تشخیص جنسیت نویسندگان از روی متون با استفاده از جنگل تصادفی بیز

هدیه ساجدی*[1] و مهناز تسلیمی[2]

[1] دانشکده ریاضی، آمار و علوم کامپیوتر، پردیس علوم، دانشگاه تهران، ایران

[2] دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی، واحد قزوین، قزوین، ایران

چکیده

امروزه استفاده زیاد کاربران از محیط‌های مجازی و ارتباط آنها از طریق شبکه‌های اجتماعی مانند فیسبوک و توییتر لزوم بررسی مطالب موجود را در فضای مجازی بیشتر از گذشته کرده است. از آنجا که بالاترین میزان تبادل اطلاعات در فضای مجازی از طریق متن صورت می‌گیرد؛ لذا تشخیص هویت کاربران از نظر سن، جنس، عقاید مذهبی و سیاسی از روی متن‌های اینترنت، پراهمیت خواهد بود. مسأله تشخیص جنسیت در حوزه‌های امنیت و بازاریابی، می‌تواند مؤثر واقع شود. در مقالهٔ حاضر به تشخیص جنسیت نویسندگان مطالب بلاگ‌ها پرداخته می‌شود و جهت تشخیص جنسیت نویسنده، ویژگی‌های نحوی، مبتنی بر واژه، مبتنی بر حروف و واژگان گرامری مورد استفاده قرار می‌گیرند. به‌علاوه نتایج نشان می‌دهد که استفاده از ویژگی‌های n-گرمی حروف در بهبود عملکرد، بسیار مؤثر است. جهت انجام عمل دسته‌بندی روش جدیدی با عنوان جنگل تصادفی بیز ارائه می‌شود. نتایج آزمایش‌ها نشان می‌دهد که این روش در مقایسه با الگوریتم‌هایی مانند الگوریتم بیز ساده، درخت بیز ساده و جنگل تصادفی، نتایج بهتری ارائه داده و دقت دسته‌بندی را تا ۸۹/۵ ٪ افزایش داده است.

واژگان کلیدی: تشخیص جنسیت نویسنده، جنگل تصادفی، درخت بیز ساده، متن‌کاوی، دسته‌بندی.

Author gender identification from text using Bayesian Random Forest

Hedieh Sajedi*[1] & Mahnaz Taslimi[2]

[1] School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Iran

[2] Computer and Electrical Engineering, Azad Islamic University, Qazvin Branch, Qazvin, Iran

Abstract

Nowadays high usage of users from virtual environments and their connection via social networks like Facebook, Instagram, and Twitter shows the necessity of finding out shared subjects in this environment more than before. There are several applications that benefit from reliable methods for inferring age and gender of users in social media. Such applications exist across a wide area of fields, from personalized advertising to law enforcement of reputation management. Text posts represent a large portion of user generated content, and contain information which can be relevant to discovering undisclosed user attributes, or investigating the honesty of self-reported age and gender. Because the highest rate of information exchanges is in text format, author identification from the aspects like age, gender, political and religious opinions from these contents will seem more considerable. Gender identification that could be useful in security and marketing, also answers the following question: given a short text document, can we identify if the author is a male or a female? This question is motivated by recent events where people faked their gender on the Internet. In this paper, author gender identification in blog's data is investigated. In this regard, four groups of features include syntactic features, word-based features, character-based features, and function words are employed. In addition,

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال: ۹۶/۰۷/۰۵ ● تاریخ پذیرش: ۹۷/۱۱/۰۶ ● نوع مطالعه: پژوهشی

character n-gram features is used for improving the accuracy of classification. For evaluation of the proposed method, 3212 texts were collected from Technorati.com and blogger.com. Experimental results demonstrate that these types of features are practical. furthermore, a new classification method called "Bayesian Random Forest" is introduced. Each tree in Bayesian Random Forest is a Bayes tree. The results of experiment show that this method attains noticeable results in comparison with other classification algorithms such as Naïve Bayes, Naïve Bayes Tree, and Random Forest and it increases accuracy of gender identification to 89.5%.

Keywords: Author gender identification, Random Forest, NBTree, Text mining, Classification.

۱- مقدمه

با وجود اینکه فناوری‌های متفاوت ارتباطی در سالیان اخیر به‌وجود آمده، ولی همچنان استفاده از متن، رایج‌ترین و پرکاربردترین راه ارتباطی در فضای مجازی است. رشد بسیار سریع شبکه‌های اجتماعی و گسترده‌شدن ارتباطات انسانی از طریق دنیای مجازی، سبب به‌وجودآمدن مقدار زیادی دادهٔ تولیدشده توسط کاربران شده است. دسترسی به این داده‌ها فرصت خوبی برای بررسی خصوصیات زبان غیر رسمی اینترنت و همچنین یافتن ویژگی‌هایی چون سن، جنس، عقاید مذهبی و سیاسی گروه‌های مختلف مردم فراهم آورده است [1].

گمنام‌بودن، یک ویژگی مهم در ارتباطات اینترنتی است؛ به‌طوری‌که کاربران محیط مجازی ملزم نیستند اطلاعات شخصی خود را به‌طور صحیح بیان کنند؛ لذا لزوم یافتن روش‌های کارامد جهت تشخیص هویت کاربران اینترنت امری ضروری خواهد بود [2]. در این مقاله به‌طور خاص به بررسی مسأله تشخیص جنسیت نویسندگان در متون نوشته‌شده در بلاگ‌ها پرداخته شده است.

اگرچه تاریخچه اینترنت به این نکته اشاره دارد که در گذشته اینترنت به‌عنوان یک محیط مجازی مختص به مردان دیده می‌شد؛ ولی امروزه با افزایش سهولت دسترسی به اینترنت و شهرت این محیط مجازی، زنان بیشتر از گذشته در این شبکه اجتماعی شرکت می‌کنند و عقاید، نگرانی‌ها و دانش خود را به اشتراک می‌گذارند [3]. درنتیجه در میان پژوهش‌گران تمایل بیشتری به بررسی تفاوت‌های جنسیتی در محیط‌های مجازی شکل گرفته است. این تفاوت‌ها به نحوهٔ استفاده زنان و مردان از اینترنت و شکل‌گیری این تفاوت‌ها بر اساس تمایلات و روحیاتشان اشاره دارد [4]. به‌عنوان مثال طبق یافته‌های اوگان، زنان علاقهٔ کمتری به بیان اعتقادات سیاسی داشته و کمتر از حالت سلطه‌جویانه در مکالماتشان استفاده می‌کنند. همچنین می‌توان گفت که توجهات زنان حول زندگی خصوصی و حوزه خانه و خانواده معطوف است و از طرف دیگر تمرکز مردان بیشتر بر موضوعات خارجی در فضای عمومی و همچنین سیاسی شامل دولت و تجارت است [5]. گیلر و دورال بیان می‌کنند که به‌طور کلی زنان تمایل دارند با دیگران موافقت و مشارکت‌های احساسی و شخصی با دیگر افراد برقرار کنند و در مقابل مردان مایل هستند با لحن سلطه‌گرانه صحبت کرده و واکنش‌های منفی در محاوراتشان به‌کار برند [6]. از طرفی استفاده از یافته‌های علم روان‌شناسی در تفاوت ساختار مکالمه و نوشتار زن و مرد توانسته است در یافتن ویژگی‌های هرگروه نقش مؤثری داشته باشد. برای مثال الگوی مکالمه مردها به‌طورمعمول دارای ویژگی استقلال بوده و تأکید بر قدرت سلسله‌مراتبی از بالا به پایین دارد؛ بنابراین بیشتر از ضمیر مفرد "من" در جملات استفاده می‌کنند و از طرفی تمایل بیشتری به بیان اعتقادات، نظریات و مباحث سیاسی دارند و در مقابل زنان به میزان زیادی درباره احساسات، نگرانی‌ها و روابط خانوادگی بحث می‌کنند [1].

شناخت تفاوت‌های جنسیتی از روی متن و اهمیت آنها بیشتر در دو حیطه امنیت و بازاریابی مورد توجه قرار می‌گیرد. در حیطه امنیت جهت ردیابی اطلاعات شخصی و بررسی تمایلات هر جنس، بررسی عقاید اصلی هرگروه و شناخت رفتارهای بالقوهٔ آنان میتواند مفید باشد. تشخیص جنسیت تبهکاران مانند قاتلان از روی نوشته آنها نیز کاربرد دیگری از تشخیص جنسیت در حوزه امنیت است. همچنین در مورد محیط‌های مجازی بازاریابی، درک بهتر تفاوت‌ها در علاقه‌مندی به محصولات متفاوتِ بین دو جنس زن و مرد، کمک می کند که چه محصولی و یا چه خدمتی توسط مردان یا زنان تأیید شده و یا رد شده است [7]. داشتن اطلاعاتی از این دست موجب تقویت هوش بازار است؛ چون که می‌تواند باعث هدفمندشدن تبلیغات و توسعه محصولات شود.

بر خلاف تجزیه و تحلیل مسأله با تألیفات سنتی، که در آن با استفاده از صدها کلمهٔ ساده اقدام به دسته‌بندی متن می‌شد، در متون اینترنتی با چالش‌هایی از جمله استفاده از عبارت‌های اختصاری غیر رسمی، شکلک‌ها، غلط املایی تصادفی و عمدی، غلط گرامری، عدم وجود ساختار مشخص و وجود عناصری چون احساسات مواجه هستیم. از آنجا که این متون هیچ شباهتی به نوشته‌های رسمی مانند کتاب، آثار

ادبی و مقالات ندارند، لذا ارائه راه‌کارهای جدید جهت دسته‌بندی این گونه متون مورد نیاز است.

