به نام خدا

درس مبانی یادگیری عمیق

گزارش پروژه پایانی

استاد درس : دکتر مرضیه داوودآبادی

دستیاران : مرتضی حاجی‌آبادی، سحر سرکار، فائزه صادقی، مهسا موفق بهروزی، الناز رضایی، پریسا ظفری، حسن حماد، سید محمد موسوی، کمیل فتحی، شایان موسوی نیا، امیررضا ویشته

دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر

نیمسال اول تحصیلی ۱۴۰۲ - ۱۴۰۳

موضوع:

تحلیل احساسات در متن فارسی

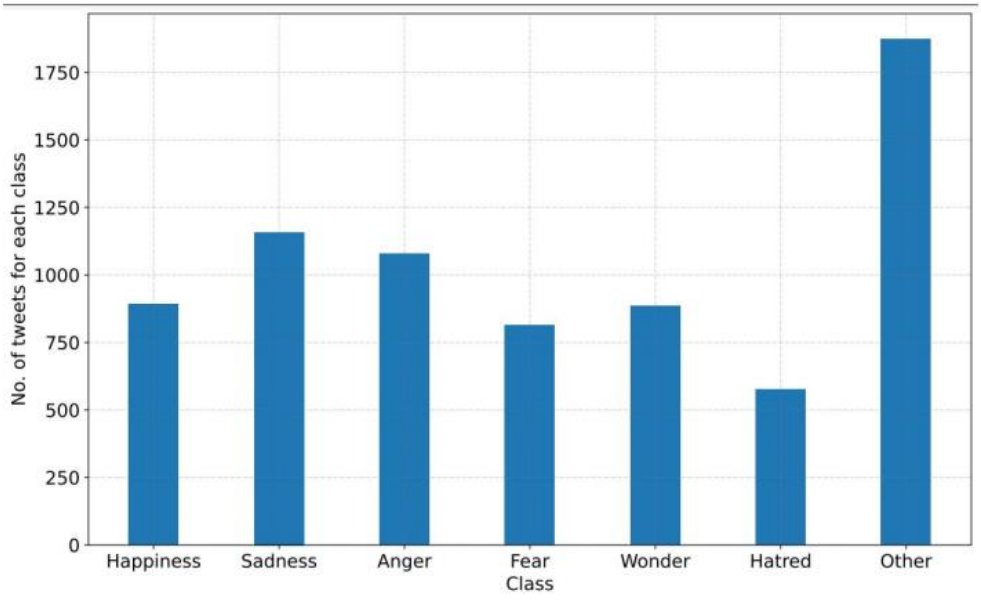
ردیف	نام ونام خانوادگی	شماره دانشجویی
۱	ملیکا محمدی فخار	99522086
۲	ستاره باباجانی	99521109

جدول ۱: مشخصات اعضای گروه

۱ شرح موضوع و مجموعه دادگان

تجزیه و تحلیل احساسات، همچنین به عنوان opinion mining شناخته می شود، یک تکنیک پردازش زبان طبیعی (NLP) است که شامل تعیین احساسات بیان شده در یک قطعه متن است. هدف این است که احساسات و نگرش هایی را که متن منتقل می کند، چه مثبت، چه منفی یا خنثی، درک کنیم. تجزیه و تحلیل احساسات به طور گسترده در برنامه های مختلف، از جمله نظارت بر رسانه های اجتماعی، تجزیه و تحلیل بازخورد مشتری، تحقیقات بازار و تعدیل خودکار محتوا استفاده می شود.

در این پروژه از مجموعه داده ArmanEmo استفاده شده است که شامل بیش از ۷۰۰۰ جمله فارسی با هفت دسته احساسات است. این مجموعه‌داده از منابع مختلف از جمله توییتر، اینستاگرام و نظرات دیجی‌کالا جمع‌آوری شده است. دسته‌بندی‌ها بر اساس شش احساس اصلی (خشم، ترس، شادی، نفرت، غم، تعجب) و یک دسته دیگر انجام شده است.



۲ پیش‌پردازش داده ها

پیش پردازش متن یک گام مهم در وظایف پردازش زبان طبیعی (NLP) برای تمیز کردن و تبدیل داده های متن خام به قالبی مناسب برای تجزیه و تحلیل و یادگیری ماشین است. روش های اصلی استفاده از پیش پردازش متن در NLP عبارتند از:

کاهش noise: حذف اطلاعات نامربوط مانند کاراکترهای خاص، علائم نگارشی، و ارقام عددی را که ممکن است به تجزیه و تحلیل کمک نکنند.

