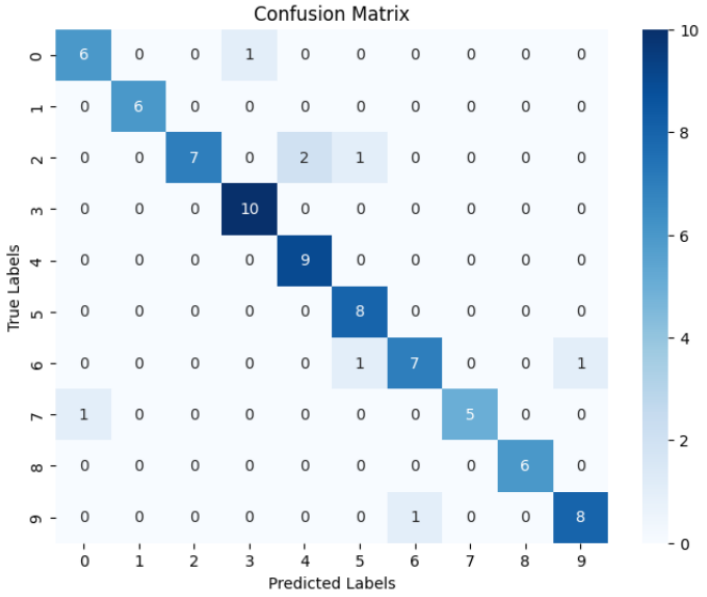
	precision	recall	f1-score	support
0	1.00	0.86	0.92	7
1	1.00	1.00	1.00	10
2	0.75	1.00	0.86	6
3	1.00	0.58	0.74	12
4	1.00	0.83	0.91	6
5	0.67	0.67	0.67	9
6	0.73	0.67	0.70	12
7	0.83	1.00	0.91	5
8	0.88	0.88	0.88	8
9	0.40	0.80	0.53	5
accuracy			0.80	80
macro avg	0.83	0.83	0.81	80
weighted avg	0.84	0.80	0.81	80