نوآوری این پژوهش ارایه جنگل تصادفی بیز به‌عنوان روشی نوین جهت دسته‌بندی و به‌کارگیری آن برای مسأله تشخیص جنسیت نویسندگان است.

در این پژوهش، علاوه‌بر به‌کارگیری ویژگی‌های جدید و مؤثر، با ارائه یک روش دسته‌بندی جدید به حل مسأله پرداخته می‌شود. در روش جدید دسته‌بندی با عنوان جنگل تصادفی بیز که حاصل ترکیب الگوریتم‌های درخت تصمیم و بیز ساده است، عمل دسته‌بندی متون انجام می‌شود. نتایج آزمایش‌ها نشان می‌دهد که این روش در مقایسه با الگوریتم‌های یادشده دقت بالاتری را تا میزان۸۹/۵٪ حاصل کرده است. این روش با ایده الگوریتم‌های تجمیعی پیشنهاد شده و توانسته است، میزان خطای دسته‌بندی را کاهش دهد. از طرفی در این مقاله به بررسی نقش ویژگی‌های n-گرمی حروف پرداخته می‌شود و نتایج حاصل نشان می‌دهد که با توجه به ویژگی‌های خاص نوشته‌های فضای مجازی، این دسته از ویژگی‌ها تأثیر به‌سزایی در دسته‌بندی این گونه متون دارند.

ادامه مقاله به این صورت سازمان‌دهی شده است. در بخش بعد مروری بر پژوهش‌های انجام‌شده در زمینه تشخیص جنسیت روی متون به زبان انگلیسی با رویکردهای مختلف انجام شده است. بخش سوم به بیان مسأله تشخیص جنسیت و استخراج ویژگی‌های متن پرداخته است. بخش چهارم به توضیح روش پیشنهادی و معرفی الگوریتم جنگل تصادفی بیز می‌پردازد. نتایج به‌دست‌آمده از آزمایش‌های انجام‌یافته در بخش پنجم ارائه شده و مورد تحلیل قرار می‌گیرند و درنهایت، بخش ششم به جمع‌بندی مقاله اختصاص دارد.

۲- مروری بر کارهای انجام‌شده

بیشتر مطالعات انجام‌شده در دسته‌بندی متن، تأکید بر شناخت نویسندهٔ متن و یا دسته‌بندی موضوعی متون داشته و مطالعات کمتری در زمینه تفکیک جنسیت نویسندگان صورت گرفته است.

تشخیص جنسیت نویسنده در سال ۲۰۱۱ توسط Cheng بر داده‌های خبرگزاری رویترز که به زبان انگلیسی است، انجام شد. خبرگزاری رویترز یکی از بزرگ‌ترین آژانس‌های خبری بین‌المللی است. داده‌های متنی این مجموعه از تاریخ ۱۹۹۶ لغایت ۱۹۹۷ جمع‌آوری شده‌اند و در قالب فایل‌های XML ، جهت مطالعات پژوهشی در دسترس قرار دارند. همچنین پایگاه داده رایانامه‌های Enron شامل رایانامه‌های کاربران، مورد استفاده قرار گرفته است [1]. در این بررسی از ویژگی‌های مستقل از متن مانند ویژگی‌های ساختاری، ویژگی‌های نحوی، ویژگی واژگان گرامری و ویژگی‌های مبتنی بر واژه و مبتنی بر حروف استفاده شده است. جهت انجام عمل دسته‌بندی از الگوریتم‌های یادگیری ماشین بردار پشتیبان، درخت تصمیم، یادگیری بیز رگرسیون Logistic، Adaboost استفاده شده است. در این بررسی در بهترین حالت با استفاده از الگوریتم ماشین بردار پشتیبان دقت دسته‌بندی ۷۶/۷۵ ٪ و ۸۲/۲۳ ٪ به‌ترتیب بر داده‌های رویترز و رایانامه به‌دست آمد. همچنین نشان داده شده که ویژگی‌های واژگان گرامری و پس از آن ویژگی‌های مبتنی بر واژه، بیشترین تأثیر را بر دسته‌بندی جنسیت نویسندگان متون خبری و رایانامه‌ها داشته‌اند [1].

پژوهش دیگری در این زمینه توسط چانگ بر محتویات انجمن‌های مربوط به زنان جامعه اسلامی که به زبان انگلیسی بود در سال ۲۰۱۱ انجام شد. در این بررسی از هر دو مجموعه ویژگی‌های مستقل از متن و ویژگی‌های وابسته به متن استفاده شد و تأثیر هر یک از این دو مجموعه ویژگی مورد بررسی قرار گرفت. در دسته ویژگی‌های مستقل از متن، ویژگی‌های نحوی و ویژگی‌های لغوی و ویژگی‌های ساختاری قرار گرفتند و در دسته ویژگی‌های وابسته به متن، ویژگی‌های ۱-گرمی و ۲-گرمی مطرح شدند. در این بررسی از الگوریتم ماشین بردار پشتیبان جهت دسته‌بندی استفاده شد. با اعمال این دسته‌بند بر ویژگی‌های مستقل از متن به‌دقت ۵۹ ٪ و بر مجموع ویژگی‌های مستقل از متن و وابسته به متن، به دقت دسته‌بندی ۶۴ ٪ دست یافته شد. پس از آن با اعمال الگوریتم انتخاب ویژگی بهره اطلاعاتی بر مجموع کل ویژگی‌های وابسته به متن و مستقل از متن به دقت دسته‌بندی ۸۶ ٪ دست یافته شد. نتایج این بررسی نشان داد که تفاوت‌های جنسیتی در مطالب انجمن‌ها نیز وجود دارد. از طرفی ویژگی‌های وابسته به متن نقش مؤثری در بهبود دقت دسته‌بندی داشته‌اند و چنانچه روش‌های انتخاب ویژگی بر این ویژگی‌ها اعمال شود، می‌تواند در به‌دست‌آوردن دقت بیشتردر دسته‌بندی متون، تأثیر به‌سزایی داشته باشد [7].

در موضوع تشخیص جنسیت نویسنده، مطالعه دیگری بر داده‌های بلاگ به زبان انگلیسی توسط Mukherjee صورت پذیرفت. در این بررسی از ویژگی‌هایی چون ویژگی F-measure، ویژگی فاکتور واژه، ویژگی‌های تجزیه‌ای[1] با طول متغیر استفاده شد. همچنین جهت انجام عمل دسته‌بندی الگوریتم‌های یادگیری مانند بیز، ماشین بردار پشتیبان، به‌کار

[1] Part of Speech(POS)

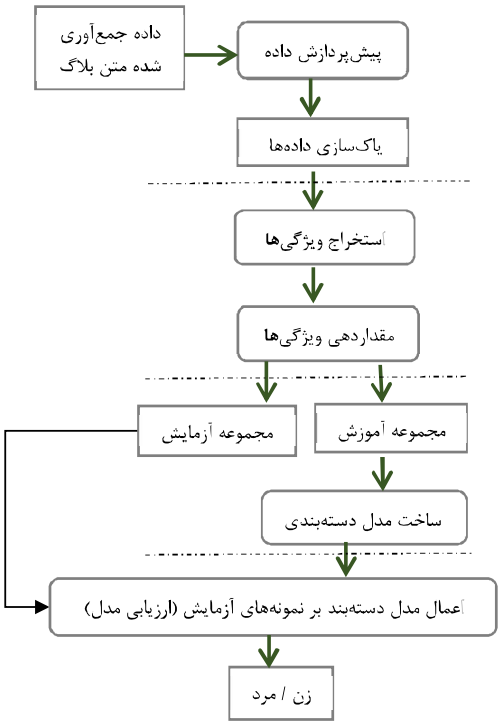
گرفته شدند. در این بررسی یک روش انتخاب ویژگی ترکیبی به نام EFS که حاصل از ترکیب روش‌های Filter و Wrapper بود، طراحی شده و مورد استفاده قرار گرفت. دلیل ترکیب این دو روش کاهش خطاهای هر یک از روش‌های Filter و Wrapper بوده است. انجام روش Wrapper در مسائلی که تعداد ویژگی‌ها زیاد باشد، غیر عملی بوده؛ چون جهت هر ویژگی که افزوده می‌شود، یک‌بار عمل دسته‌بندی صورت می‌گیرد تا مفیدبودن ویژگی بررسی شود. از طرف دیگر در روش Filter فقط از یک روش جهت وزن‌دهی ویژگی‌ها استفاده می‌شود؛ لذا این امر ممکن است موجب شود برخی از ویژگی‌های مفید و برجسته‌ای که توسط دیگر روش‌های انتخاب ویژگی، انتخاب می‌شوند، حذف شوند. روش انتخاب ویژگی EFS ایرادات هر دو رویکرد Filter و Wrapper را پوشش داد و درنهایت با اعمال دسته‌بند SVM_R به دقت ۸۸/۵۶ ٪ دست یافته شد. همچنین در این بررسی نشان داده شد که ویژگی‌های تجزیه‌ای با طول متغیر در کنار انتخاب ویژگی EFS، می‌تواند تأثیر به‌سزایی در دسته‌بندی جنسیت داشته باشد؛ به‌طوری‌که بدون استفاده از ویژگی یادشده و روش EFS، دقت دسته‌بندی به ۶۶/۱۷ ٪ کاهش یافته است [8].

