“It ain’t all Good”

Machinic Abuse Detection and Marginalization in Machine Learning

16/06/2021

Zeerak
University of Sheffield
Twitter: @zeerakw

Disclaimers

Source: Disney

Source: Disney

The problem of online abuse

Source: Shutterstock

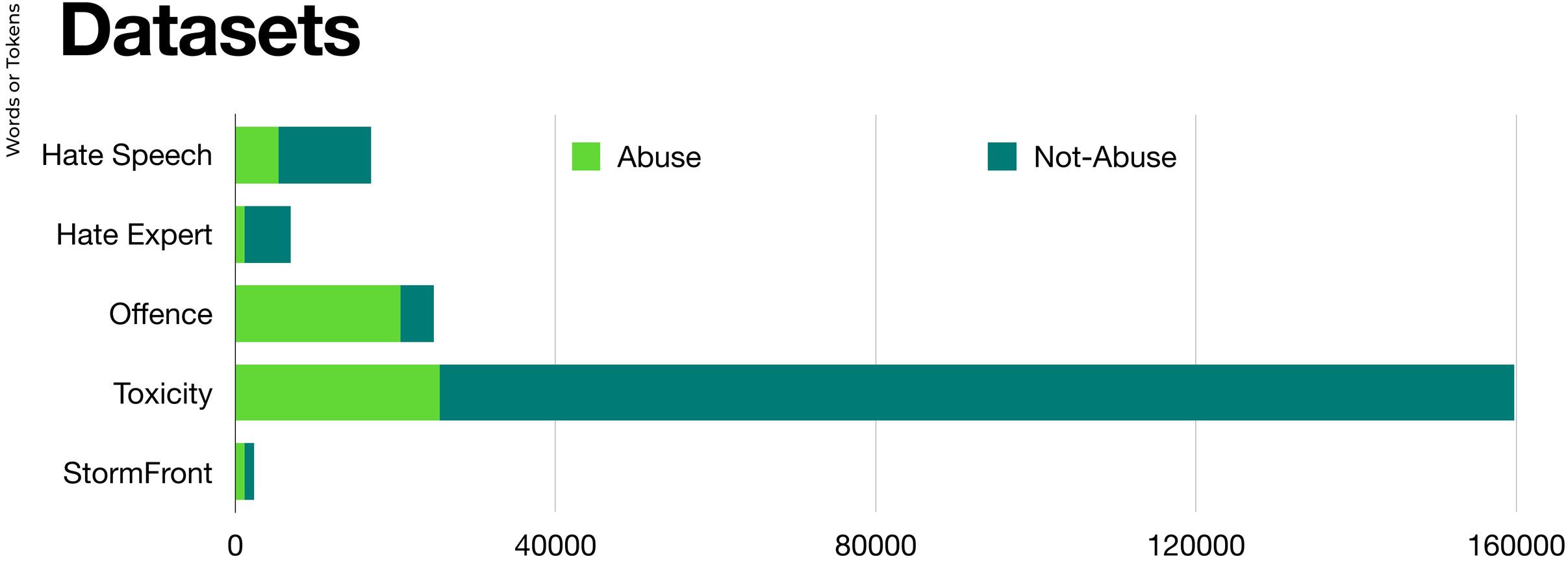
Words or Tokens
Rethinking input data representation for abuse detection

Tokenisations

- Surface Forms
- Byte-Pair Embedding Vocabulary
- LIWC
  - 92 LIWC dimensions mapping 6,400 words
    - E.g. ‘cried’ invokes sadness, negative emotion, overall affect, verbs, and past focus.

- Heinzerling, Benjamin, and Michael Strube. 2018. “BPEmb: Tokenization-Free Pre-Trained Subword Embeddings in 275 Languages.” In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1473.
- Pennebaker, James W., Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. “The Development and Psychometric Properties of LIWC2015.” Austin, Texas: University of Texas at Austin.

Datasets



Waseem, Zeerak, and Dirk Hovy. 2016. “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.” In *Proceedings of the NAACL Student Research Workshop*

Waseem, Zeerak. 2016. “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*.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. “Automated Hate Speech Detection and the Problem of Offensive Language.” in *Proceedings of IWCSM*.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. “Ex Machina: Personal Attacks Seen at Scale.” In *Proceedings of the 26th International Conference on World Wide Web*.

Gibert, Ona de, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. “Hate Speech Dataset from a White Supremacy Forum.” In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*.

