MLOffense: Multilingual Offensive Language Detection and Target Identification On Social Media Using Graph Attention Transformer Model

Problem & Objectives

Social media has become an important part of our everyday lives. However, offensive language on social media has become a serious issue. Problems include but are not limited to:

- Fear, anxiety, isolation, and mental health problems for targeted individuals
- Discrimination against certain groups of people (e.g., race, gender, ethnicity, sexual orientation)
- Negative impacts on the overall online environment
- Contributes to the spread of stereotypes and biases

Image created by AI

Tragically losing a friend to cyberbullying, I was motivated to combat the issue of offensive language on social media, leveraging my background in computational linguistics.

Challenges

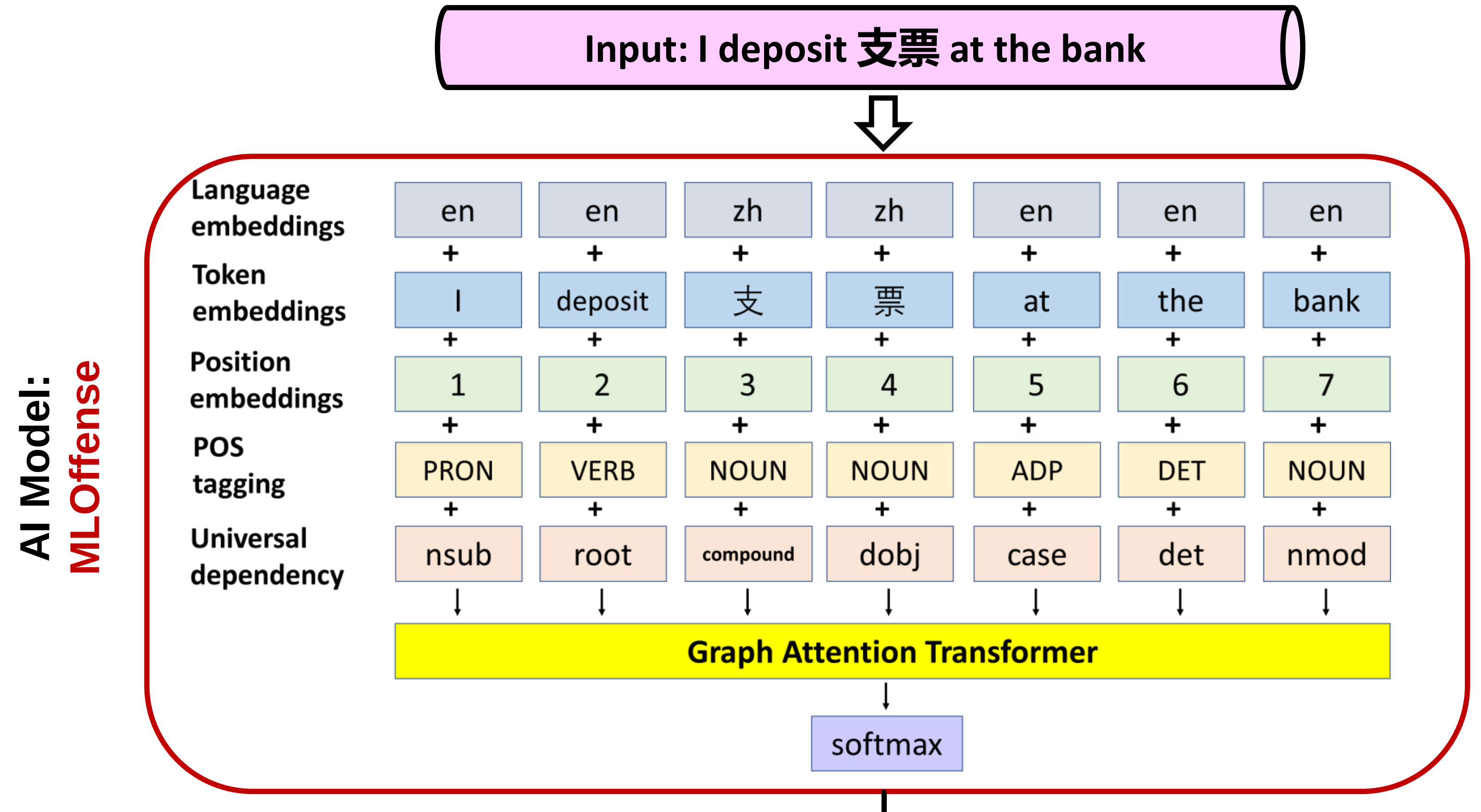
- Most existing studies are limited to English due to limited labeled training data in low-resource languages
- Different ways of expressing offense
- Diversity of languages
- No studies on identification of the target

