مجموعه‌داده فارسی برای تشخیص شخصیت در بستر توییتر

نام دانشجو: زهرا انوریان
استاد راهنما: دکتر صالح اعتمادی
دانشگاه علم و صنعت ایران
اردیبهشت ۱۴۰۰

فهرست

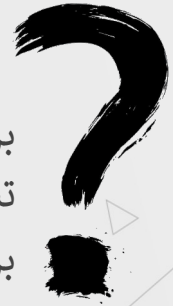
۱

مقدمه

مقدمه

- شخصیت و ویژگی‌های شخصیتی تأثیر زیادی بر زندگی ما، انتخاب‌ها، ترجیحات و خواسته‌های ما دارد.
- شخصیت، مجموعه مشخصه‌های رفتار، شناخت و الگوهای عاطفی است که از عوامل بیولوژیکی و محیطی نشأت می‌گیرد.
- اغلب این ویژگی‌های شخصیتی توسط روان‌شناسان و با استفاده از پرسشنامه بدست می‌آید.
- در دنیای امروز، وجود متون نوشته شده توسط افراد، به خصوص در فضای مجازی، این فرصت را برای روان‌شناسان و محققان فراهم نموده که با استفاده از این متون، ویژگی شخصیتی افراد را بدست آوردند.

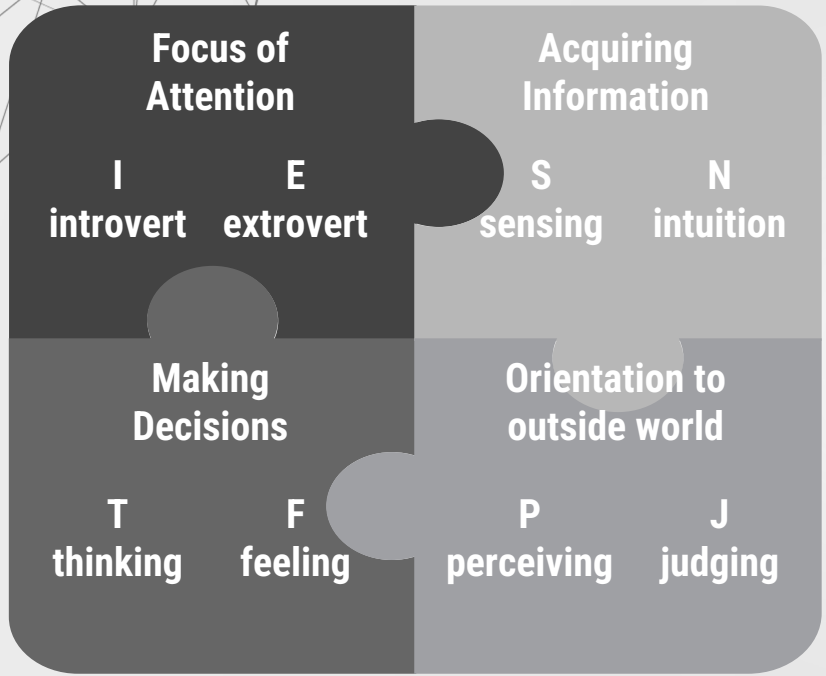
شرح مسئله



برای دستیابی به سیستمی هوشمند که ویژگی‌های شخصیتی افراد را بتواند تشخیص دهد، نیازمند مجموعه‌داده‌ای از متون نوشته شده توسط افراد با برچسب ویژگی شخصیتی‌شان می‌باشیم.

برای جمع‌آوری این مجموعه‌داده، ما برای متون از توییت‌های افراد در بستر توییتر و همچنین ویژگی شخصیتی مایرز-بریگز (MBTI) را برای برچسب داده‌ها استفاده کردیم.

