Aggression and Misogyny Detection using BERT: A Multi-Task Approach

Niloofar Safi Samghabadi*, Parth Patwa*,
Srinivas PYKL, Prerana Mukherjee,
Amitava Das, Thamar Solorio

Motivation

❖ **Social Media:** Important and influential means of communication.

❖ Some people misuse them by engaging in aggressive behavior and by spreading hateful content.

❖ This antisocial behavior causes **disharmony** in society.

❖ It is **not possible** to moderate online content manually due to the time and cost.

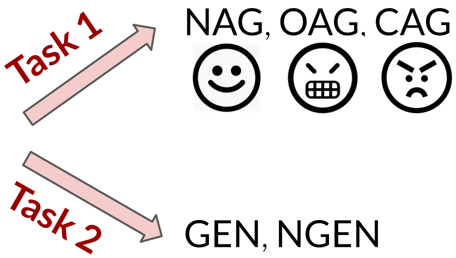
❖ **Solution:** Build automatic model to identify aggression and hate-speech.

Problem Statement

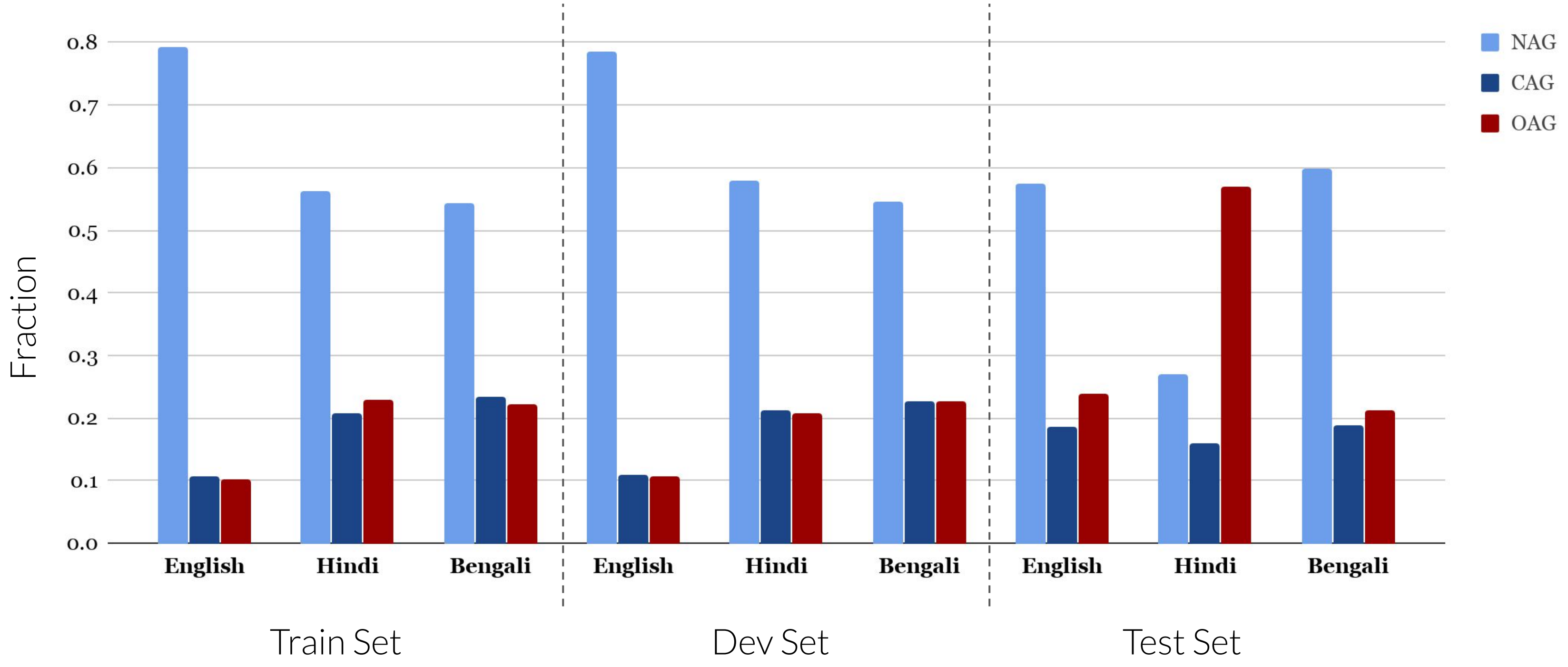
- Let's assume that $\{w_1, w_2, ..., w_n\}$ is the sequence of words in a comment.
- We aim at creating a model that given this input
  - **Task 1:** identify whether it is aggressive.
  - **Task 2:** identify whether it is gendered.