شکل ۵.۳: نتایج مدل پنجم

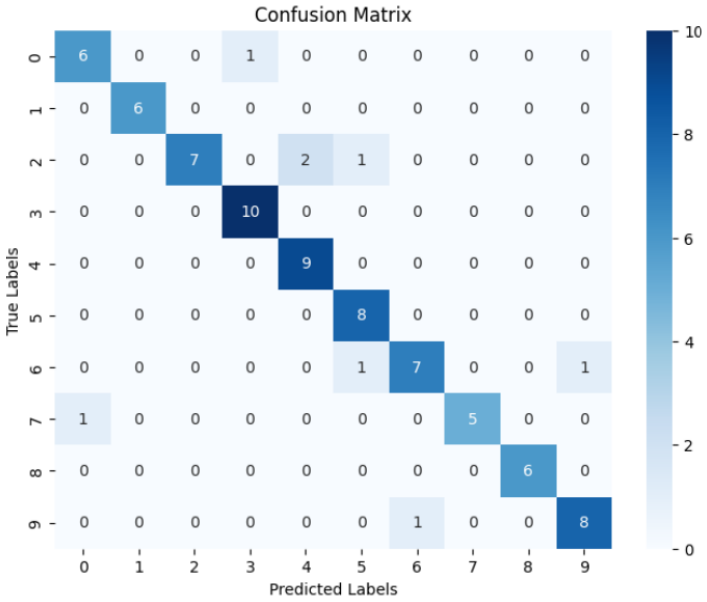
۳.۳ ماتریس درهم ریختگی



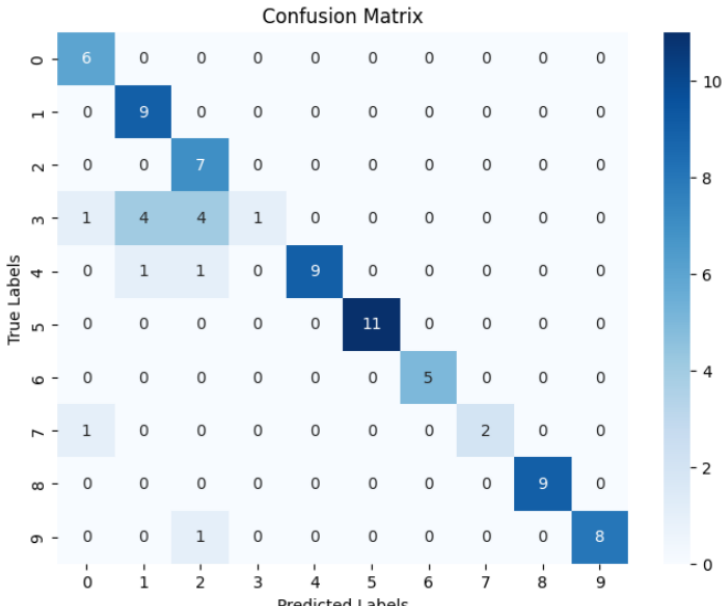
شکل ۶.۳: ماتریس درهم ریختگی مدل اول



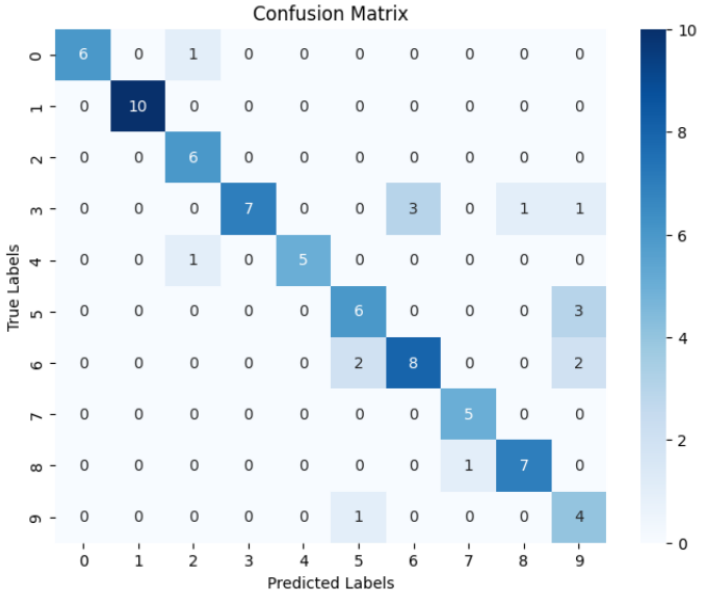
شکل ۷.۳: ماتریس درهم ریختگی مدل دوم



شکل ۸.۳: ماتریس درهم ریختگی مدل سوم



شکل ۹.۳: ماتریس درهم ریختگی مدل چهارم



شکل ۱۰.۳: ماتریس درهم ریختگی مدل پنجم

۴.۳ بررسی تاثیر نرخ یادگیری

بر اساس آزمایش‌های قبلی، مشخص شد که آموزش و ارزیابی مدل با استفاده از تنظیمات مشخص شده ممکن است زمان‌بر باشد. برای پرداختن به این مسئله و افزایش سرعت در روند آموزش، تعیین نرخ بهینه یادگیری ضروری بود. با استفاده از یک زمان‌بندی نرخ یادگیری، نرخ یادگیری به صورت پویا در طول آموزش تنظیم می‌شود، ابتدا با مقدار نسبتاً بالا شروع شده و سپس به تدریج کاهش می‌یابد. هدف این رویکرد ایجاد تعادل بین یادگیری اولیه سریع و تنظیم دقیق پارامترهای مدل است. به عبارت دیگر، اتخاذ این استراتژی منجر به همگرایی سریع‌تر و دستیابی به دقت مطلوب در تعداد قابل توجهی از دوره‌ها شده است. این رویکرد باعث ساده‌تر شدن فرآیند آموزش و افزایش کارایی مدل می‌شود.