در بررسی دیگری که در سال ۲۰۱۲ توسط Miller بر داده‌های شبکه اجتماعی توییتر به زبان انگلیسی انجام شد از ویژگی‌های n- گرمی حروف به‌عنوان ویژگی‌های وابسته به متن استفاده شد. در این مطالعه میلر از ویژگی‌های ۱-گرمی تا ۵-گرمی جهت دسته‌بندی متون استفاده کرد. از آنجا که تعداد این ویژگی‌ها بسیار زیاد بودند از شش الگوریتم انتخاب ویژگی متفاوت جهت وزن‌دهی ویژگی‌ها استفاده شد. که این امر منجر به تولید شش زیرمجموعه از ویژگی‌های برتر n-گرمی شد. درنهایت ویژگی‌هایی به‌عنوان ویژگی‌های مفید انتخاب شدند که دست‌کم توسط چهار الگوریتم از شش الگوریتم انتخاب ویژگی، برگزیده شده بودند. پس از اعمال دسته‌بندهای یادگیری بیز و پرسپترون دقت بالای ۹۰٪ به‌دست آمد. همچنین در این بررسی بیان شد که استفاده از ویژگی‌های n-گرمی در کنار روش‌های انتخاب ویژگی مناسب می‌تواند در دسته‌بندی متون مفید باشد [2].

با توجه به اهمیت و کاربرد تشخیص جنسیت از روی متون نوشته شده افراد و دقت نه‌چندان بالای روش‌های قبلی، لزوم پژوهش‌های بیشتر در زمینه استخراج ویژگی و دسته‌بندی به‌منظور افزایش دقت احساس می‌شود.

۳- مسأله تشخیص جنسیت

مسأله تشخیص جنسیت به‌صورت یک مسأله دسته‌بندی دوتایی مطرح می‌شود که در آن هر نمونه متن، به یکی از دو دستهٔ زن یا مرد اختصاص می‌یابد. در ابتدا باید مجموعه‌ای از ویژگی‌ها را به‌دست آوریم. چنانچه تعداد ویژگی‌های انتخاب‌شده را جهت عمل دسته‌بندی d در نظر بگیریم، می‌توان هر نمونه از متن را با یک بردار d بُعدی نمایش داد که d تعداد کل ویژگی‌های به‌دست‌آمده است. ریاضی‌گونه می‌خواهیم تابع $y_i = f(x_i)$ را برای n نمونه متن ارائه‌شده به‌دست آوریم به‌طوری‌که $x = \{x_1,\dots,x_n\}$ نشان‌دهنده یک مجموعه از نمونه‌ها و $y_i \in \{\pm 1\}$ نشان‌دهنده برچسب و یا دسته هر نمونه است. عمل تشخیص جنسیت در دو مرحلهٔ آموزش و آزمایش انجام می‌شود، مرحلهٔ آموزش شامل پیش‌پردازش داده، استخراج ویژگی و درنهایت ساخت مدل دسته‌بندی است و مرحلهٔ آزمایش انجام عمل دسته‌بندی بر نمونه‌های آزمایشی و اعتبارسنجی مدل را در بر می‌گیرد.



(شکل-۱): فرآیند دسته‌بندی تشخیص جنسیت نویسنده

(Figure-1): Classification process of author gender identification

به‌طور کلی فرآیند تشخیص جنسیت را می‌توان به چهار مرحله تفکیک کرد:

- انتخاب یک مجموعه مناسب از نمونه‌های متنی که به‌عنوان پایگاه داده استفاده شود.

- استخراج و انتخاب ویژگی‌هایی که بتوانند جداکننده‌های خوبی محسوب شوند.
- ساخت یک مدل دسته‌بندی جهت انجام عمل دسته‌بندی با استفاده از نمونه‌های آموزش.
- اعمال مدل ساخته‌شده بر نمونه‌های آزمایش و ارزیابی مدل.

در شکل (۱) این مراحل مشاهده می‌شود.

۳-۱- خصوصیات متون فضای مجازی

متن، به‌عنوان عمده‌ترین کانال ارتباطی وب، از منابع مهم برای ردیابی هویت در فضای مجازی است. هر چند موفقیت‌هایی در روش شناسایی جنسیت نویسنده به‌دست آمده، اما با توجه به ویژگی‌های خاص این متون، این مسأله نیاز به رویکردها و آزمایش‌های جدید دارد [6]. در مقایسه با عناصر مرسوم شناسایی نویسنده از جمله آثار ادبی و مقالات منتشرشده، یکی از چالش‌های شناسایی کاربران، طول محدود متن نوشته‌شده در اینترنت است. طول کوتاه این متون ممکن است، باعث شود برخی از ویژگی‌های شناسایی در متون عادی، بی‌اثر باشد. بنابراین چگونگی شناسایی ویژگی‌های نویسندگان با اسناد به‌نسبه کوتاه، به یک چالش تبدیل می‌شود [5]. از سوی دیگر، متن‌های اینترنتی دارای برخی ویژگی‌ها مانند غلط املایی هدفمند، غلط گرامری، استفاده از شکلک‌ها و واژگان اختصاری غیر رسمی بوده و مواردی مانند بیان احساسات در آنها به چشم می‌خورد که در واقع یک زبان غیر معمول را شکل می‌دهد [9]. یکی دیگر از مشخصه‌های مهم این‌گونه متن‌ها، ماهیت چندزبانه‌بودن آن‌هاست. اینترنت یک شبکه جهانی است و کاربران می‌توانند پیغام‌های خود را به هر زبانی در فضای مجازی پخش کنند. بنابراین قدرت پیش‌بینی ویژگی‌های نویسنده، برای زبان‌های مختلف، خود یک نگرانی دیگری است [10].

در این بررسی مسألهٔ دسته‌بندی بر اساس جنسیت در مطالب بلاگ‌ها مد نظر است. این مسأله در بسیاری از محیط‌های تجاری، موضوع حایز اهمیتی است. بلاگ‌ها به‌طورمعمول یادداشت‌های شخصی را که به‌صورت نوشتارهای غیر رسمی هستند، شامل می‌شود. با رشد سریع تعداد بلاگ‌ها ارزش آنها به‌عنوان یک منبع اطلاعاتی در حال افزایش است. مطالب بلاگ ه‌طورمعمول کوتاه و غیر ساخت‌یافته هستند و از بسیاری جملات غیر رسمی تشکیل شده است و می‌تواند دارای غلط‌های گرامری و یا املایی باشد [8,10]. بر این اساس دسته‌بندی نوشته‌های بلاگ‌ها مسأله سخت‌تری نسبت به متن‌های سنتی و رسمی است. در این بررسی، سعی شده است از انواع متفاوتی از ویژگی‌ها، جهت دسته‌بندی استفاده شود. پیچیدگی و تنوعی که در مطالب نوشته‌شده در بلاگ‌ها وجود دارد، نیاز به استفاده از انواع مختلف ویژگی‌ها را در دسته‌بندی، بیشتر می‌کند.

۳-۲- استخراج ویژگی‌ها

متون نوشته‌شده در اینترنت به‌صورت متن بدون ساختار، همراه با غلط‌های املایی و گرامری فراوان هستند. بر این اساس ویژگی‌هایی باید استخراج شوند که با وجود مشکلات بالا بتوانند در عملیات دسته‌بندی، نقش مفیدی داشته باشند. جهت به‌دست‌آوردن کارایی بالاتر در دسته‌بندی صحیح نویسندگان، انتخاب ویژگی‌های مناسب از اهمیت به‌سزایی برخوردار است [11,12]. پس از استخراج ویژگی‌ها، هر نمونه متن، به‌عنوان یک بردار از ویژگی‌ها نمایش داده می‌شود. در ادامه انواع ویژگی‌های به‌دست‌آمده در این پژوهش بیان می‌شود. ویژگی‌های مرتبط با تشخیص جنسیت نویسنده در قالب ویژگی‌های وابسته به متن و ویژگی‌های مستقل از متن مطرح می‌شوند.

۳-۲-۱- ویژگی‌های مستقل از متن

ویژگی‌های مستقل از متن به‌طور گسترده‌ای در حل مسایل تشخیص هویت نویسندگان استفاده می‌شود و در بیش‌تر مطالعات انجام‌شده از برخی از این ویژگی‌ها استفاده شده است. ویژگی‌هایی که در این بررسی استفاده شده‌اند، شامل شش دسته ویژگی: ویژگی‌های مبتنی بر حروف، مبتنی بر واژه، ویژگی‌های نحوی، ویژگی‌های فاکتور واژه، ویژگی واژگان گرامری، ویژگی نوع واژگان استفاده‌شده است که در جدول (۱) نمونه‌هایی از آنها یادشده است. در ادامه توضیحی بر هر یک از آنها ارائه می‌شود:

ویژگی‌های مبتنی بر حروف: این ویژگی‌ها به‌صورت گسترده‌ای در مسایل تشخیص نویسنده استفاده می‌شود؛ مانند تعداد نویسه‌های فاصله متن و تعداد نویسه‌های خاص مانند ٪ و & [13].

