خلاصه تحلیلی مقاله InstructABSA

اطلاعات مقاله

عنوان: Instruction Learning for Aspect Based Sentiment Analysis

نویسندگان: Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Arjun Sawant, Swaroop Mishra, Chitta Baral
دانشگاه: Arizona State University
کنفرانس: NAACL 2024 (North American Chapter of ACL)
کد: github.com/kevinscaria/InstructABSA

چکیده (Abstract):

تحلیل احساسات مبتنی بر جنبه (Aspect-Based Sentiment Analysis) نقش حیاتی در درک احساسات ریزدانه کاربران دارد. برخلاف تحلیل احساسات سنتی که یک احساس کلی برای کل متن تعیین می‌کند، ABSA احساسات را در سطح جنبه‌های مختلف استخراج می‌کند.

مثال کاربردی:
جمله: "سوشی عالی بود، ولی قیمتش گرون بود!"
جنبه ۱: سوشی احساس: مثبت نظر: عالی
جنبه ۲: قیمت احساس: منفی نظر: گرون

این مقاله InstructABSA را که یادگیری مبتنی بر دستورالعمل (Instruction Learning) است و توانایی استدلال مدل‌های زبانی بزرگ را خیلی بهبود داده است، برای زیرتسک‌های تحلیل احساسات مبتنی بر جنبه (Aspect Based Sentiment Analysis) با روش پیشنهادی و دستورالعمل محور زیر معرفی میکند:

- اضافه کردن مثال‌های مثبت، منفی و خنثی به هر نمونه آموزشی
- تنظیم دستورالعمل (Instruction Tuning) مدل Tk-Instruct برای زیرتسک‌های ABSA

یعنی: اول دستور را می‌خوانند، بعد کار را انجام می‌دهند

نوآوری اینجاست: استفاده هم‌زمان از مثال‌های مثبت، منفی و خنثی داخل instruction برای **یادگیری** مدل

چرا این موضوع مهم است؟

- ✓ کاربرد تجاری: شرکت‌ها می‌توانند بفهمند مشتریان از کدام ویژگی محصول راضی یا ناراضی هستند
- ✓ تصمیم‌گیری هوشمند: به جای خواندن هزاران نظر، سیستم خودکار جنبه‌ها و احساسات را استخراج می‌کند

RBSA بسته به Prompt ، دو نسخه زیر رو داره:

نسخه	Prompt
InstructABSA-1	Definition + 2 × Positive Example
InstructABSA-2	Definition + 2 × Positive Example + 2 × Negative Example + 2 × Neutral Example

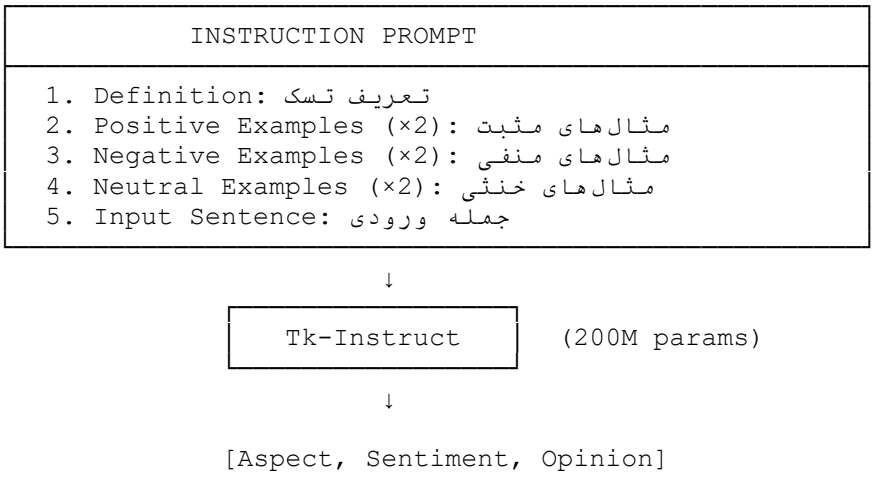
زیرتسک‌های ABSA:

مقاله زیرتسکهای مختلف ABSA را به شرح جدول ذیل بررسی کرده است:

زیرتسک	نام کامل	ورودی	خروجی
ATE	Aspect Term Extraction	جمله	جنبه‌ها
ATSC	Aspect Term Sentiment Classification	جمله +جنبه	قطبیت احساس
ASPE	Aspect Sentiment Pair Extraction	جمله	(جنبه، قطبیت)
AOOE	Aspect Oriented Opinion Extraction	جمله +جنبه	کلمه نظری
AOPE	Aspect Opinion Pair Extraction	جمله	(جنبه، نظر)
AOSTE	Aspect Opinion Sentiment Triplet Extraction	جمله	(جنبه، نظر، قطبیت)
ACOSQE	Aspect Category Opinion Sentiment Quadruple	جمله	(جنبه، دسته، نظر، قطبیت)

جزئیات آزمایش:

- مدل مورد استفاده: Tk-Instruct-base-def-pos از huggingface
- تعداد پارامترها: حدود ۲۰۰ میلیون
- معماری: T5-based
- مدل از قبل بلد است دستور بخواند و روی کارهای مختلف NLP آموزش دیده است .پس نویسندگان مدل را از صفر آموزش نمی‌دهند و فقط آن را روی ABSA fine-tune می‌کنند



- دیتاست‌ها:

دیتاست	دامنه	توضیح	Train	Test
Lapt14	لپ‌تاپ	SemEval 2014	3045	800
Rest14	رستوران	SemEval 2014	3041	800
Rest15	رستوران	SemEval 2015	1315	685
Rest16	رستوران	SemEval 2016	2000	676

- هایپرپارامترها:

GPU	1x Nvidia Tesla P40	نکته مثبت: هر آزمایش ۵ بار اجرا شده و میانگین گزارش شده. این یعنی نتایج قابل اعتمادتر هستن نکته منفی: فقط یک GPU استفاده شده. این موضوع مفیده برای reproducibility ، ولی مشخص نیست در scale بزرگتر چی میشه.
Batch Size	16 (ATE, ATSC) / 8 (other)	
Gradient Accumulation	2	
Learning Rate	5e-5	
Epochs	4	
Runs per experiment	5	

شبه‌کد:

```
def InstructABSA(sentence, task_type):
  # 1. Build instruction prompt
  prompt = task_definition[task_type]
  prompt += get_examples(positive=2, negative=2, neutral=2)
  prompt += f"Input: {sentence}"

  # 2. Feed to instruction-tuned model
  output = TkInstruct.generate(prompt)

  # 3. Parse structured output
  return parse_aspects_sentiments(output)
```

مقایسه نتایج اجرای مدل در دو تسک ATE و ATSC:

F1 Score – ATE

Model	Lapt14	Rest14	Rest15	Rest16
GPT2-med (1.5B)	82.04	75.94	-	-
GRACE	87.93	85.45	-	-
BARTABSA	83.52	87.07	75.48	
IT-MTL	76.93	-	74.03	79.41
InstructABSA-1	91.40	**92.76**	75.23	**81.48**
InstructABSA-2	**92.30**	92.10	**76.64**	80.32

Accuracy – ATSC

مدل	Lapt14	Rest14	Rest15	Rest16
ABSA-DeBERTa	82.76	89.46	-	-
LSAT	**86.31**	**90.86**	-	-
Dual-MRC	75.97	82.04	73.59	-
InstructABSA-1	80.62	86.25	83.02	89.10
InstructABSA-2	81.56	85.17	**84.50**	**89.43**

تحلیل:

- ✓ در Lapt14 و Rest14 ، InstructABSA از LSAT ضعیف‌تره ولی در Rest15 برنده‌ست!
- ✓ احتمالا LSAT مدل تخصصی برای ATSC هست اما InstructABSA عمومی‌تره و در همه تسک‌ها "خوب" عمل می‌کنه، نه "عالی" که این موضوع یک Trade-off مهم است
- ✓ GPT2-med با ۱.۵ میلیارد پارامتر، ضعیف‌ترین است. این نشون میده پارامتر بزرگتر همیشه بهتر نیست.
- ✓ InstructABSA-2 همیشه بهتر از InstructABSA-1 نیست چراکه شاید در بعضی دامنه‌ها، مثال‌های اضافی noise ایجاد کنند.
- ✓ InstructABSA در تسک ATE در همه دیتاست‌ها بهترین نتیجه فعلی رو داره (SOTA) و در تسک ATSC، برای برخی دیتاست‌ها بهترین نتیجه رو داره و در برخی رقابتی عمل میکنه

نتایج کلی ارزیابی InstructABSA:

۱. نتایج حاصل از ارزیابی مدل به صورت بین دامنه ای[1] که به معنی آموزش روی یک دامنه و تست روی دامنه دیگر است، نشون میده که مدل قابلیت تعمیم بین دامنه‌ای خوبی دارد

۲. نتایج حاصل از ارزیابی مدل در شرایط ترکیب دامنه‌ها[2] که داده‌های آموزشی دو دامنه مختلف (لپ‌تاپ‌ها و رستوران‌ها) برای آموزش مدل و داده های تست آن دو دامنه نیز برای تست مدل ترکیب می‌شوند، نشان میدهد که داده بیشتر منجر به عملکرد بهتر مدل میشود

۳. Instruction Tuning حدود ۲۰٪ بهتر از LoRA[3] که یک روش Fine-Tuning است، عمل می‌کند. چراکه LoRA برای انتقال دانش طراحی شده، نه یادگیری تسک جدید از صفر. وقتی مدل پایه (T5) درکی از ABSA نداره، LoRA نمی‌تونه کمک کنه. ولی Instruction Tuning با توضیح دادن تسک، این مشکل رو حل می‌کنه.

٤. اگر مثال‌هایی با برچسب اشتباه بدهیم، عملکرد مدل ۱۰٪ افت می‌کند. این نشان می‌دهد مدل واقعاً از مثال‌ها یاد می‌گیرد و طراحی درست instruction می‌تواند به اندازه‌ی معماری پیچیده مهم باشد

٥. InstructABSA با یک مدل M۲۰۰ پارامتری، از مدل‌های B۱.۵ (۷ برابر بزرگتر) بهتر عمل می‌کنه

[1] Cross-Domain
[2] Joint-Domain
[3] Low-Rank Adaptation

٦. کارایی نمونه(Sample Efficiency) : به صورت میانگین در تسک های مختلف ABSA، با ۵۰٪ داده، نتایج رقابتی با روش‌های دیگه می‌گیری.این یعنی نیاز به داده برچسب‌خورده کمتر میشه، هزینه annotation کاهش پیدا می‌کنه و برای زبان‌هایی با داده کم (مثل فارسی!) خیلی مفیده

محدودیت‌ها و ضعف ها

- فقط دیتاست‌های SemEval تست شده و باید روی دیتاست‌های دیگه هم تست شود
- فقط مدل M۲۰۰: شاید مدل‌های کوچکتر نتایج متفاوتی بدهند
- فقط زبان انگلیسی: قابلیت چندزبانه بررسی نشده
- در تسک‌های AOPE ، AOSTE و ACOSQE عملکرد ضعیف‌تره

ایده‌های ادامه کار (پیشنهاد برای پروژه)

- استفاده از مدل‌های جدیدتر: جایگزینی Tk-Instruct با Flan-T5 یاLlama
- تحلیل تعداد مثال‌ها: بررسی اثر تعداد مختلف مثال‌های مثبت/منفی/خنثی
- بررسی Cross-lingual Transfer : آموزش روی انگلیسی، تست روی فارسی