مدل مایرز-بریگز(MBTI)



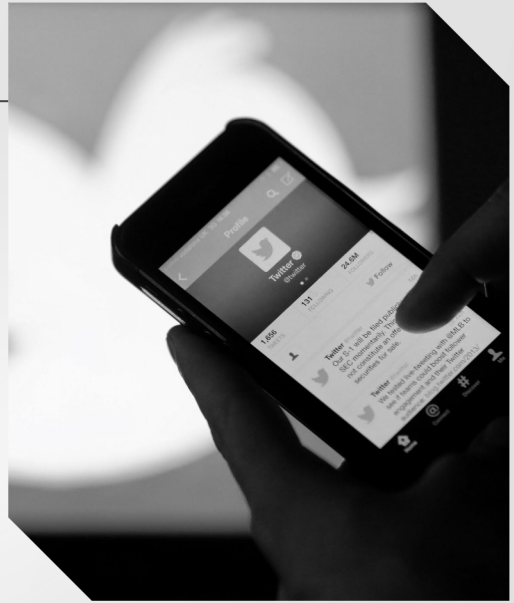
- مدل روان‌شناختی مایرز-بریگز دارای چهار ویژگی شخصیتی می‌باشد:
  - **درون‌گرا - برون‌گرا**
  - **حسی - شمی**
  - **منطقی - احساسی**
  - **ادراکی - قضاوتی**

۲

مروری بر کارهای مرتبط

مروری بر کارهای مرتبط

توییتر



- **پژوهشگران**
  - آقایان باربارا و دیرک
- **اندازه مجموعه‌داده**
  - ۱.۲میلیون توییت از ۱۵۰۰ کاربر توییتر
- **نحوه جمع‌آوری مجموعه‌داده**
  - جمع‌آوری توییت‌های کاربرانی که ویژگی شخصیتی خود را اعلام کردند.
  - جمع‌آوری توییت‌هایی که دارای یکی از ۱۶ویژگی شخصیتی MBTI بودند.

مروری بر کارهای مرتبط

ردیت

- **پژوهشگران**
  - ماتژ جورکوویچ و یان اشنایدر
- **اندازه مجموعه‌داده**
  - ۳۵۴۹۹۶ پست از ۹۸۷۲ کاربر ردیت
- **نحوه جمع‌آوری مجموعه‌داده**
  - جمع‌آوری فلیرزهایی که دارای یکی از ۱۶ویژگی شخصیتی MBTI بودند.
  - تمیز کردن اطلاعات بدست آمده از روش اول
  - جستجوی عبارت "I am an <TYPE>" در نظرهای پست‌های مربوطه و جمع‌آوری کاربران

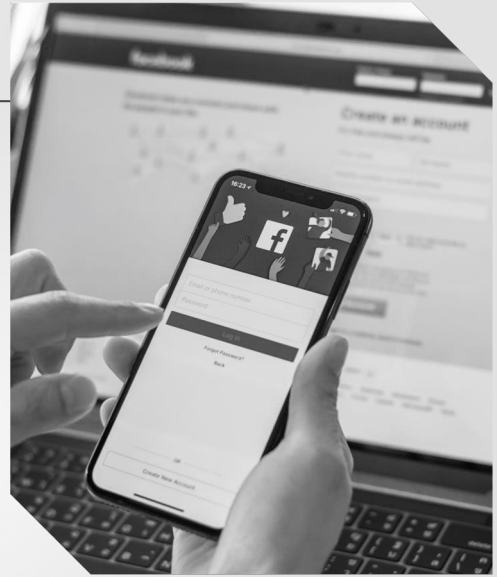
مروری بر کارهای مرتبط

پاندورا

- **پژوهشگران**
  - ماتژ جورکوویچ و همکاران
- **اندازه مجموعه‌داده**
  - ۱۷میلیون نظر از بیش از ۱۰هزار کاربر ردیت
- **نحوه جمع‌آوری مجموعه‌داده**
  - برای ویژگی شخصیتی MBTI از مجموعه‌داده MBTI9K استفاده کردند.
  - برای ویژگی شخصیتی Enneagram به طور دستی کاربرانی که ویژگی خود را در فلیرزهایشان اعلام کرده بودند را جمع‌آوری کردند.
  - برای ویژگی شخصیتی پنج‌عامله، جستجو در نظرهایی که در زیر پست‌های مربوطه از کاربران بوده، جمع‌آوری کردند.

