XCS22U: Natural Language Understanding Experimental Protocol

Dongyao Wang, Xiuyu Yan, Blake Norwick

Background

Our task is to build a model to detect offensive language. Specifically, we build a classifier to either label text as offensive or not offensive. We realize that the lack of a universal definition of offensive language complicates this task.

Whenever researchers crowdsource labels for NLP tasks, they must have a mechanism to aggregate the labels in such a way that they can be consumed by machine learning algorithms. In the context of classification, researchers often use a majority voting scheme to assign a class label to text from a set of labels provided by the annotators. When the annotators don't agree unanimously on the class to assign to some text, this raises the question: is the lack of consent a product of noise (i.e. assuming a majority class label exists and a couple of "outlier" annotator's mistakenly assigned the wrong label to the class) or a product of the ambiguity implicit in the problem (i.e. in the case of sarcasm detection, it isn't clear, without more context, whether the text presented to the annotators represents sarcasm). For the present task, offensive language detection, Leonardelli et al. (*Agree to Disagree, 2021)* saw that models designed to detect offensive language were better (i.e. higher F1 scores on the task at hand) when they were trained only on data in which the annotators agreed unanimously on the class label. However, while Leonardelli et al note models trained on data in which the annotators don't unanimously agree on the label saw worse F1 score, they also show that randomly assigning a class label to those ambiguous examples results in an even worse model, proving that ambiguous data still provides some signal to the model. We want to expose the model to that ambiguous data in a clever way such that it improves the classifier's F1 score.

Hypothesis:

Leonardelli et al. (*Agree to Disagree, 2021)* argue that there is an opportunity to make models designed to identify offensive language more robust by exposing them to more language where annotators disagree about the offensiveness of the language. It is known that models designed to detect offensive language struggle more with language that is *implicitly* offensive -- i.e. text that may not have any overtly offensive language, but instead is offensive in its content. Specifically, we hypothesize that language whose offensiveness is not unanimous in the eyes of the annotators is more likely to contain *implicitly* offensive language -- by exposing our models to more of this language, we think we can improve the ability of the model's to detect implicitly offensive language, without harming its ability to detect more overtly offensive language.

Data:

To verify our hypothesis, we must have access to text whose offensiveness was labelled by crowdworkers. Critically, we require that the crowdworkers labels are not aggregated into a single class label -- we want to use the degree to which the crowdworkers agree about the offensiveness of a piece of text to inform how we present that text to the models. We've found a couple relevant datasets:

- Our primary dataset will be the *Agreeing to Disagree* dataset, which Leonardelli et al. have released to us. It consists of more than 10k tweets, each of which is associated with labels assigned by 5 different crowdworkers, via Amazon Mechanical Turk. The tweets fall under the following domains: Covid-19, BLM, and US Elections. Below is a snippet of the dataset:

Split	Offensive_binary_label	Tweet_ID	Crowd_agreement_level	Crowd_annotations	Domain
1	1270166804166574082	A++	O++	BLM	train
0	1269049607939657728	A+	N+	BLM	train
0	1270013192329191431	A++	N++	BLM	train
0	1270410143809880065	A+	N+	BLM	train
0	1269063125145473027	A++	N++	BLM	train
1	1270184268422164481	A++	O++	BLM	train

  Legend: A++ represents a unanimous agreement among 5 annotators or classifiers, A+ represents mild agreement as 4 out of 5 annotations agreeing on the same label, and A0 represents weak agreement as 3 out of 5 annotations agreeing on the same label. Offensive tweets are marked as O++/+/0, non offensive ones are marked as N++/+/0, respectively, with the same scale as defined above.
- The wiki-detox dataset (https://github.com/ewulczyn/wiki-detox) also may be of use. In *Ex Machina: Personal Attacks Seen at Scale*, the paper associated with the wiki-detox dataset, the authors argue that the crowdworker's annotations naturally form an approximate empirical distribution (ED) over opinions of whether the comment is offensive. They show that using these ED labels (as opposed to a one hot encoding of the class label -- offensive or inoffensive) improve the ability of their classifiers to identify offensive language. This dataset actually makes the distinction between

Metrics:

We will primarily use the F1 score to evaluate our models. However, we also consider the ideas presented in *The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality*, which argue that traditional metrics like the F1 score and ROC AUC tend to overestimate the performance of classifiers on social computing tasks. We will also use the metrics Gordon et al. present in *The Disagreement Deconvolution* to assess our models.

Models:

Our baseline model will be BERT trained only on data where the annotators are in unanimous agreement on the offensiveness of the text. We will experiment with ways to expose the model to data where the crowdworkers don't agree unanimously on the label and see whether, by exposing the models to such data, we can improve their performance. As for how we will train the models on data where the crowdworkers don't agree unanimously on the label, we have the following ideas:

- Weighting the cross-entropy loss function by a parameter that takes into account the group of annotator's consensus on whether the text or not is offensive. A similar approach saw success in POS tagging (see https://aclanthology.org/E14-1078.pdf).
- Use soft labels: Thiel argues in *Classification on Soft Labels is Robust Against Label Noise* that the use of soft labels makes classifiers more resilient to label noise. While we acknowledge that a lack of agreement amongst annotators may not be label noise in the context of our task (instead, the text may just *be* ambiguously offensive), we are optimistic that this technique may allow us to better model the degree of the offensiveness of the language, which will in turn benefit our classifiers.
- We'd also like to expand upon the ideas presented in *Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection*. Essentially, the authors argue that organizing the annotators into clusters (in an unsupervised fashion), building a classifier per cluster, and then using the ensemble of classifiers to detect Hate Speech is superior to a single classifier.

General Reasoning:

The intuition behind our approach is that offensive language is nuanced and that the use of a majority voting scheme to assign class labels to text (in cases where annotators disagree on the correct label) obfuscates the nuance that is captured by the annotator's disagreement. We are optimistic that by taking annotator's disagreement

into account when training models to detect offensive language, we will build models that not only perform better on the specific task, but are also more robust, in the sense that they are better at detecting offensive language when applied to datasets that they weren't trained on, a known issue in the field.

Summary of Progress So Far:

We haven't made any progress thus far (it took us a while to settle on a specific project). Now that we've settled on a specific project, we'll get to work. We don't foresee any issues that will prevent us from implementing our simpler ideas (weighting the cross entropy function and soft labels), the ideas laid out in *Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection* might be beyond the scope of the project (at least, given time constraints).

References:

Barbara Plank, Dirk Hovy, Anders Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. https://aclanthology.org/E14-1078.pdf

Christian Thiel. Classification on Soft Labels is Robust Against Label Noise. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.616.5484&rep=rep1&type=pdf.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, Sara Tonelli Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement arXiv:2109.13563

Ellery Wulczyn, Nithium Thain, Lucas Dixon. Ex Machina: Personal Attacks Seen at Scale. https://arxiv.org/pdf/1610.08914.pdf.

Gregor Wiedemann, Seid Muhie, Chris Biemann (2020). *Fine-Tuning of Pre-Trained Transformer Networks for Offensive Language Detection.* https://arxiv.org/pdf/2004.11493.pdf

Mitchell Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, Michael Bernstein. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. http://www.kayur.org/papers/chi2021.pdf.

Sohail Akhtar, Valerio Basile, Viviana Patti. Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection. https://ojs.aaai.org/index.php/HCOMP/article/view/7473/7260