My Contributions

- ➢ Use of graph attention mechanisms
- ➢ Development of a novel multilingual model for 100 languages
- ➢ Use of transfer learning to leverage existing English resources
- ➢ Break new ground as the first study ever to identify the specific individuals or groups targeted by offensive posts.
- ➢ An app to allow social media users to filter out offensive posts, balancing users' mental health with freedom of speech

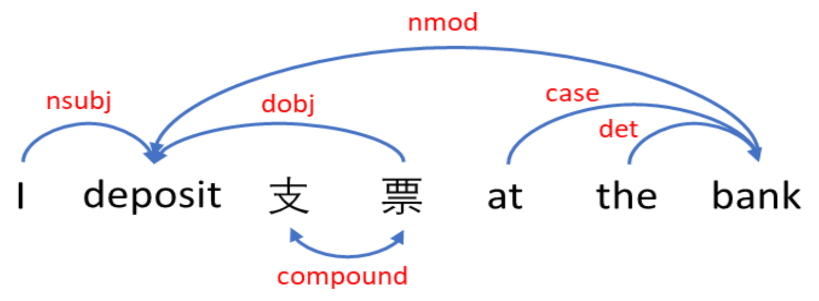
Model Design

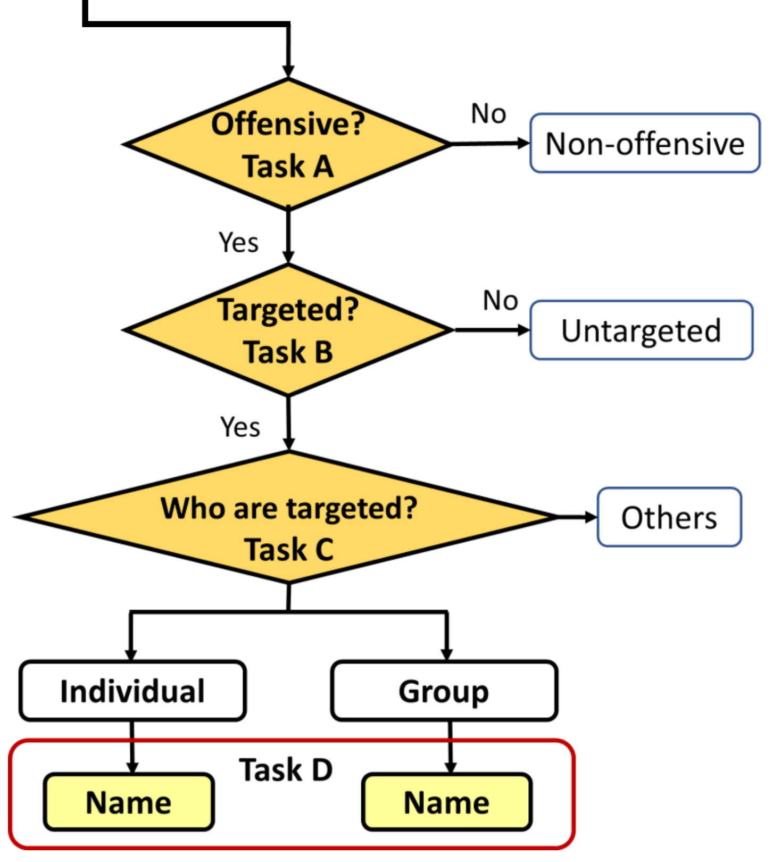
- Transformers are a type of neural network architecture for sequential data.
- “Attention" used to weigh the importance of different parts of the data differently.
- Allows for more effective learning of the context and relationships within the data.