Tokenisations

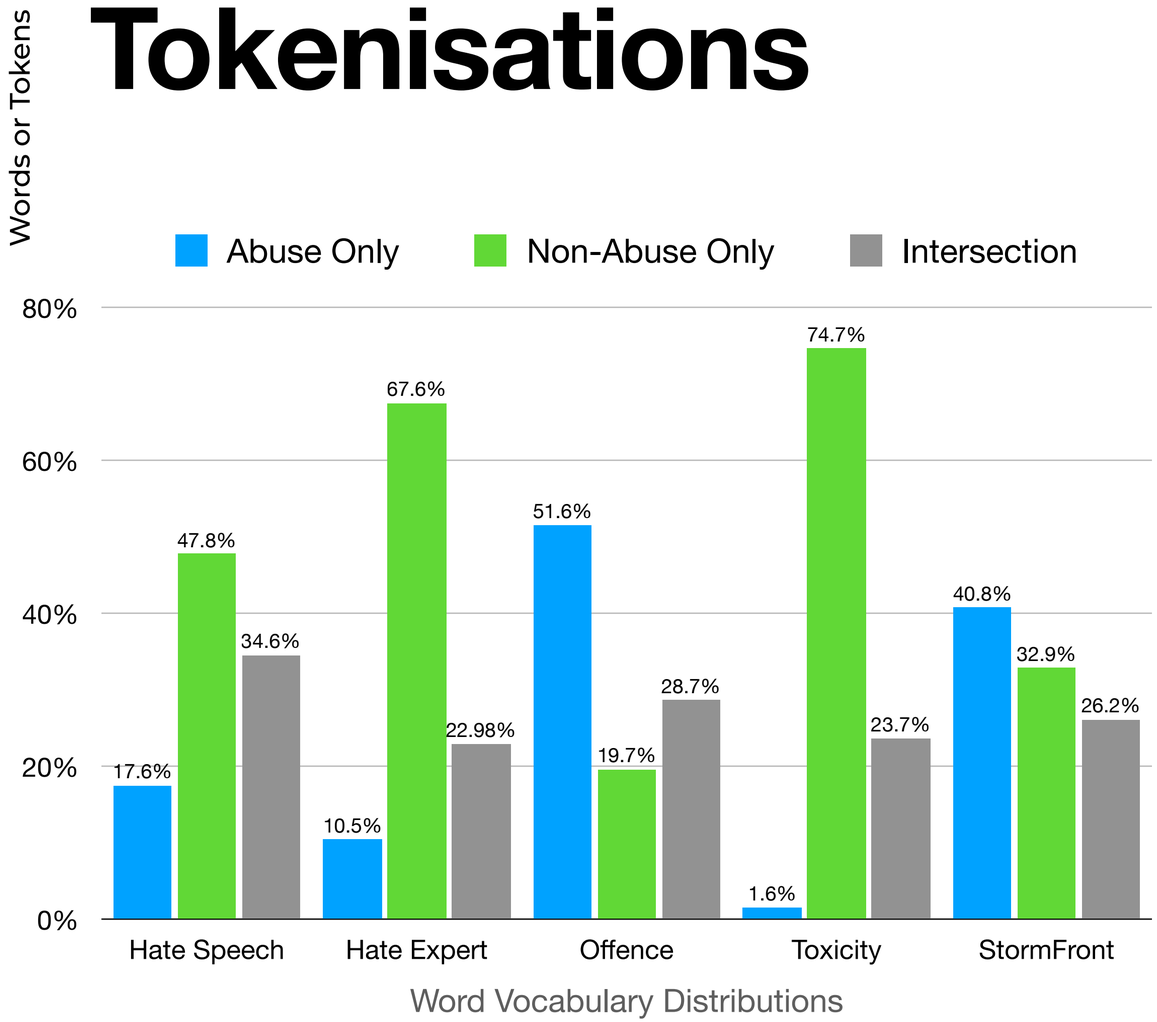
Human Traffic (1999)