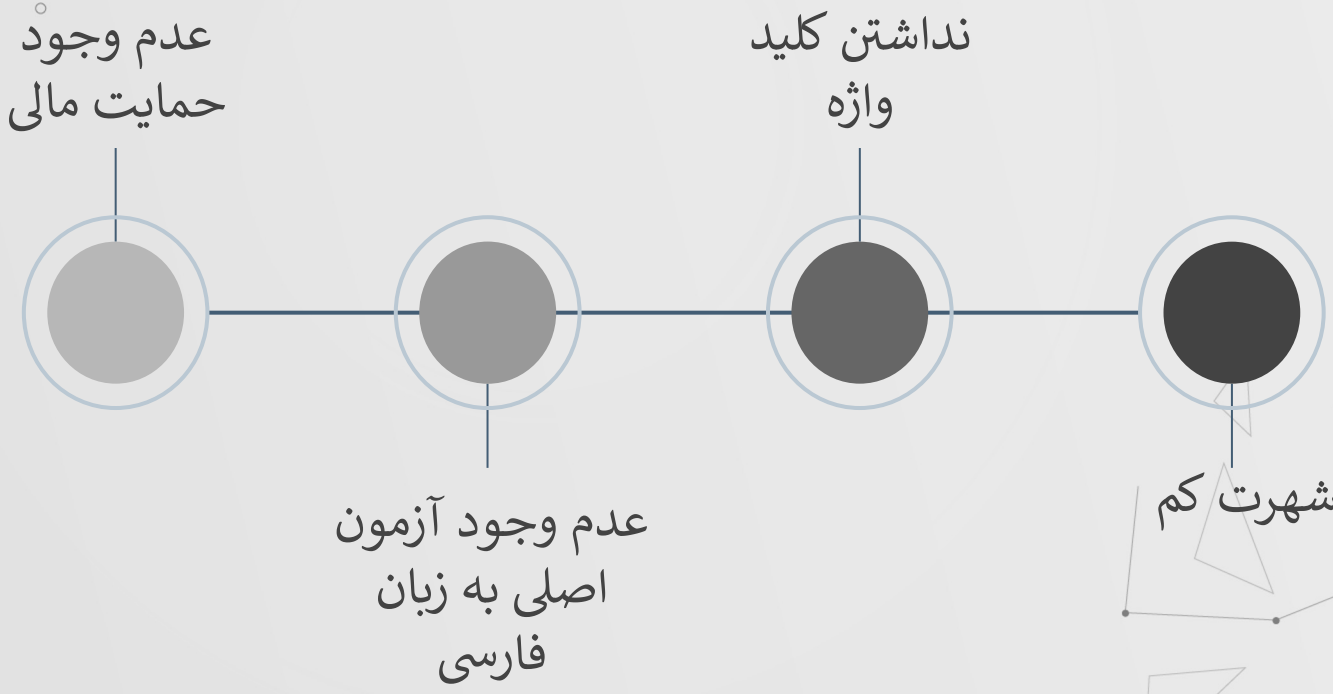
مروری بر کارهای مرتبط

فیس‌بوک



- **پژوهشگران**
  - دیوید استیلول و میشال کوسینسکی
- **اندازه مجموعه‌داده**
  - بیش از ۶ میلیون داوطلب فیس‌بوک
- **نحوه جمع‌آوری مجموعه‌داده**
  - طراحی پرسشنامه
- **در سال ۲۰۱۸ انتشار این مجموعه‌داده متوقف شد.**

