

# **EXTREME SPEECH DETECTION WITH NLP**

*An overview of extreme speech detection work,  
motivations and challenges from the perspective of NLP*

# TABLE OF CONTENTS

**01**

**ML IN NLP**

**02**

**MOTIVATION**

**03**

**CURRENT WORK**

**04**

**CHALLENGES**

**05**

**SUMMARY**

**01**

**ML IN NLP**

# MACHINE LEARNING IN NLP

- Tools are built around Machine Learning (ML) models
- These models take as input text and make predictions
- Predictions can be extreme speech severity, targets, etc.
- To train these models, annotated data is required (text + label)

these lazy greeks stealing our jobs

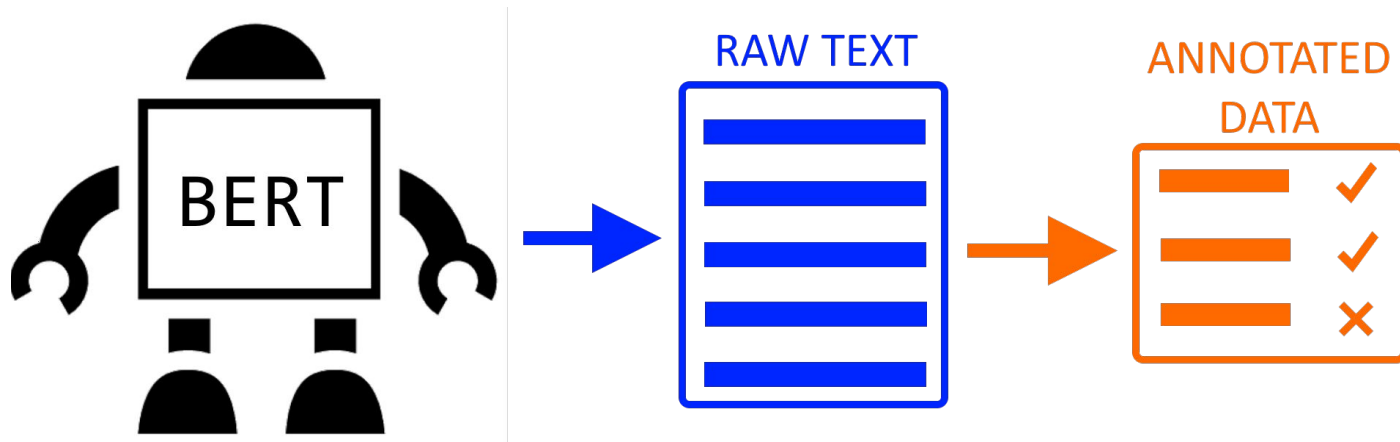
extreme speech

I was in Greece recently, I was so tanned

normal speech

# MACHINE LEARNING IN NLP

- Recently, *pretrained* models (for example, *BERT*) have gained traction. These models are generally bigger and are pretrained on raw text data before training on the annotated dataset.
- We use off-the-shelf models that have already been pretrained and we only need to further train them on our annotated data.



# MACHINE LEARNING IN NLP

- Due to their size, these models require a lot of resources to train. Not only are large quantities of data required (multiple GBs) but also expensive computing infrastructure (eg. GPUs).
- Larger models generally perform better than traditional machine learning models and have superseded them in academic circles.
- Oftentimes quantity is favoured over quality when it comes to data. It has been shown that some of the pretraining datasets can propagate existing biases. This is something we need to keep in mind, especially when dealing with sensitive topics.
- Smaller models still find use in industry applications, where computational efficiency is of importance.

**02**

**MOTIVATION**

# MOTIVATION

- Extreme speech has been flooding online platforms at an alarming rate
- Apart from harming target groups, it has become a political weapon that poisons online discourse and radicalizes people
- Manual filtering is costly and requires a lot of effort
- Automatic methods can be used either to aid humans make decisions, or filter content on their own
- Unfortunately, peddlers of extreme speech have been getting more and more evasive (introducing lingo, abbreviations, typos)



**03**

**CURRENT WORK**

# CURRENT WORK

- Earlier work examined classification of neutral, offensive and hateful speech. These are the terms most often used in NLP.
- Neutral language is normal speech, offensive language is speech that contains swearwords and/or slurs, while hateful speech is speech targeting marginalized groups.
- Data is collected via three main methods:
  - Keyword querying on social media
  - Sampling random social media posts
  - Mass collecting data from known extreme speech hotspots
- Collected data is then given to annotators, trained or untrained, for labelling

# CURRENT WORK

- Recently, work has become more broad
- Data is annotated with fine-grained labels, as well as honing in on specific issues, like misogyny.
- More aspects of extreme speech are captured:
  - targets (eg. religion, ethnicity, gender)
  - other forms of unacceptable content (eg. lewdness, call to violence).
- Through interdisciplinary collaboration, a theoretical base is being set, like the design of taxonomy trees for hateful content and foundational work on global definitions

# CURRENT WORK

- Efforts have been made for more inclusivity in the research body
- While English is still the most dominant language, a lot of data has been collected for other languages as well
- Work has been done to counter certain social imbalances. It has been found that text written in the African American English dialect is classified as extreme speech more often than Standard American English. There has been work ongoing to remove this bias exhibited in models, with more research to come.
- Recent works have started to employ trained annotators to ensure labelling is as informed as possible. Input from these annotators is considered during design of the labelling scheme.