با به‌روزرسانی‌های مدل، بهبود در دقت آموزش و اعتبارسنجی مشاهده شده است. این به‌روزرسانی‌ها نشان می‌دهند که استفاده از نرخ‌های یادگیری اولیه و پایانی کمتر به طور قابل‌ملاحظه‌ای عملکرد مدل را افزایش داده است. این یافته‌ها نشان می‌دهد که تنظیم مناسب نرخ‌های یادگیری می‌تواند بهبود قابل توجهی در عملکرد مدل داشته باشد و در نتیجه، توصیه می‌شود از این مقادیر برای بهبود عملکرد و دقت در فرآیند آموزش مدل استفاده گردد. توجه داشته باشیم که با افزایش نرخ یادگیری، ممکن است سرعت آموزش مدل افزایش یابد، به عبارت دیگر، مدل ممکن است سریع‌تر به یادگیری الگوهای داده‌های آموزشی بپردازد. اما اگر نرخ یادگیری بسیار بالا باشد، ممکن است مدل به سرعت به نقاط اشتباه همگرا شود و عملکرد آن بهبود نیابد یا حتی بدتر شود. نرخ یادگیری بالا ممکن است منجر به overfitting شود، زیرا مدل به سرعت الگوهای دقیق داده‌های آموزشی را یاد می‌گیرد و به اندازه کافی تعمیم پذیر نیست.

۵.۳ بررسی تاثیر حذف کلمات توقف

تاثیرات مثبت حذف کلمات پرتکرار در داده‌های متنی عبارت‌اند از:

کاهش نویز

حذف کلمات پرتکرار می‌تواند به کاهش نویز در داده‌ها منجر شود. این کلمات معمولاً اطلاعات کمی ارائه می‌دهند و اغلب در تشخیص الگوهای معنایی مهم کمک نمی‌کنند. با حذف این کلمات، داده‌های تمیزتر و با توجهی کمتر به جزئیات غیرضروری خواهیم داشت که می‌تواند به بهبود کیفیت و دقت مدل منجر شود.

افزایش سرعت آموزش

با کاهش تعداد کلمات در داده‌ها، زمان مورد نیاز برای آموزش مدل کاهش می‌یابد. این امر می‌تواند به زمان‌بندی موثرتر و سریع‌تری برای آموزش مدل منجر شود و امکان آموزش مدل‌های بزرگ‌تر یا پیچیده‌تر را فراهم کند.

کاهش اندازه مدل

حذف کلمات پرتکرار باعث کاهش تعداد واژگان استفاده شده در مدل می‌شود. این می‌تواند منجر به کاهش حجم مدل و اشغال حافظه شود، که می‌تواند در کاربردهایی که محدودیت‌هایی بر روی حافظه و منابع محاسباتی وجود دارد، مفید باشد.

افزایش توجه به مفاهیم مهم

با حذف کلمات پرتکرار، مدل ممکن است به مفاهیم مهم‌تر و اطلاعات با ارزش توجه بیشتری بپردازد. این می‌تواند به تقویت نیروی یادگیری مدل و افزایش دقت در تشخیص و تفسیر الگوهای معنایی کمک کند.

ساده شدن تحلیل داده‌ها

با کاهش تعداد کلمات و حجم داده‌ها، فرآیند تحلیل داده‌ها و استخراج اطلاعات می‌تواند ساده‌تر و موثرتر شود. این امر می‌تواند به تسهیل فرآیندهای تصمیم‌گیری و اجرای استراتژی‌های کسب و کار کمک کند.

برای ارزیابی تأثیر حذف کلمات توقف بر عملکرد مدل، ما از ۳ نوع داده برای آموزش مدل استفاده کرده‌ایم. دسته اول داده‌های ما شامل داده‌های خام بود که هیچ تغییری در آن‌ها صورت نگرفته است. در دسته دوم، کلمات توقفی که از معانی مهم و پرتکراری برخوردارند (کلماتی مانند "و"، "در"، "با" و غیره) از داده حذف شده‌اند. در نهایت، در دسته سوم، یک لیست گسترده از کلمات توقف، که شامل ۱۴۷۲ کلمه بود، از داده‌های اصلی حذف شده است.

