Joint Modelling of Emotion and Abusive Language Detection

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis,
Ekaterina Shutova

Introduction

- As online communication platforms arise, aggressive and abusive behavior online are more frequently witnessed
- This includes racism, sexism, personal attacks, harassment and cyber-bullying and so forth
- The paper aims to introduce the first joint model of emotion and abusive language detection, specifically on text data from Twitter

Related work

- Modeling the linguistic properties of the text = detecting *explicit* abuse
  - RNN and CNN based models
  - Character-based models
  - Graph-based learning models

(1) stop editing this, you **dumbass**.
(2) Just want to slap the **stupid** out of these **bimbos**!!!
(3) Go lick a pig you arab muslim piece of **scum**.

Wiegand, Michael, Josef Ruppenhofer, and Thomas Kleinbauer. "Detection of abusive language: the problem of biased datasets." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1,* 2019

Related work

- Promising results on explicit abuse detection, but challenging to identify implicit abuse (i.e. sarcasm, jokes, and particularly the usage of negative stereotypes)

  (4) i havent had an intelligent conversation with a woman.
  (5) Jews don't marry children. Muslims do. All the time.

- Abusive language and behaviour are also inextricably linked to the emotional and psychological state of the speaker

Research method

- The authors propose to model these two phenomena (emotion detection + abusive language detection) jointly via a multitask learning (MTL) paradigm
- Multitask learning?
    - Allows two or more tasks to be learned jointly, thus sharing information and features between the tasks
    - Is inspired by human learning
    - Successful for many NLP tasks
- **Primary task** : abuse detection ; **Auxiliary task** : emotion detection

Research method

- Abuse detection dataset
  - Since MTL models are sensitive to differences in the domain, both of the dataset are from Twitter (OffensEval 2019 & Waseem and Hovy 2016)
  - OffensEval 2019 contains 13, 240 annotated tweets, and each tweet is classified as to whether it is offensive (33%) or not (67%)
  - Waseem and Hovy 2016 contains 16, 202 tweets, and each tweet is classified as 1, 939 (12%) racism; 3, 148 (19.4%) sexism; and 11, 115 (68.6%) neither

Research method

- Emotion detection dataset
  - SemEval-2018 contains 11K tweets with 11 emotion labels (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust)

Research method

- For a baseline model, writers experiment with four different Single-Task Learning (STL) models that utilize abuse detection as the sole optimization objective
  - Model architecture : Maxpooling and MLP classifier vs BiLSTM and Attention classifier
  - Input representation : GloVe vs GloVe + ELmo
- They choose the most robust architecture for each of the abuse detection datasets, and adopt the selected model's configuration for MTL models

Research method

- Best combinations for each dataset:

STL model		P	R	F1
G	*maxpool+MLP*	76.35	**73.34**	74.24
	BiLSTM+attn	77.34	72.77	73.97
G+E	*maxpool + MLP*	77.19	72.73	73.95
	BiLSTM+attn	**77.40**	73.27	**74.40**

(a) *Twitter - OffensEval* STL results.

STL model		P	R	F1
G	***maxpool+MLP***	79.39	**78.20**	**78.33**
	BiLSTM+attn	77.97	77.57	77.49
G+E	*maxpool+MLP*	**80.66**	77.13	78.31
	BiLSTM+attn	79.08	77.93	78.16

(b) *Twitter - Waseem and Hovy* STL results.

