تحلیل احساسات جنبه‌محور از طریق تولید تصویر مصنوعی

Ge Chen, Zhongqing Wang و Guodong Zhou (دانشگاه سوچو، چین)

مسئله اصلی و اهمیت آن

مسئله اصلی این مقاله بهبود تحلیل احساسات جنبه‌محور (ABSA) است. در حالی که روش‌های سنتی تنها بر متن تکیه می‌کنند، معنای استخراج شده از داده‌های متنی صرف، محدود است و نمی‌تواند سرنخ‌های عاطفی ظریف را به طور کامل حفظ کند.

استفاده از اطلاعات تصویری (چندرسانه‌ای) می‌تواند درک بصری و عمیق‌تری از محتوا ارائه دهد. با این حال، بسیاری از متون فاقد تصویر مرتبط هستند یا جمع‌آوری آن‌ها دشوار است. این مقاله با تولید تصاویر مصنوعی که با متن و بار عاطفی آن همخوانی دارند، این شکاف داده‌ای را پر می‌کند.

ورودی‌ها و خروجی‌های مدل

ورودی: متن اصلی (مانند یک توییت) که شامل جنبه‌های مختلف (Aspects) و نظرات است.

خروجی: استخراج جنبه‌ها (Aspect Term Extraction) و تعیین قطبیت احساسی (Sentiment Polarity) برای هر جنبه (مثبت، منفی یا خنثی).

داده‌های مورد استفاده

این پژوهش از دو مجموعه داده استاندارد توئیتر استفاده کرده است:

Twitter-۲۰۱۵: شامل داده‌های آموزشی، ارزیابی و تست با توزیع‌های مختلف احساسی (مثبت، منفی، خنثی).

Twitter-۲۰۱۷: مجموعه داده بزرگتری نسبت به نسخه ۲۰۱۵ که برای ارزیابی قدرت تعمیم‌دهی مدل به کار رفته است.

تولید داده آموزشی: نویسندگان با استفاده از مدل‌های یادگیری چندرسانه‌ای (InternLM-XComposer۲) و مدل‌های تشخیص اشیاء(OWL-ViT) ، جفت‌های تصویر-متن با برچسب‌های کاذب (Pseudo-labels) ایجاد کردند تا مدل تولید تصویر را دقیق‌تر آموزش دهند.

روش پیشنهادی

روش پیشنهادی شامل سه گانه اصلی است:

تولید تصویر تحت نظارت (Supervised Generation): با استفاده از مدل Stable Diffusion و تکنیک LoRA، تصاویری تولید می‌شوند که نه تنها با کلمات متن، بلکه با بار احساسی آن نیز همسو هستند.

بهبود بصری (Visual Refinement) : برای افزایش کیفیت، از مدل SAM (برای بخش‌بندی معنایی) و نقشه‌های حرارتی توجه (Attention Heatmaps) استفاده می‌شود تا بخش‌های مهم تصویر که با احساسات مرتبط هستند (مثل چهره یا شیء مورد نظر) برجسته شده و نویزهای پس‌زمینه حذف شوند.

مدل چندرسانه‌ای (Multi-modal Integration) : در نهایت، متن اصلی و تصویر مصنوعی بهبود یافته به یک مدل زبانی بصری (MLLM) داده می‌شوند تا تحلیل نهایی انجام گیرد.

۱. مدل چندرسانه‌ای اصلی (InternLM-XComposer۲)

این مدل یک (**MLLM مدل زبانی-تصویری بزرگ**) است.

- نویسندگان از این مدل به عنوان بدنه اصلی برای ترکیب متن و تصویر استفاده کرده‌اند. همچنین در مرحله آماده‌سازی داده، برای تولید برچسب‌های توصیفی (Description) برای عکس‌ها از آن بهره برده‌اند.

۲. تشخیص اشیاء متن-محور (OWL-ViT / OWLv۲)

این مدل ساخته گوگل است و می‌تواند اشیاء را بر اساس متن (مثلاً "همبرگر" یا "چهره خندان") در تصویر پیدا کند.

- نویسندگان از این مدل برای **Visual Refinement** استفاده کرده‌اند. وقتی تصویر مصنوعی تولید می‌شود، این مدل اشیاء مرتبط با جنبه (Aspect) مورد نظر را در تصویر مکان‌یابی می‌کند تا مدل بتواند روی همان بخش تمرکز کند و نویزهای تصویر را نادیده بگیرد.

۳. تولید تصویر (Stable Diffusion v۱-۴)

این مشهورترین مدل تولید تصویر از متن است.

- هسته اصلی نوآوری مقاله اینجاست. نویسندگان این مدل را با تکنیک **LoRA** بازآموزی کرده‌اند تا تصاویری بسازند که فقط "زیبا" نباشند، بلکه "بار احساسی (Sentiment) "متن را هم نشان دهند.

نویسندگان در بخش Methodology و Experiments جزئیات فنی کافی را ارائه داده‌اند. برای بازسازی این فرآیند، شما باید مراحل زیر را طی کنید:

گام اول: استفاده از stable-diffusion-v۱-۴ و نوشتن یک اسکریپت برای آموزش LoRA با استفاده از کتابخانه‌هایی مثل.(diffusers)

گام دوم: استفاده از OWL-v۲ برای استخراج باکس‌های اشیاء (Bounding Boxes) بر اساس کلمات کلیدی موجود در متن(Aspects).

گام سوم: ترکیب این‌ها در مدل internlm-xcomposer۲ که قابلیت دریافت تصویر و متن همزمان را دارد.

شبه‌کد فرآیند:

Python

 # ۱. Image Generation

synthetic_image = StableDiffusion_LoRA(input_text, sentiment_label)

 # ۲. Refinement

masks = SAM(synthetic_image)

attention_map = CLIP_Attention(masks, textual_query)

refined_image = synthetic_image * attention_map

 # ۳. ABSA Task

final_prediction = MLLM_Encoder_Decoder(input_text, refined_image)

نتایج، محدودیت‌ها و ایده‌های آینده

نتایج اصلی: مدل پیشنهادی (Ours) با امتیاز F۱ معادل ۷۰.۰ در Twitter-۱۵ و ۷۳.۵ درTwitter-۱۷ ، از تمام مدل‌های قدرتمند متنی مانند Llama-۳ و GPT-۴o و مدل‌های چندرسانه‌ای قبلی پیشی گرفته است.

محدودیت‌ها: پیچیدگی محاسباتی مدل بالاست که کارایی آن را در مقیاس‌های بسیار بزرگ با چالش مواجه می‌کند. همچنین مدل عمدتاً بر روی داده‌های انگلیسی ارزیابی شده است.

ایده‌های ادامه

- طراحی معماری‌های سبک‌تر برای کاهش هزینه‌های پردازشی.
- آزمایش مدل بر روی زبان‌های دیگر (مانند چینی) یا حوزه‌های تخصصی‌تر.
- استفاده از استراتژی‌های همجوشی (Fusion) دقیق‌تر برای ترکیب بهتر ویژگی‌های متن و تصویر.