تحلیل نتایج حاکی از آن است که دقت مدل در هنگام آموزش با داده‌های دسته دوم، بالاتر از دو دسته دیگر است. این نتیجه نشان می‌دهد که حذف کلمات توقفی مهم و پرتکرار از داده‌ها می‌تواند بهبود قابل توجهی در عملکرد مدل داشته باشد. این تأثیر مثبت احتمالاً به دلیل این است که حذف این کلمات باعث کاهش نویز و تمرکز بیشتر مدل بر کلمات کلیدی و معنی‌دار در متن می‌شود، که در نهایت به دقت بیشتر و عملکرد بهتر مدل منجر می‌شود. همچنین می‌توان نتیجه گرفت که با حذف تعداد زیادی از کلمات توقف، دقت مدل کاهش می‌یابد. این امر ممکن است به دلیل حذف کلمات معنی‌دار و اساسی از داده باشد که می‌تواند به از دست رفتن اطلاعات مهم و کلیدی در متن منجر شود. بنابراین، باید توازنی مناسب بین حذف کلمات توقف و حفظ کلمات اساسی و مهم در متن داشته باشیم تا دقت مدل را حفظ کنیم و بهبودهای معنادار در عملکرد مدل داشته باشیم.

۶.۳ بررسی تاثیر طول اسناد

تحلیل اینکه چگونه طول سند می‌تواند بر عملکرد مدل تأثیر بگذارد، امری حیاتی است که درک درستی از رفتار مدل در برابر طول متون مختلف را فراهم می‌کند. ممکن است سندهای بلند چالش‌هایی مانند بار اطلاعات زیاد، پیچیدگی محاسباتی افزایش یافته و از دست رفتن محتوا به

دلیل بلندی متن را به همراه داشته باشند، در حالی که سندهای کوتاه ممکن است اطلاعات کافی برای مدل برای انجام پیش‌بینی‌های دقیق را فراهم نکنند.

در آزمایش‌های ما مشاهده کردیم که با افزایش طول سند، عملکرد مدل در ابتدا بهبود می‌یابد، که نشان می‌دهد متون بلند، اطلاعات بیشتری را برای مدل برای یادگیری فراهم می‌کنند. با این حال، بیش از یک حد معین، عملکرد شروع به کاهش می‌یابد، احتمالاً به دلیل دشواری مدل در پردازش متن‌های بیش از حد بلند یا مواجه شدن با اطلاعات غیرمرتبط این نتیجه حاصل می‌شود.

برای کاهش تأثیر طول سند بر عملکرد مدل، می‌توان از تکنیک‌هایی مانند خلاصه‌سازی سند، قطع کردن متن و یا مکانیسم‌های توجه پویا استفاده کرد. این رویکردها به دنبال یافتن تعادلی مناسب بین گرفتن اطلاعات اساسی از سندهای بلند در حالی که از دست رفتن جزئیات غیرضروری مانند جزئیات هستند. به علاوه، آزمایش با معماری‌ها و پارامترهای مدل مختلفی که به طور خاص برای کنترل طول متون مورد تنظیم قرار می‌گیرند، می‌تواند عملکرد را در سطوح مختلف طول متون بهینه کند.

در این پروژه، ما طول اسناد را به صورت ثابت و برابر با ۵۰۰ در نظر گرفته‌ایم. این انتخاب طول ثابت برای اسناد به دلایلی همچون اصلاح پیچیدگی مدل و استانداردسازی فرآیند آموزش مدل انجام شده است. با این حال، در جریان آزمایشات، یکبار مدل را با استفاده از اسنادی به طول ۴۰۰ آموزش دادیم و مشاهده کردیم که عملکرد مدل با این اسناد کمتر از حالتی بود که از اسناد ۵۰۰ کلمه‌ای استفاده می‌شد. این نتایج نشان می‌دهد که انتخاب طول ثابت برای اسناد، باعث حفظ اطلاعات مهم و اصلی در اسناد و افزایش دقت مدل می‌شود. از طرف دیگر، کاهش طول اسناد ممکن است باعث از دست رفتن اطلاعات مهم و موجود در اسناد شود، که این موضوع می‌تواند منجر به کاهش دقت مدل شود. بنابراین، انتخاب طول مناسب برای اسناد در طراحی مدل و انجام آزمایشات بسیار حیاتی است و باید با دقت و توجه به نیازهای ویژه هر پروژه صورت گیرد.

فصل ۴

مقایسه مدل با رویکردهای سنتی ML

در مقایسه مدل ما با رویکردهای یادگیری ماشینی سنتی (ML)، عوامل متعددی مطرح می‌شوند که در ادامه به آن‌ها می‌پردازیم.

