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

مطالعه مقایسه‌ای روش‌های مبتنی بر یادگیری ماشین در تشخیص نویسنده فارسی زبان بر اساس سبک نوشتاری

زینب فرهمند پور[۱]، هومان نیک مهر[۲]، محرم منصوری زاده[۳]، امید طبیب‌زاده قمصری[۴]
[۱] کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشگاه بوعلی سینا، همدان
zeinab.farahmandpoor@gmail.com
[۲] استادیار، گروه مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان
nikmehr@eng.ui.ac.ir
[۳] استادیار، گروه مهندسی کامپیوتر، دانشگاه بوعلی سینا، همدان
mansoorm@basu.ac.ir
[۴] دانشیار، گروه زبان‌شناسی همگانی، دانشگاه بوعلی سینا، همدان
o.tabibzadeh@basu.ac.ir

چکیده

تشخیص نویسنده، تلاشی است برای نشان دادن خصوصیات نویسنده‌ی تکه ای از اطلاعات زبانی به طوری که نهایتا بتوان بین متون مختلفی که توسط افراد گوناگون نوشته شده‌اند، تمایز معنی داری قائل شد. پیشرفت سریع ارتباطات اینترنتی، ابزارهای اینترنتی با هویت ناشناس مانند ایمیل و وبلاگ را به روش‌های ارتباطی محبوبی برای مرتکبین اعمال غیرقانونی تبدیل کرده و مسائل امنیتی خاصی را بوجود آورده است. زبان فارسی به علل مختلفی همچون سیاسی، اجتماعی و مذهبی مورد توجه افراد و سازمان‌های مختلفی قرار دارد. در این مقاله روشهای هوشمند writeprint **که به شناسایی نویسنده فارسی زبان و بر اساس سبک نوشتاری او کمک می‌نماید، معرفی و مقایسه شده‌اند. در این تحقیق، پس از جمع آوری دو پایگاه داده، از چهار مجموعه ویژگی شامل واژگانی، نحوی، معنایی و وابسته به کاربرد برای استخراج اطلاعات سبکی استفاده شده و مقایسه‌ای روی انواع مختلف روش‌های دسته‌بندی مانند** KNN، Delta، **شبکه عصبی، درخت تصمیم‌گیری*** **و تحلیل** Linear Discriminant† **روی این پایگاه ها انجام گردیده است. بررسی‌های این تحقیق نشان می دهد که روشهای تحلیل** Linear Discriminant **و** KNN **به ترتیب رتبه یکم و دوم دقت را بین روش‌های بررسی شده، در دست دارند.**

کلمات کلیدی

تشخیص هویت نویسنده، سبک نوشتاری، writeprint

* Decision tree
† تحلیل تفکیکی خطی