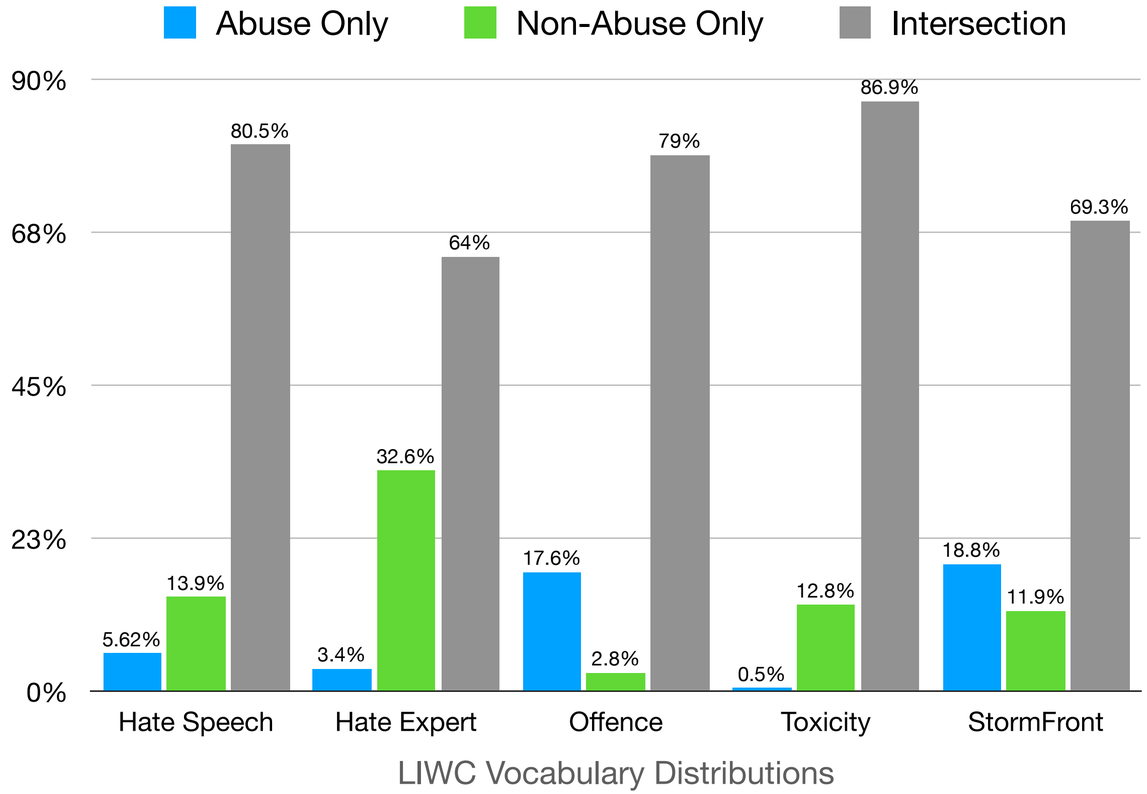
Tokenisations

"Fucking nice one bruva!"
- John Simm (Human Traffic, 1999)

	Surface	BPE	LIWC
Fucking	Fucking	_Fucking	AFFECT_SEXUAL_BIO_INFORMAL_NEGEMO_ANGER_ADJ_SWEAR
nice	nice	__nice	POSEMO_AFFECT_ADJ
one	one	__one	NUMBER
bruva	bruva	_bru va	<UNK>
!	!	!	<UNK>

Tokenisations





Tokenisations

	Surface	BPE	LIWC
Hate Speech	14,730	14,834	837
Hate Expert	9,110	9,181	739
Offence	16,768	16,663	851
Toxicity	95,710	95,712	1,022
StormFront	5,566	5,510	622

Vocabulary Sizes in raw counts

Model Designs

- Models
  - MLP
  - CNN
  - LSTM
- Architectural details
  - No pre-trained embeddings

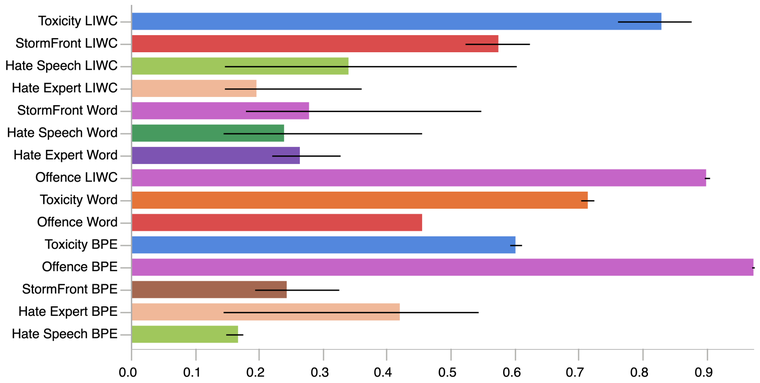
	MLP	CNN	LSTM
Word	27,009,400	8,686,464	40,884,200
BPE	6,144,920	19,756,936	19,461,400
LIWC	169,000	256,768	707,600

Model size by parameters for models optimized on Toxicity (Wulczyn et al. 2017)

