دسته بندی اسناد فارسی با استفاده از یادگیری عمیق

نام دانشجو: رامین عقیقی

نام استاد راهنما: دکتر حسن بشیری

گروه مهندسی کامپیوتر، دانشگاه صنعتی همدان
ایمیل دانشجو: ramin_aghighi@yahoo.com

گروه مهندسی کامپیوتر
دانشگاه صنعتی همدان

چکیده

دسته‌بندی داده‌های متنی در کاربردهای مختلف بسیاری مورد توجه قرار گرفته و پژوهشگران زیادی بر روی آن کار کردند. در ابتدا از روش‌های سنتی مانند نایو بیز، ماشین بردار پشتیبان و نزدیک‌ترین همسایه استفاده می‌شد که بخاطر وابستگی به استخراج ویژگی‌ها و مشکل زمان‌بر بودن و پرهزینه بودن استخراج ویژگی، از سال ۲۰۱۰ اکثر پژوهشگران و مقالات به سمت روش‌های یادگیری عمیق رفتند که نیاز به مهندسی ویژگی ندارد. روش Text CNN تعداد مرجع زیادی را در مدل یادگیری عمیق دارد که در آن یک شبکه‌ی عصبی کانولوشن (CNN) برای حل مشکل طبقه‌بندی متن برای اولین بار معرفی شده است. حالا در این پایان‌نامه، قصد پیاده‌سازی الگوریتم‌های به‌روز یادگیری عمیق مانند CNN، RNN، LSTM و GRU را بر روی داده‌های همشهری داریم که در زبان فارسی از داده‌های معتبر به حساب می‌آید.

واژه های کلیدی : دسته بندی اسناد، شبکه های عصبی عمیق، شبکه های عصبی کانولوشن، دا همشهری

مقدمه

در دهه‌های گذشته، مدل‌های سنتی نظیر NB، KNN، SVM و Random Forest (RF) به عنوان دسته‌بندهای متنی مورد استفاده قرار گرفتند. از سوی دیگر، مدل‌های مبتنی بر یادگیری عمیق مانند Text CNN و LSTM نیز به دلیل قابلیت بهتر در عملکرد، جایگاه خود را برتر یافته‌اند.

در زمینه دسته‌بندی اسناد به زبان فارسی، با وجود پیشرفت‌های گسترده در حوزه زبان‌شناسی و پردازش زبان فارسی، تعداد محدودی از تحقیقات در این زمینه انجام شده است. این تحقیقات نشان می‌دهند که چالش‌های خاصی وجود دارد که از جمله آن‌ها تفاوت‌های فرهنگی، اصطلاحات خاص به زبان فارسی و کمبود داده‌های آموزشی متنوع را می‌توان نام‌برد. به منظور حل این چالش‌ها، تصمیم گرفته‌ایم از مدل‌هایی که در زبان انگلیسی پاسخ‌های قابل قبولی ارائه داده‌اند، بهره‌مند شویم و دقت این مدل‌ها را بر روی داده‌های فارسی اندازه‌گیری نماییم. این رویکرد نوآورانه امکان استفاده از تجربیات موفق در زبان انگلیسی را برای دسته‌بندی دقیق‌تر اسناد به زبان فارسی ایجاد می‌کند.

در این پژوهش، هدف اصلی ما بررسی و پیاده‌سازی الگوریتم‌های یادگیری عمیق از جمله CNN، RNN، LSTM و GRU بر روی داده‌های همشهری فارسی است. با اجرای این الگوریتم‌ها، تلاش می‌شود تا بهبودهای قابل توجهی در دقت و کارآیی دسته‌بندی اسناد به زبان فارسی حاصل گردد.