# CURRENT WORK

	Davidson et al. (2017)	Founta et al. (2018)	Sap et al. (2019)
DistilRoBERTa	73.8	74.2	52.8

Work by Marc-Anthony Bauer

- We evaluate a BERT variant model on three datasets, reporting the F1 score over three labels (neutral/offensive/hateful).
- Performance is quite low across the board, with no model scoring higher than 75. This is an indication of how challenging this task is.
- Performance is even lower when traditional ML models are used

# **04 CHALLENGES**

# CHALLENGES

- Subjectivity: This is a very heavily subjective topic, with differing opinions and perspectives on what constitutes extreme speech.
- Varying definitions: Even though there has been work to settle definitions (for example, taxonomy trees), there is still little common ground between works.
- World Knowledge: In a lot of cases, to infer the extremity of speech, sociopolitical knowledge and expertise is needed.
- Representation: Demographics targeted by extreme speech are usually not in the annotator pool. For example, in Founta et al. 66% of annotators are male and in Sap et al. 82% are white.

# CHALLENGES

- False Negatives: A lot of hate speech examples are getting classified as offensive language. This occurs because of the overlap in vocabulary between the two labels. For example, the “n-word” is a slur that has been associated with hate speech, while in some instances it can be acceptable.

False negatives are dangerous, since failing to filter hate speech may lead to further harm against target groups.

True label	hate	0.25	0.66	0.09
	offensive	0.01	0.96	0.03
	neither	0.01	0.04	0.95
			hate	offensive
		Predicted label		

BERT performance on Davidson et al.



# CHALLENGES

- Generalization: A model trained on one dataset does not necessarily work well on other datasets and external data.

Much like different datasets exhibit differences in definitions, data collection, and other features, different social media platforms face different challenges as well.

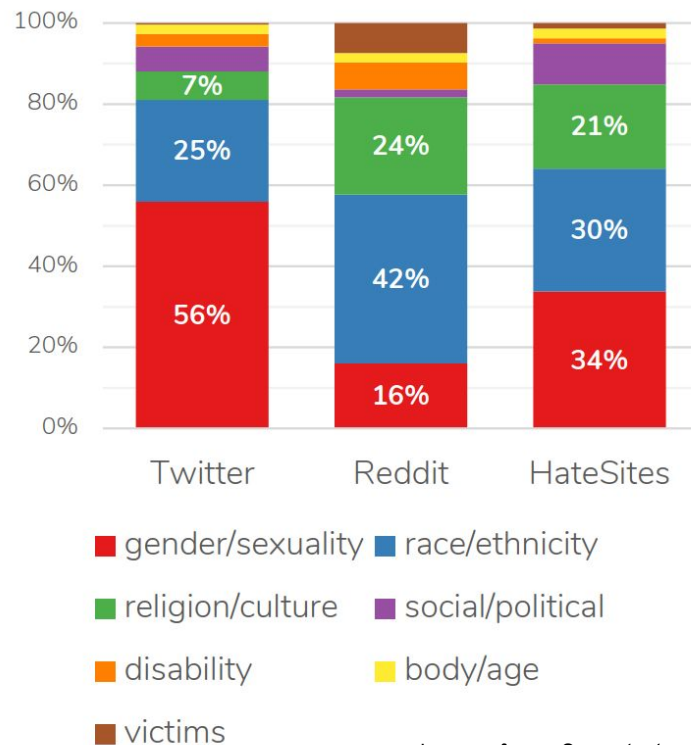
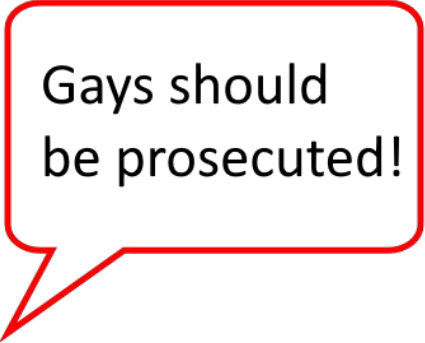



Image from Sap et al.  
(2020)

# CHALLENGES

- Little Context: Examples are presented in an isolated, out-of-context form, making it harder to infer their severity.
- Negative Bias: Certain innocuous words/terms appear in negative contexts and the model learns to associate them with extreme speech. For example, “jew” or “gay” are terms that lead models to flag the text as extreme speech even when it is not.



Gays should  
be prosecuted!



It doesn't  
matter if  
they are gay

**05**

**SUMMARY**

# SUMMARY

- A lot of effort has been invested in building NLP tools to battle extreme speech found online
- These tools employ Machine Learning models, with more recent models being very large and requiring a lot of resources
- NLP research in the area has focused on curation of datasets from online platforms, as well as theoretical groundwork
- The community faces challenges related to both data and modelling, but if current zeal is an indicator of future success, we should be expecting more breakthroughs in the near future

**THE END**

# REFERENCES

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding
- Antigoni-Maria Founta et al. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. 2018.
- Maarten Sap et al. Social Bias Frames: Reasoning about Social and Power Implications of Language. 2019.
- Jae Yeon Kim et al. Intersectional Bias in Hate Speech and Abusive Language Datasets. 2020.
- Paula Fortuna and Sergio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text.
- Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: the case of the european refugee crisis.
- Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny.