Experimental Design

Problem Statement

Cyberbullying detection suffers from many dataset quality issues, the largest of which has traditionally been class imbalance. This problem is exacerbated when discussing the fine-grained problem. Wang et. al[1] tried to address this problem by using a dynamic query expansion procedure, but upon close inspection this procedure has resulted in a large amount of tweets being mislabeled. Preliminary testing on the dataset reveals anywhere from 10-15% of all examples are labeled incorrectly. Mislabeled data is present throughout all 6 classes, but is most prevalent around the "other" and "not cyberbullying" classes. See Dataset Quality Issue for more information.

Another issue in the dataset is the conflation of bullying traces with instances of cyberbullying. Cyberbullying[2] definitions conflict, but primarily agree that it involves the "willful and repeated harm inflicted through the use of computers, cell phones, or other electronic devices." Bullying is a multifaceted action often involving parties beyond the perpetrator and victim, as can be seen in the diagram from Xu et. al[3]. Participants in a bullying episode often post on social media creating **bullying traces**. These can include reporting a bullying instance, accusing someone as a bully, revealing self as victim, and cyberbullying direct attacks. The following example displays how bullying traces have been misclassified as cyberbullying. Imagine trying to voice your pain, and then getting taken down for harassment.

Revealing self as victim, label - **age**

> new class and then this girl just interrupted me and said "you're gay" in front of everybody I never got bullied or anything and yet that ONE MOMENT was the turning point for ruining the first half of high school for me. kids are so mean and for WHAT?!?!

The domains of hate speech detection, abusive language detection, and detecting bullying traces are adjacent but distinct tasks from cyberbullying detection. Mislabeled data is the single greatest issue, and even manual annotation is flawed[4]. Zeerak Waseem used expert analysis to find serious issues with mislabeled data in his own dataset. The original Waseem dataset also happens to be the most significant dataset used in Wang's paper (greatest

number of tweets). The annotator bias problem is neatly summed with the following graphic. Cyberbullying classification is a hard problem even for humans.

Another key issue with the FGCD dataset is the lack of diversity in examples, and mislabeling of retweets as instances of cyberbullying themselves. The following screenshot of the dataset neatly illustrates my point. This lack of diversity if prevalent throughout all classes in the dataset. This is most likely a byproduct of the dynamic query expansion process. A common approach in the literature is to simply remove retweets and quote tweets from the dataset.

In a few words, cyberbullying detection faces sparse data, data mislabeling by humans and machines, and noisy data. Inherent problems with datasets cannot be solved by more advanced models. They will simply learn to misclassify better.

Potential Solutions

The need for a data-centric approach to cyberbullying classification was apparent to us even in the early stages of this project, but most of our early focus was on removing noise (tweets from other languages, spam tweets, etc.). While this approach yielded mild improvements in model performance it does not resolve the key issues of data. Knowing what we know, and trying to publish a paper where the very foundations are shaky is not how I want to begin my academic career. I layout three potential solutions to key issues, and at the end I will give my recommendation for project direction and experimental design.

Reducing/Eliminating Label Noise

If mislabeling is the key issue, then we should work to build systems to actively detect mislabeled examples. Unsupervised detection of corrupted labels is a topic that has exploded recently, particularly in the subfield of image classification. Many techniques[5] have shown promise in improving model accuracy even with a high degree of mislabeled data. Much of this research has not been applied to NLP tasks, creating a niche contribution for us to leverage.

There are many approaches to the noisy learning problem, the first being estimating T, the noise transition matrix. Let X represent the feature space, $Y, \widetilde{Y}$ represent the true and noisy label space respectively.

An example x is considered to be mislabeled if the given label $y_i \neq Y$ the true label. For a given c-class classification problem we can estimate the probability that that an example with class label c_i will we inaccurately flipped to class label c_j. This is known as the transition probability, it can be read as the probability that class label c_i will be mislabeled as c_j. The transition matrix T consists of all such probabilities.

$$T_{ij} = p(\widetilde{y} = c_j, y = c_i|x)$$

For a binary classification problem we may have a $[0, 1]$ label corresponding to not cyberbullying and cyberbullying respectively. We may have a sample transition matrix

$$T = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix}$$

In this example 10% of not cyberbullying tweets are mislabeled as cyberbullying & 20% of cyberbullying tweets are labeled as not cyberbullying. Note that the noise transition matrix is not symmetrical. Theory has shown that knowing the noise transition matrix in combination with reweighting the loss is equivalent to learning from a clean dataset, albeit with less examples. Unfortunately datasets do not come with the noise transition matrix as a csv file. Some notable methods for estimating the noise transition matrix not mentioned in the survey (footnote 5) include HOC[6], Feature based Methods[7], and Confident Learning[8].

Other methods for learning from noisy labels include adversarial training to improve resistance to noise, choosing loss functions robust to noise, loss adjustment by weighting noisy examples lower, sample selection to choose the cleanest examples for learning, collaborative learning, and model distillation.