دلیل عدم انتخاب مدل پنج‌عامله






۳

جمع‌آوری مجموعه‌داده

جمع‌آوری مجموعه‌داده

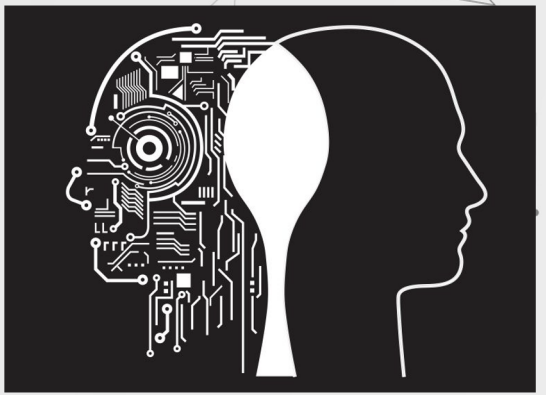
۱ جستجو کلید واژه

۲ پرسشنامه

۳ مقایسه روش‌های پیشنهادی

جستجو کلید واژه

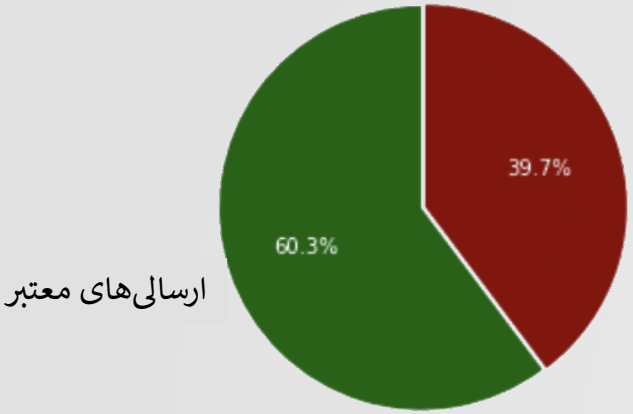
- از طریق دو منبع اصلی که افراد ویژگی شخصیتی خود را ذکر کردند.
  - بیو
  - توییت‌ها
- نحوه جمع‌آوری داده از طریق **بیو**
  - جمع‌آوری چند کاربر که دارای دنبال‌کننده و دنبال‌شوندگان زیادی هستند، به عنوان گره‌های اصلی
  - شروع جستجو از گره‌های اصلی و بررسی بیوهایی که دارای یکی از ۱۶ ویژگی شخصیتی MBTI هستند.
- نحوه جمع‌آوری داده از طریق **توییت**
  - جستجو ۱۶ ویژگی شخصیتی MBTI با استفاده از ویژگی جستجو پیشرفته توییتر و جمع‌آوری توییت‌های مربوطه

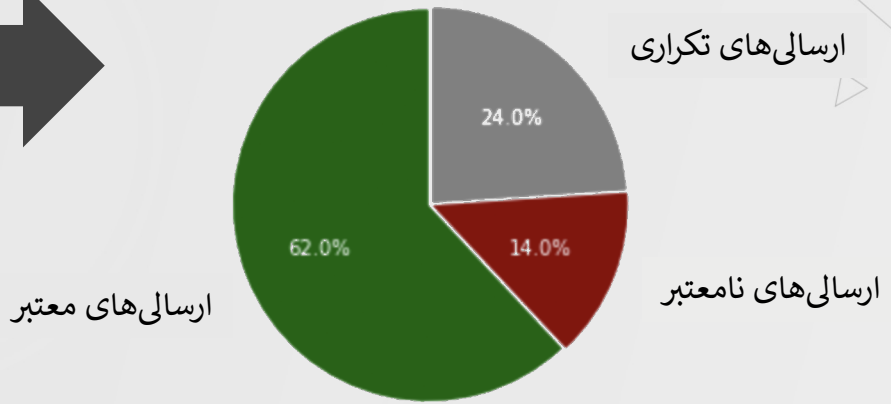


پرسشنامه

- طراحی پرسشنامه و توزیع آن در کانال‌های دانشگاهی و گروه‌های متفرقه تلگرامی
- شرط تکمیل پرسشنامه، داشتن حساب کاربری عمومی توییتر
- اطمینان دادن به افراد برای حفظ اطلاعات شخصی آن‌ها مانند آیدی حساب کاربری
- **چالش:**
  - دریافت حجم قابل توجهی از داده‌های نامعتبر به دلیل عدم مطالعه کامل توضیحات داده شده در پرسشنامه توسط افراد
- **راه‌حل:**
  - طراحی پرسشنامه‌ای جدید با منطق انشعاب با استفاده از فرم‌های مایکروسافت
  - تغییر محیط توزیع پرسشنامه به بستر توییتر

نتایج پرسشنامه





آمار کلی روش‌های جمع‌آوری داده‌ها

	روش ۱		روش ۲	
	بیو	توییت	پرسشنامه	کل
شمار توییت‌ها	۳۰۹۳۶۴	۱۱۳۴۲۹۴	۱۰۸۸۷۴	**۱۵۵۲۵۳۲**
شمار کاربران	۲۱۰	۶۵۳	۷۵	**۹۳۸**

مقایسه روش‌های پیشنهادی

چالش: یافتن جامعه هدفی که علاقه‌مند به این موضوع هستند ویا آگاهی و دانشی در این باره دارند.

راه‌حل: درخواست توزیع مجدد پرسشنامه از افرادی که دارای دنبال‌کنندگان بالایی هستند.