ویژگی‌های مبتنی بر واژه: در طول چهار دهه گذشته پژوهش‌گران پژوهش‌هایی داشته‌اند که نشان می‌دهد سلامت ذهنی و فیزیکی افراد به‌طورکامل با واژگانی که استفاده می‌کنند، مرتبط است. تحلیل متن صورت‌گرفته روی این مطالعات نشان می‌دهد کسانی که تمایل بیشتری در استفاده از واژگان احساسی مثبت دارند، (مانند love, nice, sweet) و استفاده از واژگان احساسی منفی را تعدیل می‌کنند (مانند hurt, ugly, nasty) و از طرفی تعداد بیشتری واژگان

شناختی به‌کار می‌برند (مانند cause, know) از مزایای شخصیتی بالاتری برخوردارند. ما با شمارش تعداد دفعات تکرار واژگان مورد نظر و تقسیم آن بر تعداد کل واژگان، مقدار ویژگی‌های مبتنی بر واژه را به‌دست می‌آوریم.

ویژگی‌های نحوی: این دسته از ویژگی‌ها ساختار نوشته‌های نویسنده را ارزیابی می‌کند. ویژگی‌های نحوی در برگیرنده علامت‌هایی نظیر ' : , ; و ؟؟؟ !!! است. چون ممکن است، نویسندگان در شرایط خاص از چندین علامت سؤال یا علامت تعجب جهت تأکید بر منظورشان استفاده کنند. قدرت ویژگی‌های نحوی که ناشی از علامت‌های به‌کاررفته در متن است، از تفاوت‌های عادات نوشتاری زن و مرد در استفاده از علامت‌گذاری ناشی می‌شود [8].

ویژگی‌های واژگان گرامری: واژگان گرامری واژگانی هستند که دارای معنای لغوی کمتر و یا دارای معنای مبهمی هستند، اما برای بیان ارتباط واژگان دیگر و یا جملات استفاده می‌شوند و می‌توانند تمایل و حالت فرد گوینده را مشخص کنند. واژگان گرامری را می‌توان به‌عنوان یک زیرمجموعهٔ مهم از ویژگی‌های مبتنی بر واژه در نظرگرفت؛ زیرا این دسته از ویژگی‌ها نقش مهمی درتعیین ویژگی‌های شخصیتی نویسنده دارند.

ویژگی نوع واژگان استفاده شده: این دسته ویژگی، شامل ویژگی‌هایی هستند که توسط Corney جهت دسته‌بندی جنسیت مورد استفاده قرار گرفت. این ویژگی‌ها از مطالعات متفاوتی که در زمینه طرز بیان زن و مرد صورت گرفته، به‌دست آمده است. همچنین بررسی‌ها نشان می‌دهد که زنان تمایل به استفاده از قیدهای احساسی مانند so, terribly, awfully داشته و در محاوراتشان به‌طورمعمول در قالب موافقت، درک و حمایت دیگران صحبت می‌کنند [12,14].

ویژگی فاکتور واژه: تحلیل و بررسی فاکتور واژه، به معنای یافتن گروهی از واژگان مشابه که در متن‌ها به چشم می‌خورند، است. این فرایند در قالب بررسی واژگان هم معنی و کنار هم قراردادن آنها در یک مجموعه، شکل می‌گیرد. در پژوهشی که توسط Argamen انجام شده است، واژگان تحلیل شدند و در قالب بیست فاکتور - که هر فاکتور نشان‌دهنده یک ویژگی بود - قرار گرفتند [8]. در این بررسی از این بیست فاکتور به‌عنوان بیست ویژگی استفاده شده است. استخراج همه ویژگی‌های یادشده از متن بلاگ‌ها، منجر به تولید ۴۶۵ ویژگی شد که در دسته‌بندی جنسیت از آنها استفاده شده است. در ادامه این ویژگی‌ها را با عنوان مجموعه F_1 بیان می‌کنیم.

(جدول -۱): نمونه‌هایی از ویژگی‌های مستقل از متن
(Table-1): Samples of text independent features

ویژگی	توصیف ویژگی
ویژگی‌های مبتنی برحروف	تعداد کل نویسه‌ها
	تعداد نویسه‌های حرفی (a,b,…z)
	تعداد نویسه‌های با حروف بزرگ
	تعداد نویسه‌های عددی
	تعداد نویسه‌های Space استفاده شده
	تعداد نویسه Tab Space
	تعداد نویسه‌های خاص $, %
ویژگی‌های مبتنی بر واژه	تعداد کل واژگان استفاده شده درهر متن
	متوسط طول واژگان (به نویسه)
	غنای واژگان
	واژگان بزرگتر از ۶ نویسه
	تعداد واژگان کوتاه (۱ تا ۳ نویسه)
ویژگی‌های نحوی	تعداد نویسه گیومه تک '
	تعداد نویسه کاما ,
	تعداد نویسه نقطه .
	تعداد نویسه کولون :
	تعداد نویسه سمی کولون ;
	تعداد نویسه علامت سؤال ؟
	تعداد علامت سؤال چندتایی ؟؟؟
	تعداد علامت تعجب !
	تعداد علامت تعجب چندتایی !!!
	تعداد سه نقطه ...
	تعداد واژگان a , an , the
	تعداد واژگان شبه‌جمله مانند yes, okey
ویژگی‌های گرامری	تعداد واژگان ضمایر مانندanyone, each, her (۷۴ ویژگی)
	تعداد افعال کمکی مانند are, can ,is (۴۷ ویژگی)
	تعداد حروف ربط[1] مانند and, or, yet (تعداد ۲۲ ویژگی)
	تعداد واژگان حرف اضافه[2] مانند-after-about-at of (۱۲۴ ویژگی)
ویژگی‌های نوع واژگان	واژگانی که به able ختم می‌شوند.
	واژگانی که به al ختم می‌شوند.
	واژگانی که به full ختم می‌شوند.
	واژگانی که به ible ختم می‌شوند.
	واژگانی که به ic ختم می‌شوند.
	واژگانی که به ive ختم می‌شوند.
	واژگانی که به less ختم می‌شوند.
	واژگانی که به ly ختم می‌شوند.
	واژگانی که به ous ختم می‌شوند.

[1] conjunction words
[2] Adposition words

۲-۲-۳- ویژگی‌های وابسته به متن

در این بررسی از ویژگی‌های n-گرمی حروف به‌عنوان ویژگی وابسته به متن استفاده شده است. از آنجا که تعداد بالای n، پیچیدگی مسأله را بیشتر می‌کند، ویژگی‌های ۱-گرمی تا ۴-گرمی برای هر نمونه از متن استفاده شد. تنها دلیل کاهش استفاده از n- گرمی‌های بالا این است که با افزایش n تعداد ویژگی‌ها به‌صورت نمایی افزایش می‌یابد [2].

ابتدا ویژگی‌های n-گرمی با طول دو استخراج شدند که تعداد آن شامل ۷۲۱ ویژگی بود. سپس ویژگی‌های ۳-گرمی به تعداد ۸۷۴۶ و ویژگی‌های ۴-گرمی به تعداد ۳۴۷۸۴ استخراج شدند. از آنجا که تعداد ویژگی‌های n-گرمی به‌طورمعمول خیلی زیاد هستند و تعداد زیاد آنها موجب عدم کارایی الگوریتم دسته‌بندی می‌شود، لذا ویژگی‌های یادشده پس از اعمال روش‌های انتخاب ویژگی مورد استفاده قرار می‌گیرند. در اینجا تعداد ویژگی‌های ۳-گرمی پس از اعمال انتخاب ویژگی به ۲۷۷۳ و تعداد ویژگی‌های ۴-گرمی به ۶۹۵۶ کاهش یافته و مورد استفاده قرار گرفتند.

۳-۳- مقداردهی ویژگی‌ها

پس از استخراج ویژگی‌ها می‌توان هر نمونه متن را توسط یک بردار که دارای ۴۶۵ بعد است و هر یک از این بعدها نشان‌دهندهٔ مقدار یک ویژگی از آن متن است، نمایش داد. در این پژوهش، جهت مقداردهی ۴۶۵ ویژگی، از نسبت هر ویژگی استفاده شده است، به این معنی که ویژگی‌هایی که مرتبط به حروف هستند، تقسیم بر تعداد کل حروف شدند و ویژگی‌هایی که مرتبط با واژه هستند، تقسیم بر تعداد کل واژگان شدند.

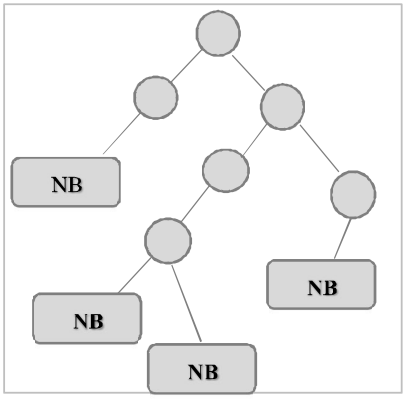
جهت اطمینان از این‌که تمام ویژگی‌ها به نسبت یکسان در تعیین جنسیت نقش ایفا می‌کنند با استفاده از رابطه (۱)، روش هنجارسازی MAX-MIN، مقادیر تمام ویژگی‌ها را به بازهٔ صفر تا یک تبدیل می‌کنیم:

$$Normalized\left(X_{ij}\right)=\frac{X_{ij}-\min\left(X_j\right)}{\max\left(X_j\right)-\min\left(X_j\right)} \qquad (۱)$$

که در اینجا x_{ij} نشان‌دهندهٔ ویژگی j ، از نمونه i است. همچنین $\min(x_j)$ کمترین مقدار برای j امین ویژگی و $\max(x_j)$ بیشترین مقدار برای j امین ویژگی است.