عملکرد

مدل‌های یادگیری عمیق مانند مدل‌های مبتنی بر BERT اغلب در انجام وظایف پردازش زبان طبیعی مانند شناسایی نویسنده نسبت به الگوریتم‌های سنتی یادگیری ماشین عملکرد بهتری دارند. این به این دلیل است که مدل‌های یادگیری عمیق قادرند الگوها و ارتباطات پیچیده در داده‌های متنی را به دقت کشف کنند که برای الگوریتم‌های سنتی ممکن است سخت باشد.

مهندسی ویژگی

الگوریتم‌های سنتی یادگیری ماشین معمولاً نیاز به مهندسی دستی ویژگی دارند که در آن دانش حوزه برای استخراج ویژگی‌های مرتبط از داده استفاده می‌شود. در مقابل، مدل‌های یادگیری عمیق به طور خودکار نمایش‌های سلسله مراتبی از داده را یاد می‌گیرند، که نیازی به مهندسی ویژگی گسترده ندارند. این می‌تواند مزیتی مهمی باشد، به ویژه در وظایف پردازش زبان طبیعی که مهندسی ویژگی

می‌تواند زمان‌بر و خطاگیر باشد.

اندازه داده

مدل‌های یادگیری عمیق اغلب نیاز به مقادیر زیادی از داده برای عملکرد مناسب دارند، در حالی که الگوریتم‌های سنتی یادگیری ماشین ممکن است برای وظایف با مجموعه داده‌های کوچکتر مناسب‌تر باشند. با این حال، با وجود وجود مدل‌های پیش‌آموزش دیده مانند BERT که بر روی مجموعه‌های بزرگی از داده متنی آموزش داده شده‌اند، این نیاز کمتری برای مدل‌های یادگیری عمیق وجود دارد.

تفسیر پذیری

الگوریتم‌های سنتی یادگیری ماشین معمولاً قابل تفسیرتر از مدل‌های یادگیری عمیق هستند. این به این معناست که اغلب آسان‌تر است که تصمیماتی که توسط الگوریتم‌های سنتی اتخاذ می‌شود را درک و تفسیر کرد که می‌تواند در برخی برنامه‌هایی که تفسیریته مهم است، مفید باشد.

منابع محاسباتی

مدل‌های یادگیری عمیق به طور معمول نیاز به منابع محاسباتی بیشتری مانند GPU و زمان آموزش بیشتری نسبت به الگوریتم‌های سنتی یادگیری ماشین دارند. این می‌تواند یک محدودیت باشد، به ویژه در محیط‌های محدود منابع یا زمانی که نیاز به استقرار سریع مدل وجود دارد.

به طور خلاصه، در حالی که مدل‌های یادگیری عمیق مانند BERT عملکردی بهتر را در حل مسائل پردازش زبان طبیعی مانند شناسایی نویسنده ارائه می‌دهند، الگوریتم‌های سنتی یادگیری ماشین همچنان جایگاه خود را دارند، به ویژه در صورتی که مجموعه داده‌ها کوچک‌تر باشد یا زمانی که تفسیرپذیری اهمیت دارد. انتخاب بین روش‌های یادگیری عمیق و الگوریتم‌های سنتی یادگیری ماشین در نهایت به نیازها و محدودیت‌های خاص وظیفه مورد نظر بستگی دارد.

