توضیح راه حل:

تسک داده شده به نحوی می تواند موضوع یک پایان نامه باشد بستگی دارد چقدر بخواهیم وارد جزئیات شده و نتایج را بهبود دهیم. برای این تسک من حدود ۳ روز وقت در نظر گرفتم و با توجه به این زمان سعی در ارائه با کیفیت ترین خروجی را داشتم.

روز اول به آشنایی با مقالات درباره متون Xenophobic و Racist گذشت. شاید به طور خلاصه بتوان گفت متون نژادپرستانه وقتی درباره مهاجران و کسانی که بومی منطقه ای نباشند گفته شود آن متون Xenophobic هستند تا جایی که معادل anti-immigration برای کلمه Xenophobic نیز در مقالات علمی مشاهده گردید.
در این باره سوالاتی مطرح می شود که:
آیا هر جمله Xenophobic در عین حال Racist هم هست یا خیر. یا اینکه آیا نظرات منفی درباره ادیان و فرهنگ و رسوم مهاجران نیز شامل دسته Xenophobic می شود یا خیر.

از سوالات موشکافانه که بگذریم فعلا میتوان در حد یک تسک استخدامی در نظر بگیریم که یک مدلی می خواهیم که تشخیص دهد یک متن آیا جز متون مورد نظر ما هست یا خیر و با چه احتمالی. برای این کار نیاز به داده های متنی داریم و مقداری برنامه نویسی و کمی تجربه.

روز دوم به پیدا کردن داده های متنی مرتبط با موضوعات بالا گذشت. اولین و دم دست ترین و تر و تمیز ترین دیتاست از سایت Hugging face بود. بقیه دیتاست ها با گشتن و دنبال کردن لینک های مختلف پیدا شد و لزوما داده های مورد نیاز ما را ندارند ولی به نحوی به درد میخورند و نیاز به بررسی دارند.

به طور کلی می شد برنامه ای هم نوشت که توییت هایی که هشتگ های نژاد پرستانه و ضد مهاجران و اقلیت ها (مثلا *AntiImmigration*, *DeportThemALL*) را دارد ذخیره و آنان را دستی نشانه گذاری کرد مثل همان کاری که اکثر مقالات نیز انجام داده بودند.

من در روز دوم نیز تلاش کردم دیتابیس های مختلف را جمع آوری کنم ولی وقت نکردم همه آن ها را بررسی کنم و آن ها را یکی کنم. دیگر دیتاست هایی که به نحوی در این تسک کاربرد دارند و من تا کنون پیدایشان کرده ام

دیتاست اول

Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter

- Link to publication: https://www.aclweb.org/anthology/N16-2013

- Link to data: https://github.com/ZeerakW/hatespeech
- Task description: 3-topic (Sexist, Racist, Not)
- Details of task: **Racism**, Sexism
- Size of dataset: 16,914
- Percentage abusive: 0.32
- Language: English
- Level of annotation: Posts
- Platform: Twitter
- Medium: Text

دیتاست دوم

The Gab Hate Corpus: A collection of 27k posts annotated for hate speech

- Link to publication: https://psyarxiv.com/hqjxn/
- Link to data: https://osf.io/edua3/
- Task description: Binary (Hate vs. Offensive/Vulgarity), Binary (Assault on human Dignity/Call for Violence – sub task on message delivery, binary: explicit/implicit), Multinomial classification: Identity based hate (**race**/ethnicity, nationality/regionalism/**xenophobia**, gender, religion/belief system, sexual orientation, ideology, political identification/party, mental/physical health)
- Details of task: Group-directed + Person-directed
- Size of dataset: 27,665
- Percentage abusive: 0.09 Hate, 0.06 Offensive/Vulgar
- Language: English
- Level of annotation: Post
- Platform: Gab
- Medium: Text
- Reference: Kennedy, B., Araria, M., Mostafazadeh Davani, A., Yeh, L., Omrani, A., Kim, Y., Koombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatain, A., Hussain, A., Lara, A., Olmos, G., Omary, A., Park, C., Wang, C., Wang, X., Zhang, Y. and Dehghani, M., 2018, The Gab Hate Corpus: A collection of 27k posts annotated for hate speech. PsyArXiv.

دیتاست سوم

Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter

- Link to publication: https://pdfs.semanticscholar.org/3eeb/b7907a9b94f8d65f969f63b76ff5f643f6d3.pdf
- Link to data: https://github.com/ZeerakW/hatespeech
- Task description: Multi-topic (Sexist, **Racist**, Neither, Both)
- Details of task: **Racism**, Sexism
- Size of dataset: 4,033
- Percentage abusive: 0.16
- Language: English
- Level of annotation: Posts
- Platform: Twitter
- Medium: Text
- Reference: Waseem, Z., 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In: Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science. Copenhagen, Denmark: Association for Computational Linguistics, pp.138-142.