۴-۳- درخت بیز ساده[1]

ساخت درخت بیز ساده همانند دیگر درخت‌های تصمیم شروع می‌شود؛ به‌طوری‌که در هرمرحله ویژگی که بیشترین تأثیر را در تفکیک دسته‌ها داشته باشد، انتخاب می‌شود. تا زمانی‌که میزان سودمندی گره جدید در جداسازی، از میزان سودمندی گره پدر بیشتر باشد، ساخت درخت ادامه می‌یابد. چنانچه این شرط برقرار نشود و به‌عبارتی میزان ناخالصی گره جدید از میزان ناخالصی گره پدر بیشتر باشد، گره جدید تبدیل به برگ می‌شود و نمونه‌هایی که به این برگ رسیده‌اند، توسط الگوریتم یادگیری بیز دسته‌بندی شده و دستهٔ نمونه مورد نظر تشخیص داده می‌شود. درخت بیز ساده برخلاف درخت‌های دیگر، به نمونه‌هایی که به برگ درخت رسیده‌اند، دسته خاصی نسبت نمی‌دهد؛ بلکه جهت دسته‌بندی این نمونه‌ها از الگوریتم بیز استفاده می‌کند. جهت جلوگیری از زیادشدن تقسیم در درخت، مشخص شد درصورتی تقسیم ادامه یابد که کاهش تقریبی خطا بیشتر از ۵٪ بوده و تعداد نمونه‌ها در گره مورد نظر دست‌کم سی نمونه باشد [15]. شکل (۲) این درخت را نمایش می‌دهد.



(شکل-۲): درخت بیز ساده
(Figure-2): Naïve Bayes tree

۴- روش پیشنهادی

در این بخش روش پیشنهادی دسته‌بندی نمونه‌ها با عنوان جنگل تصادفی بیز معرفی می‌شود. از آنجا که الگوریتم یادگیری بیز در بیش‌تر مسایل دسته‌بندی متن به‌عنوان دسته‌بند مناسبی استفاده شده است، در روش پیشنهادی از این الگوریتم در تعیین دسته نمونه‌ها استفاده شده است. در این قسمت به بیان روش تولید جنگل تصادفی بیز می‌پردازیم.

[1] NB Tree

۱-۴- جنگل تصادفی بیز[1]

از آنجا که در برخی مسایل داده‌کاوی شکستن مسأله به زیرمجموعه‌های کوچکتر و ساخت مدل دسته‌بندی بر زیرمجموعه‌های حاصل، منجر به یادگیری بهتر الگوریتم دسته‌بندی می‌شود، در حل مسایل از الگوریتم‌های تجمیعی که اساس کارشان بر این اصل استوار است، استفاده می‌شود. ایده ساخت جنگل تصادفی بیز نیز بر همین اصل بنا شده است. این روش می‌تواند به‌عنوان یک الگوریتم دسته‌بندی تجمیعی مناسب که از ترکیب درخت تصمیم و بیز ساده و جنگل تصادفی تشکیل شده است، در مسایل دسته‌بندی استفاده شود.

پارامترهای این الگوریتم در جدول (۲) نمایش داده شده است. مقدار پارامترهای تعداد ویژگی انتخابی و تعداد نمونه‌های آموزشی با روش سعی و خطا به‌دست آمده‌اند. در این قسمت به توضیح مراحل ساخت مدل پرداخته می‌شود.

ابتدا از نمونه‌های مجموعه آموزشی جهت مرحله ساخت مدل استفاده می‌شود و مراحل یک تا پنج بر این مجموعه از نمونه‌ها به شرح زیر اعمال می‌شود:

۱- تعداد k ویژگی به‌طور تصادفی انتخاب می‌شود.

۲- از زیر مجموعه به‌دست‌آمده مرحله یک، به میزان j٪ از نمونه‌ها، با استفاده از روش نمونه‌گیری با جای‌گذاری انتخاب می‌شوند.

۳- الگوریتم یادگیر درخت بیز روی زیرمجموعه به‌دست‌آمده در مرحلهٔ دو اعمال می‌شود.

۴- یک مدل دسته‌بندی با توجه به ویژگی‌ها و نمونه‌های انتخاب‌شده ایجاد و ذخیره می‌شود.

۵- مراحل یک تا چهار به تعداد دفعات n مرتبه تکرار شده و n مدل دسته‌بندی، تشکیل می‌شود.

در مرحله بعد، مدل‌های ساخته‌شده بر نمونه‌های آزمایشی اعمال می‌شوند؛ لذا بر نمونه‌های مجموعه آزمایش مراحل زیر اعمال می‌شود:

۶- کلیه نمونه‌های آزمایشی به مدل دسته‌بند نخست داده می‌شود و این مدل، دستهٔ هر نمونه را مشخص می‌کند.

۷- کلیه نمونه‌های آزمایشی به مدل دسته‌بند دوم داده می‌شود و این مدل دستهٔ هر نمونه را مشخص می‌کند.

۸- این عملیات تا مدل n ام ادامه می‌یابد؛ درنتیجه بابت هر نمونه، n مرتبه دسته M یا F مشخص می‌شود و درنهایت با درنظرگرفتن "رأی اکثریت" دسته نهایی هر نمونه تشخیص داده می‌شود.

در شکل (۳) مراحل ساخت و آزمایش روش پیشنهادی، نمایش داده شده است. دلایل ارائه این روش را می‌توان با موارد زیر بیان کرد:

- شکستن مسأله به مسأله‌های کوچکتر که موجب می‌شود الگوریتم دسته‌بندی عمل یادگیری مدل را بهتر انجام دهد.
- استفاده از الگوریتم درخت تصمیم که جزییات بیشتری را در زمان یادگیری لحاظ می‌کند.
- تعیین دسته هر نمونه با استفاده از الگوریتم بیز که با وجود سادگی، در دسته‌بندی متون به‌طورمعمول دقت خوبی ارائه می‌دهد.
- دسته هر نمونه توسط چندین مدل تعیین می‌شود و در این‌صورت به‌دلیل استفاده از رأی‌گیری اکثریت احتمال خطا در تشخیص دسته نمونه کاهش می‌یابد.
- این روش در مسائل دسته‌بندی با ابعاد زیاد می‌تواند مفید باشد.

(جدول-۲): پارامترهای الگوریتم پیشنهادی جنگل تصادفی بیز
(Table-2): Parameters of proposed Bayes Random Forest

پارامتر	مقدار
روش انتخاب ویژگی‌ها	تصادفی
تعداد ویژگی‌ها(*k*)	۹۰ ویژگی
روش انتخاب نمونه‌ها	با جای‌گذاری
نسبت تعداد نمونه‌ها (*j*)	٪۹۰ مجموعه آموزش
دفعات تکرار	*n*
الگوریتم یادگیر	درخت بیز ساده
مجموعه آموزش	۷۰ ٪ کل داده

۲-۴- ویژگی‌های n-گرمی حروف

پیش‌بینی جنسیت از طریق متن در گذشته، در درجهٔ نخست به‌وسیله ساختار جمله و نقطه‌گذاری و یا تعداد واژه بررسی می‌شد. با این حال، آماده‌سازی یک فرهنگ لغت از ویژگی‌های متمایز در یک محیط مجازی که در آن معانی به‌شدت فشرده و استفاده از واژگان اختصاری، شکلک‌ها، و غلط املایی وجود دارد، به‌طورتقریبی غیر ممکن است [2]؛ به همین دلیل با بهره‌گیری از ویژگی‌های n- گرمی مبتنی برحروف و انتخاب ویژگی‌های برتر اقدام به پیش‌بینی جنسیت نویسندگان شده است.