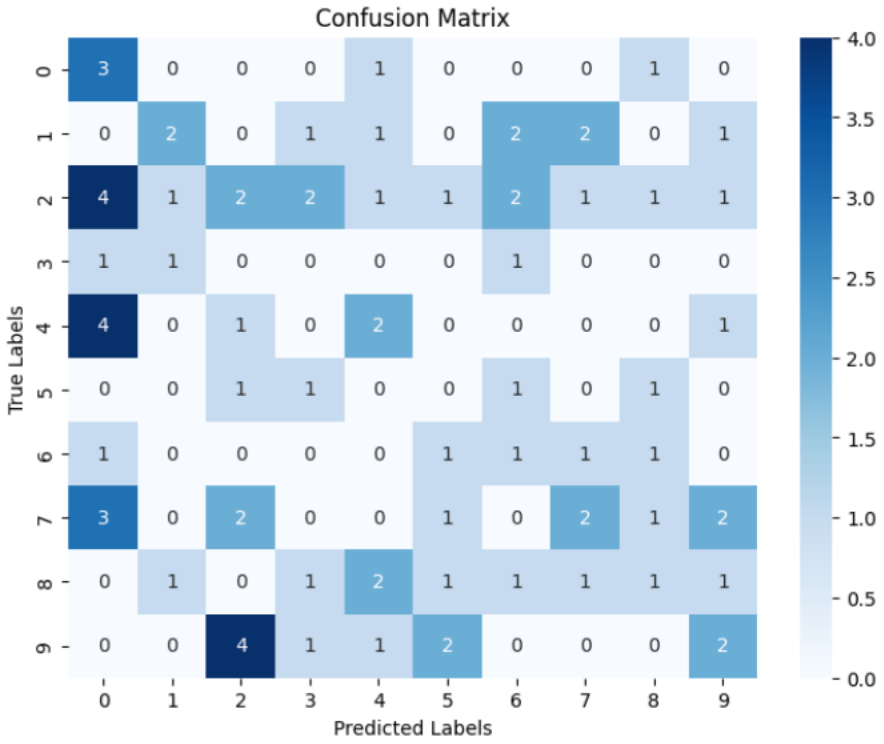
در این پروژه، ما آموزش مجموعه داده‌ها با استفاده از یک معماری شبکه عصبی مانند CNN را

ارائه کرده‌ایم. این معماری با یک لایه ورودی که شامل ۵۰۰ گره است شروع می‌شود، و پس از آن یک لایه فشرده با ۲۵۶ واحد و فعال‌سازی ReLU دنبال می‌شود که ظرفیت مدل را برای یادگیری الگوهای پیچیده افزایش می‌دهد. در ادامه، یک لایه dropout با نرخ dropout برابر با ۰.۱ وارد می‌شود تا از بیش‌برازش با غیرفعال کردن تصادفی ۱۰ درصد از واحدها در طول آموزش جلوگیری کند. یک لایه دیگر با ۵۰ واحد به این معماری اضافه شده و توانایی مدل در ضبط روابط پیچیده در داده‌ها را بهبود می‌بخشد. در نهایت، لایه خروجی از تابع فعال‌سازی softmax استفاده می‌کند تا احتمالات برای هر کلاس تولید شود. این راه‌اندازی یادآور روش‌های سنتی یادگیری ماشینی است که قبلاً به کار می‌رفتند، و هدف ما ارزیابی عملکرد مدل‌های از پیش آموزش‌دیده در این چارچوب برای اطمینان از هرگونه بهبود دقت بالقوه است.

در ادامه به معیارهای ارزیابی عملکرد و ماتریس درهم ریختگی این مدل می‌پردازیم.

	precision	recall	f1-score	support
0	0.19	0.60	0.29	5
1	0.40	0.22	0.29	9
2	0.20	0.12	0.15	16
3	0.00	0.00	0.00	3
4	0.25	0.25	0.25	8
5	0.00	0.00	0.00	4
6	0.12	0.20	0.15	5
7	0.29	0.18	0.22	11
8	0.17	0.11	0.13	9
9	0.25	0.20	0.22	10
accuracy			0.19	80
macro avg	0.19	0.19	0.17	80
weighted avg	0.22	0.19	0.19	80

شکل ۱.۴: معیارهای ارزیابی مدل مبتنی بر رویکرد سنتی ML



شکل ۲.۴: ماتریس درهم ریختگی مدل مبتنی بر رویکرد سنتی ML

فصل ۵

نتیجه‌گیری

در این پروژه، ما با استفاده از مدل‌های زبانی پیش‌آموزش‌دیده مانند BERT برای شناسایی نویسنده در ادبیات فارسی کار کردیم. ابتدا یک مجموعه داده گسترده و نمونه‌گیری شده را جمع‌آوری کردیم و سپس مدل‌های BERT را بر روی این مجموعه داده‌ها آموزش دادیم و عملکرد آن‌ها را ارزیابی کردیم. نتایج نشان داد که مدل‌های زبانی پیش‌آموزش‌دیده مانند BERT از عملکرد بسیار خوبی برخوردارند و توانایی بالایی در شناسایی نویسنده دارند، به خصوص زمانی که به درستی پارامترهای مدل تنظیم شوند.