***OffensEval*: $STL_{BiLSTM+attn}$**

concat(Glove embeddding, ELMo embedding)

Stacked BiLSTM Encoder

BiLSTM layer

Attention layer

Linear Output Layer

***Waseem&Hovy*: $STL_{maxpool+MLP}$**

Glove embedding

Stacked BiLSTM Encoder

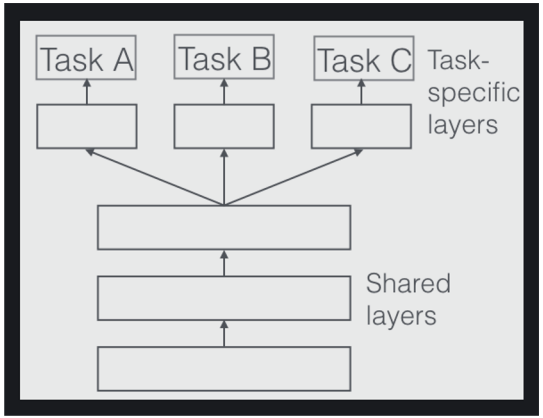
Maxpooling layer

MLP layer

Linear Output Layer

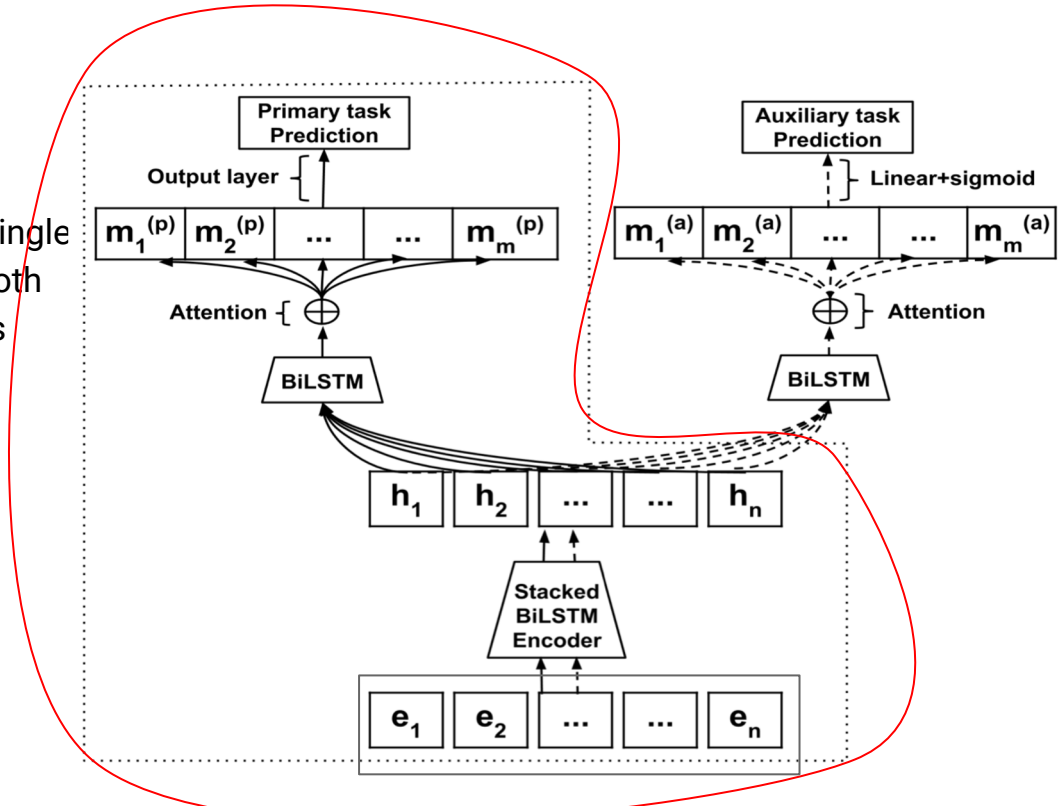
Research method

- For a multi-task learning model, the authors propose MTL models that contain two network branches – one for the primary task and one for the auxiliary task – connected by a shared encoder which is updated by both tasks alternately
    - Hard Sharing Model
    - Double Encoder Model