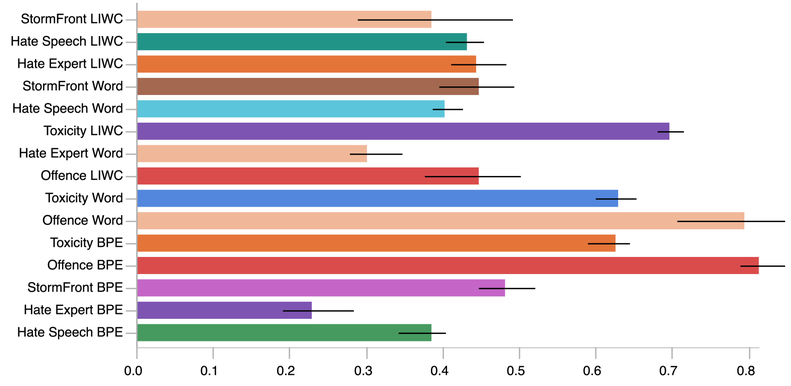
Words or Tokens

Results

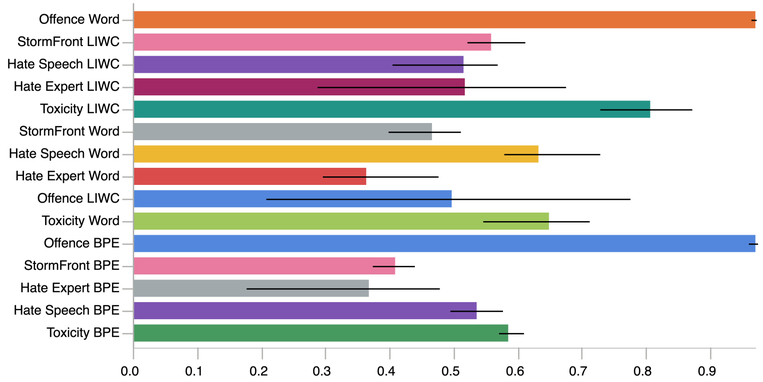
Classification Results



Offence MLP F1



Offence LSTM F1



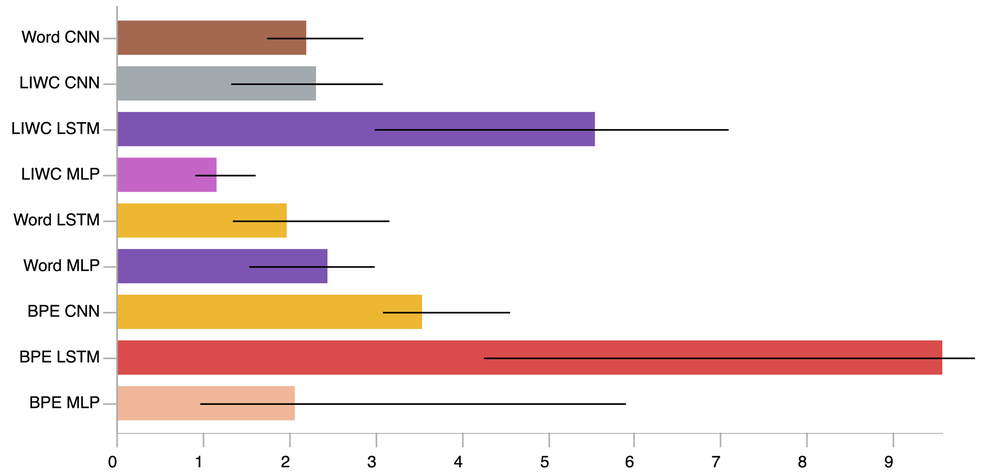
Offence CNN F1

Words or Tokens

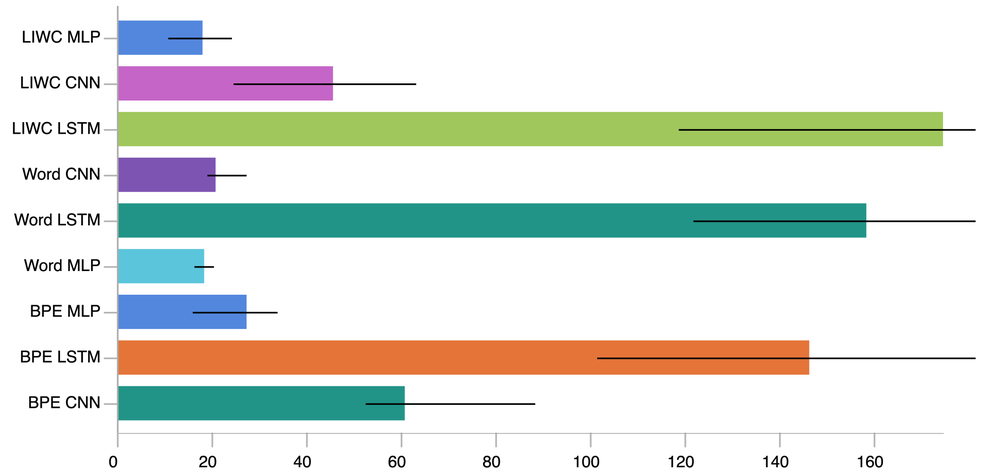
Results

Optimisation Time

- Measured in terms of minutes
- Measurements are on optimization time, not inference.



