CS344 Progress Report: Fine-grained Toxicity Detection of Social Media Posts via Domain Adaptation

Stella Jaquiss
2003869

December 5, 2022

1 Introduction

This report will cover the progress that has been made on this project since the submission of its specification. It should also provide clarification on areas previously left open to interpretation or not previously addressed, as to not only provide an improved criteria against which to mark the success of this project upon completion, but also to provide a clear outline and plan for the project going ahead into term 2.

2 Elaboration of Methodology

Here, the steps taken in order to meet the objectives, outlined in the specification, will be clarified and elaborated on in further detail.

– 1. DATASET SELECTION –

This project will center around the *Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments* [McC], with the aim of refining the grain of its annotations through domain adaptation.

The set from which data will be using to extend the domain aforementioned will be sourced from the paper *Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts* [Tsva]. Structurally as you can see in Figure 1, the taxonomy of this dataset is very distinct from [McC], and as a dataset, aims to classify various traits of toxic comments *("Attributive","Institutionalised","Teaming","Othering")* in contrast to identifying the type of comment *("Homophobic/Transphobic pejoratives or derogation","Homophobic/Transphobic threatening language","Non-anti-LGBT counter speech","non-anti-LGBT hope speech","None")*.

A third dataset has been selected, which is sourced from *SemEval-2023: Task 10: Towards Explainable Detection of Online Sexism* [Vid]. This set will provide something to compare the aforementioned datasets against to

aid in statistical analysis further along in the project progression. For example, having a baseline distribution of the frequency of toxic comments in various classes in these three selected sets would help set a guideline for an appropriate distribution in the new fine-grained set being constructed. Having this third dataset helps standardise data later on.
Furthermore, it helps provide a blueprint structurally for how our new fine-grained set is to be constructed. Its taxonomy is very much structurally what would be desired. This is as for the first two levels in Figure 2, the class system is almost a direct parallel to that in Figure 3 (from the dataset we want to extend and refine), yet it also provides further insight and nuance to each of the discrete classes.

It's worth noting here that the datasets referenced above are from reliable sources that have gone through several checks and been cross referenced before publish. Cornell, the ACL anthology and SemEval are all reputable sources from which to conduct research.

A clear definition of *toxicity* is also yet to be declared and bias introduced by extending [McC]'s domain is to be discussed.

– 2. EXTENSION OF MODEL –

The better part of this project will be navigating how to adjoin the two aforementioned datasets [McC] [Tsva](introducing a new class system) in order to construct a set that is fine-grained in nature.

When utilised in training classification models, this set will aim to accurately classify more elusive toxic comments than models trained with other sets such as [McC]. The F1 score of the classification model implemented will be used as a metric to compare the performance of the newly constructed set against its predecessors. (see (1), NB the values for the true positives **TP**, false positives **FP** and false negatives **FN** can be sourced from the classification model's confusion matrix.)

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The focus of this project is on optimising existing corpora of toxic comments, rather than focusing on optimising a specific classification model to then go on to classify toxicity within speech online. In this way, any classification models implemented in this project will only be used to: peer review any results from papers relating to selected datasets (to ensure reproducibility going forward); or provide a way to compare datasets (a classification model trained on datasets *1 and 2* with similar grains - where the model trained on *1* has a greater F1 score than having been trained on *2* - indicates that dataset *1* is generally *more representative* of toxic comments in general

than *2.*)

In the paper [Tsva] exploring the detection of micro-agressions in speech, linear SVMs were trained for each class in a one vs all set up. In this way, outsourcing and using a pre-made model such as **sklearn.svm.LinearSVC** from a standard library like scikit-learn [sl] would be appropriate.

– 3. FURTHER EXTENSIONS –

If a fine-grained classification model is successfully constructed, then the model will be applied on the *Explainable Detection of Online Sexism* dataset. The performance of the model on this dataset can then be quantified and compared against that of the previous test data.

An evaluation should also be provided of bias introduced or identified throughout the project, and how that could be mitigated in the future. A review of other ways to improve the contructed/adapted dataset (if not already implemented) should also be provided. One way to improve or expand our final dataset to help classifiers identify more elusive toxic language is provided in the following paper: *Fortifying Toxic Speech Detectors Against Veiled Toxicity* [Tsvb].

3 Project Management

Table 1 below assigns deadlines to sub tasks within this project. Though this timeline is similar in nature to that presented in the specification, it has been revised to allow more time dedicated to adapting the domain [McC] and extending to construct a set that has a finer grain of classification of various types of toxicity.