	Train	Dev
English	4,263	1,066
Hindi	3,984	997
Bengali	3,826	957

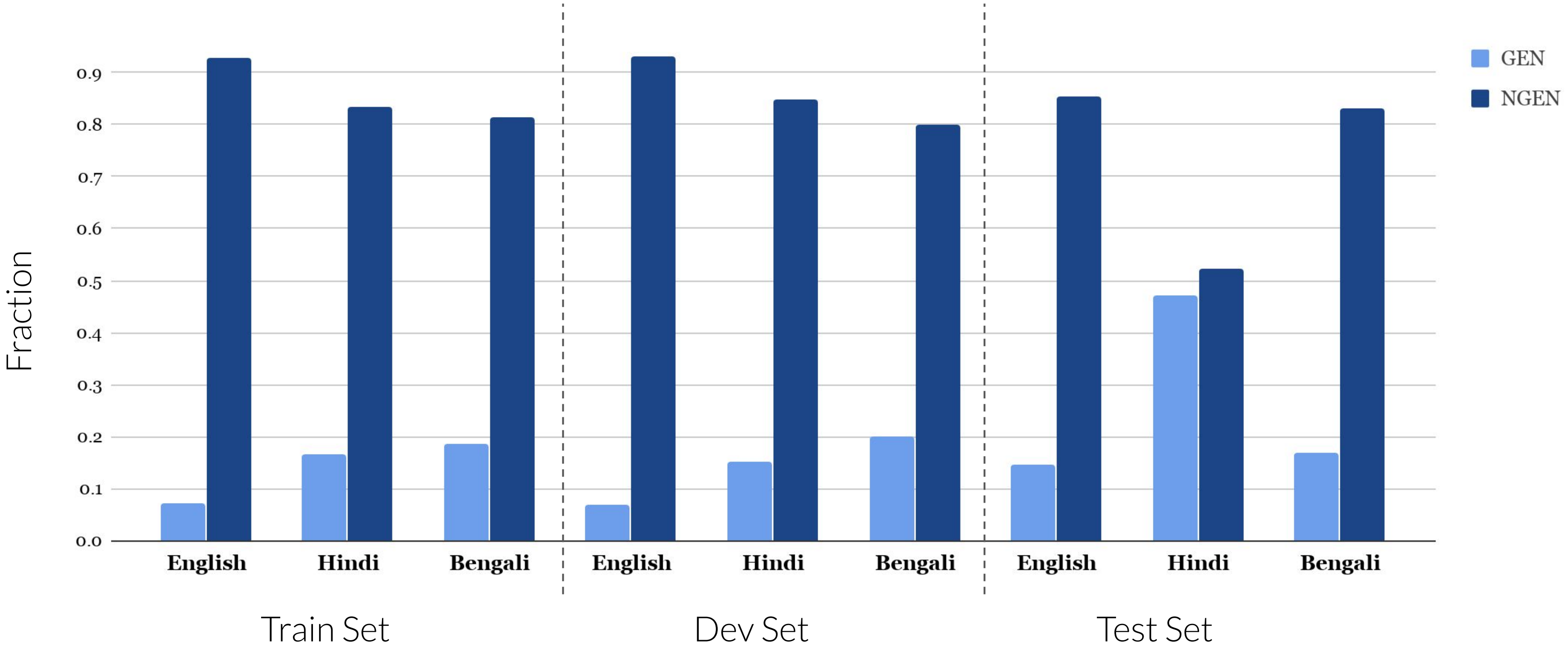
	Test
English	1,200
Hindi	1,200
Bengali	1,188



