**COUNTER SPEECH**

* Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: A Comparative Study
  + **CN data collection.** The existing studies for collecting CN datasets employ four main approaches. Crawling consists in automatically scraping websites, starting from an HS content and searching for possible CNs among the responses (Mathew et al., 2018, 2019). With crowdsourcing CNs are written by non-expert paid workers as responses to provided hate content (Qian et al., 2019). Niche sourcing relies on a niche group of experts for data collection (De Boer et al., 2012), and it was employed by Chung et al. (2019) for CN collection using NGO’s operators. Hybrid approaches use a combination of LMs and humans to collect data (Wallace et al., 2019; Dinan et al., 2019; Vidgen et al., 2020). Studies on CN collection are presented in more detail by Tekiro˘glu et al. (2020); Fanton et al. (2021).

**HATE SPEECH**

* DEEP LEARNING FOR HATE SPEECH DETECTION: A COMPARATIVE STUDY (My conclusion and future research ideas from this paper):
  + **There is a need to develop a model which will detect hate speech with high accuracy.**
  + **We don’t have model that is trained on all the different categories of hate speech, and for that we don’t have a large dataset which has hate speech of all categories which can be used for training the model.**
  + **We don’t have a large dataset that gives us the whole contextual information before identifying whether a statement is hateful or not.**
  + **Hence we should have a large dataset having both contextual information and different categories of hate speech for building a model which efficiently and effectively detects hate speech.**
* Gupta and Waseem (2017): The results showed that domain-specific word embedding has a desirable performance and is suitable for unbalanced classes datasets.
* Liu (2018): employed domain-specific word embedding model trained on the articles from hate speech websites and high centrality users’ tweets to reach to the semantics of code words used in hate speech. They experimented on CNN, and LSTM models and concluded that CNN performed better than LSTM on tweets due to the length of tweets. The achieved f1-score is 78% given that they experimented on the previous tweet-length 180 characters.
* Mozafari et al. (2020) studied the performance of the BERT language model on hate speech detection as a multi-class problem in a recently released work. They employed a BERT basis and a variety of classifiers, including CNN, which provided the highest score that reached to 92% f1-score.
* Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model: However, domain-specific and word embedding overcomes BERT model in that it can detect hate terms abbreviations and intentionally misspellings meaning.
* **NB: (Just a random Idea) Can we have each cell of neural network a ML classifier like LR, SVM, Random Forest, etc. and hence can we combine the power of both Deep neural network along with ML models.**