[Offence] Optimisation time (M)

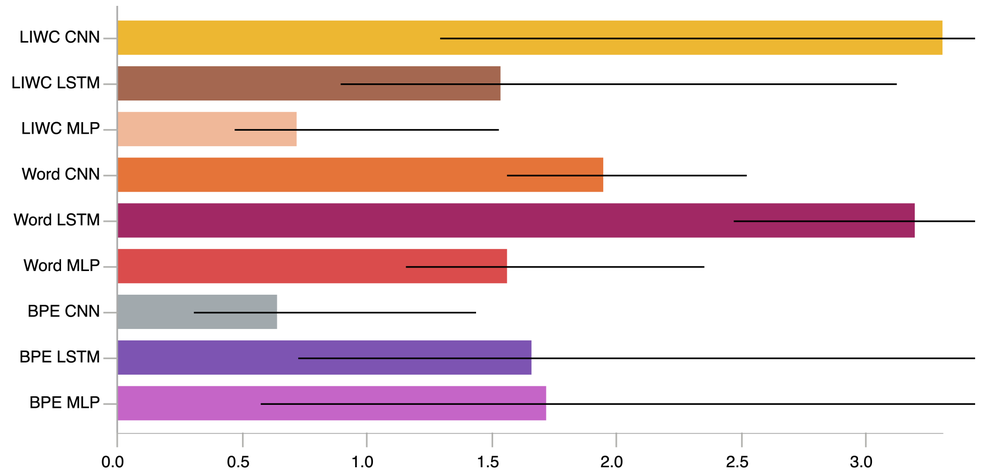


[Toxicity] Optimisation time (M)

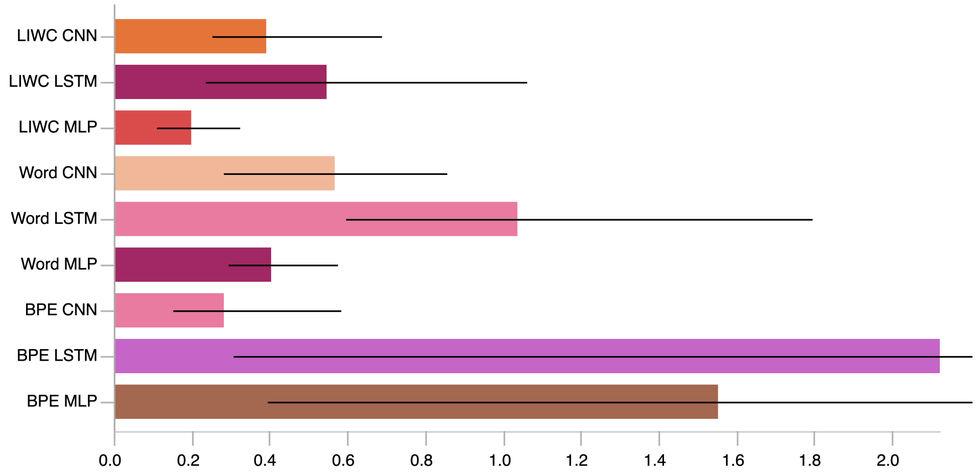
Words or Tokens

Results

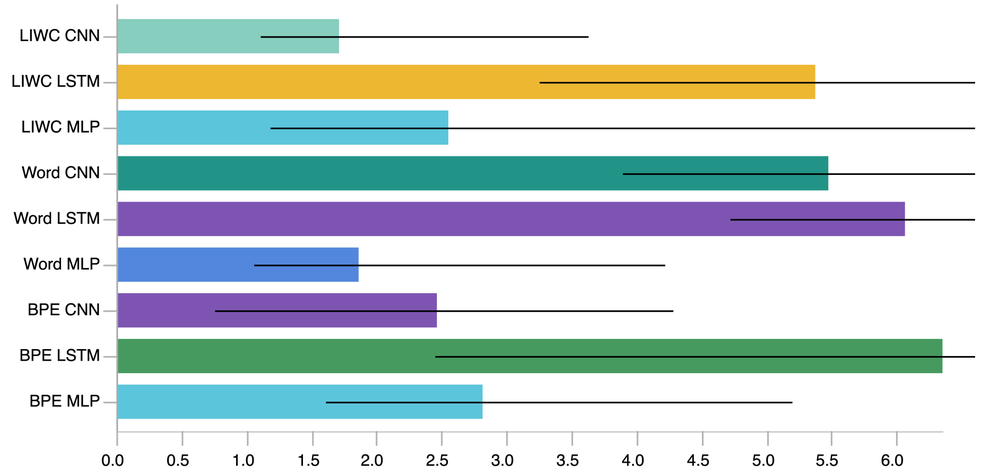
Optimisation Time



[Hate Expert] Optimisation time (M)



[StormFront] Optimisation time (M)



[Hate Speech] Optimisation time (M)