Data Statistics: Sub-task A

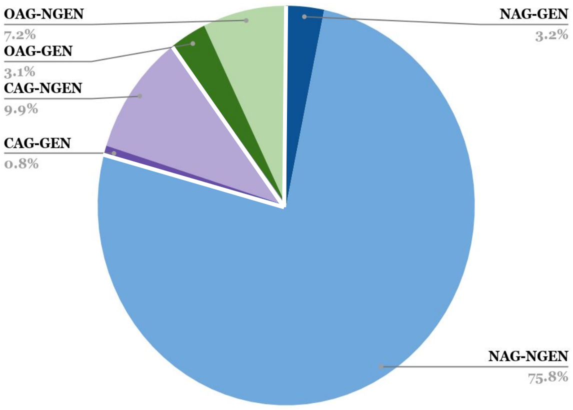


Data Statistics: Sub-task B

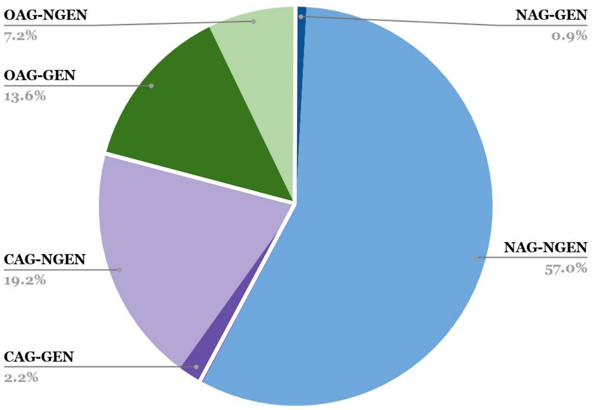


Co-occurrence of Sub-task Labels

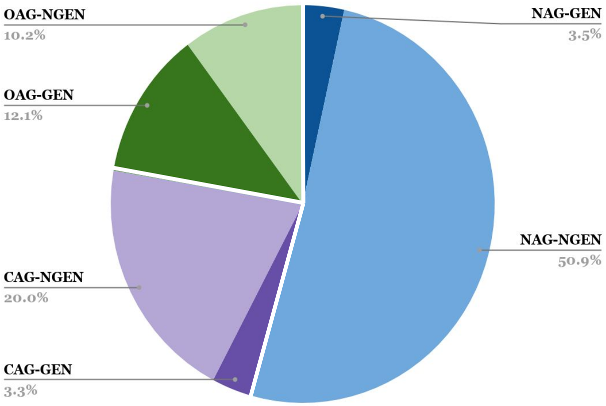
- The probability that an example belonging to an aggression class is also GEN increases as directness of aggression increases (P(GEN | NAG) < P(GEN | CAG) < P(GEN | OAG)).
- Hence, the two sub-tasks are related.



