بسم تعالی

چکیده

امروزه با افزایش محبوبیت و همچنین نرخ استفاده از شبکه‌های اجتماعی، حجم بسیار گسترده‌ایی از اطلاعات به وجود آمده است. اهمیت و ارزشمندی این اطلاعات زمانی مشخص می‌شود که بتوان تحلیل‌های مناسبی را در مسائل گوناگون، با استفاده از این اطلاعات انجام داد. از این‌رو محققان به این شبکه‌های اجتماعی نگاه ویژه‌ایی به موجب چالش‌های گوناگون موجود در آنها دارند. توییتر، یکی از محبوب‌ترین، این شبکه‌های اجتماعی است که نرخ استفاده از آن روبه افزایش بوده و اقشار بسیار مختلفی از آن استفاده می‌کنند. یکی از موضوعات داغی که امروزه می‌توان برای بررسی آن از اینگون اطلاعات استفاده کرد، تشخیص شخصیت افراد است که ارتباط میان فعالیت افراد در شبکه‌های اجتماعی با شخصیت آنها در کارهای گذشته اثبات شده است. همچنین شخصیت‌شناسی از این جهت که اهمیت آن در تصمیم‌گیری‌های فرد، روابط عاطفی و سلامت روان فرد تاثیر دارد کاملا مشخص است. هدف اصلی در این پژوهش ساخت یک سیستم هوشمند تشخیص شخصیت با استفاده از فعالیت‌هایی است که یک فرد در شبکه اجتماعی توییتر انجام می‌دهد.

واژه‌های کلیدی: داده‌کاوی، MBTI، تیپ شخصیتی مایرز– بریگز، Tweeter Mining، Personally Detection

۱ مقدمه

دنیای امروز مملو از افردی است که بیشترین زمان خود را برای ارتباط با دیگران در اینترنت و شبکه‌های اجتماعی مختلف، صرف می‌کنند. یکی از قابلیت‌های مهم شبکه‌های اجتماعی، پخش هر نوع اطلاعاتی در ابعاد مختلف است که این امر باعث به وجود آمدن حجم بسیار گسترده‌ایی از اطلاعات در مورد همه زمینه‌ها شده است. این حجم از اطلاعات در زمینه‌های گوناگونی نظیر پیش‌بینی انتخابات، تشخیص شایعات و یا حتی پیش‌بینی شخصیت هر فرد نیز می‌تواند مفید واقع شده و از آنها بهره‌برداری کرد. شخصیت هر فرد با رفتار وی ارتباط مستقیمی داشته و از این رو که هر فرد خود واقعیش را در شبکه‌های اجتماعی بروز میدهد، بنابراین می‌توان به درستی نتایج به دست آمده از اطلاعات و رفتار کاربران در شبکه‌های اجتماعی بیشتر تکیه کرد. کاربردهای مشخص بودن شخصیت افراد نه تنها در زمینه‌های تحقیقات جامعه شناختی بلکه در دیگر زمینه‌هایی همچون سیستم‌های پیشنهاد دهنده، دستیارهای هوشمند مورد استفاده در تلفن‌های همراه و یا حتی بازارهای کسب و کار مورد استفاده قرار گیرد. در علم روانشناسی چندین نظریه در مورد شخصیت افراد وجود دارد که دو مورد از معروف‌ترین آنها نظریه BigFive [۱] و نظریه MBTI [۲] است. با این حال پس از بررسی‌های گوناگون و مرور ادبیات، در این پژوهش نظریه MBTI به دلیل وجود محبوبیت بالا و دقت خوب آن در میان روانشناسان انتخاب شده است.

مجموعه داده استفاده شده در این پژوهش شامل ۱۱۷۸ کاربر انگلیسی زبان، از شبکه‌اجتماعی توییتر است که به صورت خود اظهاری تیپ شخصیت خود را در این شبکه

اجتماعی اعلام کرده‌اند. پس از به دست آوردن اطلاعات کاربری این افراد در این شبکه‌اجتماعی نوبت به مرحله توییت‌های این افراد می‌رسد که در این مرحله توییت‌ها آنها جمع‌آوری شده است. در انتها برای ساخت سیستم هوشمند تشخیص شخصیت افراد بر اساس محتوای تولید شده توسط آنها در توییتر از یک الگوریتم *** استفاده شده است.

2 پیش زمینه و کارهای گذشته

در این قسمت به تشریح نظریه MBTI که اساس آزمایشات موجود در این پژوهش خواهد بود پرداخته و در ادامه در مورد پژوهش‌های انجام شده در این حوزه و کاربردهای آن بحث خواهیم کرد.

