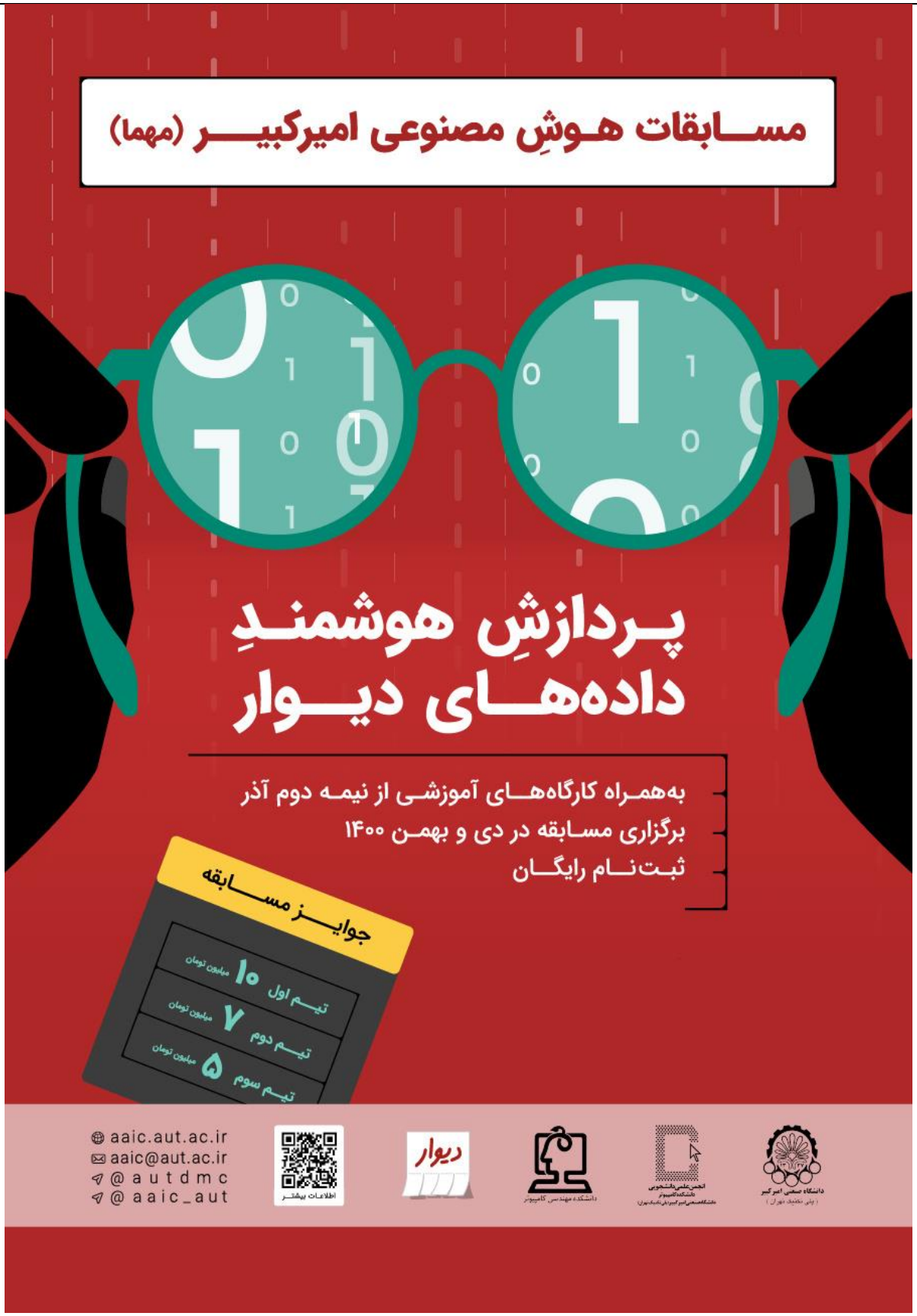
مسابقات هوشِ مصنوعی امیرکبیر (مهما)
پردازشِ هوشمندِ
داده‌های دیوار
به‌همراه کارگاه‌های آموزشی از نیمه دوم آذر
برگزاری مسابقه در دی و بهمن ۱۴۰۰
ثبت‌نام رایگان
جوایز مسابقه
تیم اول ۱۰ میلیون تومان
تیم دوم ۷ میلیون تومان
تیم سوم ۵ میلیون تومان
aaic.aut.ac.ir
aaic@aut.ac.ir
@autdmc
@aaic_aut
اطلاعات بیشتر
دیوار