a) English

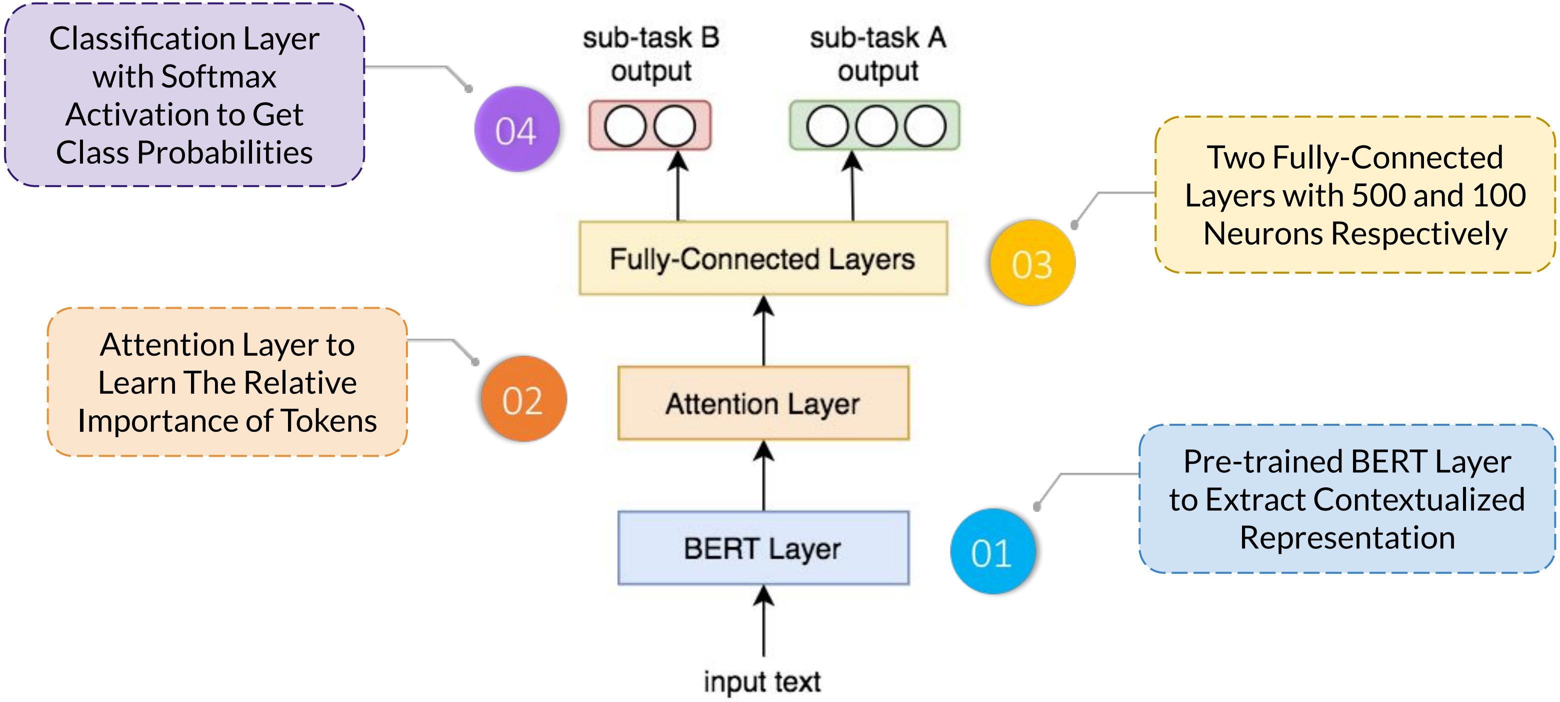


b) Hindi



c) Bengali

Model Architecture



Experimental Setup

- We used pre-trained BERT models (which are not fine-tuned):
  - English: bert_base_uncased
  - Hindi & Bengali: bert_base_multilingual_cased
- Binary cross entropy loss (sum for task A and B).
- Class weights used in loss function to address data imbalance.
- Adam optimizer.
- Learning rate: 10^{-5}.
- Run for 200 epochs, save on best validation F1.
- Trained on Tesla P40 GPU, Approx 1.5min/epoch.

Results: Sub-tasks

- Best rank on English-B (3rd out of 15).
- Misogyny (2 classes) is relatively easier to detect than Aggression (3 classes).
- System lags behind the winner on English-B (0.8715 F1), and Bengali-B (0.9365 F1) by 0.0136 and 0.0159, which makes it competitive.

Sub-task	English	Hindi	Bengali
A	0.7143	0.7183	0.7369
B	0.8579	0.8008	0.9206

Table 1: Weighted F1 scores for all sub-tasks

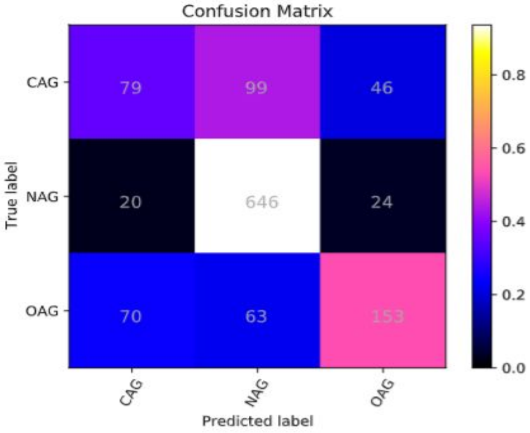
Results: Class-wise

- CAG - least score hence most challenging aggression class.
- English - least OAG, CAG scores due to higher data imbalance (79% train examples NAG).
- Max difference in NGEN F1 and GEN F1 on English due to higher data imbalance (93% train examples NGEN).