در این پروژه، ما از روش 5 fold validation-cross استفاده کردیم، که نتایج دقیق و قابل اعتمادی را ارائه می‌دهد. علاوه بر این، با بررسی عیارهای ارزیابی عملکرد مدل‌ها مانند دقت، صحت، پوشش و F۱ Score و ماتریس درهم‌ریختگی، ما توانستیم عملکرد هر مدل را به صورت جامع و دقیق بررسی کنیم.

همچنین، بررسی تاثیر نرخ یادگیری و اثرات آن بر عملکرد مدل‌ها، به ما اطلاعات ارزشمندی درباره بهینه‌سازی مدل‌ها و بهبود عملکرد آنها ارائه کرد. همچنین، بررسی تاثیر طول اسناد و حذف کلمات توقف به ما کمک کرد تا عوامل مختلفی که ممکن است بر عملکرد مدل‌ها تاثیرگذار باشند را شناسایی و ارزیابی کنیم.

با توجه به نتایج به‌دست‌آمده از این تحلیل‌ها و آزمایش‌ها، می‌توان نتیجه گرفت که مدل‌های زبانی

پیش‌آموزش‌دیده مانند BERT دارای عملکرد بسیار خوبی در شناسایی نویسنده در ادبیات فارسی هستند و می‌توانند به عنوان ابزار قدرتمندی برای این مسئله مورد استفاده قرار بگیرند.
علاوه بر این، ما مقایسه‌ای بین عملکرد مدل‌های زبانی پیش‌آموزش‌دیده و روش‌های سنتی یادگیری ماشینی انجام دادیم و مشاهده کردیم که مدل‌های BERT اغلب عملکرد بهتری نسبت به روش‌های سنتی ارائه می‌دهند. این نتایج نشان می‌دهد که استفاده از مدل‌های زبانی پیش‌آموزش‌دیده می‌تواند بهبود قابل توجهی در عملکرد و دقت مدل‌های شناسایی نویسنده داشته باشد.

فصل ۶

بهبود و گسترش

بررسی نتایج آزمایشات انجام‌شده نشان می‌دهد که فرآیند fine-tuning مدل‌ها، با تغییرات کوچک در پارامترها، بهبود چشمگیری در عملکرد مدل‌ها داشته است. به عنوان مثال، با تغییراتی در نرخ یادگیری می‌توان به سرعت بهبود در عملکرد مدل‌ها دست یافت. همچنین، از طریق افزایش تعداد epoch ها، می‌توان با افزایش دقت مدل‌ها، پارامترهای بهتری را به‌دست آورد.

در لایه‌های fine-tuning مدل‌ها، استفاده از dropout یا افزودن dense-layer های بیشتر، می‌تواند به افزایش دقت و عملکرد کلی مدل‌ها کمک کند. همچنین ارزیابی نمودارهای train و validate نشان می‌دهد که هنوز مدل‌ها به همگرایی نرسیده‌اند. بنابراین، ممکن است با افزایش تعداد epoch ها، بهبود دقت مدل‌ها دست‌یافته شود و پارامترهای بهتری به‌دست آمده و در نتیجه عملکرد مدل‌ها بهبود یابد.

لازم به ذکر است که علاوه بر مدل‌های BERT مورد استفاده در این پروژه، مدل‌های BERT دیگری نیز وجود دارند که می‌توانند برای تحلیل مسئله مورد بررسی استفاده شوند. بر اساس نتایج تست‌های انجام‌شده، تشابه قابل‌ملاحظه‌ای در عملکرد مدل‌ها به‌دست آمده و بهتر است تمام مدل‌های BERT موجود بررسی شوند.

همچنین، می‌توان با بهینه‌سازی داده‌ها و حذف کلماتی که رابطه چندانی با مسئله مورد بررسی ندارند، عملکرد مدل‌ها را بهبود بخشید.

واژه‌نامه فارسی به انگلیسی

شناسایی نویسنده Author Identification
دقت Accuracy
صحت Precision
پوشش Recall
نیم‌فاصله Half Space
یونیکد Unicode
کلمات توقف Stop Word
تنظیم دقیق Fine Tuning
داده برچسب‌خورده Supervised Data
نرخ یادگیری Learning Rate
تابع ضرر Loss Function
یادگیری عمیق Deep Learning
فعال ساز Activation