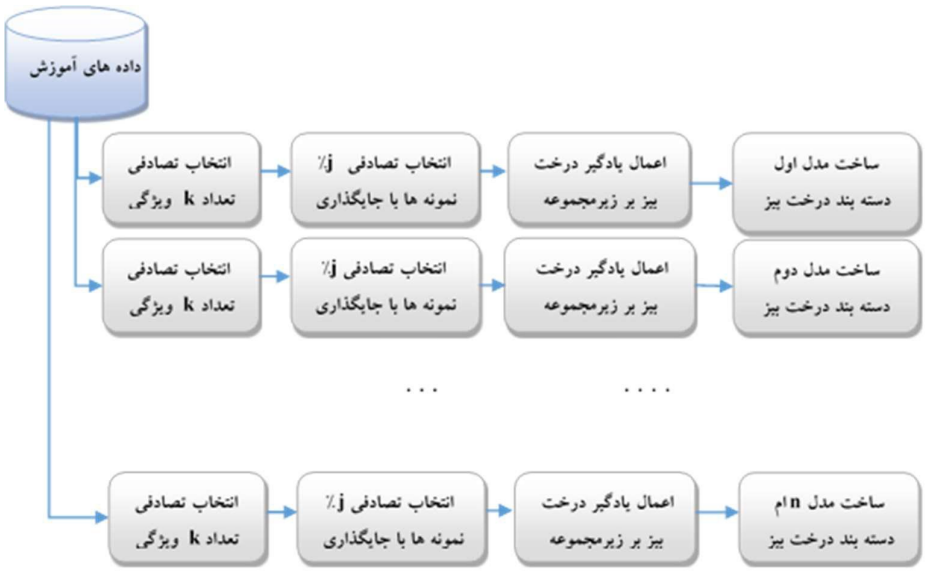
برای نشان‌دادن هر متن با استفاده از بردار ویژگی، از n- گرمی حروف استفاده شد که مجموعه‌ای از n حرف متوالی است. به‌منظور کاهش تعداد ویژگی‌های ۱-گرمی، از مجموعه کامل ۲۵۶ حروف اسکی تعداد سی مورد که در پایگاه داده بلاگ استفاده شده است، انتخاب شدند. هرکدام از –

[1] Bayesian Random Forest

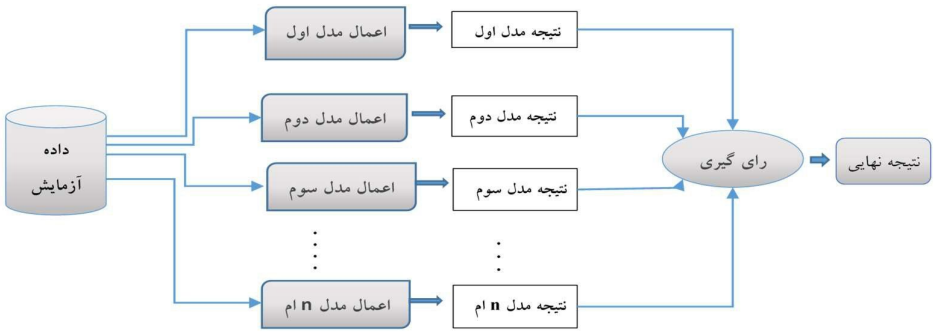
nگرمی‌های به‌دست آمده به‌عنوان یک ویژگی مورد استفاده قرارگرفت.

از آنجا که تعداد بالاتر n ، پیچیدگی مسأله را بیشتر می‌کند، ویژگی‌های ۱-گرمی تا ۴-گرمی برای هر نمونه از متن استفاده شدند. تنها دلیل کاهش استفاده از n-گرمی‌های بالا این است که با افزایش n تعداد ویژگی‌ها به‌صورت نمایی افزایش می‌یابد. به‌عبارت دیگر اگر سی ویژگی در ۱-گرمی استخراج شود بنابراین تعداد 30^2 که برابر ۹۰۰ است در ۲-گرمی استخراج خواهد شد و به همین صورت تا انتها ادامه خواهد یافت. لذا تنها n- گرمی‌هایی که در مجموعه آموزشی مشاهده شدند، به‌عنوان ویژگی در نظر گرفته شدند. در این بررسی ویژگی‌های ۳-گرمی و ۴-گرمی در افزایش دقت دسته‌بندی نقش مؤثری داشته‌اند؛ لذا در دو دسته مجزا، جهت بررسی دقت هر یک، در نظر گرفته شدند.

الف-



ب-



(شکل-۳): مراحل تولید الگوریتم پیشنهادی و اعمال آن بر نمونه‌های آزمایش. الف- مراحل ساخت جنگل تصادفی بیز ب- به‌کارگیری جنگل تصادفی به‌منظور دسته‌بندی

(Figure-3): Production steps of the proposed method and its running on the test samples (a) Production steps of Bayes Random Forest (b) Employing Random Forest for classification

۵- آزمایش‌ها

نتایج به‌دست‌آمده از آزمایش‌های انجام‌شده در این بخش بررسی می‌شود.

۵-۱- پایگاه داده بلاگ

پایگاه داده استفاده‌شده در این پژوهش مربوط به اطلاعات نویسندگان بلاگ است. هر نوشته‌ای از یک بلاگ به‌عنوان یک نمونه در نظر گرفته شده است. جهت جمع‌آوری متن‌های غیر

ساخت‌یافته‌ای که در بلاگ‌ها نوشته شده است، از بسیاری از سایت‌هایی که توسط کاربران مورد استفاده قرار می‌گیرند، مانند blogger.com و Technorati.com استفاده شده است. این مجموعه شامل ۳۲۱۲ بلاگ بوده است. نوشته‌های هر بلاگ با یک برچسب زن/ مرد مشخص شده که این برچسب با بررسی صفحه پروفایل فرد به‌دست آمده است. همچنین در پروفایل‌هایی که جنسیت به‌طور واضح مشخص نشده بود از تصاویر نویسنده موجود در پروفایل جهت تعیین جنسیت استفاده شده است [8].

از تعداد کل ۳۲۱۲ متن موجود تعداد ۱۶۷۳در حدود ٪۵۲ توسط مردان و تعداد ۱۵۳۹ در حدود ٪۴۸ توسط زنان نوشته شده و متوسط طول متن‌ها تعداد ۲۵۰ واژه برای مردان و ۳۳۰ واژه برای زنان بوده است.

پس از انجام مراحل استخراج ویژگی و مقداردهی ویژگی‌ها، نمونه‌های متن داده‌شده به دو زیرمجموعهٔ آموزش و آزمایش تقسیم می‌شود. در این بررسی مقدار ٪۷۰ نمونه‌ها به‌منظور آموزش الگوریتم دسته‌بندی در نظر گرفته و ٪۳۰ نمونه‌ها جهت مرحله آزمایش مدل دسته‌بندی لحاظ شدند.

۲-۵- بررسی تأثیر جنگل تصادفی بیز

با توجه به نتایج آزمایش‌های جنگل تصادفی بیز در دسته‌بندی و مقایسه با نتایج دسته‌بندهای بیز، درخت بیز و جنگل تصادفی مشاهده می‌شود این روش در بهبود دقت دسته‌بندی نقش مؤثری داشته و بیش از ٪۱۰ افزایش دقت نسبت به هر یک از الگوریتم‌های مذکور داشته است؛ لذا می‌توان گفت این روش می‌تواند در کاهش خطای دسته‌بندی نقش به‌سزایی داشته باشد. در ادامه به بررسی آزمایش‌ها و نتایج آنها می‌پردازیم.

با درنظرگرفتن ویژگی‌های مستقل از متن شامل شش دسته ویژگی استخراج‌شده، که شامل ۴۶۵ ویژگی و تعداد ۳۲۱۲ نمونه بوده است، آزمایش‌ها انجام می‌شود. جهت انجام آزمایش‌ها از نرم‌افزار Rapid Miner استفاده می‌شود.

در مرحلهٔ نخست، الگوریتم یادگیری بیز بر نمونه‌ها اعمال شد. همان‌طور که در شکل (۴) مشاهده می‌شود، دقت به‌دست آمده با استفاده از اعمال الگوریتم بیز ٪۶۴/۱۱ بوده است.

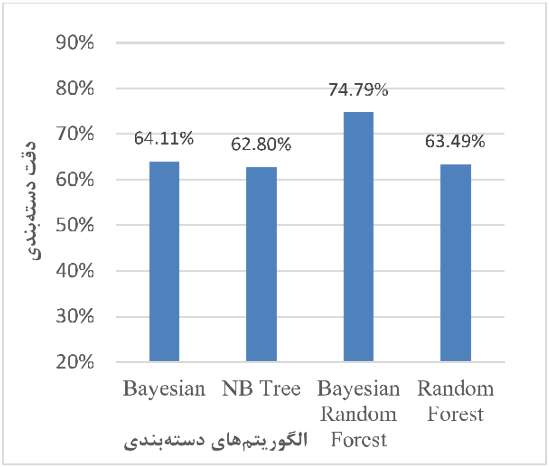
در مرحلهٔ بعدی آزمایش، تأثیر هر دسته از ویژگی‌های مستقل از متن مد نظر خواهد بود؛ لذا هر نوع از ویژگی‌ها را جداگانه بررسی کرده و دسته‌بند بیز را بر آنها اعمال کردیم. نتایج به‌دست‌آمده در جدول (۳) نمایش داده شده است. همان‌طور که از نتایج بر می‌آید، بالاترین تأثیر در تفکیک نمونه‌ها را ویژگی‌های واژگان گرامری با دقت ٪۶۰/۳۷ داشته است و پس از آن ویژگی فاکتور واژه با دقت ٪۵۴/۲۵ بهترین نتیجه را در دسته‌بندی با الگوریتم بیز داشته است.

در مرحلهٔ بعد آزمایش‌ها را با استفاده از روش پیشنهادی جنگل تصادفی بیز ادامه می‌دهیم. ولی از آنجا که از درخت بیز ساده در این روش استفاده شده، ابتدا عمل دسته‌بندی با دسته‌بند درخت بیز انجام داده شد و دقت ٪۶۲/۸۰ مشاهده شد. در شکل (۴) نتایج این دسته‌بندی نشان داده شده است.