توضیح مسئله:

روزانه حجم زیادی آگهی توسط کاربران به منظور انتشار بر روی سایت دیوار ارسال می‌شود و به منظور حفظ کیفیت آگهی‌های دیوار، هر یک از آگهی‌های ارسال شده قبل از انتشار توسط ناظران بررسی می‌شوند تا اطمینان حاصل کنیم که آگهی مورد نظر کیفیت و استاندارد لازم جهت نمایش به کاربران را داشته باشد.
بررسی این میزان آگهی روزانه، مستلزم صرف زمان و هزینه بسیار زیادی خواهد بود همچنین احتمال بروز خطای انسانی در این دست مسائل امری ناگزیر است. به همین دلیل صورت مساله‌ی ما در این چالش این خواهد بود که سعی کنیم با استفاده از تحلیل‌ها و مدل‌های دیتایی در هر چه بهتر شدن این فرآیند کمک کنیم.

معرفی دیتاهای مربوط به مسئله:

در این چالش بخشی از دیتاهای برچسب‌گذاری شده توسط ناظران در دسته‌ی خودروی سواری در اختیار شما قرار گرفته شده است که در آن ناظران بعد از بررسی اطلاعات مربوط به هر آگهی مشخص کرده‌اند که آگهی مربوطه باید قبول یا رد شود. همچنین از ایشان خواستیم که در صورت رد کردن آگهی یکی از دلایل رد آن را مشخص نمایند. (دقت کنید که ممکن است برخی از آگهی‌ها چند تا از قوانین و شرایط انتشار را نقض کرده باشند اما ناظران فقط یکی از این دلایل را برای ما مشخص کرده‌اند) شما در این چالش باید با استفاده از این داده‌ها، مدلی را آموزش بدهید که بتواند در نهایت رد یا تایید شدن آگهی را پیش‌بینی کند.

نام و توضیح مختصر دیتاها:

- دیتاست آموزش (train.parquet)

این دیتاست شامل چهار ستون با نام‌های post_id, post_data، reject_reason_id , review_label است. می‌توانید از این دیتاست به جهت آموزش مدل استفاده کنید.

- دیتاست ارزیابی اولیه(validation.parquet)

این دیتاست تنها ستون post_id, post_data را دارد و از ابتدای چالش نیز قابل مشاهده است. در طول فاز اول چالش شما می‌توانید پیش‌بینی‌های مدل خود در رابطه با رد یا تایید شدن هر کدام از آگهی‌ها را به دست آورده و آن را بر روی سایت بارگذاری کنید و پس از بارگذاری می‌توانید روی برد امتیازات دقت مدل خود را مشاهده کنید و با دقت بقیه‌ی گروه‌ها مقایسه کنید.

- دیتاست ارزیابی نهایی (final_test.parquet)

فرمت این دیتا نیز مانند دیتای ارزیابی اولیه است با این تفاوت که تنها در سه روز آخر چالش(فاز دوم) روی سایت قابل مشاهده خواهد بود و رتبه بندی نهایی گروه ها بر اساس این دیتا محاسبه خواهد شد. در طول این سه روز شما می‌توانید پیش‌بینی‌های خود را بر روی سایت بارگذاری و ارسال کنید و آخر هر روز دقت ارزیابی آخرین ارسال همه‌ی تیم‌ها حساب خواهد شد و بر روی سایت قابل مشاهده خواهد بود. همچنین دقت کنید که برای رتبه‌بندی و امتیاز نهایی تنها آخرین نتیجه‌ی ارسالی شما بر روی سایت در نظر گرفته خواهد شد و به نتایج ارسالی قبلی شما توجهی نخواهد شد.

- دیتاست توضیح مختصر دلایل رد: (reject_reasons_info.csv)

در این فایل می‌توانید id مربوط به هر دلیل رد را در کنار توضیح مختصری از آن دلیل مشاهده کنید.