اهداف و روش پژوهش

طبقه‌بندی متن در تجزیه و تحلیل احساسات(SA)، برچسب‌گذاری موضوع (TL)، طبقه‌بندی اخبار (NC)، پاسخگویی به سؤال (QA) و غیره مورد توجه بسیاری قرار گرفته است. همینطور مدل‌های یادگیری عمیق نیز در دهه اخیر برای طبقه‌بندی متن نسبت به سایر مدل‌های یادگیری ماشین مانند SVM و جنگل‌های تصادفی و KNN، به شدت مورد استقبال قرار گرفته است. البته با وجود کارهای بسیار در زبان‌های دیگر، متاسفانه در حوزه زبان فارسی کارهای کمی صورت گرفته است که همین نشان از اهمیت پیاده‌سازی متدهای روز شبکه‌های عصبی مانند RNN، CNN، LSTM و GRU در زبان فارسی می‌باشد که بر روی دیتاست همشهری با ده هزار سند مورد بررسی قرار خواهد گرفت و در ادامه دقت آن با سایر روش‌ها مورد قیاس قرار خواهد گرفت.

یافته های پژوهش

در پیاده‌سازی الگوریتم‌های یادگیری عمیق بر روی داده‌های همشهری فارسی، نتایج حاصل از تحقیق نشان داد که از میان الگوریتم‌های مورد بررسی، روش‌های LSTM و CNN توانستند دقت مطلوبی را در دسته‌بندی اسناد به زبان فارسی به‌دست آورند. با اعمال این دو روش، دقت حاصل از مدل بر روی داده‌های آزمایشی به اندازه قابل توجهی ارتقاء یافت. این نتایج نشان دهنده قابلیت بالای روش‌های LSTM و CNN در حل مسائل دسته‌بندی اسناد در زبان فارسی می‌باشد. این پیشرفت‌ها در دستیابی به دقت مطلوب، اساساً بهبود عملکرد مدل در تفکیک و دسته‌بندی متون فارسی را نشان می‌دهند و این امر می‌تواند اساسی در توسعه راهکارهای یادگیری عمیق برای زبان فارسی باشد.

راهبردهای پیشنهادی

در پژوهش حاضر، توجه به روش‌های مدرن از قبیل BERT و GPT، همچنین روش‌های مبتنی بر گراف مورد بررسی قرار نگرفته است. با این وجود، این روش‌ها از اهمیت بسزایی برخوردارند و می‌توانند در بهبود دقت دسته‌بندی اسناد به زبان فارسی تأثیرگذار باشند. به همین دلیل، در ادامه تحقیقات و به منظور ارتقاء کارایی مدل‌ها، پیشنهاد می‌شود که به روش‌های زیر توجه شود:

1) بررسی مدل‌های پیش‌آموزش‌دیده: ارزیابی و بررسی مدل‌های پیش‌آموزش‌دیده نظیر BERT و GPT در زمینه دسته‌بندی اسناد به زبان فارسی. این مدل‌ها به عنوان یکی از جدیدترین رویکردها در حوزه یادگیری عمیق و پردازش زبان طبیعی شناخته شده‌اند.
2) استفاده از روش‌های مبتنی بر گراف: بررسی و اعمال روش‌های مبتنی بر گراف برای دسته‌بندی اسناد. این رویکرد می‌تواند ویژگی‌ها و ارتباطات پیچیده‌تری را در متون فارسی مدل کند.
3) تحقیق در زمینه انتقال یادگیری: بررسی امکانات انتقال یادگیری از مدل‌های آموزش‌دیده بر روی داده‌های انگلیسی به مدل‌های دسته‌بندی اسناد فارسی.
4) افزایش حجم داده‌های آموزشی: تلاش برای جمع‌آوری و افزایش حجم داده‌های آموزشی به منظور افزایش تعمیم‌پذیری مدل‌ها.
5) تجزیه و تحلیل خطاها: تحلیل دقیق خطاها و نقاط ضعف مدل‌ها به منظور بهینه‌سازی و اصلاح راهبردها و پارامترها.

با اعمال این راهبردها، می‌توان به بهبود کارایی و دقت مدل‌ها در دسته‌بندی اسناد به زبان فارسی دست یافت و در جهت افزایش تنوع و کارآیی مدل‌ها قدم برداشت.

منابع

[1] Q. Li *et al.*, "A survey on text classification: From traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1-41, 2022.

[2] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

تایید استاد راهنما

نام و امضا استاد راهنما :

تایید تحصیلات تکمیلی :

تایید امور پژوهشی: