

فهرست

قوانین.....	۱
پرسش ۱. تشخیص هرزنامه.....	۳
۱-۱. مجموعه داده.....	۳
۲-۱. پیش پردازش داده ها.....	۳
۳-۱. نمایش ویژگی.....	۴
۴-۱. ساخت مدل.....	۴
۵-۱. ارزیابی.....	۵
۵-۱. امتیازی.....	۵
پرسش ۲ - پیش بینی ارزش نفت.....	۶
۱-۲. مقدمه.....	۶
۲-۲. مجموعه دادگان و آماده سازی.....	۶
۳-۲. پیاده سازی مدل ها.....	۷
۴-۲. ARIMA.....	۷

پرسش ۱. تشخیص هرزنامه

هدف از این سوال آشنایی با وظیفه تشخیص هرزنامه بر روی متن فارسی می‌باشد. تشخیص هرزنامه شامل شناسایی و دسته‌بندی پیام‌ها به دو دسته‌ی هرزنامه و غیر هرزنامه است. برای آشنایی با روند سوال، این مقاله را مطالعه کنید.

۱-۱. مجموعه داده

(۱۰ نمره)

برای شروع، [این مجموعه داده](#) را از Kaggle دریافت کنید. توجه داشته باشید که این مجموعه داده با داده مورد استفاده در مقاله متفاوت است. در صورت عدم دسترسی، میتوانید از فایل فشرده این مجموعه داده که پیوست شده است استفاده کنید. کلاسهای موجود در ستون label و تعداد نمونه‌های هر کلاس را به کمک یک نمودار میله‌ای نمایش دهید.

۱-۲. پیش‌پردازش داده‌ها

(۲۰ نمره)

پیش‌پردازش متن در پردازش زبان طبیعی (NLP) یک مرحله‌ی رایج برای آماده‌سازی داده‌ها است که معمولاً برای هماهنگ‌سازی فرمت داده‌ها و حذف نویزها انجام می‌شود. در این تمرین، شما باید مراحل زیر را به‌منظور بررسی تأثیر پیش‌پردازش بر عملکرد مدل اجرا کنید:

مراحل پیش‌پردازش:

برای پیش‌پردازش داده‌ها، مراحل زیر باید اعمال شود:

- حذف URL ها (لینک‌ها): حذف لینک‌های موجود در متن به دلیل اینکه اغلب ارتباط مستقیمی با موضوع تحلیل ندارند.
- حذف آدرس‌های ایمیل: حذف ایمیل‌ها از متن به دلیل اینکه معمولاً اطلاعات معنایی خاصی برای مدل فراهم نمی‌کنند.
- حذف شماره‌های تلفن: حذف اعداد طولانی یا شماره‌های تماس.
- حذف تکرار حروف: کاهش تکرار حروف به یک حرف ساده (مثلاً "عاللی" به "عالی").

- حذف کلمات توقف: حذف کلماتی مانند "و"، "به"، "از" که به تنهایی اطلاعات معنایی خاصی ندارند.

۳-۱. نمایش ویژگی

(۲۰ نمره)

برای این بخش:

۱. داده‌های متنی پیش‌پردازش شده را با استفاده از توکن‌ساز **ParsBERT** به اعداد تبدیل کنید.
۲. **Padding**: تمام سطرها باید طول یکسانی داشته باشند. طول جملات را برابر با ۳۲ در نظر بگیرید.
۳. بردار تعبیه: با استفاده از مدل از پیش آموزش دیده **ParsBERT** بردار تعبیه (**Embedding**) را برای ورودی‌ها به دست آورید.
۴. ابعاد بردار تعبیه را به ۱۲۰ کاهش دهید.

سؤال:

- ابعاد پیش‌فرض بردار تعبیه در **ParsBERT** چقدر است؟
- تعداد ابعاد این بردار بیانگر چیست؟
- مفهوم بردار تعبیه را توضیح دهید و بیان کنید کدام کلمات موجود در مجموعه داده ممکن است تعبیه‌ای نزدیک به هم داشته باشند؟

۴-۱. ساخت مدل

(۳۵ نمره)

وظایف:

۱. داده‌ها را با نسبت ۷۰-۳۰ به دو دسته‌ی آموزش و تست تقسیم کنید. از ۲۰٪ داده‌های آموزش به عنوان مجموعه اعتبارسنجی استفاده کنید.
- الگوریتم جستجوی حریصانه برای یافتن بهترین ترکیب هایپرپارامترها برای مدل **CNN-LSTM** در فضای زیر اعمال کنید:

- **batch_sizes = [8, 64]**
- **learning_rates = [0.001, 0.0001]**

- optimizers = [Adam, SGD]

۲. با استفاده از بهترین هایپرپارامترها، مدل CNN-LSTM را بسازید و آموزش دهید.