Humans-In-The-Loop Systems/Active Learning

Active learning is a subset of humans-in-the-loop (HITL) systems. The motivation for this approach is simple: humans alone are biased annotators, so multiple annotators or subject matter experts are needed to label data. Machines, especially large language models do best with large amounts of available data. Active learning seeks to shift from a supervised to semi-supervised/self-supervised approach. We let models choose which examples are most important to be labeled to get the most bang for our buck. This is similar to the idea of "anchor points" found in many of the noisy learning methods. Active learning literature is abundant, but less well known. Here is an old review from 2009[9]. The most recent literature review[10] is from 2020.

This approach makes the most inherent sense for the cyberbullying problem. Most social media platforms use a combination of AI and labor for content moderation. The intuition is clear - a large portion of tweets have obvious classifications: one with the n-word is (probably) clear racism, a positive tweet about *The Great British Baking Show* is the furthest thing from cyberbullying. The examples on the margins are the most difficult to classify, and that is where the oracle (usually a human, or team of humans) comes into play.

Active learning theory focuses on finding the optimal **query strategy** for streaming examples to be labeled. The simplest strategies revolve around random sampling, and get consistently more complicated. Querying new examples for labeling based on active learning techniques resembles a dynamic query expansion

procedure with HITL. Optimizing query strategy with another DNN is known as learning to learn or *meta-learning*.

HITL/AL systems have shown to improve model accuracy beyond naively labeling mass amounts of data, or using increasingly overparametrized networks. Some noisy learning techniques even leverage HITL to relabel examples, known as label restoration. From a research direction perspective very few HITL/AL papers have been written with transformers or on hate speech classification.

Rich Tweet Embeddings by Leveraging Graph Neural Networks

Tweets are not just short pieces of text - more often than not they are parts of complex human interactions generating network data. Decontextualizing tweets to text loses critical context for classification. Retweet propagation, speed of spread, user characteristics, time, and geospatial data are all parts of a single tweet. Earlier this year we were discussing the problem of fake news, and difficulties in identifying the ground truth. Lauren mentioned research showing that fake news spread faster than real news, and that these characteristics were useful in classification. Some researchers leveraged rich tweet network data including user interactions, word embeddings, and retweet propagation in combination with graph neural networks (GNN) for fake news classification. Rich tweet data improved fake news classification by nearly 20 accuracy points[11].

I strongly believe that applying such techniques to the cyberbullying problem would produce impressive results. User information and interactions are key elements to differentiate targeted cyberbullying from other negative sentiment tweets. Is being mean to a president really cyberbullying? (I honestly don't know the answer this question, showing the inherent nuance and difficulty of this problem.)

Proposal for Research Direction

Each of the 3 solutions I've mentioned could be their own project. Unfortunately our time and resources are limited. I believe that building systems to learn from noisy tweet labels would be the best project we can accomplish. The experimental design is as follows:

- Pick 3-4 promising techniques to learn with noisy labels
- Find 2-3 clean hate speech/cyberbullying detection datasets, and test technique strength by injecting synthetic label noise into the dataset
- Leverage the best technique on real twitter data and observe real-world performance

It seems so simple when I write it out like that :)

---

1. SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection
2. E. Englander, E. Donnerstein, R. Kowalski, C. A. Lin, and K. Parti, "Defining Cyberbullying," *Pediatrics*, vol. 140, no. Supplement_2, pp. S148–S151, Nov. 2017, doi: 10.1542/peds.2016-1758U.
3. J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from Bullying Traces in Social Media," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, Jun. 2012, pp. 656–666. Accessed: Jun. 10, 2022. [Online]. Available: https://aclanthology.org/N12-1084
4. Z. Waseem, "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter," in *Proceedings of the First Workshop on NLP and Computational Social Science*, Austin, Texas, Nov. 2016, pp. 138–142. doi: 10.18653/v1/W16-5618.
5. H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from Noisy Labels with Deep Neural Networks: A Survey." arXiv, Mar. 09, 2022. doi: 10.48550/arXiv.2007.08199.
6. Z. Zhu, Y. Song, and Y. Liu, "Clusterability as an Alternative to Anchor Points When Learning with Noisy Labels," Feb. 2021, doi: 10.48550/arXiv.2102.05291.
7. Z. Zhu, Z. Dong, and Y. Liu, "Detecting Corrupted Labels Without Training a Model to Predict," arXiv, arXiv:2110.06283, Jan. 2022. doi: 10.48550/arXiv.2110.06283.
8. C. Northcutt, L. Jiang, and I. Chuang, "Confident Learning: Estimating Uncertainty in Dataset Labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, Apr. 2021, doi: 10.1613/jair.1.12125.
9. Active Learning Literature Survey, https://minds.wisconsin.edu/handle/1793/60660
10. P. Ren *et al.*, "A Survey of Deep Active Learning," Aug. 2020, doi: 10.48550/arXiv.2009.00236.
11. Y.-J. Lu and C.-T. Li, "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media." arXiv, Apr. 24, 2020. doi: 10.48550/arXiv.2004.11648.