Tasks that Matter
Multi-task Learning for Abusive Language Detection

Main Plot

Selecting the main task

- Abusive Tasks
  - Hate Speech
  - Offence
  - Toxicity

An auxiliary story
Selecting tasks

- Abusive Tasks
  - Hate Speech
  - Hate Expert
  - Offence
  - Toxicity
- Non-abusive tasks
  - Sarcasm detection
  - Fact/Feeling-based argument identification
  - Moral Sentiment detection

Rajamanickam, Santhosh, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. "Joint Modelling of Emotion and Abusive Language Detection." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
Waseem, Zeerak, James Thorne, and Joachim Bingel. 2018. "Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection." In *Online Harassment*, edited by Jennifer Golbeck, Springer International Publishing
Oraby, Shereen, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. "Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue." In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.
Oraby, Shereen, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. 2015. "And That's A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue." In *Proceedings of the 2nd Workshop on Argumentation Mining*,
Hoover, Joe, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, et al. 2020. "Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment." *Social Psychological and Personality Science* 11 (8).
Röttger, Paul, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2020. "HateCheck: Functional Tests for Hate Speech Detection Models." *To appear at ACL 2021*.

Tokenisations

- Surface Forms
- Byte-Pair Embedding Vocabulary

Model Designs

- Baseline models
  - Ensemble classifier
  - Single-task SVM
  - Single-task MLP
- Experimental Models
  - Multi-task MLP
- Architectural details
  - No pre-trained embeddings

Results

	Accuracy	Precision	Recall	F1-score
Linear	**0.8871**	**0.6997**	0.6789	**0.6850**
Ensemble	0.7729	0.4241	0.3948	0.3946
MLP	0.8790	0.5625	**0.9163**	0.5721

[Offence] Baseline Model scores.

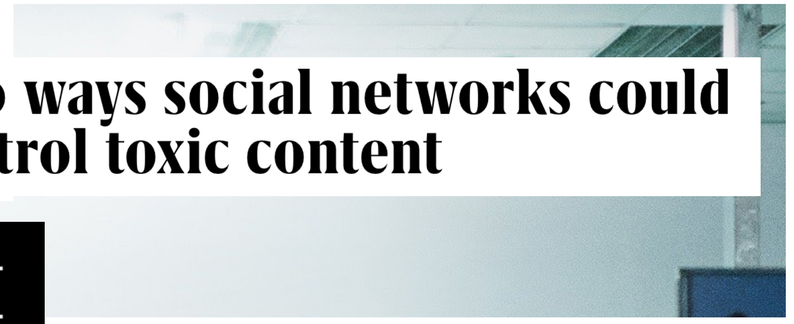
	Accuracy	Precision	Recall	F1-score
Hate Speech	0.8970 (σ 0.0028)	0.6952 (σ 0.0240)	0.7685 (σ 0.0339)	0.7119 (σ 0.0138)
Toxicity	0.8916 (σ 0.0036)	0.6829 (σ 0.0176)	0.7659 (σ 0.0074)	0.7099 (σ 0.0159)
Hate Expert	0.8871 (σ 0.0048)	0.6794 (σ 0.0274)	0.7552 (σ 0.0110)	0.7028 (σ 0.0193)
Sarcasm	0.8933 (σ 0.0029)	0.6790 (σ 0.0303)	0.7594 (σ 0.0141)	0.6987 (σ 0.0271)
Argument Basis	0.8973 (σ 0.0045)	0.6740 (σ 0.0238)	0.7775 (σ 0.0202)	0.6847 (σ 0.0207)
Moral Sentiment	0.8956 (σ 0.0045)	0.6860 (σ 0.0344)	0.7541 (σ 0.0267)	0.6938 (σ 0.0356)
Hate Speech | Toxicity	0.8860 (σ 0.0046)	0.6961 (σ 0.0157)	0.7294 (σ 0.0211)	0.7033 (σ 0.0135)
Sarcasm | Hate Speech	0.8891 (σ 0.0080)	0.6561 (σ 0.0239)	0.7757 (σ 0.0157)	0.6806 (σ 0.0200)
Sarcasm | Toxicity	0.8977 (σ 0.0011)	0.6723 (σ 0.0126)	**0.7899** (σ 0.0063)	0.6811 (σ 0.0227)
Sarcasm | Toxicity | Hate Speech	0.8923 (σ 0.0028)	0.6877 (σ 0.0213)	0.7601 (σ 0.0079)	0.7091 (σ 0.0151)
Argument Basis | Hate Speech	0.8864 (σ 0.0036)	0.7043 (σ 0.0320)	0.7386 (σ 0.0173)	0.7155 (σ 0.0139)
Argument Basis | Toxicity	0.8917 (σ 0.0052)	0.6929 (σ 0.0203)	0.7552 (σ 0.0049)	0.7143 (σ 0.0139)
Argument Basis | Sarcasm	0.9010 (σ 0.0022)	0.6939 (σ 0.0170)	0.7791 (σ 0.0110)	0.7105 (σ 0.0190)
Argument Basis | Hate Speech | Toxicity	0.8972 (σ 0.0036)	0.6861 (σ 0.0089)	0.7723 (σ 0.0085)	0.7064 (σ 0.0083)
Argument Basis | Hate Speech | Sarcasm	**0.9025 (σ 0.0007)**	**0.7107 (σ 0.0090)**	0.7820 (σ 0.0093)	**0.7291 (σ 0.0072)**
Argument Basis | Toxicity | Sarcasm	0.8925 (σ 0.0044)	0.6973 (σ 0.0218)	0.7472 (σ 0.0082)	0.7138 (σ 0.0125)
Argument Basis | Hate Speech | Toxicity | Sarcasm	0.8948 (σ 0.0016)	0.6979 (σ 0.0048)	0.7652 (σ 0.0102)	0.7100 (σ 0.0049)

