

تفکیک متنون تولید شده توسط هوش مصنوعی از
متنون انسانی در محتواهای HTML های استخراج
شده از TOR



بازیابی اطلاعات

استاد تاجبخش

کیانا نصیری

۱۴۰۴ پاییز

فهرست مطالب

۱. چکیده
۲. مقدمه
۳. ابزارها و تکنولوژی‌های مورد استفاده
۴. پیاده‌سازی
 - ۴.۱. استخراج داده
 - چالش‌های استفاده از TorBot
 - استخراج لینک از موتور جستجوی Ahmia
 - خرشن (Crawling) و گردآوری مجموعه داده
 - ۴.۲. پیش‌پردازش داده‌ها
 - پاک‌سازی HTML و استخراج متن اصلی
 - ۴.۳. معماری مدل تشخیص
 - Score Pipeline و Label
۵. نتایج و آمار نهایی
۶. تحلیل نتایج
 - امار متون AI و Human
۷. منابع
 - عملکرد معماری RoBERTa در شناسایی الگوهای آماری
 - تفاوت بردارهای احتمالی کلمات در متون ماشین و انسان

۱. چکیده

در این پروژه، تشخیص محتوای تولید شده توسط مدل‌های زبانی و یا انسانی مورد بررسی قرار گرفته است. هدف اصلی، تعیین درصدی از دیتای جمع آوری شده از شبکه TOR توسط انسان یا AI تولید شده است. برای تحلیل فایل‌های HTML و دسته‌بندی آن‌ها با استفاده از ترانسفورمرها است. در این راستا، از کتابخانه `transformers` و مدل‌های RoBERTa برای تعیین برچسب گروه تعیین شده (AI یا Human) استفاده شده است.

۲. مقدمه

با گسترش ابزارهای تولید متن مبتنی بر هوش مصنوعی (مانند ChatGPT)، تشخیص اینکه محتوا با چه چیزی تولید شده به یک چالش تبدیل شده است. تشخیص تفاوت‌های ساختاری و آماری بین متن انسانی و ماشینی، نیازمند مدل‌های پیچیده یادگیری عمیق است که بتوانند الگوهای احتمالی توزیع کلمات را تحلیل کنند.

- استخراج داده: استفاده از کتابخانه BeautifulSoup برای خواندن برچسب‌های HTML.
- مدل‌سازی: RoBERTa که برای تشخیص ویژگی‌های آماری متن‌های استخراج شده.

۳. ابزارها و تکنولوژی‌های مورد استفاده

برای اجرای این پروژه از ابزارهای زیر استفاده شده است:

- Python 3.10.15
- TOR
- لایبری‌ها:
 - .transformers
 - tf-keras و tensorflow
 - beautifulsoup4
 - scikit-learn

۴. پیاده‌سازی

۴.۱. استخراج داده

ابتدا تلاش شد تا از ریپو [TorBot](#) استفاده شود، اما به دلیل بروز خطا در تعیین عمق جستجو و عدم شناسایی لینک‌های لایو در تور، برای حل این مشکل، فایل‌های ایندکس موتور جستجوی Ahmia، تعداد ۱۰۰ آدرس [onion](#). لایو و منحصر به فرد استخراج گردید. سپس با توسعه یک اسکریپت Requests و با تعریف عمق ۳، خوش بر روی این لینک‌ها اجرا شد. این اسکریپت با پاکسازی پارامترهای ریدایرکت و استخراج آدرس‌های واقعی، در نهایت منجر به گردآوری مجموعه‌دادهای شامل ۲۰۰ HTML گردید.

[لینک HTML‌های استخراج شده](#)

۲.۴. پیش‌پردازش داده‌ها

در این بخش، با تابعی که فایل‌های HTML را خوانده و با حذف نویزهای کدنویسی، متن اصلی را استخراج می‌کند. این مرحله برای جلوگیری از کاهش دقیق مدل است.

۲.۵. معماری مدل تشخیص

از Pipeline تشخیص هوش مصنوعی استفاده شده است که خروجی آن شامل دو پارامتر اصلی است:

1. **Label:** LABEL_1 (AI معادل) یا LABEL_0 (Human).
2. **Score:** میزان قطعیت مدل در مورد برچسب اختصاص داده شده.

۵. نتایج و آمار نهایی

بر اساس اجرای کد بر روی HTML ها موجود، آمار نهایی به شرح زیر است:

- متون تشخیص داده شده برای هوش مصنوعی: ۲۱
- متون تشخیص داده شده برای انسان: ۱۴۰

نکته:

۱. ای‌های که طول تکست آنها کمتر از ۱۰۰ بود بررسی نشدند.

۲. همه‌ی score های تخمین، بالای ۹۰٪ بودند.

```
Device set to use mps:0
Analyzing 168 files...
File: site_92.html | Prediction: Human (99.99%)
File: site_147.html | Prediction: Human (98.56%)
File: site_110.html | Prediction: Human (98.67%)
File: site_84.html | Prediction: Human (99.99%)
File: site_151.html | Prediction: Human (99.98%)
File: site_10.html | Prediction: Human (99.97%)
File: site_47.html | Prediction: Human (99.93%)
File: site_51.html | Prediction: AI (53.64%)
File: site_184.html | Prediction: Human (99.30%)
File: site_71.html | Prediction: Human (99.98%)
File: site_26.html | Prediction: Human (96.75%)
File: site_67.html | Prediction: Human (98.96%)
File: site_126.html | Prediction: Human (95.90%)
File: site_171.html | Prediction: Human (99.98%)
File: site_167.html | Prediction: Human (99.73%)
File: site_188.html | Prediction: AI (88.38%)
File: site_130.html | Prediction: Human (99.91%)
File: site_131.html | Prediction: Human (98.75%)
File: site_166.html | Prediction: Human (99.88%)
File: site_170.html | Prediction: Human (99.98%)
File: site_127.html | Prediction: Human (99.98%)
File: site_66.html | Prediction: Human (99.99%)
File: site_5.html | Prediction: AI (88.16%)
File: site_89.html | Prediction: AI (50.17%)
File: site_31.html | Prediction: AI (99.92%)
```

۶. تحلیل نتایج

بنابر نتایج، امار گزارش شده نشان می‌دهند که مدل‌های ترانسفورمر قدرت بالایی در شناسایی الگوهای تکراری و ساختارهای نحوی خاص هوش مصنوعی دارند. عملکرد مدل roberta بر اساس تحلیل توزیع توکن‌ها و

شناسایی ویژگی‌های ساختاری موجود در معماری ترانسفورمرها است. در واقع، متونی که توسط LLM‌ها تولید می‌شوند، علی‌رغم ظاهر طبیعی، دارای بردارها و الگوهای احتمالی مشخصی هستند که توسط معماری hidden در لایه‌های RoBERTa ترین شده شناسایی می‌شوند. از آنجایی که این مدل بر روی داده‌های خروجی GPT-2، بهینه شده است، توانایی بالایی در درک یکنواختی معنایی یا الکوهای بردار های کلمات دارد؛ ویژگی خاصی که در آن کلمات با احتمال وقوع بالا در کنار یکدیگر قرار می‌گیرند و نوسانات متون انسانی را ندارند.

۷. منابع

Source code and results	.1
Transformers: roberta-base-openai-detector model	.2
TOR	.3
Gemini	.4
TorBot Github Repo	.5