۳. مدل‌های ساده‌ی CNN و LSTM را نیز با هایپرپارامترهای بهینه ایجاد و آموزش دهید.

سؤال:

- نقاط قوت و ضعف هر یک از مدل‌ها CNN و LSTM چیست؟
- ادغام این دو مدل با چه هدفی انجام می‌شود؟

۱-۵. ارزیابی

(۱۵ نمره)

داده‌های تست را به کمک معیارهای ارزیابی ذکر شده در مقاله ارزیابی کنید و یک جدول مشابه جدول ۳ مقاله برای مدل‌های LSTM، CNN و CNN-LSTM چاپ کنید.

۱-۵. امتیازی

(۵ نمره)

از روش کیسه کلمات (Bag of Words) برای نمایش ویژگی استفاده کنید.

از بین مدل‌های سنتی یادگیری ماشین که مقاله در جدول ۳ به آن‌ها اشاره کرده ۴ مورد را به انتخاب خود با استفاده از کتابخانه‌ی sklearn آموزش دهید و روی داده‌های تست ارزیابی کرده و نتایج را به جدول بخش قبل اضافه کنید.

پرسش ۲ - پیش‌بینی ارزش نفت

۲-۱. مقدمه

از رایج‌ترین کاربرد شبکه‌های حافظه‌دار می‌توان به پیش‌بینی سری‌های زمانی اشاره کرد. در این سوال با نحوه‌ی پیش‌بینی ارزش نفت خام^۱ با استفاده از چهار روش متفاوت آشنا خواهید شد.

۲-۲. مجموعه دادگان و آماده‌سازی

مجموعه داده‌ی مورد استفاده در این سوال، $CL=F$ ، را از سال ۲۰۱۰ تا کنون را از Yahoo Finance دانلود کنید. از بین ویژگی‌های داده شده ستون Adj Close به عنوان ویژگی اصلی مد نظر قرار دهید. در برخی روزها داده‌ای ثبت نشده است که به عنوان داده‌ی null تلقی میشوند.

- علاوه بر داده‌های null موجود، ده درصد داده‌های ثبت شده را به صورت رندم حذف کنید.
- سپس، روش‌هایی برای جایگزینی داده‌های ناموجود ارائه دهید و داده‌ها را تکمیل کنید. (۱۰)

(نمره)

- طبق نسبت موجود در مقاله داده‌ها را به دو دسته‌ی آموزشی و آزمایشی تقسیم کرده و نرمال کنید. (۵ نمره)

- مشابه شکل ۶ داخل مقاله، هیستوگرام توزیع قیمت را نمایش دهید. (۵ نمره)

¹ <https://www.ijournalse.org/index.php/ESJ/article/view/2149>

۳-۲. پیاده‌سازی مدل‌ها

در مقاله ی داده شده پیش‌بینی سری زمانی توسط سه مدل LSTM، Bi-LSTM و GRU انجام شده‌است. ضمن در نظر گرفتن میانگین مربعات خطا^۱ به عنوان تابع خطا، طبق هایپرپارامترهای جدول ۴ موجود در مقاله این مدل‌ها را آموزش دهید و موارد خواسته شده را گزارش کنید. (۳۰ نمره)

Table 4. Hyperparameters of LSTM, GRU, and Bi-LSTM modelling

Learning Rate	0.0010
Batch Size	100
Optimizer	Adam
Epochs	50
Units	512 (LSTM & GRU); 1024 (Bi-LSTM)

– برای هر سه مدل داده شده، نتایج پیش‌بینی شده را همراه مقادیر واقعی نمایش دهید. (۱۵

نمره)

– ابتدا به طور مختصر در مورد معیارهای داخل مقاله، MAE، RMSE، R-Squared و MAPE

توضیح دهید. سپس مقادیر را گزارش کرده و نتایج را تحلیل و مقایسه کنید. (۱۵ نمره)

۴-۲. ARIMA

(۲۵ نمره)

در این قسمت از سوال با مدل کلاسیک ARIMA^۲ و SARIMA^۳ آشنا خواهید شد. در ابتدا

تفاوت این دو مدل را بیان کنید.

– مزایا و محدودیت‌های مدل ARIMA را مختصر ذکر کنید.

¹ Mean Square Error

² Autoregressive Integrated Moving Average

³ Seasonal ARIMA

- مدل ARIMA پارامترهایی دارد، مفهوم ریاضی این مدل را با ذکر پارامترها شرح دهید.
- پارامترهای بهینه‌ی این مدل را بدست آورده و گزارش کنید.
- ضمن ارائه‌ی جدولی مشابه جدول شماره ۶، نتایج را با نتایج داخل مقاله مقایسه کنید.