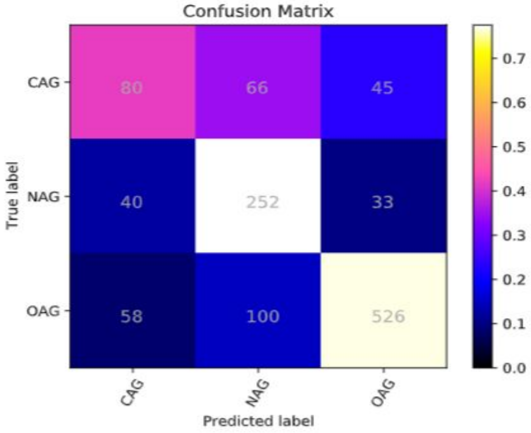
Language	Sub-task A			Sub-task B	
	NAG	CAG	OAG	GEN	NGEN
English	0.86	0.40	0.62	0.53	0.91
Hindi	0.68	0.43	0.82	0.77	0.83
Bengali	0.84	0.45	0.71	0.75	0.96

Table 2: Class-wise F1 scores for all sub-tasks.

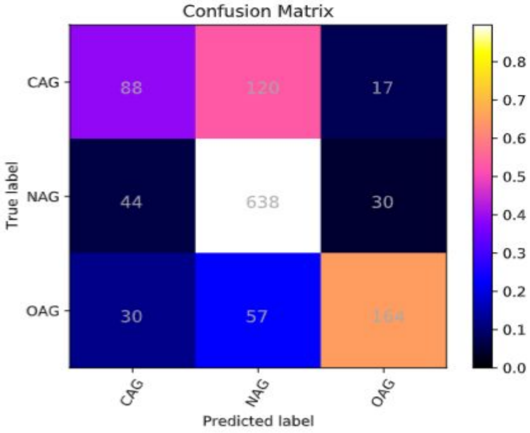
Confusion Matrices : Sub-task A



(a) English sub-task A



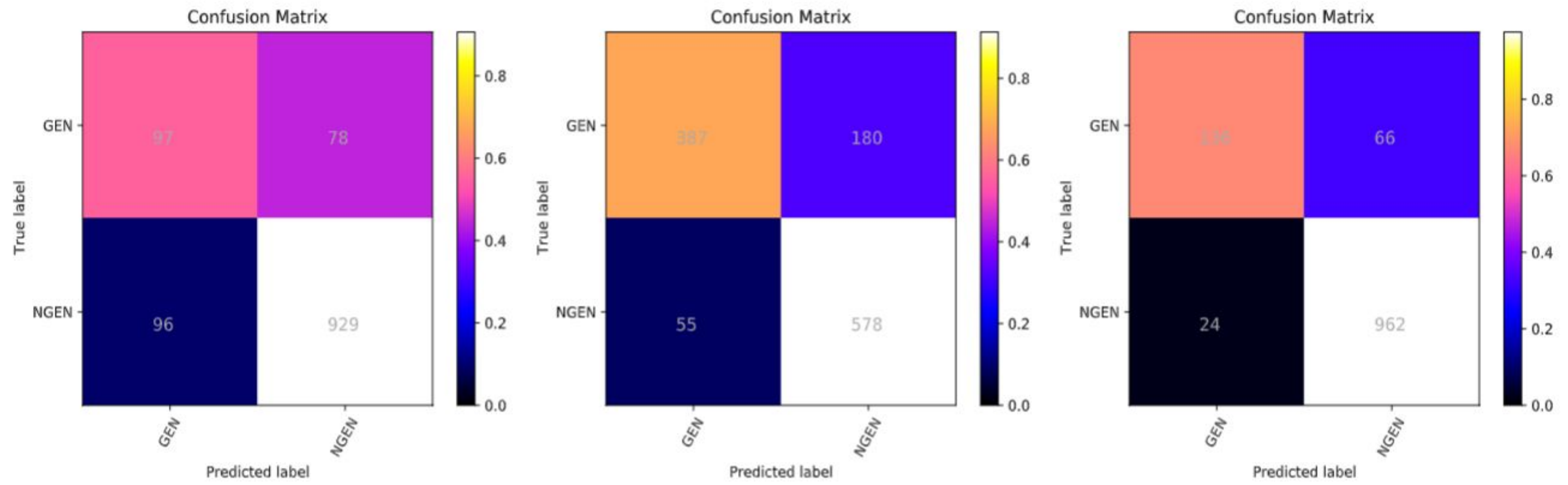
(b) Hindi sub-task A



(c) Bengali sub-task A

- ❖ CAG more likely to be wrongly predicted as NAG than OAG, due to lack of abusive/explicit words in CAG.
- ❖ In Hindi, OAG-NAG confusion (100) is high, as majority of the train instances are NAG (56.35%), whereas the majority of the test instances are OAG (57.00%).