نام و توضیح ستون‌ها:

post_id: شناسه‌ی آگهی.
post_data: اطلاعات یک آگهی که قبول یا رد انتشار یک آگهی توسط ناظران بر اساس آن صورت می‌گیرد. این فیلد خود شامل اطلاعات مختلفی از آگهی مثل عنوان، توضیحات، قیمت، سال ساخت و ... می‌شود که در حقیقت یک دیکشنری از این اطلاعات بوده و با json.dumps به فرمت

string تبدیل شده است. به جهت آشنایی بیشتر با محتوای این قسمت می‌توانید به این لینک مراجعه کنید و با اسکیمای مربوط به هرکدام از این فیلدها در آگهی‌های این دسته بیشتر آشنا شوید.
reject_reason_id: در صورتی که مقدار این فیلد برابر صفر باشد به این معنی است که آگهی مورد نظر قابلیت انتشار بر روی دیوار را داشته است مقادیر دیگر این فیلد بیانگر این موضوع است که آگهی بررسی شده رد شده است و دلیل رد آگهی چه بوده است. توجه کنید که ما در هنگام برچسب‌گذاری از ناظران خواستیم که فقط یکی از دلایل رد را برای هر آگهی مشخص کنند. برای مثال ممکن است یک آگهی به سه دلیل مختلف امکان انتشار بر روی دیوار را نداشته است اما ما از ناظران می‌خواهیم که فقط یکی از این موارد را در فرآیند برچسب‌گذاری مشخص کنند.
review_label: مشخص می‌کند که آگهی بررسی شده توسط ناظران رد یا تایید شده است. در صورت تایید مقدار این ستون برابر ۱ خواهد بود و در غیر این صورت مقداری برابر صفر خواهد داشت. در حقیقت شما باید مقادیر این ستون را پیش‌بینی کنید. دقت کنید که مدل شما در حقیقت احتمال تایید شدن آگهی را پیش‌بینی خواهد کرد و باید شامل مقادیر بین صفر و یک باشد. (لزومی ندارد که فقط صفر یا یک پیش‌بینی کند.)

متریک ارزیابی:

برای ارزیابی مدل‌ها و رتبه‌بندی و مقایسه‌ی تیم‌ها از متریک AUC استفاده خواهیم کرد. برای آشنایی بیشتر با این متریک میتوانید به این لینک مراجعه کنید. (دقت کنید که در اینجا ما یک مساله‌ی binary داریم و مدل شما تنها باید احتمال رد یا تایید شدن آگهی را پیش‌بینی کند.)
فرمت فایل ارسالی:
شما باید یک فایل csv با دقیقا تعداد سطرهای ورودی (تعداد سطرهای فایل validation.parquet یا test.parquet) به اضافه یک ردیف سرتیتر ارسال کنید. اگر ستون‌های اضافی (به غیر از post_id, predictions) یا ردیف‌های اضافی داشته باشید، ارسال شما با خطا مواجه می‌شود. (دقت کنید که ترتیب سطرها دقیقا مشابه سطرهای فایل‌های مرجع باشند.)

فایل ارسالی باید دقیقا 2 ستون داشته باشد:

- post_id: شناسه‌ی مربوط به آگهی که احتمال رد یا تایید شدن آن را پیش‌بینی می‌کنید.
- predictions: پیش‌بینی مدل شما از احتمال تایید شدن آگهی.

قوانین شرکت در مسابقه

- شما می‌توانید در تیم‌های **یک الی سه** نفره در مسابقه شرکت کنید.
- تمامی مکاتبات با **مسئول تیم** صورت می‌گیرد. لذا از مسئولین محترم تیم‌ها خواهشمندیم که به صورت مرتب ایمیل و سایر راه‌های ارتباطی خود را چک کنند.
- هر فرد تنها می‌تواند عضو یک تیم شرکت‌کننده در این مسابقه باشد.
- حضور حداقل یک نفر از اعضای هر تیم در مراسم افتتاحیه که سوم دی برگزار خواهد شد **الزامیست**. لینک اتاق برگزاری افتتاحیه از طریق ایمیل در اختیار شما قرار خواهد گرفت.

با آرزوی موفقیت
تیم برگزاری مسابقه