

تعیین شاعر به کمک روش‌های یادگیری ماشین

امیرشهاب شاهمیری و محمدرضا مطش بروجردی

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران.

amir@shahmiri.ir, borujerm@aut.ac.ir

چکیده - شناسایی شاعر یکی از گرایش‌های دانش شناسایی نویسنده است که از مسایل مهم در دسته‌بندی متن و پردازش زبان‌های طبیعی به‌شمار می‌رود. این نوشتار دست‌آورد پژوهشی با هدف تعیین هوشمند اشعار چهار شاعر بزرگ پارسی «حافظ، خیامی، فردوسی و مولوی» به کمک روش‌های یادگیری ماشین «درخت تصمیم و شبکه عصبی مصنوعی» است. برای سنجش کارایی شبکه، دقت تصمیم‌گیری و تعیین توسط انسان نیز آزموده و نتایج آنها با بازدهی روش‌های یادگیری ماشین مقایسه شده است. همچنین در هنگام استخراج ویژگی‌های گوناگون اشعار برای ارزیابی توسط ماشین، کوشش شده تا شیوه تفکر انسان برای تشخیص اشعار شاعران شبیه‌سازی گردد و بدین منظور، نزدیک به ۵۰ ویژگی گوناگون از اشعار ایرانی استخراج شده و در سه دسته فیزیکی، مفهومی و آوایی دسته‌بندی شده است. در این پژوهش - به‌طور میانگین - دقت تعیین شاعر به کمک درخت تصمیم تا ۹۴٪ و با شبکه عصبی تا ۹۵/۹٪ درستی نیز رسیده، در حالی که دقت تشخیص انسان، در جامعه آماری مورد مطالعه، تنها ۳۴/۴٪ بوده است.

کلید واژه- تعیین شاعر، یادگیری ماشین، شبکه عصبی مصنوعی، متن‌کاوی، طبقه‌بندی متن، درخت تصمیم، حافظ، خیامی، فردوسی، مولوی بلخی.

۱- مقدمه

شناسایی نویسنده از روی نثر، سبک و شیوه نوشتاری یا به‌عبارت دیگر ویژگی‌های نهفته در متون نوشته شده توسط وی، یکی از مباحث جدید در زمینه هوش مصنوعی و پردازش زبان طبیعی^۱ به‌شمار می‌آید. در این مبحث کوشش می‌شود تا با استخراج ویژگی‌هایی از درون متن و پردازش و تحلیل آن به کمک انواع روش‌های هوش مصنوعی، نویسنده متن شناسایی گردد، که به این سلسله امور متن‌کاوی^۲ گفته می‌شود. شناسایی شاعر نیز یکی از شاخه‌های تخصصی تعیین نویسنده^۳ است که در آن می‌توان افزون بر ویژگی‌های مربوط به رشته حروف، ویژگی‌های دیگری مانند وزن، لحن و آهنگ جملات را نیز پردازش کرد.

ایده شناسایی نویسنده از مبحث طبقه‌بندی متن^۴ - که خود سرفصلی از دانش فهم زبان‌های طبیعی است [۱] - گرفته شده [۲] و در آن کوشش می‌شود تا با تجزیه و تحلیل

واژگان، دستورزبان و مفهوم یک جمله و نیز با کمک گرفتن از دانش مربوط به واژگان، معنای آن جمله برای ماشین درک گردد [۳]. البته برای شناسایی نویسنده و شاعر لزوماً تکنیک‌های مورد استفاده در پردازش زبان‌های طبیعی به‌کار نمی‌رود، اما از آنجا که عکس‌العمل درست ماشین به هر یک از ویژگی‌های سیگنال، آوا، واژه، دستور، معنا یا مفهوم جمله‌ها، درک زبان طبیعی به‌شمار می‌آید [۴]، شناسایی شاعر نیز در این شاخه از دانش هوش مصنوعی می‌گنجد.

مهم‌ترین عواملی که موجب پیدایش و گسترش دانش طبقه‌بندی متن (که گاهی دسته‌بندی متن^۵ نیز خوانده می‌شود) در چند سال گذشته شده است، تولید فزاینده دانش و اطلاعات در جهان، گسترش چشمگیر اینترنت و ارتباطات و تبادل اطلاعات و نیز پیشرفت‌های چشمگیر در فناوری حافظه‌های جانبی بوده است. این موارد موجب گردیده که کاربر انسانی در طبقه‌بندی اطلاعات نوشتاری -