Referencing Figure 1 in the project specification, it is worth noting that though datasets on which to build this research project have been selected [Tsva] [Vid] [McC], I am yet to research or cite varying definitions of toxicity (both within and outside of the queer community), and am still yet to start the long and arduous process of tagging and restructuring data from [McC] and [Tsva]. Deadlines in table 1 have been adjusted to account for this fall behind schedule.

References

[McC] Bharathi Raja Chakravarthi; Ruba Priyadharshini; Rahul Ponnusamytft; Prasanna Kumar Kumaresantft; Kayalvizhi Sampath; Durairaj Thenmozhi; Sathiyaraj Thangasamytl; Rajendran Nallathambi1r; John Phillip McCrae. **Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments**. https://doi.org/10.48550/arXiv.2109.00227. Associated Datasets can be found at: https://codalab.lisn.upsaclay.fr/competitions/5310#participate.

Completion Date	Time Assigned	Task
25/12/22	3 weeks	Statistical Analysis of each of datasets [McC] [Tsva] [Vid].
8/1/23	5 weeks	Re-annotating old datasets and combining to make one cohesive set.
5/2/23	4 weeks	Implement classification model to compare results and peer review performance scores mentioned in papers [McC] [Tsva] [Vid].
26/2/23	3 weeks	Implement classification model on new dataset to compare against performance from other papers [McC] [Tsva] [Vid].
12/3/23	2 weeks	Produce in depth comparison of each of datasets, from those initially referenced to the set constructed, from analysis derived so far.
17/3/23	1 week	Prepare Oral Presentation
9/4/23	3 weeks	Any necessary catchup or extensions implemented.
2/5/23	6 weeks	Prepare Final Report

Table 1: Adjusted project timetable

[sl] scikit learn. scikit-learn: Machine learning in python: Multiclass and multioutput algorithms. https://scikit-learn.org/stable/modules/multiclass.html#multiclass-classification.

[Tsva] Luke M. Breitfeller; Emily Ah; David Jurgens; Yulia Tsvetkov. **Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts, ACL Anthology**. https://aclanthology.org/D19-1176/. Supplementary material can be as attachment at cited url, Associated Datasets can be found at: https://drive.google.com/drive/folders/1bKf8PQuuOk7z3ehgAcmTLjmK5Cb86ZTz.

[Tsvb] Xiaochuang Han; Yulia Tsvetkov. Fortifying toxic speech detectors against veiled toxicity. https://doi.org/10.48550/arXiv.2010.03154.

[Vid] Hannah Rose Kirk; Wenjie Yin; Paul Röttger; Bertie Vidgen. **SemEval-2023: Social Attitudes: Task 10: Towards Explainable Detection of Online Sexism**. https://codalab.lisn.upsaclay.fr/competitions/7124#learn_the_details-overview. Original list of SemEval Tasks for 2023 can be found at: https://semeval.github.io/SemEval2023/tasks.html.

