استخراج و تحلیل احساسات نظرات کاربران

محمد سینا محلاتی
بوت کمپ هوش مصنوعی مولد

۱۵ آبان ۱۴۰۴

۱ استخراج داده‌ها از دیجی‌کالا :

در این مرحله یک خزنده (crawler) اختصاصی پیاده‌سازی و اجرا شد که از روی صفحات دسته «گوشی موبایل» در دیجی‌کالا، اطلاعات محصولات برند سامسونگ و نظرات کاربران را جمع‌آوری می‌کند. برای هر محصول، حداکثر ۵۰۰ نظر استخراج شده است. محصولات ناموجود و همچنین موارد توقف تولید نیز در استخراج لحاظ شده‌اند.
کدهای مربوط به خزش در فایل `digikala_crawl.py` قرار دارند.

اسکریپت دو فایل CSV تولید می‌کند :

`Digikala_products.csv` :

ویژگی ها :
`id, title_fa, title_en, selling_price, rrp_price, rating_avg, rating_count`

`Digikala_comments.csv` :

ویژگی ها :
`product_id, product_title, comment_id, created_at, rating, comment_text`

خروجی نهایی دارای ابعاد زیر است :

```
Products: (426, 7)
Comments: (72474, 6)
```

در نهایت بیشتر از 70000 متن نظر استخراج شد .

۲ برچسب‌گذاری خودکار احساس با ParsBERT :

برای برچسب گذاری از مدل از پیش‌آموزش‌دیده HooshvareLab/bert-fa-base-uncased-sentiment-deepsentipers-binary استفاده شد.

چرا از Rating استفاده نکردیم ؟ ویژگی Rating به‌تنهایی معیار دقیقی برای احساس متن نیست مواردی مثل امتیاز اشتباه، تفاوت سلیقه کاربران، یا تضاد بین متن و امتیاز باعث خطا می‌شود. بنابراین، برای تولید برچسب‌های قابل‌اعتمادتر، از پیش‌بینی‌های مدل احساس فارسی استفاده شد .

۳ ساخت امبدینگ کامنت‌ها با ParsBERT :

در این مرحله بردار های امبدینگ برای کامنت ها با استفاده از ParsBERT ساخته شد .
خروجی و ابعاد داده برداری :

- تعداد بردارها : 59464 نمونه.
- بعد هر بردار : 768
- شکل نهایی ماتریس امبدینگ: **59464 × 768**.

٤ آموزش مدل SVM (روی امبدینگ‌های ParsBERT) :

به‌جای جست‌وجوی شبکه‌ای، برای سرعت و سادگی، از تنظیمات زیر استفاده شد؛ این تنظیمات روی امبدینگ‌های متراکم ۷۶۸ بعدی معمولا عملکرد پایدار و خوبی می‌دهند:

`rbf` **Kernel**

`2.0` **C**

`"balanced"` **class_weight**

`False` **probability**

`True` **shrinking**

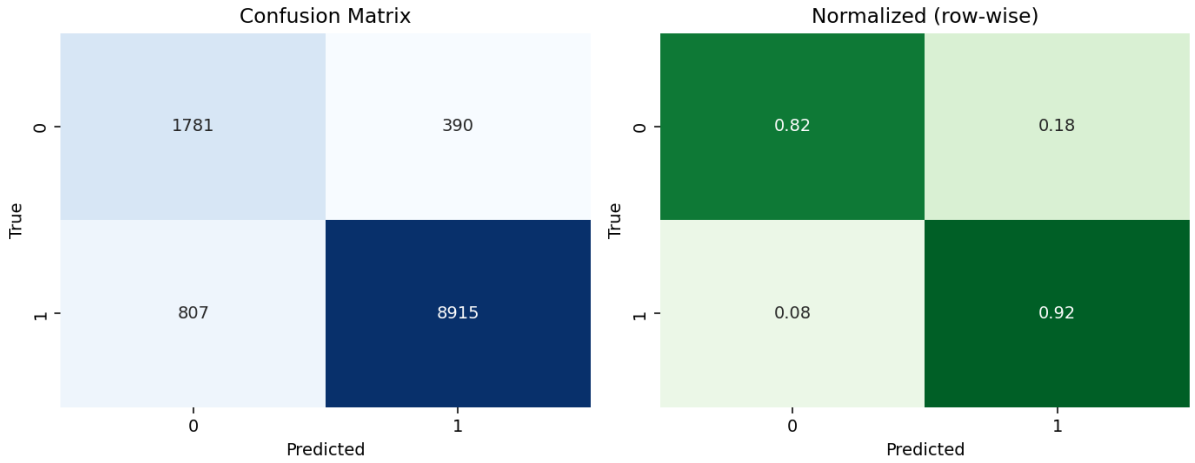
`1000` **cache_size** (MB)

`1e-3` **tol**

`-1` **max_iter**

۵ Shekar → ParsBERT → SVM :

در این بخش یک ماژول یکپارچه پیاده‌سازی شد که متن خام کاربر را به‌صورت خودکار به برچسب احساس تبدیل می‌کند. ابتدا متن با Shekar نرمال می‌شود تا نویز زبانی کاهش یابد. سپس با ParsBERT توکن‌سازی انجام شده و با Mean Pooling روی خروجی لایه آخر و L2-Norm ، یک بردار ۷۶۸ بعدی پایدار برای هر کامنت ساخته می‌شود (با `max_length=256` و پردازش دسته‌ای برای کارایی). در گام پایانی، همین بردارها وارد Pipeline ذخیره‌شده SVM می‌شوند اسکیل داخل پایپلاین انجام می‌گیرد و خروجی نهایی به‌صورت برچسب دودویی یا برچسب متنی ارائه می‌شود. این طراحی، مسیر را کاملا تکرارپذیر، ماژولار و آماده استفاده می‌سازد.


Confusion Matrix
1781
390
807
8915
True
Predicted
0
1
Normalized (row-wise)
0.82
0.18
0.08
0.92