(جدول-۳): دقت دسته‌بند بیز بر اساس انواع ویژگی‌های مستقل از متن

(Table-3): Accuracy of Bayes classifier based on different text independent features

انواع ویژگی‌ها	دقت دسته‌بندی بیز
ویژگی‌های مبتنی بر حروف	٪ ۵۱/۹۷
ویژگی‌های مبتنی بر واژه	٪ ۵۲/۰۷
ویژگی‌های نحوی	٪ ۵۲/۹۰
ویژگی واژگان گرامری	٪ ۶۰/۳۷
ویژگی فاکتور واژه	٪ ۵۴/۲۵
ویژگی نوع واژگان	٪ ۵۲/۳۹



(شکل-۴): دقت دسته‌بندی با الگوریتم‌های بیز، درخت بیز، جنگل تصادفی، جنگل تصادفی بیز

(Figure-4): Classification accuracy using Bayes, Bayes tree, Random Forest and Bayes random Forest

در ادامه، اقدام به ساخت جنگل تصادفی بیز با پارامترهای مرتبط آن با استفاده از روش سعی و خطا شد. به‌طوری‌که در ابتدا ساخت هر درخت بیز با ده ویژگی (مقدار k) انجام شد. از آنجا که نتیجه مناسبی حاصل نشد ساخت درخت با مقادیر متفاوت k از جمله پنجاه، شصت، هفتاد و نود ویژگی ادامه یافت، که بهترین نتایج مربوط به نود ویژگی بوده است. همچنین آزمایش‌ها با دو مقدار j (نسبت انتخاب

تصادفی نمونه‌های آموزش) به میزان ۷۰٪ و ۹۰٪ انجام داده شد. نتایج آزمایش منجر به انتخاب مقدار نود برای پارامتر j شد؛ لذا هر درخت بیز ساده با استفاده از نود ویژگی و ۹۰ ٪ نمونه‌های آموزشی ساخته شد.

در ادامه جنگل تصادفی بیز با استفاده از ۲۱ درخت بیز ساخته و بر نمونه‌های آزمایش اعمال شد. در این آزمایش به دقت ۷۲/۵۱ ٪ دست یافتیم. جهت بهبود دقت دسته‌بندی مسأله، ساخت جنگل تصادفی را با تعداد متفاوت درخت بیز ادامه دادیم که پس از آزمایش‌های متعدد جنگل تصادفی حاصل از ۳۱ درخت بیز، دقت ۷۴/۷۹٪ را نتیجه داد. آزمایش‌ها نشان داد که با افزایش تعداد درختان به ۴۱ و ۵۱ درخت، روند کاهشی در دقت دسته‌بندی خواهیم داشت. جدول (۴) دقت دسته‌بندی با تعداد مختلف درخت بیز را نمایش می‌دهد.

(جدول-۴): دقت رأی‌گیری اکثریت بر جنگل تصادفی درختان بیز ساخته‌شده از نود ویژگی

(Table-4): Accuracy of Majority vote in Bayes random forest using 90 features

دقت نهایی "رأی اکثریت"	جنگل تصادفی درخت بیز حاصل از نود ویژگی
٪ ۷۱/۶۸	۱۱ درخت بیز
٪ ۷۲/۵۱	۲۱ درخت بیز
٪ ۷۴/۷۹	۳۱ درخت بیز
٪ ۷۲/۸۲	۴۱ درخت بیز
٪ ۷۲/۲۰	۵۱ درخت بیز

جهت بررسی کارایی روش پیشنهادی در شکل (۴) مقایسه‌ای بین این روش با الگوریتم‌های دسته‌بندی بیز، درخت بیز ساده، جنگل تصادفی صورت گرفته است. همان‌طور که نتایج نشان می‌دهند، استفاده از مدل جنگل تصادفی بیز توانسته است، افزایش دقت خوبی را نسبت به الگوریتم بیز و همچنین درخت بیز داشته باشد.

(جدول-۵): مقایسه دقت دسته‌بندی روش پیشنهادی و جنگل تصادفی

(Table- 5): Comparison of proposed method and Random Forest based on accuracy measure

	۱۱ درخت	۲۱ درخت	۳۱ درخت	۴۱ درخت	۵۱ درخت
جنگل تصادفی	۶۱/۳۱	۶۳/۵۹	۶۳/۴۹	۶۳/۰۷	۶۳/۹۰
روش پیشنهادی جنگل	۷۱/۶۸	۷۲/۵۱	۷۴/۷۹	۷۲/۸۲	۷۲/۲۰

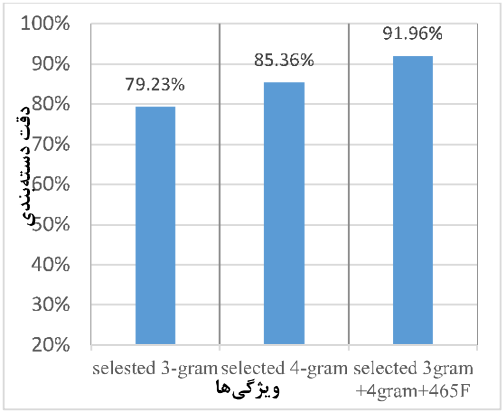
جهت بررسی دقیق‌تر، در جدول (۵) مقایسه‌ای بین الگوریتم جنگل تصادفی و روش پیشنهادی جنگل تصادفی بیز در حالت "تعداد مشابه درخت و تعداد نود ویژگی استفاده‌شده در ساخت درخت"، نمایش داده شده است و مشاهده می‌شود که روش پیشنهادی در همه حالت‌ها دقت بالاتری در دسته‌بندی داشته است.

۳-۵- بررسی تأثیر ویژگی‌های n-گرمی

به‌دلیل تعداد بسیار زیاد ویژگی‌های n-گرمی و اینکه این ویژگی‌ها تمام جزییات متون را استخراج می‌کند، می‌توانند قدرت جداکنندگی خوبی در دسته‌بندی متون داشته باشند. از آنجا که نوشته‌های فضای مجازی به میزان زیادی غلط املایی داشته و در آن شکلک‌ها و واژگان اختصاری عامیانه بسیاری به‌کار می رود، لذا استفاده از ویژگی‌های n-گرمی در مقیاس حروف، مفید و مؤثر خواهد بود. این بررسی نشان داده است که استفاده از این نوع ویژگی دقت دسته‌بندی را تا میزان ۹۱/۹۶ ٪ افزایش می‌دهد. در ادامه به بررسی این آزمایش‌ها می‌پردازیم.

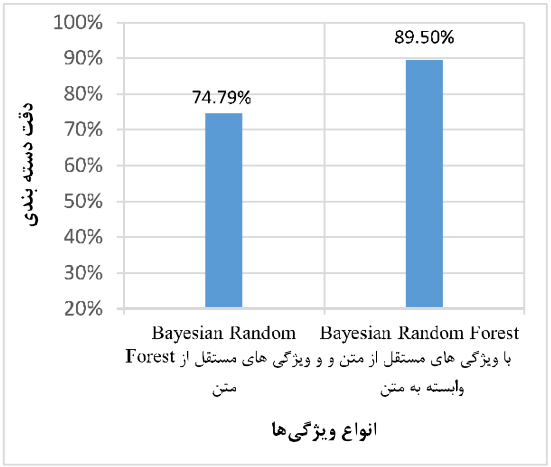
در این مرحله از ویژگی‌های وابسته به متن n- گرمی حروف جهت دسته‌بندی استفاده شد. از آنجا که آزمایش‌های انجام‌شده نشان داد n-گرمی‌های با طول یک و دو تأثیر چندانی در تفکیک دسته‌ها نداشتند؛ لذا ویژگی‌ها با طول یک تا سه در یک مجموعه و ویژگی‌ها با طول چهار در مجموعه دیگری در نظر گرفته شدند. ابتدا ویژگی‌هایی n- گرمی تا طول سه، استخراج شد که شامل تعداد ۸۷۴۶ ویژگی است و الگوریتم انتخاب ویژگی نسبت بهره اطلاعاتی بر آنها اعمال شد که تعداد ویژگی‌های انتخاب‌شده را به حدود ۲۷۰۰ ویژگی کاهش داده است. نتایج دسته‌بندی توسط الگوریتم بیز، ۷۹/۲۳ ٪ بوده است. در ادامه اقدام به استخراج ویژگی‌های ۴-گرمی کردیم که پس از استخراج به تعداد ۳۴۷۸۴ ویژگی در قالب۴-گرمی رسیدیم و سپس با اعمال الگوریتم انتخاب ویژگی نسبت بهره اطلاعاتی تعداد آنها به ۶۹۵۶ کاهش یافت و سپس با اعمال دسته‌بند بیز به دقت ۸۵/۳۶ ٪ دست یافته شد. در ادامه ترکیب ویژگی‌های مستقل از متن و وابسته به متن برای عمل دسته‌بندی استفاده شدند. چنانچه مجموعه ۴۶۵ ویژگی مستقل از متن را با F_1 نمایش دهیم، مجموعه $+F_1$-۴-گرمی + ۳-گرمی شامل ۱۰۱۹۴ ویژگی خواهد بود که با اعمال الگوریتم انتخاب ویژگی نسبت بهره اطلاعاتی تعداد آنها به ۷۱۳۶ ویژگی کاهش یافت. پس از انجام عمل انتخاب ویژگی، الگوریتم دسته‌بندی جنگل تصادفی بیز

را اعمال کرده و دقت دسته‌بندی ۹۱/۹۶٪ به‌دست آمد. شکل (۵) نتایج به‌دست‌آمده بر این نوع از ویژگی‌ها را نمایش می‌دهد.



