پروژه درس مبانی پردازش زبان و گفتار

یگانه مرشدزاده

نیم‌سال دوم ۱۳۹۹-۱۴۰۰

موضوع داده

رابطه‌ی بین اضطراب و افسردگی با توجه به عبارات بیان شده توسط کاربران معمولی. در این داده‌ها تمرکز بر روی کشف تشابهات بین این دو مشکل روانی است.

زبان داده؟ آیا داده به زبان محاوره‌ای است یا رسمی؟

زبان داده به انگلیسی است.

متن‌های داده توسط کاربران Reddit نوشته شده است و لذا عموما به زبان محاوره‌ای است.

دسته‌بندی داده (حداقل ۲ دسته) به چه شکل است؟

دو دسته کلی مورد بررسی اضطراب و افسردگی است.

موارد بسیاری به مانند اختلال دوقطبی، شخصیت ضداجتماع، عدم قبول ظاهر و فیزیک بدنی، خودکشی و... نیز می‌تواند در ذیل این دسته‌ها نیز قرار گیرد و بررسی شوند.

داده از چه جهت برای شما جذاب است؟

در زندگی هر فرد عوامل و اتفاقات متفاوتی باعث به وجود آمدن اضطراب و استرس و همچنین افسردگی و سایر پیامدهای ناشی از آن می‌شود. این حالات می‌تواند باعث به وجود آمدن مشکلاتی در روابط بین خانوادگی، کار و ... شود. اینکه گفته می‌شود روان ناخودآگاه انسان مثل فیلی است که لزوما فیل‌سوار، که خودآگاه فرد است، نمی‌تواند آن را به هر مسیری که می‌خواهد هدایت کند نشان‌دهنده‌ی اهمیت این مبحث است.

آیا فرضیه‌ای در رابطه با داده دارید؟

به نظر می‌رسد که اختلال اضطراب تاثیر بسیار زیادی بر روی افسردگی و بیماری‌های حاصل از آن (مانند اختلال دو قطبی) دارد. به عبارتی کلمات پرتکرار اضطراب به احتمال زیاد در داده‌های مربوط به افسردگی نیز مشاهده خواهد شد و این بدین معنی می‌تواند باشد که بین افسردگی و اضطراب رابطه بسیار معنی‌داری وجود دارد و این دو بر روی هم تاثیر گذار هستند.

نحوه جمع‌آوری داده؟

برای جمع‌آوری داده‌ها قصد این است که از متن‌هایی که مستقیما توسط عامه‌ی مردم نوشته می‌شود استفاده شود. به همین منظور از داده‌های Facebook, Twitter, Reddit می‌توان استفاده کرد. به دلیل اینکه در Reddit تعدادی Community وجود دارند که از آن‌ها label های متون را می‌توان دقیق‌تر اختصاص داد، از این منبع استفاده خواهد شد. Label زدن بر روی داده‌های Facebook و Twitter در این مرحله کار کم‌تر دقیقی به نظر می‌رسد.

آیا برای جمع‌آوری داده از کتابخانه یا api خاصی استفاده خواهید کرد و دسترسی به این کتابخانه/api دارید؟

از api خود Reddit برای این کار استفاده می‌شود که پست‌های مربوط به community ها را می‌توان توسط آن به صورت فایل json استراج کرد.

چه محدودیتی برای جمع‌آوری داده دارید؟ پیش‌بینی می‌کنید چه حجم از داده بتوانید جمع‌آوری کنید؟

به نظر می‌رسد این است که دسته بندی داده‌های جمع آوری، تشخیص و حذف تا حد امکان داده‌های کمتر دقیق، که شبکه را گمراه می‌کنند، محدوددیت و سختی‌ای است که به احتمال بالا با آن روبرو خواهم شد. همچنین انتخاب کردن label ها و community های مرتبط بهم تا حدی چالش‌برانگیز در نظر می‌آید. به نظر می‌رسد برای جمع‌آوری ۵۰ هزار داده خام به دلیل وجود Community ها با پست‌های زیاد به مشکل نخورم. اما، از طرف دیگر آماده‌سازی این داده برای پردازش‌های آتی بسیار زمان‌بر تخمین زده می‌شود.