2.1 نظریه MBTI

نظریه MBTI [۲] انسان‌ها را نظر شخصیت به ۱۶ نوع تقسیم‌بندی می‌کند، که مبنای آن متشکل از ۴ ویژگی اصلی است که در ادامه در مورد آن بحث خواهیم کرد:

- انرژی: دومین شاخص که همان حسی در مقابل شهودی است مشخص می‌کند که فرد به چه صورتی اطلاعات را از محیط پیرامونش دریافت می‌کند که می‌تواند به صورت حسی و یا شهودی باشد. به این ترتیب افرادی که بیشتر به اطلاعاتی توجه داشته باشند که از طریق حواس پنج گانه به دست بیاید حسی در غیر اینصورت شهودی خواهد بود.
- ذهنی: این جنبه مشخص می‌کند که فرد درک خود از پیرامون را از کدام منبع تامین می‌کند (درونی یا بیرونی) و ضمنا جهت مسیر تمرکز او بیشتر به کدام سمت است. فرد برون‌گرا در مجموع بر دنیای بیرون از خودش متمرکز است حال آنکه فرد درون گرا در مجموع بر دنیای درونی خودش متمرکز است و انرژی این فرد بیشتر در دنیای درونی نمود دارد.
- تصمیم گیری: این شاخص که فکری در مقابل احساسی است، بیان‌کننده آن است که فرد مورد نظر اطلاعاتی را که به دست می‌آورد چگونه تحلیل خواهد کرد. مبتنی بر منطق و یا احساسات.
- سبک زندگی: دیگر شاخص این نظریه در مورد روش اجرای تصمیمات فرد است. به این ترتیب فرد می‌تواند یا به صورت ساختارگرا و یا منعطف در مقابل اجرای تصمیمات خود قرار بگیرد.

2.2 کارهای گذشته

از مهمترین شبکه‌های اجتماعی که در سال‌های گذشته از طرف محققان نگاه ویژه‌ایی به آنها شده است توییتر و فیسبوک هستند، دلیل این امر بالا بودن و متنوع بودن جامعه آماری کاربران این شبکه‌ها است[۳]. به همین دلیل در این پژوهش از شبکه‌اجتماعی توییتر استفاده شده است. در بیشتر کارهای گذشته برای ساخت سیستم هوشمند تشخیص شخصیت با استفاده از متون تولید شده توسط کاربران بیشتر از ویژگی‌هایی مثل LIWC[1] MRC و یا DLA بوده است.

- محققان با استفاده از این دو ابزار ویژگی‌هایی را که می‌توان با استفاده از آنها، دقت خوبی برای

[1] Linguistic Inquiry and Word Count

پیش‌بینی شخصیت افراد به دست آورد، جمع‌آوری می‌کنند و سپس به ساخت این سیستم می‌پردازند[۴].

- Navonil Majumder و همکارانش [۵] با استفاده از نظریه BigFive نسبت به ساخت یک سیستم هوشمند به جهت تشخیص شخصیت نویسنده یک متن با استفاده از الگوریتم یادگیری عمیق اقدام کرده‌اند. آنها در این پژوهش برای هر یک ویژگی‌های شخصیتی موجود در این نظریه یک کلاس‌بند باینری طراحی کرده و در انتها از آنها استفاده کرده اند.
- TG Henkel و همکارانش [۶] در این پژوهش بر پایه نظریه MBTI و همچنین BigFive نسبت به بررسی ویژگی‌های شخصیتی هر مدیر پروژه برای مدیریت هرچه بهتر یک پروژه پرداخته‌اند. در این پژوهش از ۲۰۴ مدیر پروژه که تیپ شخصیتی خود را در هر دو نظریه اعلام کرده‌اند استفاده شده است.
- Srilakshmi Bharadwaj و همکارانش [۷] در یک پژوهش با استفاده از ابزار LIWC و EmoSenticNet بر روی متون تولید شده توسط کاربران در شبکه اجتماعی توییتر توانستند یک سیستم هوشمند تشخیص شخصیت ایجاد کنند. در این پژوهش برای دسته بندی افراد از نظر شخصیتی از نظریه MBTI و همچنین برای ساخت این سیستم از الگوریتم SVM استفاده شده است.

3 متودولوژی

3.1 مجموعه‌داده

مجموعه‌داده‌ای که در این پژوهش مورد استفاده قرار گرفته است، شامل ۱۱۷۸ کاربر شبکه‌اجتماعی توییتر بوده که در ۱۶ کلاس مختلف بر اساس نظریه MBTI تقسیم‌بندی شده است. برای هر کدام از افراد موجود در این مجموعه‌داده حداکثر ۳۰۰۰ توییت اخیر آنها جمع آوری شده که در نهایت این مجموعه‌داده چیزی بالغ بر ۳ میلیون توییت در خود جای داده است. در Figure *1* توزیع کاربران در این مجموعه‌داده قابل مشاهده است.