[Offence] Experimental model scores.

Don’t you know that you’re toxic? The Politics of Toxicity in Content Moderation Infrastructures

Clean up the Internet

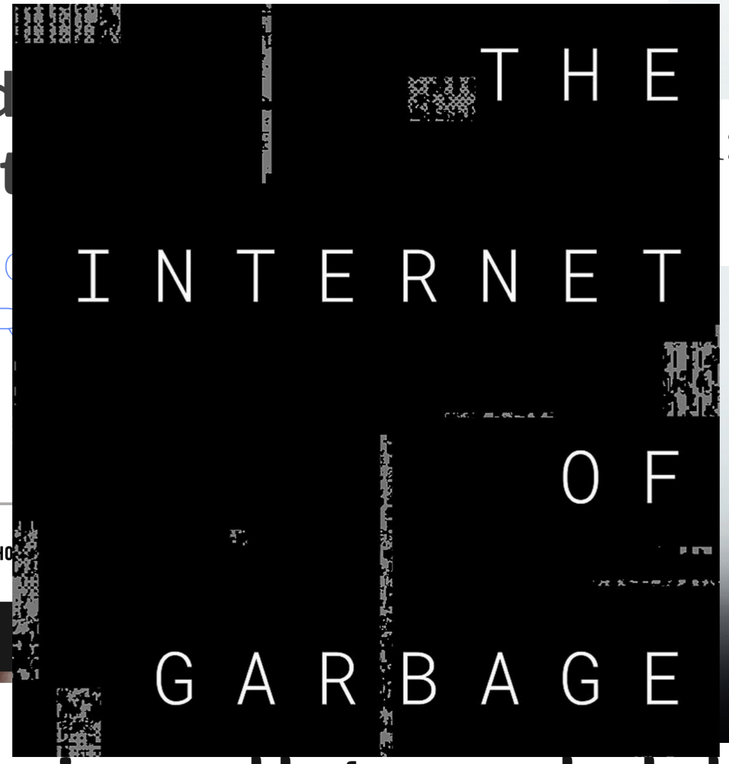
Clean up the internet is an independent, UK-based organisation concerned about the degradation in online discours
its implications for democracy. We campaign for evidence-based action to increase civility and respect online, and to
online bullying, trolling, intimidation, and misinformation.



Two ways social networks could control toxic content

Unilever warns social med
to clean up "toxic" content

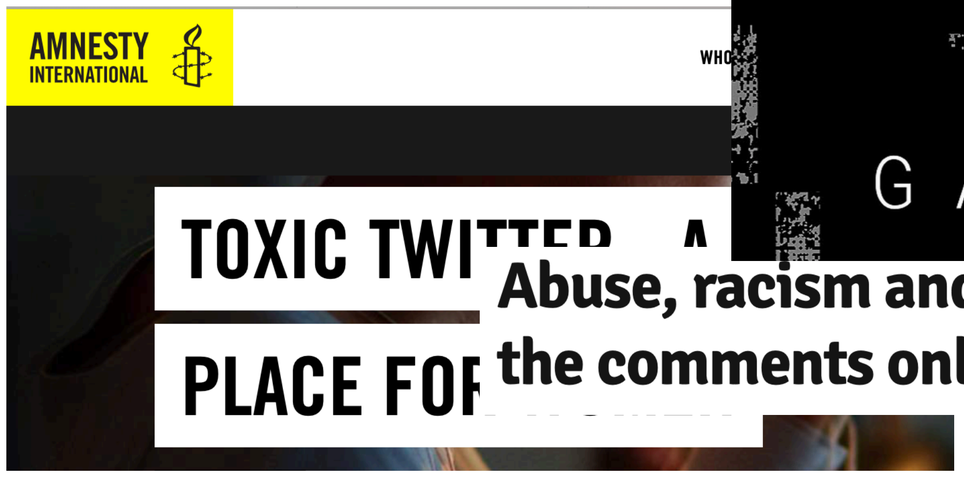
Toxic Content on Digital Pl
the rise. How Can Brands R
Consumer Trust?