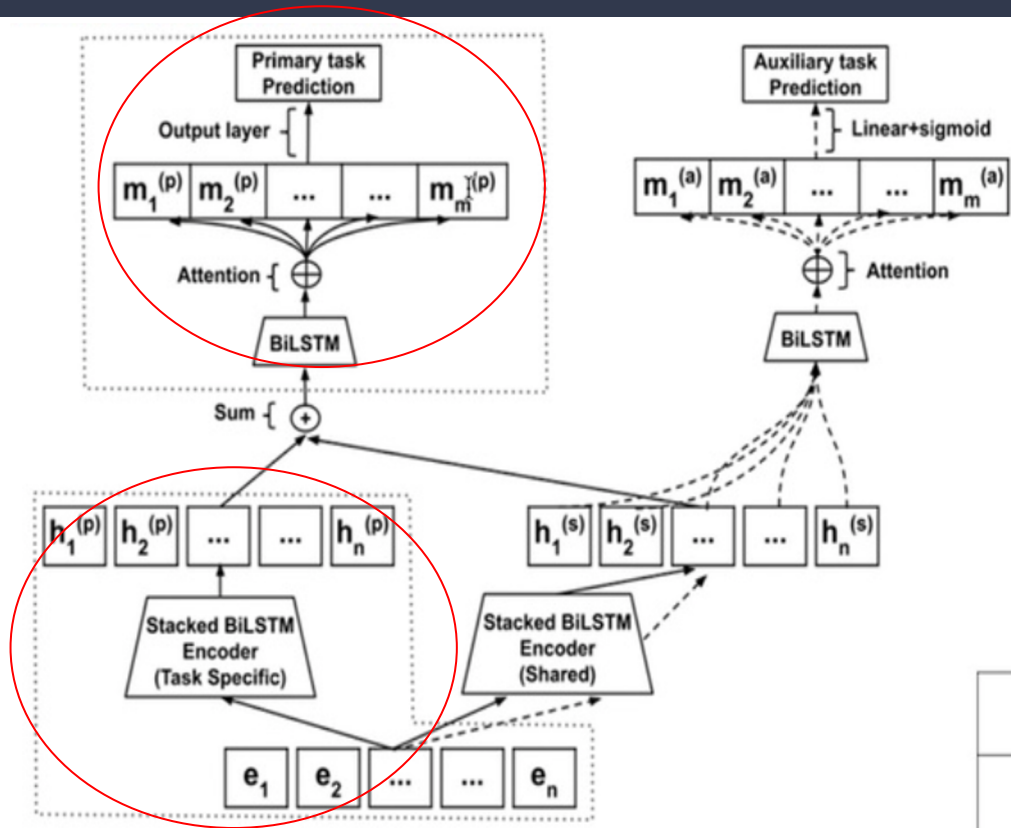
Research method

- Hard Sharing Model
  - *hard parameter sharing*: it consists of a single encoder that is shared and updated by both tasks, followed by task-specific branches
  - e : embedding layer output / h : encoder output / m : intermediate representation



