نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

ارائه یک سیستم پیشنهادی به‌منظور کشف تقلب در متون فارسی با پیچیدگی ساده بر اساس تکنیک‌های پردازش تصویر و سیستم‌های هوشمند مصنوعی

محمد فیوضی[۱و*]، جواد حدادنیا[۲]
[۱و*] بخش مهندسی پزشکی، دانشکده مهندسی برق و کامپیوتر، دانشگاه حکیم سبزواری، سبزوار، ایران، mohammad.fiuzy@yahoo.com
[۲] بخش مهندسی پزشکی، دانشکده مهندسی برق و کامپیوتر، دانشگاه حکیم سبزواری، سبزوار، ایران، haddadnia@sttu.ac.ir

چکیده

با توجه به گسترش روز افزون تقلب[۱] در متون مختلف از قبیل بانکداری، ارتباطات، تجارت و .... در سال‌های اخیر تکنیک‌های گوناگونی برای کشف تقلب ارائه شده است. تشخیص تقلب در متون به طور چشمگیری باعث کاهش این پدیده می‌شود. پردازش تصویر، دقیق، ارزان و موثر برای تشخیص این پدیده است. شاید بتوان از مهم ترین مشکلات روش‌های تشخیص برابری ظاهر اسناد و متون را دانست، که گاهاً با چشم غیر مسلح قابل تشخیص نیست. به همین خاطر امروزه تلاش می‌شود تا تشخیص نهایی به صورت هوشمند و ماشینی انجام گیرد. متاسفانه در این روش‌ها عدم در نظر داشتن ویژگی‌های مناسب در آزمایشات، روش‌های ماشینی و هوشمند مساله‌ایی مهم است حتی منجر به عدم تشخیص صحیح این پدیده می‌شود. در این تحقیق سعی شد تا با استفاده از ترکیب روش‌های هوشمند و پردازش تصویر به‌منظور افزایش کیفیت تصاویر متون فارسی، الگوریتم‌های طبقه بندی[۲] (FCM[3]) به‌منظور جدا سازی قسمت‌هایی از متون مورد نظر، الگوریتم تحلیل‌مولفه‌های اصلی (PCA[4]) بمنظور کاهش ویژگی‌ها و شبکه‌های عصبی مصنوعی (NN[5]) برای مدل‌سازی، شناسایی ساختار و شناسایی پارامترهای متون دارای تقلب روشی نو در جهت تشخیص صحیح و دقیق این پدیده در متون فارسی پیشنهاد شده است. روش پیشنهادی بر روی بانک تهیه شده شامل ۷۶ نمونه (دستخط) (۴۵ نمونه متن بدون تقلب و ۳۱ نمونه متن دارای تقلب) پیاده شد، که در نهایت به دقت شناسایی " بیش از ۹۹%" دست یافت. که در نوع خود با توجه به هوشمند، غیر تهاجمی، سریع و دقیق بودن می‌تواند به عنوان سیستم تشخیص کارآمدی به‌منظور کشف تقلب در متون فارسی با پیچیدگی ساده معرفی،بررسی و مورد استفاده قرار گیرد.

کلمات کلیدی

متون فارسی، تقلب، ویژگی، پردازش تصویر، هوش مصنوعی.

۱- مقدمه

بازشناسی حروف و ارقام دستنویس همواره یکی از موضوعات مورد علاقه برای تحقیق بوده است[4-1]. در آینده مستندات الکترونیکي به عنوان اصلي‌ترين ابزار ارتباطات نوشتاري مطرح خواهند شد و کتاب‌ها و مجلات کاغذي بخشي از تاريخ خواهند بود.[5] .آنچه امروزه از اهميت بسيار زيادي برخوردار گرديده، کمبود يا نبود اطلاعات نيست بلکه کمبود روش‌هايي در جهت استخراج و بهره‌برداري از اطلاعات در دسترس، به صورت مطلوب است .يك ويراستار انساني، تنها به وسيله دنبال کردن دقيق همه منابع متني مي‌تواند متوجه وقوع يك‌مساله جديد شود و يا اسناد را دسته‌بندي کند .اين روش‌کار براي حجم بالاي اطلاعات در سيستم‌هاي اطلاعاتی با پيچيدگي زياد، نامناسب است[6]. روش‌های پویا به دلیل دسترسی به ویژگی‌های زمانی از دقت و حساسیت بالاتری در تصدیق افراد برخوردار هستند [7]. بازشناسی نوری حروف یکی از مباحث مهم بازشناسی الگو می‌باشد. همه ساله تحقیقات گسترده‌ای برای ارایه سیستم‌های OCR[7] با کارایی و سرعت بهتر انجام می‌شود. یکی از اساسی ترین مراحل در تحلیل اسناد و تشخیص متون، بازشناسی قلم است، که شامل تعیین مواردی چون زبان متن، اندازه قلم، استیل قلم و نوع قلم می‌تواند کارایی سیستم درک سند را به مراتب بالا ببرد چرا که با استفاده از اطلاعات قلم می‌توان پارامترهای سیستم را بطور خاصی تنظیم کرد [8]. در یک تحقیق، از ویژگی‌های گشتاوری و طبقه بندی کننده نیز برای بازشناسی حروف دستنویس فارسی استفاده شده است[9]. روش‌هاي يادگيري آماري زيادي در زمينه دسته بندي متن‌ها در سال‌هاي اخير بکار برده شده است، که شامل مدل‌هاي برگشتي[۸]، دسته بندي نزديکترين همسايه KNN[9]، شبکه‌هاي بيز[۱۰]، درخت تصميم گيري، الگوريتم‌هاي يادگيري بر اساس قانون، شبکه‌هاي عصبي مصنوعي[۱۱] و تکنيك‌هاي يادگيري استنتاجي[۱۲] مي‌باشد [10, 11, 12]. ظهور و کشف پديده تقلب در متون از مسائل و