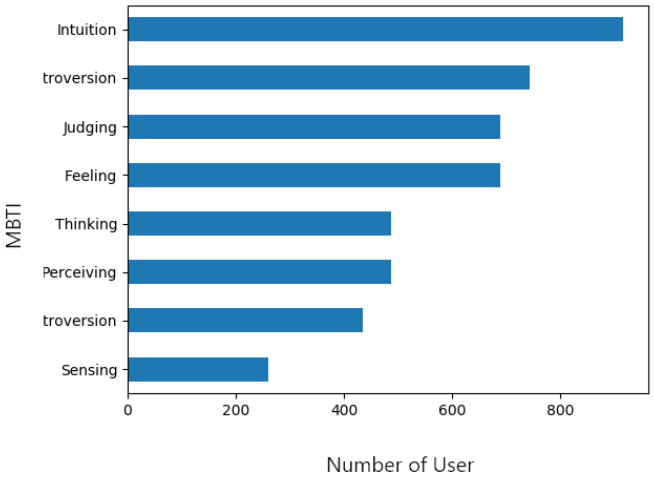

Figure 1 : Distribution of Dataset

3.2 پیش‌پردازش و ویژگی‌ها

یکی از مهم‌ترین بخش‌ها برای حل هر مسئله کلاس‌بند انجام پیش‌پردازش بر روی مجموعه‌داده آن مسئله است، تا به وسیله آن بتوان به بهترین کیفیت دست یافت. به همین منظور در این بخش مجموعه از فرآیندها برای رسیدن به این هدف انجام شده است که در Figure *2* مشخص است.

- حذف URLها
- حذف Mentionها
- حذف Hashtag ها
- Tokenizing: فرآیندی که طی آن کلماتی که باهم تشکیل یه جمله داده‌اند، شکسته شوند. جداکردن کلمات یا براساس فاصله و یا دیگر جداکنندگان انجام می‌شود. همچنین در این فرآیند تمام اعداد و نشانه‌گذاری‌هایی که هیچ معنی ندارند حذف شده و در نهایت تمام کلمات به حروف کوچک تبدیل می‌شوند.
- حذف کلمات توقف: کلماتی که به وفور در متن آمده به عنوان کلمات ارتباط ظاهر می‌شوند. این کلمات به دلیل اینکه ممکن است در فرآیند کلاس‌بند، تاثیر مخرب داشته باشند، حتما باید حذف شوند.
- Steaming: فرآیندی که طی آن یک کلمه به حالت اصلی خود اصطلاحا ریشه برمی‌گردد. در این فرآیند تمام پیش‌وندها، میان‌وندها و پس‌وندها از کلمه مورد نظر حذف خواهد شد.
- وزن‌دهی به کلمات: در این پژوهش برای وزن‌دهی از روش TF / IDF[۲] استفاده شده است [۸] (مرجع اصلی TF / IDF). وزن‌دهی فرآیندی است که براساس آن به هر کلمه یک ارزش عددی طبق میزان تکرار آن در اسناد داده می‌شود.

برای محاسبه این معیار برای هر کلمه باید از معادله ۱ استفاده کرد.

$$W = TF(t, d) * IDF(t, D) \quad (۱)$$

$$TF(t.d) = log(1 + freq(t.d)) \quad (۲)$$

$$IDF(t, D) = log(\frac{N}{count(d \in D.\ t \in d)}) \quad (۳)$$

که در آن w وزنی است که به هرکلمه داده خواهد شد و N برابر تعداد کل متن‌ها خواهد بود. در معادله ۲ فراوانی کلمه مورد نظر در یک سند در مقیاس لوگاریتم به دست آمده و همچنین در معادله ۳ میزان فراوانی کلمه مورد نظر در کل اسناد در مقیاس لوگاریتم به دست می‌آید.