Novel Graph Attention:

- Replace self-attention in conventional transformer with graph attention
- Pay more attention to words closer in syntactic distance (see example)

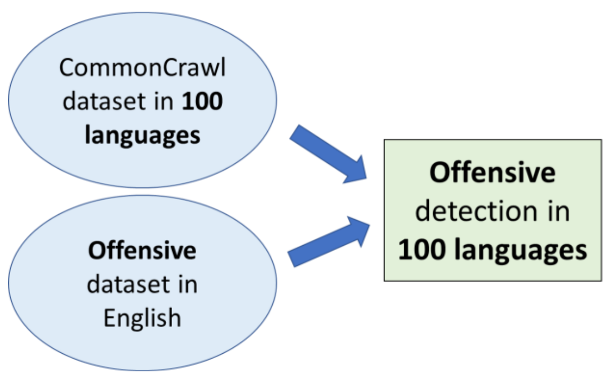




First time target’s name is identified

Model Training

- Learn to understand or “speak” different languages: pretrain in **100 languages** with general content CommonCrawl datasets
- Learn to **detect offensive content even w/o offensive words**
- Zero-shot cross-lingual transfer learning with benchmark offensive datasets in **English**
- Training datasets:
  OLID + HASOC + TweetNER7


CommonCrawl dataset in 100 languages
Offensive dataset in English
Offensive detection in 100 languages

Data Analysis & Results

MLOffense was statistically evaluated on languages with available datasets for four tasks in terms of F1 scores:

Task A: Offensive or Not

		Macro	Weighted	Offensive	Non-Offensive
English	**MLOffense**	**0.8461**	**0.8531**	**0.8252**	**0.8670**
	DeepOffense	0.8126	0.8228	0.7818	0.8433
Arabic	**MLOffense**	**0.7848**	**0.7862**	**0.7793**	**0.7904**
	DeepOffense	0.7148	0.7198	0.6941	0.7356
Chinese	**MLOffense**	**0.8361**	**0.8388**	**0.8192**	**0.8530**
	DeepOffense	0.7743	0.7794	0.7426	0.8060
Marathi	**MLOffense**	**0.6924**	**0.6936**	**0.6614**	**0.7233**
	DeepOffense	0.6087	0.6088	0.6070	0.6104
Spanish	**MLOffense**	**0.8348**	**0.8365**	**0.8228**	**0.8468**
	DeepOffense	0.7788	0.7801	0.7692	0.7883
Italian	**MLOffense**	**0.8198**	**0.8211**	**0.8123**	**0.8272**
	DeepOffense	0.7616	0.7619	0.7601	0.7632
German	**MLOffense**	**0.8333**	**0.8344**	**0.8291**	**0.8375**
	DeepOffense	0.7650	0.7656	0.7629	0.7672
Hindi	**MLOffense**	**0.7937**	**0.7991**	**0.7690**	**0.8183**
	DeepOffense	0.7051	0.7084	0.6902	0.7200
Code-mixing	**MLOffense**	**0.8081**	**0.8081**	**0.8082**	**0.8080**
	DeepOffense	0.7123	0.7167	0.6755	0.7491

Task B: Targeted or Not

Weighted F1 Scores	English	Arabic	Chinese	Marathi
MLOffense	**0.7963**	**0.7127**	**0.7828**	**0.6394**
DeepOffense	0.7609	0.6583	0.7359	0.5780