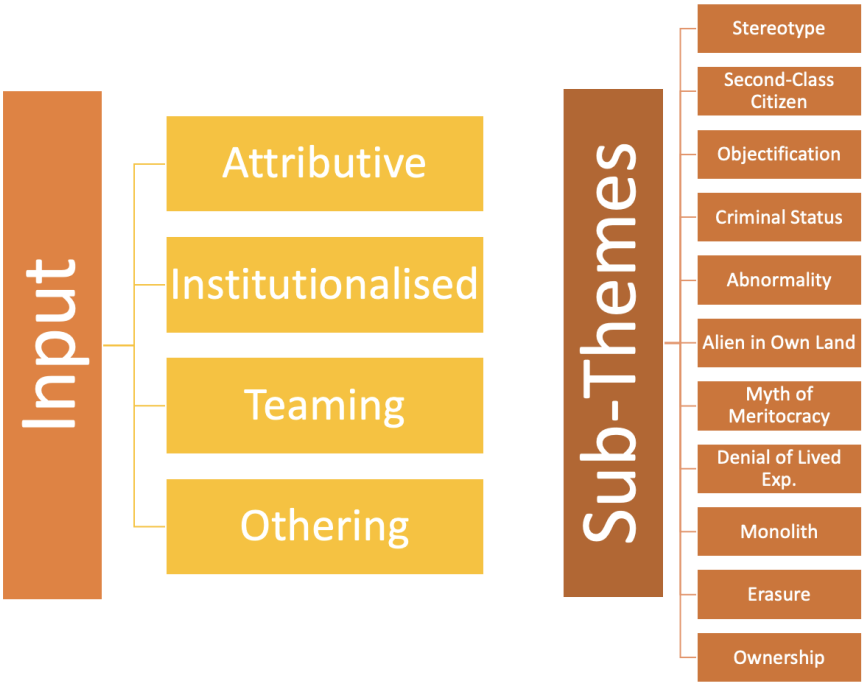

Figure 1: Taxonomy of Micro-agressions dataset

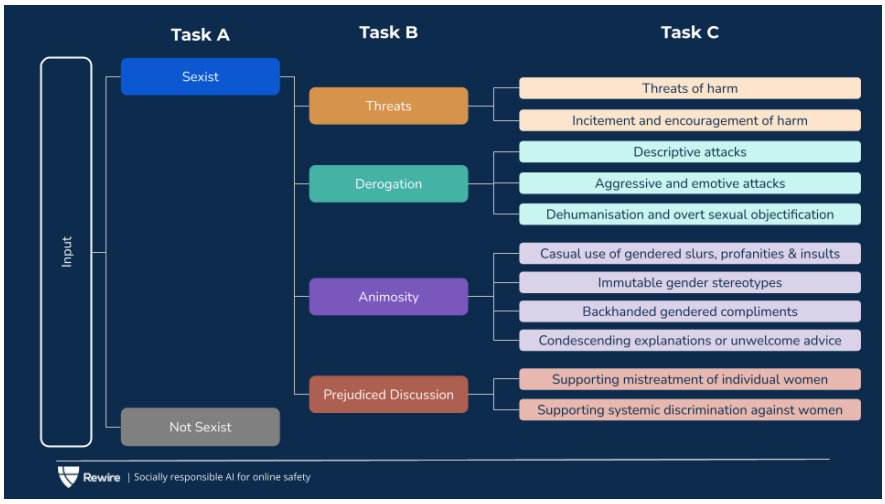

Figure 2: Taxonomy of Sexism dataset

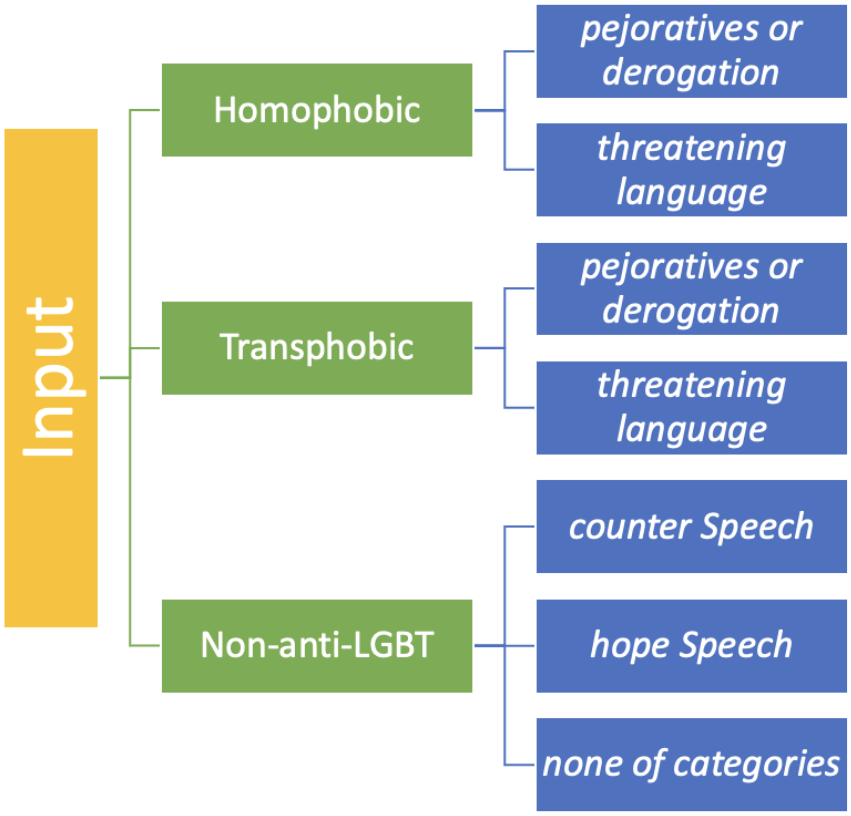

Figure 3: Taxonomy of Youtube Comments dataset