Normalization: تبدیل متن به حروف کوچک و جلوگیری از تکرار کلمات.

توکن سازی: تقسیم متن به کلمات جداگانه.

Stemming: حذف پیشوندها یا پسوندها برای به دست آوردن ریشه کلمات.

Spell Checking: اصلاح اشتباهات املایی و تایپی.

حذف تگ های HTML و کاراکترهای خاص: حذف تگ های HTML و URL ها و هشتگ ها از متن.

Vectorization: تبدیل کلمات یا عبارات به بردارهای عددی.

در این پروژه از روش های کاهش نویز، نرمالسازی، توکن سازی و حذف تگ های html برای پیش پردازش داده استفاده شده است.

۳ انتخاب مدل

مدل های موجود برای تحلیل احساسات به شرح زیر هستند:

BERT (Bidirectional Encoder Representations from Transformers): این مدل توسط گوگل توسعه داده شده است، یک مدل مبتنی بر transformer ها است که به نتایج پیشرفته‌ای در کارهای مختلف NLP، از جمله تجزیه و تحلیل احساسات دست یافته است.

GPT (Generative Pre-trained Transformer): در حالی که مدل‌های GPT، مانند GPT-۳، در درجه اول برای تولید متن شناخته شده‌اند، این مدل برای تحلیل احساسات نیز قابل تنظیم است.

XLNet: یکی دیگر از مدل‌های مبتنی بر transformer است که نقاط قوت مدل‌های auto-regressive و auto-encouding را ترکیب می‌کند. برای درک context و وابستگی های داخل متن، استفاده میشود.

RoBERTa (Robustly optimized BERT approach): نسخه بهبود یافته BERT است که برخی از محدودیت‌های آن را برطرف می‌کند و به نتایج بهتری می‌رسد.

DistilBERT: یک نسخه مقطر BERT است که بیشتر عملکرد خود را حفظ می کند اما تعداد پارامترها به میزان قابل توجهی کاهش یافته است. این برای استنتاج سریعتر و نیازهای کمتر منابع و در عین حال حفظ دقت طراحی شده است.

در این پروژه، با توجه به منابع محدود و نیاز به دقت بالا از مدل پیش آموخته DistilBert استفاده شده است و خروجی آن به یک pre-classifier داده شد و پس از آن از تابع فعالسازی tanh استفاده شد. سپس با استفاده از روش dropout منظم سازی صورت گرفت و در انتها، یک لایه classifier خطی قرار داده شد تا ابعاد خروجی به 7 کلاس احساسات برسد.

۴ اقدامات انجام شده

مقاله ArmanEmo بر روی تشخیص احساسات از متن فارسی تمرکز دارد .این مطالعه مجموعه دادگان «ArmanEmo»را معرفی می‌کند که یک مجموعه داده جدید شامل بیش از ۷۰۰۰ جمله فارسی برچسب‌گذاری شده توسط انسان که به هفت احساس بر اساس مدل Ekman و یک دسته «دیگر» طبقه‌بندی شده‌اند. این مقاله به جزئیات جمع‌آوری داده‌ها و فرآیند annotation می‌پردازد، که از منابعی مانند نظرات توییتر، اینستاگرام و دیجی کالا استفاده میکند و ترکیبی از روش‌های annotationدستی و خودکار را به کار می‌برد.

این مطالعه همچنین چندین مدل پایه را برای طبقه‌بندی احساسات با تمرکز بر مدل‌های زبان مبتنی بر transformer بررسی می‌کند. مدل با بهترین عملکرد به میانگین امتیازF1 ، 75.39% در کل مجموعه داده آزمایشی دست یافت.

در نهایت، مقاله چالش‌ها و پیچیدگی‌های تشخیص احساسات مبتنی بر متن، از جمله دشواری طبقه‌بندی دقیق احساسات مختلط و محدودیت‌های مدل‌های فعلی در گرفتن حالات هیجانی ظریف را برجسته می‌کند.

برای آشنایی بیشتر با کتابخانه Parsivar از این لینک استفاده شده است که نحوه استفاده توابع آن را با مثال آموزش داده است.