Task C: Individual, Group, or Other

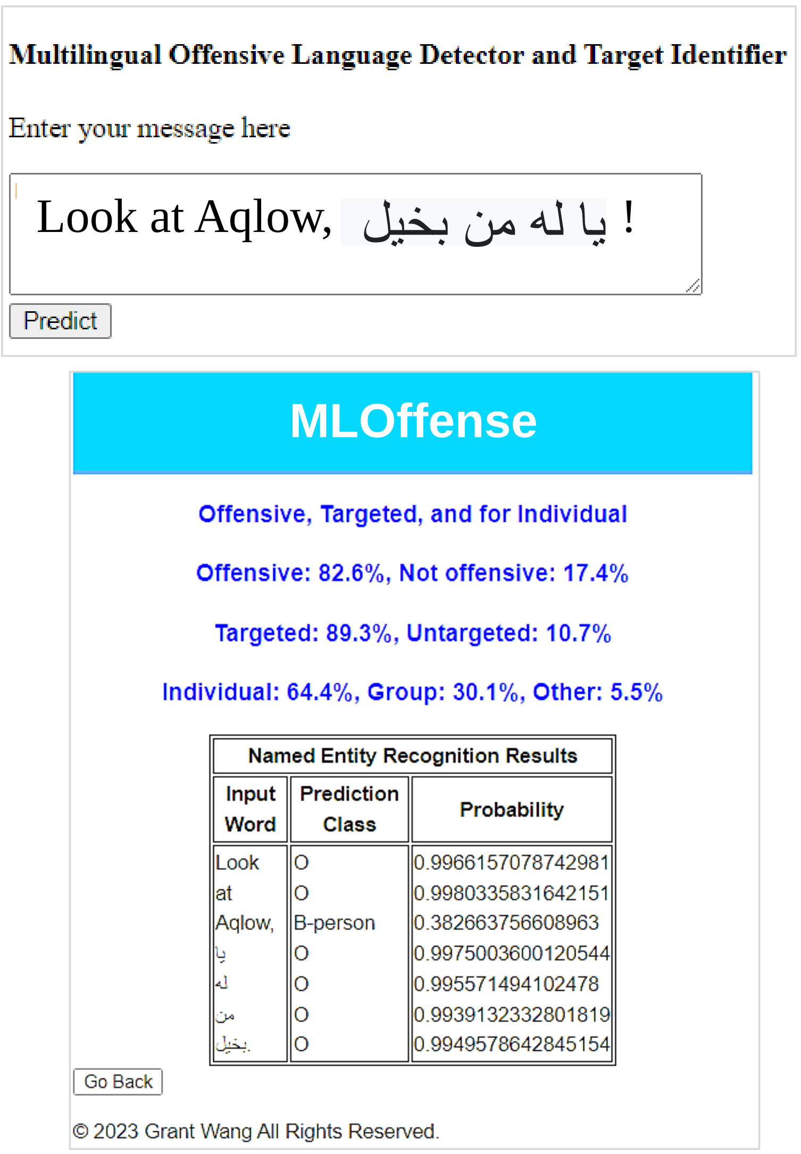
Weighted F1 Scores	English	Arabic	Chinese	Marathi
MLOffense	**0.7468**	**0.7091**	**0.7262**	**0.6117**
DeepOffense	0.7289	0.6522	0.6950	0.5568

Task D: Name Recognition

Weighted F1 Scores		English	Arabic	Chinese
MLOffense	Person	**0.8118**	**0.7426**	**0.7696**
	Group	**0.7312**	**0.6621**	**0.6807**

Demonstrations

Demo #1 – App to detect multilingual offensive language on social media



Multilingual Offensive Language Detector and Target Identifier

Enter your message here

Look at Aqlow, يا له من بخيل !

