پروپوزال پروژه درس NLP

درس: **پردازش زبان طبیعی**

مدرس: **دکتر سید صالح اعتمادی**

کاری از: **غزاله محمودی**

نیمسال دوم سال تحصیلی ۱۳۹۹-۱۴۰۰

دانشگاه علم و صنعت ایران

موضوع پروژه

بررسی متن نوشته شده توسط افراد و پیدا کردن لغات و اصطلاحات و عبارات مشترک بین افراد هر گروه. دو گروه مورد بررسی احتمالا افراد استرس دار(و یا افسرده) در مقابل افراد شاد (و یا موفق) هستند.

دیتاست پروژه

در رابطه با جمع آوری داده دو رویکرد داشتم.

رویکرد اول استفاده از متن های نوشته شده کاربران در توییتر فارسی است که به دلیل مشکل در لیبل زدن به آن ها فعلا از این داده ها صرف نظر کردم.

رویکرد دوم استفاده از سایت reddit بود که در حال حاضر بهترین منبع لیبل دار در دسترس است. دیتاست به زبان انگلیسی و تقریبا محاوره ای می باشد. با توجه به موارد گفته شده درباره لیبل زدن داده های توییتر فارسی در حال حاضر استفاده از داده reddit با لیبل را ترجیح می دهم.

به نظر می رسد با توجه به منابع موجود حجم قابل قبولی از دیتا در دسترس باشد. همچنین api های مورد نیاز برای استخراج موجود است.

چالشی که انتظار آن را دارم زمان بر بودن کار با حجم بالا دیتاست و سختی های این مورد است.

اهمیت موضوع پروژه

به دلیل پیشگیری از خطرات و آسیب هایی که این افراد ممکن است به خود(خودکشی) یا اطرافیان(آزار دیگران) بزنند سعی شده با این روند شناسایی افراد راحت تر باشد.

در ادامه انتظار داریم در متن های افراد افسرده کلمات نا امید کننده و با بار منفی بیشتر دیده شود. همچنین در گذشته سیر کردن و شاکی بودن دائمی از شرایط موجود احتمالا از ویژگی های مورد انتظار از افراد افسرده و پر استرس است.

در مقابل شادی و امید به زندگی و توجه به موفقیت ها و تعریف آن برای دوستان احتمالا در متن های افراد شاد و موفق بیشتر به چشم می خورد.

از نظر اینکه با تحلیل متن ها و کلمات مورد استفاده افراد بتوان ویژگی های شخصیتی آن ها را مورد بررسی قرار داد، علاقه مند هستم روی این موضوع کار کنم.