گزارش کار پروژه درس پردازش زبان طبیعی

موضوع پروژه:

تحلیل احساسات روی دیتاست IMDb با Fine-Tuning مدل DistilBERT

استاد: دکتر اسلامی

دانشجو: محمد گنجی

شماره دانشجویی: 404223412

پاییز 1404

فهرست مطالب

چکیده

تحلیل احساسات یکی از مسائل کلیدی در پردازش زبان طبیعی است که هدف آن تشخیص نگرش متن نسبت به یک موضوع می‌باشد. در این پروژه، مسئله تحلیل احساسات دودویی (مثبت/منفی) بر روی دیتاست IMDb پیاده‌سازی شد. برای این منظور از مدل DistilBERT که نسخه سبک‌تر و سریع‌تر BERT است استفاده گردید. مدل با روش Fine-Tuning بر روی داده‌های متوازن آموزش داده شد و با معیارهایی نظیرAccuracy ، Precision، Recallو F1-score ارزیابی شد. علاوه بر این، ماتریس درهم‌ریختگی (Confusion Matrix) و نمودارهای روند آموزش جهت تحلیل رفتار مدل ارائه گردید. نتایج نشان دادند که مدل DistilBERT می‌تواند احساسات نقدهای فیلم را با دقت قابل توجهی تشخیص دهد و عملکرد مناسبی روی داده‌های دیده‌نشده داشته باشد.

مقدمه

در سال‌های اخیر حجم عظیمی از داده‌های متنی در شبکه‌های اجتماعی، سایت‌های نقد و بررسی، و پلتفرم‌های فروش آنلاین تولید می‌شود. تحلیل احساسات به عنوان یکی از کاربردهای مهم پردازش زبان طبیعی، نقش مهمی در استخراج نگرش کاربران از متن ایفا می‌کند. بررسی رضایت مشتریان از محصولات، تحلیل دیدگاه کاربران نسبت به فیلم‌ها و سریال‌ها، تحلیل بازخورد در شبکه‌های اجتماعی و استفاده در سیستم‌های پیشنهاد دهنده از جمله کاربردهای تحلیل احساسات می باشد. در این پروژه تمرکز بر تحلیل احساسات نقد فیلم‌ها در سایت IMDb است.

بیان مسئله و اهداف

هدف پروژه، طراحی یک مدل یادگیری ماشین برای طبقه‌بندی یک متن نقد فیلم به یکی از دو کلاس زیر است:

- Positive (1)
- Negative (0)

اهداف اصلی پروژه:

1. پیاده‌سازی یک سیستم طبقه‌بندی متن مبتنی بر Transformer
2. Fine-Tuning مدل از پیش آموزش‌دیده DistilBERT
3. ارزیابی عملکرد مدل روی داده‌های Validation وTest

4. تحلیل رفتار مدل با استفاده از نمودارها و Confusion Matrix

معرفی دیتاست

IMDb Dataset: دیتاست IMDb یکی از استانداردترین دیتاست‌ها برای تحلیل احساسات است. این دیتاست شامل نقدهای کاربران در مورد فیلم‌ها است که هر نقد به صورت دودویی برچسب‌گذاری شده است.

دیتاست IMDb شامل دو بخش اصلی است:

- Train: 25000 نمونه
- Test: 25000 نمونه

در این پروژه، به منظور کاهش زمان آموزش و همچنین حفظ تعادل کلاس‌ها، از یک زیرمجموعه‌ی متوازن استفاده شد.

روش پیشنهادی

روش پیشنهادی شامل مراحل زیر است:

1. دریافت دیتاست IMDb
2. استخراج زیرمجموعه‌ی متوازن از داده‌ها
3. تقسیم داده‌ها به Train و Validation
4. Tokenization داده‌ها با tokenizer مدل DistilBERT
5. Fine-Tuning مدل DistilBERT برای طبقه‌بندی دودویی
6. ارزیابی مدل با معیارهای استاندارد

پیش‌پردازش داده‌ها

متوازن‌سازی داده‌ها: با توجه به اینکه عدم تعادل کلاس‌ها می‌تواند باعث بایاس مدل شود، داده‌ها به صورت متوازن انتخاب شدند.

- برای Train: 2000 نمونه مثبت + 2000 نمونه منفی
- برای Test: 500 نمونه مثبت + 500 نمونه منفی

این کار باعث می‌شود مدل در آموزش نسبت به هیچ‌کدام از کلاس‌ها برتری غیرواقعی پیدا نکند.