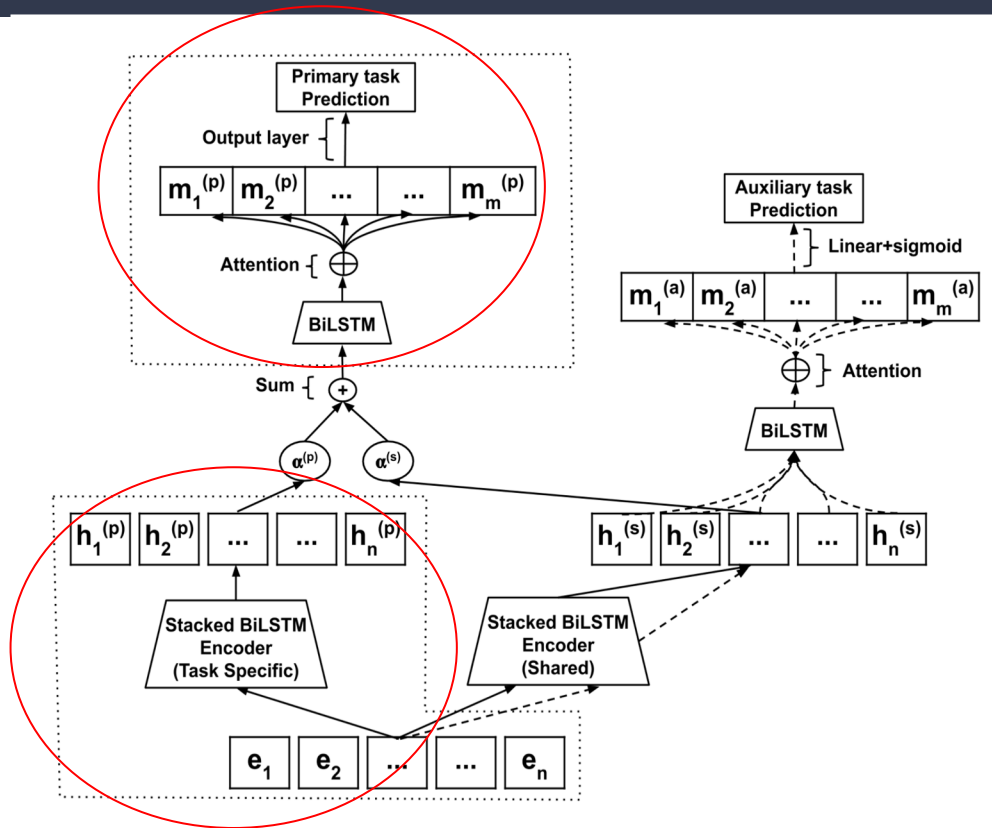
Research method

- Double Encoder Model
  - an extension of the previous model that now has two BiLSTM encoders: a task-specific two-layered BiLSTM encoder for the primary task, and a shared two-layered BiLSTM encoder
  - e : embedding layer output / h : encoder output / m : intermediate representation



Research method

- Gated Double Encoder Model
  - an extension of a double encoder model, but is different in a way it obtains the post representation m
  - e : embedding layer output / h : encoder output / m : intermediate representation / a : learnable parameter
  - Learnable parameters are to control information flow



Result

STL model		P	R	F1
G	*maxpool+MLP*	76.35	**73.34**	74.24
	BiLSTM+attn	77.34	72.77	73.97
G+E	*maxpool + MLP*	77.19	72.73	73.95
	BiLSTM+attn	**77.40**	73.27	**74.40**

(a) *Twitter - OffensEval* STL results.

STL model		P	R	F1
G	***maxpool+MLP***	79.39	**78.20**	**78.33**
	BiLSTM+attn	77.97	77.57	77.49
G+E	*maxpool+MLP*	**80.66**	77.13	78.31
	BiLSTM+attn	79.08	77.93	78.16

(b) *Twitter - Waseem and Hovy* STL results.

Result

- MTL vs Transfer learning :
  - best performing MTL model vs train firstly on the auxiliary task and then train the primary task
  - MTL achieves higher performance than transfer learning

Dataset	Method	P	R	F1
OE	MTL	**77.46**	**75.27**†	**76.03**†
	Transfer	76.81	73.71	74.67
W&H	MTL	80.12	**79.60**†	**79.55**
	Transfer	**81.28**	77.72	79.07

Result

Sample	STL	MTL	Gold Label	Predicted Emotion
Shut up Katie and Nikki... That is all :) #HASHTAG	*neither*	*sexism*	*sexism*	*disgust*
MTN I m pretty sure you are not too bad yourself...**thanks** for a lil bit of **sweetness** on this brutal world	*Offensive*	*NotOffensive*	*NotOffensive*	*joy, optimism*

Conclusion

- Experiments demonstrate that MTL with emotion detection is beneficial for the abuse detection task in the *Twitter* domain
- They also suggest the superiority of MTL over STL for abuse detection
- MTL’s performance is reported to be powerful than transfer learning

Thanks