- عدم حمایت مالی
- کمبود ابزار مناسب
- تجربه ناکافی

چالش: متقاعد کردن افراد برای تکمیل کردن پرسشنامه

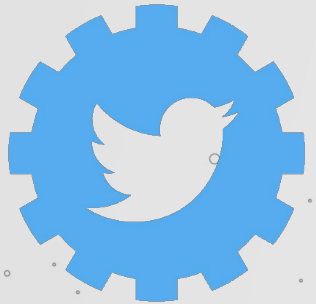
راه‌حل: استفاده از روش‌های انگیزشی و گاهاً اجباری ✖

۴

آماده‌سازی مجموعه‌داده

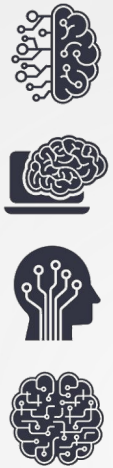
جمع‌آوری توییت‌ها

- فیلتر کردن کاربران نامعتبر
  - نداشتن حساب کاربری عمومی
  - نداشتن بیشتر از ۱۰۰ توییت
- استفاده از **سلنیوم**
  - خلاف قوانین توییتر بودن
  - استفاده از روش reCAPTCHA هر چند دقیقه یکبار
    => مانع جمع‌آوری توییت‌ها
  - زمان‌بر بودن
- استفاده از **API توییتر**



تمیز کردن داده‌ها

- باید توییت‌های جمع‌آوری شده را تمیز و تاحدودی ناشناس و قابل استفاده کنیم.
  - جایگزینی تمام حساب‌های کاربری با کلمه منحصر به فرد <"USERNAME">
  - جایگزینی تمام لینک‌ها با کلمه منحصر به فرد <"LINK">
  - جایگزینی تمام هشتگ‌ها با کلمه منحصر به فرد <"HASHTAG">
  - استخراج توییت‌های فارسی زبان با استفاده از ویژگی API توییتر
- در نظر داشتیم که تمام شکلک‌های موجود در متن توییت‌ها را با کلمه منحصر به فرد <"EMOJI"> جایگزین کنیم اما به دلیل اهمیت بالای آن در متن این کار را به عهده‌ی استفاده‌کننده از این مجموعه داده گذاشتیم با بر اساس نیاز خود و پژوهش مورد نظرش تصمیم گیری کند.

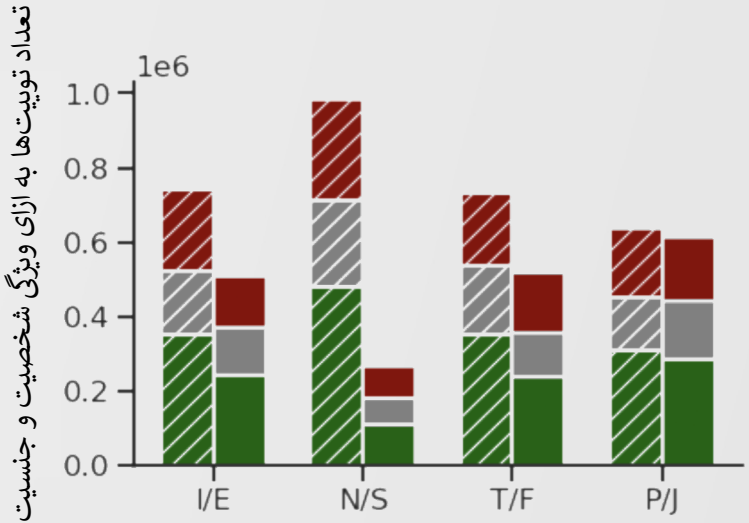


۵

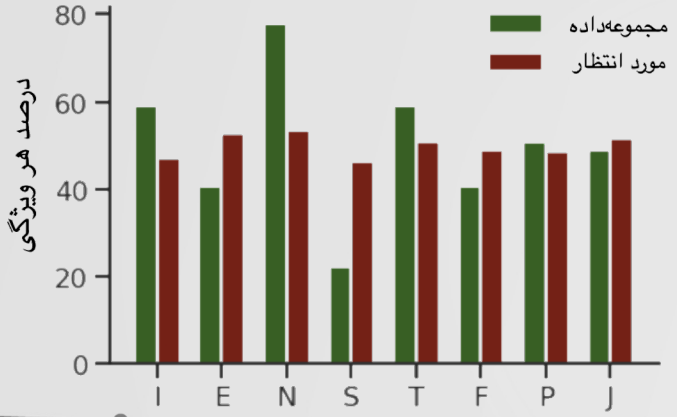
تحلیل و ارزیابی مجموعه‌داده

تحلیل مجموعه‌داده

- ۹۳۸ کاربر و توییت‌های جمع‌آوری شده ۱۵۵۲۵۳۲ و به طور متوسط هر کابر دارای ۱۶۵۵ توییت
- برچسب‌ها را به ۴ دسته تقسیم کردیم **I/E , N/S , T/F , P/J** :
- با این فرض که این چهار ویژگی مستقل از هم هستند.