خروجی کد بعد از متوازن سازی داده ها:

```
Training dataset size: 25000
Test dataset size: 25000
   Training class distribution (first 10k):
   Positive (1): 0
   Negative (0): 10000
Balanced training size: 4000
Balanced test size: 1000
   Balanced class distribution:
   Positive (1): 2000
   Negative (0): 2000
```

تقسیم Train و Validation

پس از انتخاب داده‌ها، داده‌های آموزش به نسبت 20/80 تقسیم شدند:

- Train: 3200 نمونه
- Validation: 800 نمونه

خروجی کد:

```
Final splits:
  Training: 3200 samples
  Validation: 800 samples
  Test: 1000 samples
```

معماری مدل و Tokenization

مدلDistilBERT : نسخه‌ی فشرده‌شده‌ی BERT است که تعداد پارامترهای کمتر دارد و سرعت آموزش بالاتر دارد همچنین دقتی نزدیک به BERT ارائه می‌دهد

Tokenization: برای تبدیل متن به ورودی قابل فهم برای مدل، از tokenizer رسمی DistilBERT استفاده شد. در این مرحله متن به توکن‌های عددی تبدیل شد و طول متن به 128 محدود شد سپس یکسان شدن طول توالی‌ها اعمال شد در نهایت Attention mask برای مشخص کردن توکن‌های واقعی تولید شد.

روش آموزش

در این پروژه از AdamW استفاده شد در واقع AdamW برای Transformer استاندارد است و وزن‌ها را به صورت پایدارتر از Adam بهینه می‌کند.

از `get_linear_schedule_with_warmup` استفاده شد چون در ابتدای آموزش مدل نیاز به warmup دارد سپس کاهش خطی learning rate باعث همگرایی بهتر می‌شود

برای جلوگیری از انفجار گرادیان، از clipping با مقدار 1.0 استفاده شد.

تنظیمات آموزش

- Batch size: 16
- Epochs: 3
- Max length: 128
- Learning rate: 2e-5

معیارهای ارزیابی

برای بررسی عملکرد مدل از معیارهای زیر استفاده شد:

Accuracy : درصد پیش‌بینی صحیح نسبت به کل نمونه‌ها.

Precision : درصد پیش‌بینی‌های مثبت صحیح از بین تمام پیش‌بینی‌های مثبت.

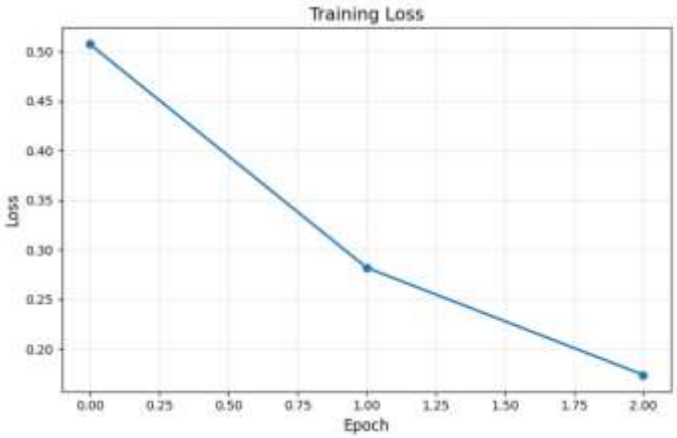
Recall : درصد نمونه‌های مثبت صحیح تشخیص داده‌شده از بین کل نمونه‌های مثبت واقعی.

F1-score : میانگین هماهنگ Precision و Recall

Confusion Matrix : برای تحلیل دقیق خطاها استفاده شد.

نتایج و تحلیل نمودارها

نمودار Training Loss

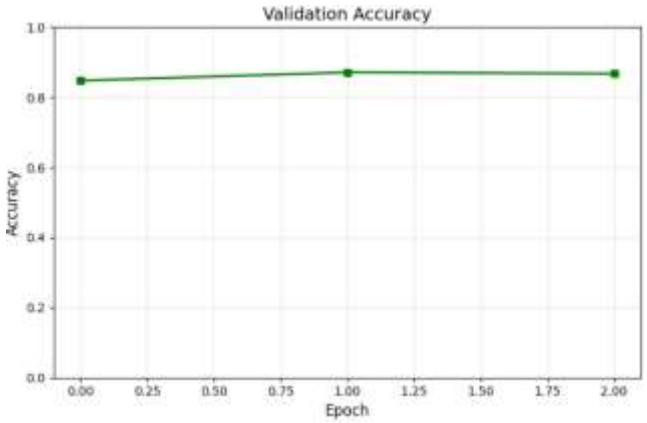


توضیح نمودار:

Training lossنشان می‌دهد مدل در طول آموزش چقدر در حال یادگیری است. انتظار می‌رود مقدار loss با افزایش epoch کاهش پیدا کند. کاهش loss به این معناست که مدل در حال تطبیق وزن‌های خود برای کمینه کردن خطا در داده‌های آموزشی است.

کاهش یکنواخت loss نشانه یادگیری پایدار است. اگر loss پس از چند epoch ثابت بماند، نشان می‌دهد مدل به همگرایی رسیده است. اگر loss نوسان شدید داشته باشد، ممکن است learning rate زیاد باشد.

نمودار Validation Accuracy



توضیح نمودار:

Validation accuracy معیار اصلی برای بررسی عملکرد مدل روی داده‌های دیده‌نشده است. افزایش validation accuracy در طول epoch ها نشان می‌دهد مدل علاوه بر یادگیری روی train ، توانایی generalization دارد.

اگر validation accuracy بالا برود و سپس افت کند، احتمال overfitting وجود دارد. اگر validation accuracy پایین بماند، احتمال underfitting وجود دارد. بهترین مدل معمولاً در epoch با بیشترین validation accuracy انتخاب می‌شود.

Confusion Matrix (Validation)

```
Confusion Matrix (Validation):
          Predicted
          Neg  Pos
True Neg  320   46
True Pos   60  374
```

توضیح ماتریس:

Confusion matrix نشان می‌دهد مدل چند نمونه را درست و چند نمونه را اشتباه طبقه‌بندی کرده است:.

- True Negative (TN): منفی‌های درست
- True Positive (TP): مثبت‌های درست
- False Positive (FP): منفی‌هایی که اشتباه مثبت تشخیص داده شدند

- False Negative (FN):مثبت‌هایی که اشتباه منفی تشخیص داده شدند

- مقدار FP زیاد ⟶ مدل در تشخیص منفی ضعیف است.
- مقدار FN زیاد ⟶ مدل در تشخیص مثبت ضعیف است.

نتایج نهایی Test

```
 Confusion Matrix (Test):
        Predicted
        Neg  Pos
True Neg  409   91
True Pos   52  448
```

توضیح:
این ماتریس نشان‌دهنده عملکرد نهایی مدل روی داده‌های کاملاً unseen است.

نتیجه‌گیری

در این پروژه، یک سیستم تحلیل احساسات برای نقدهای IMDb با استفاده از مدل DistilBERT پیاده‌سازی شد. مدل با Fine-Tuning آموزش داده شد و نتایج نشان دادند که Transformer ها برای طبقه‌بندی متن بسیار کارآمد هستند. ارزیابی روی داده‌های Validation و Test نشان داد مدل قابلیت generalization مناسبی دارد و می‌تواند نقدهای مثبت و منفی را با دقت بالا تشخیص دهد.

```
Classification Report:
              precision    recall  f1-score   support

    Negative       0.89      0.82      0.85       500
    Positive       0.83      0.90      0.86       500

    accuracy                           0.86      1000
   macro avg       0.86      0.86      0.86      1000
weighted avg       0.86      0.86      0.86      1000
```

چند نمونه از خروجی کد برای تحلیل احساسات بر روی نظرات کاربران:

```
Sample 1:
  Text: This movie was absolutely fantastic! The acting was superb and the sto...
  Prediction: Positive
  Confidence: 98.80%
  Positive prob: 98.80%
  Negative prob: 1.20%

Sample 2:
  Text: One of the best movies I've seen in years! Highly recommended to every...
  Prediction: Positive
  Confidence: 98.73%
  Positive prob: 98.73%
  Negative prob: 1.27%

Sample 3:
  Text: An amazing cinematic experience with breathtaking visuals and emotiona...
  Prediction: Positive
  Confidence: 98.71%
  Positive prob: 98.71%
  Negative prob: 1.29%

Sample 4:
  Text: I loved every minute of it! The characters were well-developed and the...
  Prediction: Positive
  Confidence: 98.64%
  Positive prob: 98.64%
  Negative prob: 1.36%
```