دیتاست چهارم

Exploring Hate Speech Detection in Multimodal Publications

- Link to publication: https://arxiv.org/pdf/1910.03814.pdf
- Link to data: https://drive.google.com/file/d/1S9mMhZFkntNnYdO-1dZXwF_8XIiFcmlF/view
- Task description: Multimodal Hate Speech Detection, including six primary categories (No attacks to any community, Racist, Sexist, Homophobic, Religion based attack, Attack to other community)
- Details of task: Racism, Sexism, Homophobia, Religion-based attack

- Size of dataset: 149,823
- Percentage abusive: 0.25
- Language: English
- Level of annotation: Posts
- Platform: Twitter
- Medium: Text and Images/Memes
- Reference: Gomez, R., Gibert, J., Gomez, L. and Karatzas, D., 2020. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1470-1478).

دیتاست پنجم

A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research

- Link to publication: https://arxiv.org/pdf/1802.09416.pdf
- Link to data: https://github.com/Mrezvan94/Harassment-Corpus
- Task description: Multi-topic harassment detection
- Details of task: Racism, Sexism, Appearance-related, Intellectual, Political
- Size of dataset: 24,189
- Percentage abusive: 0.13
- Language: English
- Level of annotation: Posts
- Platform: Twitter
- Medium: Text
- Reference: Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V. and Sheth, A., 2018. A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research. ArXiv,.

دیتاست ششم

ETHOS: an Online Hate Speech Detection Dataset (Binary)

- Link to publication: https://arxiv.org/pdf/2006.08328.pdf
- Link to data: https://github.com/intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset
- Task description: Binary (Hate, Not)
- Details of task: Gender, Race, National Origin, Disability, Religion, Sexual Orientation

- Size of dataset: 998
- Percentage abusive: 0.43
- Language: English
- Level of annotation: Posts
- Platform: Youtube, Reddit
- Medium: Text
- Reference: Mollas, I., Chrysopoulou, Z., Karlos, S., and Tsoumakas, G., 2021. ETHOS: an Online Hate Speech Detection Dataset. Complex & Intelligent Systems, Jan. 2022

و تعداد بسیار زیاد دیگری دیتاست از این لیست [Hate Speech Dataset Catalogue | hatespeechdata]
که برای جلوگیری از طولانی شدن این سند به آن ها اشاره ای نمی شود.

دیتاست های بعدی از لینک های دیگری یافت شده اند ولی ممکن است در سایت بالا نیز پیدا شوند

SemEval-2019 Task5 - Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter
https://aclanthology.org/S19-2007/
Best model in the competition was **HateBERT**
Data & Code of HateBERT: OSF | HateBERT

سایت زیر نیز لیست انواع دیتاست های زیر مجموعه hate speech را هم دارد که اشاره به آن خالی از لطف نیست.
https://github.com/aymeam/Datasets-for-Hate-Speech-Detection?tab=readme-ov-file

پیاده سازی راه حل

دیتاست را دانلود کرده و آن را از کاراکتر هایی که نیاز نداریم پاک می کنیم. همچنین با نگاهی به داده ها می توان دید که داده ها با @username شروع می شوند که می توان آن ها را نیز با عبارت منظم استخراج و پاک نمود.
بعد از پیش پردازش داده ها ان ها را با توکنایزر مناسب مدل انتخابی مان به توکن ها تبدیل می کنیم.
برای پیاده سازی مدل کلسیفیکیشن این تسک به سراغ fine tune کردن یک مدل pretrain شده رفتم. به مدل اینکودر BERT-base-uncased دو لایه داخلی به همراه dropout برای کاهش overfitting اضافه کردم و با داشتن یک لایه نهایی softmax احتمال اینکه متن ورودی در دسته مورد نظر ما باشد یا نباشد (باینری کلسیفیکیشن) بر می گرداند.
برای ارزیابی راه حل از K-fold cross validation استفاده کردم و نتیجه دقت بالای ۹۰ درصد مدل بعد از ۳ epoch آموزش می باشد.

در صورتی که دقت بالاتری برای این تسک مورد نیاز باشد می توان از ورژن های دیگر BERT که بزرگتر و حجیم تر بوده و امکان تفکیک بین بزرگ و کوچکی حروف و ... را دارند نیز استفاده نمود.
اگر سرعت بالاتر مورد نیاز باشد نیز می توان از SVM استفاده کرد که مدلی بسیار سریع و نسبتا دقیق است.

همچنین مشکل بالانس نبودن کلاس ها در دیتاست نیز وجود داشت که با تکنیک های زیر قابل رفع است

- حذف داده هایی که امبدینگ آن ها شبیه یکدیگر است از کلاس بزرگتر
- دیتا اگمنتیشن برای کلاس کوچکتر
- استفاده از LLM نظیر Llama3-7b به صورت لوکال جهت تولید داده های بیشتر برای کلاس کوچکتر