تحلیل مجموعه‌داده



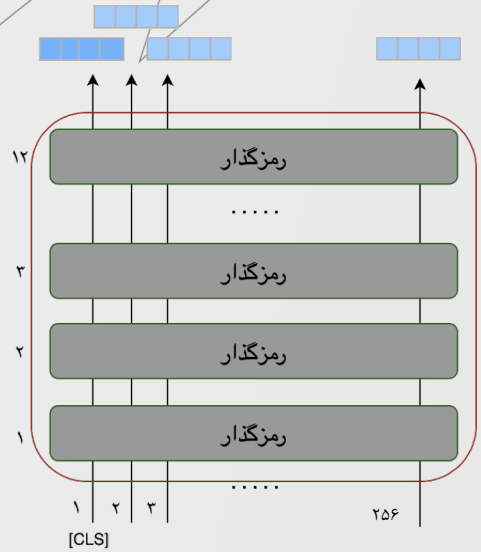
- مقایسه با نتایج بدست آمده از جمعیت ایران
  - جمعیت ایران دارای بیش از ۸۲ میلیون نفر، ۶۲۵۱۹ نفر در سایت آزمون MBTI شرکت کرده‌اند.
  - ما مقایسه‌ای میان درصد تشکیل‌دهنده هر یک از چهار ویژگی جمعیت ایران و مجموعه‌داده‌ی جمع‌آوری شده انجام دادیم.
  - افراد درون‌گرا کمتر از شبکه‌های مجازی در جامعه حضور دارند زیرا راحت‌تر می‌توانند ابراز احساسات در فضای مجازی کنند.
  - در بستر توییتر افراد به طور غیر مستقیم و از کنایه و تشبیه در صحبت‌هایشان استفاده می‌کنند.

آماده‌سازی مجموعه‌داده برای ارزیابی

- تقسیم داده‌ها به چند بخش به دلیل محدودیت شمار توکن‌های قابل پردازش مدل برت
- برای تقسیم داده‌ها به ۲ دسته: آموزش و آزمایش، باید همه توییت‌های یک کاربر در یک دسته قرار داشته باشند.
- روند تقسیم داده‌ها را ۵ بار با تقسیم‌بندی تصادفی مختلف تکرار کردیم تا از **تأثیرگذاری ترتیب داده‌ها** بر روی ارزیابی مدل جلوگیری کنیم.
- برای **کاهش تأثیر عدم تعادل** برخی از دسته‌ها بر تقسیمات، ابتدا از روش stratified استفاده نمودیم و در هنگام آموزش و آزمایش داده‌های هر برچسب را با استفاده از subsampling یکسان نمودیم.
- در نهایت تصمیم گرفتیم:
  - حداکثر طول برای هر ورودی مدل برت را ۲۵۶ توکن در نظر بگیریم.

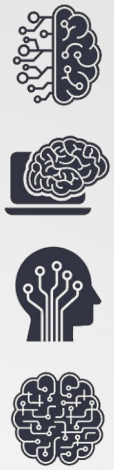
معرفی مدل پایه

- مدل پارس‌برت
  - مبتنی بر مدل برت
  - از قبل بر روی متون فارسی آموزش دیده
- استفاده از یک دسته‌بند لاجستیک رگرسیون در بالای رمزگذاری [CLS]
- استفاده از subsampling به صورت تصادفی حجم ورودی مدل را به قدری کاهش دادیم که مدت اجرای هر آموزش به ۱۵ الی ۲۰ دقیقه برسد.