(شکل-۵): دقت حاصل از ویژگی‌های n-گرمی و دسته‌بند بیز

(Figure-5): Accuracy of using n-gram features and Bayes classifier



(شکل-۶): دقت الگوریتم پیشنهادی بر ویژگی‌های مستقل از متن و وابسته به متن

(Figure-6): Accuracy of the proposed method with text dependent features and text independent features

به‌منظور اینکه در جنگل تصادفی بیز از ویژگی‌های وابسته به متن نیز استفاده شده باشد، سه مدل ساخته‌شدهٔ بالا، که با استفاده از ویژگی‌های وابسته به متن به‌دست آمده‌اند، به مدل‌های موجود در جنگل تصادفی بیز اضافه شدند؛ و از آنجا که این سه مدل با دقت بالایی دسته‌بندی جنسیتی را انجام داده‌اند، هنگام رأی‌گیری در جنگل تصادفی بیز، با وزن دو برابر در فرایند رأی‌گیری شرکت داده شدند. در ادامه دوباره عمل دسته‌بندی بر نمونه‌های آزمایشی با استفاده از هر دو مجموعه ویژگی مستقل از متن و وابسته به متن انجام و این بار عمل دسته‌بندی جنسیت نویسندگان بلاگ با دقت ۸۹/۵ ٪ انجام شد. در شکل (۶) نمودار دقت الگوریتم جنگل تصادفی بیز با هر دو مجموعه ویژگی نمایش داده شده است.

از آنجا که تشخیص جنسیت نویسنده در این پژوهش بر پایگاه داده بلاگ صورت گرفته، لذا نتایج به‌دست‌آمده از این پژوهش با نتایج به‌دست‌آمده در گذشته، در جدول (۶) مقایسه شده است. همان‌طور که مشاهده می‌شود در هر دو روش استفاده‌شده، شامل جنگل تصادفی بیز و همچنین استفاده از ویژگی‌های n-گرمی حروف در کنار دسته‌بند بیز، دقت بالاتری نسبت به بررسی انجام‌شده در گذشته به‌دست آمده است.

(جدول-۶): مقایسه پژوهش حاضر با بررسی انجام‌شده در گذشته بر متن‌های بلاگ

(Table-6): Comparison of proposed method with other text blog researches

	ویژگی‌های مورد استفاده	الگوریتم دسته‌بندی	دقت دسته‌بندی
پژوهش انجام‌شده توسط لو در سال ۲۰۱۰	ویژگی‌های مستقل از متن و ویژگی‌های تجزیه‌ای با طول متغیر	بیز ساده	٪ ۷۳/۵۷
		ماشین بردار پشتیبان	٪ ۸۶/۲۴
		SVM_R	٪ ۸۸/۵۶
روش پیشنهادی	ویژگی‌های مستقل از متن و n-گرمی حروف	جنگل تصادفی بیز	۸۹٪/۵
	ویژگی‌های n-گرمی حروف		٪ ۹۱/۹۶

در جدول (۷) مقایسه زمانی جنگل تصادفی بیز با روش ماشین بردار پشتیبان و روش بیز ساده نشان داده می‌شود. در این جدول زمان آموزش مدل و متوسط زمان آزمون برای هر نمونه نمایش داده می‌شود.

(جدول-۷): مقایسه زمانی روش جنگل تصادفی بیز

(Table-7): Time comparison of Bayes Random Forest

روش دسته‌بندی	زمان آموزش(ثانیه)	زمان تست(ثانیه)
جنگل تصادفی بیز	۴	۲
ماشین بردار پشتیبان	۵	۱
بیز ساده	۹	۱

۶- نتیجه‌گیری

در این مقاله یک الگوریتم جدید دسته‌بندی با عنوان جنگل تصادفی بیز ارائه شد که در دسته‌بندی متون با استفاده از ویژگی‌های مستقل از متن توانسته است، بهبود خوبی در دقت دسته‌بندی نسبت به الگوریتم‌های مشابه داشته باشد.

در این روش از ایدهٔ الگوریتم‌های تجمیعی و همچنین روش بیز و درخت بیز ساده استفاده شده است. در ارائه این روش، شکستن مسأله به مسأله‌های کوچکتر که موجب می‌شود، الگوریتم دسته‌بندی عمل یادگیری مدل را بهتر انجام دهد و از طرفی تشخیص دستهٔ هر نمونه با استفاده از الگوریتم بیز که با وجود سادگی، در دسته‌بندی متون به‌طورمعمول دقت خوبی ارائه می‌دهد و همچنین کاهش میزان خطای دسته‌بندی که یکی از دلایل استفاده از الگوریتم‌های تجمیعی است، مد نظر بوده است. با توجه به دقت ۸۹/۵ ٪ که با استفاده از این روش در حل مسأله تشخیص جنسیت به‌دست آمده است، می‌توان نتیجه گرفت که تفاوت‌های جنسیتی در نوشته‌های فضای مجازی بلاگ‌ها نیز وجود دارند. از طرفی با بررسی نقش هر یک از انواع ویژگی‌های مستقل از متن، مشخص شد ویژگی‌های واژگان گرامری بیشترین تأثیر در تفکیک نمونه‌ها را داشته و پس از آن ویژگی‌های فاکتور واژه در دسته‌بندی مؤثر بوده‌اند. همچنین به‌دلیل وجود غلط املایی و استفاده از واژگان اختصاری غیر رسمی در متون فضای مجازی، ویژگی‌های وابسته به متن، به‌طور خاص nگرمی حروف با طول یک تا چهار، استخراج شدند، و نشان داده شد این نوع ویژگی، در دسته‌بندی جنسیت متن بلاگ‌ها، و افزایش دقت دسته‌بندی تأثیر به‌سزایی دارد.

۷- مراجع 7- References

[1] N. Cheng, R. Chandramouli , and K.P. Subbalakshmi, "Author gender identification from text," *Elsevier. Digital investigation*, vol. 8, pp. 78-88, 2011.

[2] Z. Miller, B. Dickinson, and W. Hu, "Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features," *International Journal of Intelligence Science*, 2012.

[3] K. Mita, A. Mukesh, "Automatic Classification of Unstructured Blog Text," *Journal of Intelligent Learning Systems and Applications*, vol. 5, pp. 108-114, 2013.

[4] S. Argamon, M. Koppel, J. Fine, and A. Shimoni, "Gender, Genre and Writing Style in Formal Written Texts," *Dept of Computer Science. Illinois Institute of Technology*, pp. 321-346, 2003.

[5] G. Murugaboopathy, S. Hariharasitaraman, and N. Sankarram, "Appropriate Gender Identification from the Text," *International Journal of Emerging Research in Management &Technology*, 2013.

[6] A. Narayanan, H. Paskov and N. Z. Gong, "On the Feasibility of Internet-Scale Author Identification," *IEEE Symposium on Security and Privacy*, vol. 46, 2012.

[7] Y. Zhang, Y. Dang and H. Chen, "Gender classification for Web Forums," *IEEE Trans*. On Systems, vol. 41, no. 4, 2011.

[8] A. Mukherjee and B. Liu, "Improving Gender Classification of Blog Authors," *Conference on Empirical Methods in Natural Language Processing*, 2010, pp. 207-217.

[9] S. Nowson and J. Oberlander, "The identity of bloggers: Openness and gender in personal weblogs," *in proc. AAAI Spring Symposia Comput. Approaches Analyzing Weblogs*, Stanford,CA, 2006.

[10] S. Hota, S. Argoman, M. Koppel, "performing gender Automatic stylistic analysis of shakespeare's characters," *in Proc. Digital IIumanit. Conf*, 2006, pp. 100-106.

[11] R.S. Forsyth and D.I. Holmes, "Feature finding for text classification," *Literary Linguistic Compute*., vol. 11, No. 4, pp. 163-174, 1996.

[12] M. Koppel, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, 2002.

[13] N. Cheng, X. Chen, R. Chandramouli and K.P. Subbalakshmi, "Gender Identification from E-mails," *computational intelligence and data mining*, pp. 154-158, 2009.

[14] M. Corney, "Gender-preferential text mining of e-mail discourse," *18th Annual Computer Security applications Conference*, 2002.

[15] R. Kohavi, "Scaling Up the Accuracy of NaiveBayes Classifers a Decision Tree Hybrid," *Second International Conference on Knoledge Discovery and Data Mining*, 1996, pp. 202-207.

هدیه ساجدی درجهٔ دکترای خود را در رشتهٔ مهندسی کامپیوتر-هوش مصنوعی در سال ۱۳۸۹ از دانشگاه صنعتی شریف اخذ کرده و هم‌اکنون عضو هیأت علمی گروه علوم کامپیوتر در دانشگاه تهران است.

زمینه‌های پژوهشی ایشان عبارت است از: یادگیری ماشین، داده‌کاوی و پردازش تصویر.

نشانی رایانامه ایشان عبارات است از:

hhsajedi@ut.ac.ir

مهناز تسلیمی درجه کارشناسی ارشد خود را در رشته مهندسی کامپیوتر در سال ۱۳۹۴ از دانشگاه آزاد، واحد قزوین اخذ کرده و زمینه‌های پژوهشی ایشان عبارت یادگیری ماشین و داده‌کاوی است

نشانی رایانامه ایشان عبارت است از:

mahnaz_taslimi@yahoo.com