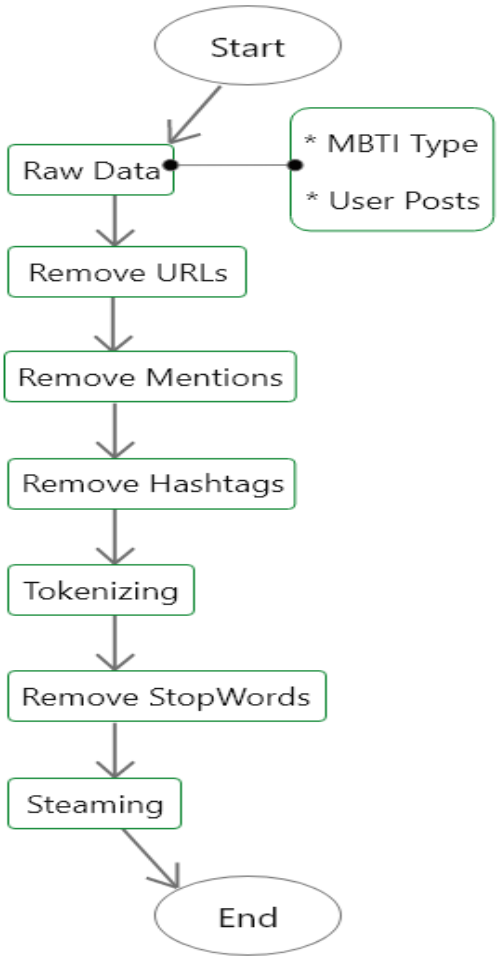

Figure 2 : Preprocessing flowchart

[۲] term frequency-inverse document frequency

3.3 مدل کلاس‌بند

در این پژوهش مسئله کلاس‌بند تشخیص شخصیت با ۱۶ کلاس به ۴ مسئله کلاس‌بند دودویی نگاشت داده شده است. به این صورت که به ازای هر کدام از ۴ بعد اصلی موجود در این نظریه یک کلاس‌بند دودویی با استفاده از الگوریتم XGboost ساخته شده است. در معادله ۴ نگاشت صورت گرفته قابل مشاهده است.

$$\hat{y} = argmax_{k \in \{1.K\}} f_k(x) \qquad (۴)$$

که در آن k برابر تعداد برچسب‌ها که برای این مسئله ۴ درنظر گرفته شده است و fk لیستی از کلاس‌بندها موجود برای مسئله است.

برای ساخت سیستم هوشمند تشخیص شخصیت افراد از الگوریتم‌های XGboost [۹], Logistic Regression و همچنین Random Forest که در sklearn پیاده‌سازی شده است همراه با پارامترهای استاندارد استفاده شده، که نتایج آن در *Tabl e* ۱ قابل مشاهده است.

Tabl e ۱ : Result for personality classifications

	Task	R	P	F
XGboost	I-E	۶۲/۷۲	۳۹/۲۸	**۶۲/۷۲**
	S-N	۷۸/۴۱	۲۲/۴۵	**۷۸/۴۱**
	T-F	۵۹/۱۳	۴۳/۳۱	**۵۹/۱۳**
	J-P	۵۹/۱۳	۴۳/۰۰	**۵۹/۱۳**
Logistic Regression	I-E	۶۶/۰۷	۴۰/۵۵	۲۶/۶۷
	S-N	۷۸/۶۶	۲۲/۰۳	۴/۶۰
	T-F	۵۹/۹۰	۴۳/۲۰	۳۲/۷۶
	J-P	۶۱/۹۵	۴۳/۵۴	۳۶/۷۵
Random Forest	I-E	۶۵/۸۱	۴۰/۲۱	۲۵/۷۰
	S-N	۷۸/۹۲	۲۲/۴۳	۴/۶۵
	T-F	۵۷/۸۴	۴۱/۰۹	۲۱/۱۵
	J-P	۵۹/۳۸	۴۱/۳۴	۳۰/۷۰

باتوجه به نتایج موجود در *Tabl e* ۱ بهترین الگوریتم از نظر f1-score برای هر ۴ کلاس‌بند طراحی شده، XGboost است. پیش بینی شخصیت کاری دشوارتر است ، اما مطالعه ما نتایج امیدوار کننده ای را نشان می دهد.

4 جمع‌بندی

[1] L. R. Goldberg, “Analternative “description of personality”. The Big Five factor structure,” *Journal of Personality and Social Psychology,* vol. 59, no. 6, pp. 1216-1229, 1990.
[2] I. Briggs-Myers, and P. B. Myers, “Gifts differing: Understanding personality type,” 1995.
[3] S. Kumar, F. Morstatter, and H. Liu, *Twitter data analytics*: Springer, 2014.
[4] T. Tandera, D. Suhartono, R. Wongso, and Y. L. Prasetio, “Personality prediction system from facebook users,” *Procedia computer science,* vol. 116, pp. 604-611, 2017.
[5] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, “Recent trends in deep learning based personality detection,” *Artificial Intelligence Review,* pp. 1-27, 2019.
[6] T. G. Henkel, G. Haley, D. T. Bourdeau, and J. Marion, “An Insight to Project Manager Personality Traits Improving Team Project Outcomes,” *Graziadio Business Review,* vol. 22, no. 2, pp. 1, 2019.
[7] S. Bharadwaj, S. Sridhar, R. Choudhary, and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator (MBTI)-A Text Classification Approach." pp. 1076-1082.
[8] J. Ramos, "Using tf-idf to determine word relevance in document queries." pp. 133-142.
[9] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, “Xgboost: extreme gradient boosting,” *R package version 0.4-2,* pp. 1-4, 2015.