Confusion Matrices : Sub-task B



(a) English sub-task B (b) Hindi sub-task B (c) Bengali sub-task B

❖ GEN-NGEN confusion for Hindi (180) is higher than that in other languages, as the distribution of classes across the test data (47% GEN) is significantly different from that in training data (17% GEN).

Error Analysis: CAG-NAG

- Due to the indirect/sarcastic nature and lack of profanity in CAG, it is often confused with NAG :
  - “Fat shaming is good. Why not?'”
  - “They have no right to live”
  - “Inko hospital bejo..ye mentally hille hue log han” (Send them to hospital, they are mentally disturbed people.)

Error analysis: Noise in the Data

Sub-task	Text	Annotated	Predicted
English-A	"Also Veere Di Wedding Fake Feminist Piece Of Shit.."	NAG	OAG
Hindi-A	"Mujhe bhi jand lagi movie lakin maine chutiyo ke samne jaban nahi kholi or nahi kholuga" (*I also found this movie stupid, but I didn't open my mouth in front of idiots and won't do so.*)	NAG	OAG
English-B	"kapil why are u listening to these chutiaasssss....give them shut upcall...insane idiots"	GEN	NGEN
Hindi-B	"Kaunsi charas ya afeem phoonk ke aayi hai ye. Gandee aurat. Aurat ke naam pe dhabba." (*Which weed or poppy has she smoked? Dirty lady. Blot on the name of a woman.*)	NGEN	GEN

Table 3: Instances where predicted labels seem more likely to be correct than annotated labels.

Conclusion and Future Work

- **Conclusion**
  - Sub-tasks A and B are related.
  - CAG is the most difficult class to detect and is often confused with NAG.

- **Future Work**
  - Finetune BERT.
  - More features for better identification of CAG.

Thank You

Contact: nsafisamghabadi@uh.edu , parthprasad.p17@iiits.in
Paper link: http://panlingua.co.in/trac-2/pdf/2020.trac2-1.20.pdf
Code and model weights: https://github.com/NiloofarSafi/TRAC-2

Classes

- **Sub-task A**
  - NAG (Not Aggressive) - No aggression in text. E.g. "hats off brother".
  - CAG (Covertly Aggressive) - Indirect aggression, sarcasm, no explicit words. E.g., "You are not wrong, you are just ignorant.".
  - OAG (Overtly Aggressive) - Direct attack, explicit words. E.g., "Liberals are retards".
- **Sub-Task B**
  - Gen (Gendered) - Targets a person or a group based on gender, sexuality, or lack of fulfillment of stereotypical gender roles. E.g., "Homosexuality should be banned"
  - NGEN (Not Gendered) - Texts that are not gendered. E.g.. "you are absolutely true bro...but even politicians supports them"

Related Research

- NLP community has shown interest in aggression detection and related areas.
- Several related workshops and share tasks have been conducted:
  - Abusive Language online (ALW) [1]
  - SemEval shared task on Identifying Offensive Language in Social Media (OffensEval) [2]
- Deep learning has become popular for hate speech identification. [3,4]
- Sexism, a subset of hate-speech has been analyzed and further categorized. [5,6]
- The first Shared Task on Aggression Identification aimed to identify aggressive social media posts and provided datasets in Hindi and English. [7]

References

[1] Sarah T. Roberts, et al., editors. (2019). Proceedings of the Third Workshop on Abusive Language Online. Association for Computational Linguistics.

[2] Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P.,Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z.,and Coltekin, c. (2020). SemEval-2020 Task 12: Multi-lingual Offensive Language Identification in Social Media (OffensEval 2020). In Proceedings of SemEval.

[3] Zhang, Z., Robinson, D., and Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRUBased Deep Neural Network. In Lecture Notes in Computer Science. Springer Verlag.

[4] Dadvar, M. and Eckert, K.(2018).Cyberbullying detection in social networks using deep learning based models; a reproducibility study. arXiv preprint arXiv:1812.08046.

[5] Jha, A. and Mamidi, R. (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In Proceedings of the Second Workshop on NLP and Computational Social Science.

[6] Sharifirad, S. and Matwin, S. (2019). When a tweet is actually sexist. A more comprehensive classification of different online harassment categories and the challenges in NLP. CoRR, abs/1902.10584

[7] Ritesh Kumar, et al., editors, Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). Association for Computational Linguistics.,2018

Model Architecture