: The Internet's Great Spring-

AMNESTY INTERNATIONAL

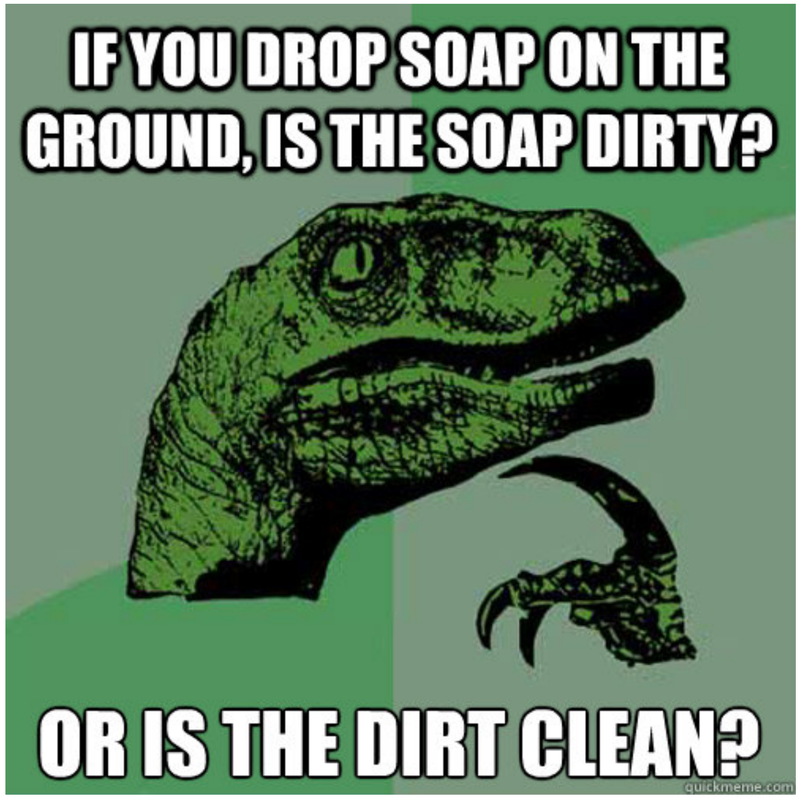
WHO



Abuse, racism and hate speech - help us clean up the comments online

Moderating Pollution

- The “toxicity” of words
- Definitional power of “toxic” discourses
- Downstream effects



State of the Art White Supremacy On Disembodiment in the Machine Learning Pipeline

“No one ever accused the God of monotheism of objectivity, only of indifference”

Donna Haraway (1988),
Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective.

The Ghost in the Model

- The Designer
- The annotation process
- Data selection
- Modelling

Futures

References

- Kulynych, Bogdan, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020. “POTs: Protective Optimization Technologies.” In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 177–88. Barcelona Spain: ACM.
- Heinzerling, Benjamin, and Michael Strube. 2018. “BPEmb: Tokenization-Free Pre-Trained Subword Embeddings in 275 Languages.” In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1473.
- Pennebaker, James W., Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *“The Development and Psychometric Properties of LIWC2015.”* Austin, Texas: University of Texas at Austin.
- Waseem, Zeerak, and Dirk Hovy. 2016. “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.” In Proceedings of the NAACL Student Research Workshop, 88–93. San Diego, California: Association for Computational Linguistics.
- Waseem, Zeerak. 2016. “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, 138–42. Austin, Texas: Association for Computational Linguistics.
- Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. “Automated Hate Speech Detection and the Problem of Offensive Language.” ICWSM 2017
- Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. “Ex Machina: Personal Attacks Seen at Scale.” In Proceedings of the 26th International Conference on World Wide Web, 1391–99. Perth Australia: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3038912.3052591.
- Gibert, Ona de, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. “Hate Speech Dataset from a White Supremacy Forum.” In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 11–20. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5102.
- Rajamanickam, Santhosh, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. “Joint Modelling of Emotion and Abusive Language Detection.” In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- Waseem, Zeerak, James Thorne, and Joachim Bingel. 2018. “Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection.” In Online Harassment, edited by Jennifer Golbeck, Springer International Publishing.
- Hoover, Joe, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, et al. 2020. “Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment.” Social Psychological and Personality Science 11 (8).
- Oraby, Shereen, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. 2015. “And That’s A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue.” In Proceedings of the 2nd Workshop on Argumentation Mining, Association for Computational Linguistics.
- Oraby, Shereen, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. “Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue.” In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics.
- Röttger, Paul, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2020. “HateCheck: Functional Tests for Hate Speech Detection Models.” To appear at ACL 2021.
- Haraway, Donna. 1988. “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective.” Feminist Studies 14 (3).