Predict

MLOffense

Offensive, Targeted, and for Individual

Offensive: 82.6%, Not offensive: 17.4%

Targeted: 89.3%, Untargeted: 10.7%

Individual: 64.4%, Group: 30.1%, Other: 5.5%

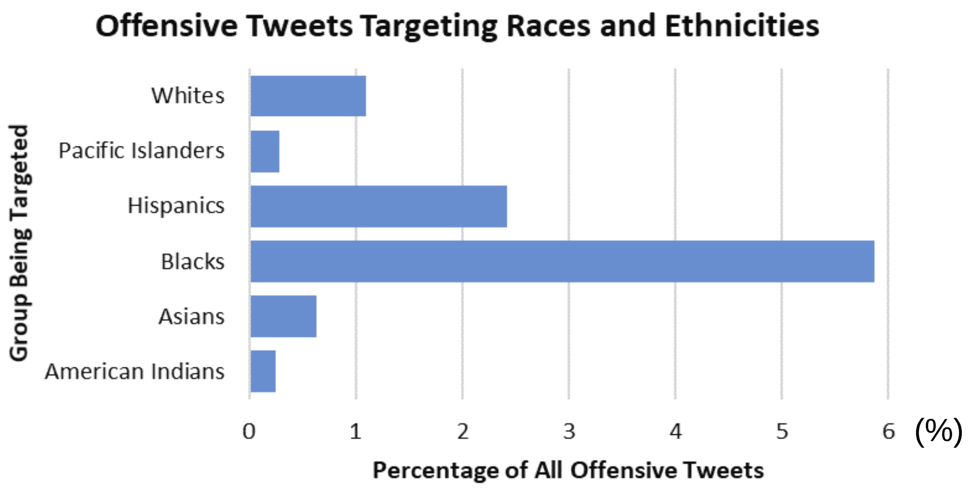
Named Entity Recognition Results

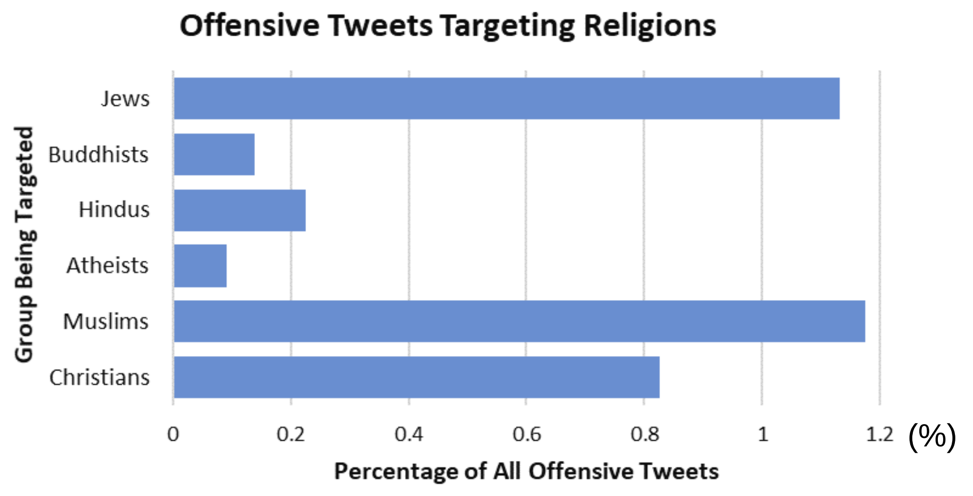
Input Word	Prediction Class	Probability
Look	O	0.9966157078742981
at	O	0.9980335831642151
Aqlow,	B-person	0.382663756608963
يا	O	0.9975003600120544
له	O	0.995571494102478
من	O	0.9939132332801819
بخيل.	O	0.9949578642845154

Go Back

© 2023 Grant Wang All Rights Reserved.

Demo #2 – Example to extract data from X for behavioral and social science research





Data randomly selected: 0.1% of all the tweets from every day 02/2022-01/2023

Conclusions

- ➢ Useful for detection and potentially prevention of online offensive language
- ➢ Optional social media plug-in for users to filter out offensive posts
- ➢ Extracts data for behavioral and social science research
  - Analyze prevalence and causes
  - Identify and support victims

Future Work

- ➢ To leverage emerging large language models like GPT-4
- ➢ To classify a spectrum of offensiveness
- ➢ More training data with the ever-changing nature of language
- ➢ To incorporate linguistics and psychology knowledge
- ➢ To perform comprehensive social science studies

All tables, graphs, and figures created by the student researcher unless otherwise noted.