نتایج بدست آمده بر حسب معیار f1-score

شمار تکرار	I/E	N/S	T/F	P/J
۱	۵۶.۶۹	۵۷.۷۵	۵۶.۹۳	۵۸.۱۵
۲	۵۶.۲۷	۵۸.۱	۵۷.۵۱	۵۶.۳۱
۳	۵۵.۳۲	۵۵.۹۲	۵۵.۹۳	۵۵.۱۴
۴	۵۷.۲۱	۵۷.۷۸	۵۷.۲۴	۵۷.۹۳
۵	۵۸.۴۱	۵۶.۹۷	۵۵.۱۲	۵۸.۴۸
میانگین	**۵۶.۷۶**	**۵۷.۳**	**۵۶.۵۵**	**۵۷.۲**



۶

نتیجه‌گیری و کارهای آینده

نتیجه‌گیری

- باتوجه به جالب بودن و مورد توجه قرار گرفتن این موضوع یعنی تشخیص شخصیت افراد، در حوزه پردازش زبان طبیعی، ما نیز در این پژوهش سعی کردیم تا اولین مجموعه‌داده فارسی این زمینه را جمع‌آوری کنیم تا دیگر محققان بتوانند از آن استفاده و مدل‌هایی با مجموعه‌داده فارسی طراحی کنند و به نتایج مناسب برسند ویا افراد علاقه‌مند به این زمینه با استفاده از روش‌ها و تحلیل‌های انجام شده و همچنین چالش‌های مطرح شده در این پژوهش، راه آسان‌تری برای جمع‌آوری مجموعه‌داده فارسی در پیش داشته باشند.

کارهای آینده

- ما قصد داریم برای بهبود داده‌های جمع‌آوری شده، عکس‌های افراد را به مجموعه‌داده جمع‌آوری شده اضافه کنیم تا از ویژگی‌های صورت نیز برای بهبود مدل‌سازی و درصد دقت بدست آمده، استفاده کنیم. علاوه‌براین غلبه بر چالش‌های موجود در جمع‌آوری داده‌ها با استفاده از آزمون‌های شناخته شده شخصیتی پنج‌عامله، می‌تواند دروازه‌ای برای پیشرفت بیشتر در عملکرد مدل‌سازی ویژگی‌های روانشناختی یک متن به زبان فارسی در نظر گرفته شود و همچنین جمع‌آوری داده‌های بیشتر به بهبود عملکرد مجموعه‌داده کمک می‌کند.

با سپاس از توجه شما

منابع

- Balmaceda, J. M., Schiaffino, S., and Godoy, D. How do personality traits affect communication among users in online social networks? *Online Information Review* (2014).
- Corr, P. J., and Matthews, G. *The Cambridge handbook of personality psychology*. Cambridge University Press, 2020.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.
- Farahani, M., Gharachorloo, M., Farahani, M., and Manthouri, M. Parsbert: Transformer-based model for persian language understanding. *ArXiv abs/2005.12515* (2020).
- Gjurković, M., and Šnajder, J. Reddit: A gold mine for personality prediction. in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media* (New Orleans, Louisiana, USA, June 2018), Association for Computational Linguistics, pp. 87–97.
- Jason Huggins, Paul Gross, J. T. W. Selenium automates browsers. https://www.selenium.dev/, 2004.
- Jurafsky, D., and Martin, J. H. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009.
- Kosinski, M., Stillwell, D., and Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences 110*, 15 (2013), 5802–5805.
- Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Wicaksana, A. S., Born, M. P., and König, C. J. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. in *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018, pp. 197–253.
- Martin, C. R. *Looking at type: The fundamentals*. Center for Applications of Psychological Type, 1997.
- Matz, S. C., Kosinski, M., Nave, G., and Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences 114*, 48 (2017), 12714–12719.
- Mehta, Y., Majumder, N., Gelbukh, A., and Cambria, E. Recent trends in deep learning based personality detection. *Artificial Intelligence Review* (2019), 1–27.
- Myers, I. B. The myers-briggs type indicator: Manual (1962).
- Plank, B., and Hovy, D. Personality traits on twitter—or—how to get 1,500 personality tests in a week. in *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (2015), pp. 92–98.
- Plank, B., and Hovy, D. Personality traits on twitter—or—how to get 1,500 personality tests in a week. pp. 92–98.
- Poria, S., Cambria, E., Hazarika, D., and Vij, P. A deeper look into sarcastic tweets using deep convolutional neural networks. *Proceedings of COLING* (10 2016).

  Rothmann, S., and Coetzer, E. P. The big five personality dimensions and job performance. *SA Journal of Industrial Psychology 29*, 1 (2003).
- team, G. What is recaptcha? https://www.google.com/recaptcha/about/.
- Yang, H.-C., and Huang, Z.-R. Mining personality traits from social messages for game recommender systems. *Knowledge-Based Systems 165* (2019), 157–168.