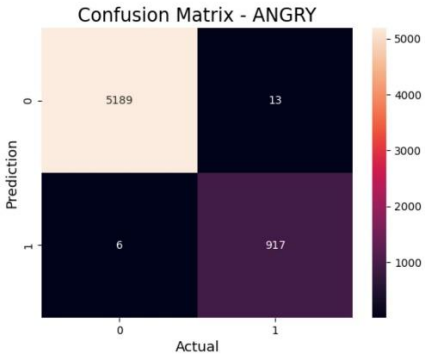
اقدامات انجام شده به شرح زیر هستند:

۱. پیش پردازش با روش های ذکر شده در بخش پیش پردازش داده ها
۲. خواندن فایل های آموزش و تست و اجرای پیش پردازش بر روی آنها
۳. تبدیل هر یک از برچسب های احساسات به یک one-hot vector
۴. توکن سازی داده آموزش
۵. مقداردهی اولیه ابرپارامترها
۶. استخراج id, mask, token_type_id از دیتا آموزش و تست
۷. طراحی مدل طبق توضیحات ذکر شده در بخش انتخاب مدل
۸. انتخاب تابع ضرر و تابع بهینه ساز
۹. آموزش مدل بر روی داده آموزش و تست آن روی داده آزمون
۱۰. تغییر فرمت خروجی اصلی و خروجی به دست آمده برای مقایسه
۱۱. محاسبه دقت و confusion matrix
۱۲. پیش بینی خروجی به ازای 5 داده تصادفی تست
۱۳. تعریف تابع پیش بینی خروجی به ازای یک متن

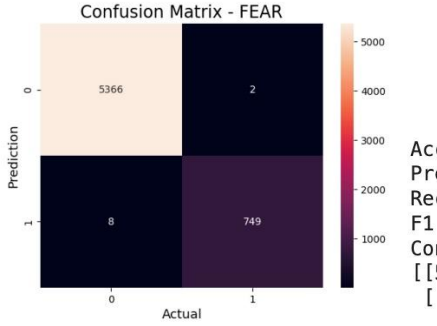
۵ ارزیابی مدل

از توابع recall, precision, f1-score در کنار confusion matrix که مربوط به کتابخانه sklearn.metrics هستند، استفاده شد.confusion matrix امکان نمایش بصری performance مدل با مقایسه برچسب ها و مقادیر پیش بینی شده را میدهد.

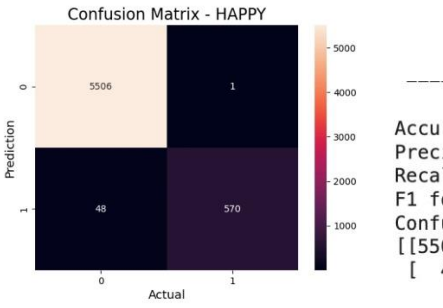
تصاویر زیر خروجی معیارها بر روی برخی کلاس ها و دقت کل مدل بر روی داده های آزمون میباشد:




```
Accuracy for ANGRY: 0.9968979591836735
Precision for ANGRY: 0.986021505376344
Recall for ANGRY: 0.9934994582881906
F1 for ANGRY: 0.9897463572584996
Confusion Matrix for ANGRY:
[[5189   13]
 [   6  917]]
```




```
Accuracy for FEAR: 0.9983673469387755
Precision for FEAR: 0.9973368841544608
Recall for FEAR: 0.9894319682959049
F1 for FEAR: 0.993368700265252
Confusion Matrix for FEAR:
[[5366    2]
 [   8  749]]
```

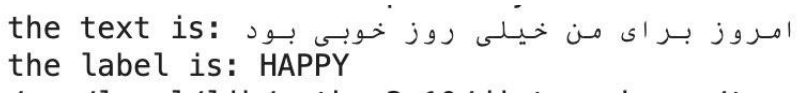



```
Accuracy for HAPPY: 0.992
Precision for HAPPY: 0.9982486865148862
Recall for HAPPY: 0.9223300970873787
F1 for HAPPY: 0.95878889823381
Confusion Matrix for HAPPY:
[[5506    1]
 [  48  570]]
```

Total accuracy is 0.9958483965014577

سپس از میان داده های آزمون، ۵ داده زیر را بصورت تصادفی انتخاب کرده و با استفاده از مدل، احساس موجود در آنن این چنین پیش بینی شده است:

خروجی مدل به ازای متن ورودی زیر، این چنین میباشد:



۶ بخش امتیازی

۷ مراجع

۱. مقاله ARMANEMO: A PERSIAN DATASET FOR TEXT-BASED EMOTION DETECTION

۲. data-hub

۳. Confusion matrix(geeks for geeks)

پیاده‌سازی این پروژه با همکاری گروه آیسا میاهی‌نیا و ریحانه شاهرخیان صورت گرفته است.