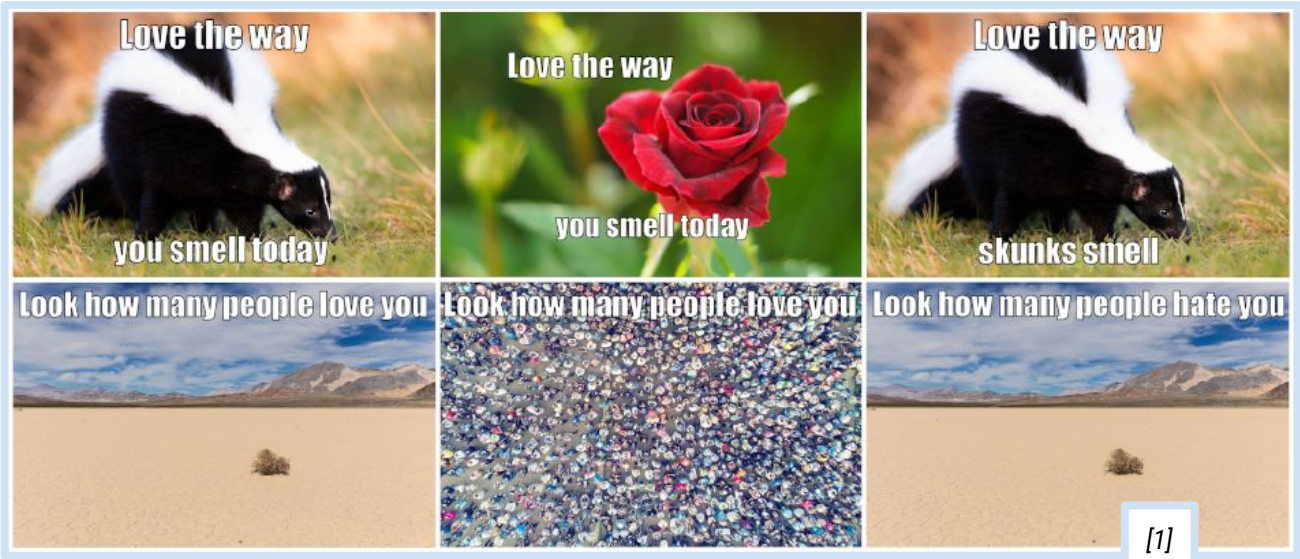
MEMES AS HATE SPEECH

Exploring Multimodality in Offensive Online Culture

TABLE OF CONTENTS

INTRO: MEMES AS HATE SPEECH



[1]

MATERIALS

Hate Speech in Pixels [2]

- Data:
  - 5020 memes, 66% non-hateful, 34% hateful
  - Reddit Memes dataset
  - Google Images queries
  - Not available
- Model:
  - Image encoding – VGG16
  - Text encoding – BERT
  - Concatenate → MLP
  - Allows using both modalities or just one
- Validation accuracies: 83.3% (img+text), 83% (img), 76.1% (text)

The Hateful Memes Dataset [1]

- Data:
  - Launched as part of challenge
  - 10k memes, train ~63% non-hateful, ~37% hateful; validation balanced; test unclear.
  - 3 annotators – if no agreement, experts were consulted
  - Does not contain slurs – unimodal
  - Each meme given at least 1 benign confounder, if possible
  - Aim to force classifiers to be multimodal
- Best performing (pre-trained) model:
  - Visual BERT COCO, test accuracy ~64%

METHOD

Version	*Hid. size*	*Epochs*	*Batch size*	*Balanced*	*Img*	*Text*	*Val. acc.*	*Test acc.*
1	50	100	25	N	Y	Y	70.4	70.1
1b	50	100	25	Y	Y	Y	65.6	62.4
2	50	50	50	N	Y	Y	69.5	67.6
2b	50	50	50	Y	Y	Y	65.3	62.3
3	20	40	50	N	Y	Y	69.4	68.6
3b	20	40	50	Y	Y	Y	65.1	61.9
3b10	20	40	50	Y	Y	N	57.7	59.5
3b01	20	40	50	Y	N	Y	65.9	65.4

RESULTS

Version	*Hid. size*	*Epochs*	*Batch size*	*Balanced*	*Img*	*Text*	*Val. acc.*	*Test acc.*
1	50	100	25	N	Y	Y	70.4	70.1
1b	50	100	25	Y	Y	Y	65.6	62.4
2	50	50	50	N	Y	Y	69.5	67.6
2b	50	50	50	Y	Y	Y	65.3	62.3
3	20	40	50	N	Y	Y	69.4	68.6
3b	20	40	50	Y	Y	Y	65.1	61.9
3b10	20	40	50	Y	Y	N	57.7	59.5
3b01	20	40	50	Y	N	Y	65.9	65.4

RESULTS

Version	*Hid. size*	*Epochs*	*Batch size*	*Balanced*	*Img*	*Text*	*Val. acc.*	*Test acc.*
1	50	100	25	N	Y	Y	70.4	70.1
1b	50	100	25	Y	Y	Y	65.6	62.4
2	50	50	50	N	Y	Y	69.5	67.6
2b	50	50	50	Y	Y	Y	65.3	62.3
3	20	40	50	N	Y	Y	69.4	68.6
3b	20	40	50	Y	Y	Y	65.1	61.9
3b10	20	40	50	Y	Y	N	57.7	59.5
3b01	20	40	50	Y	N	Y	65.9	65.4

RESULTS

Version	*Hid. size*	*Epochs*	*Batch size*	*Balanced*	*Img*	*Text*	*Val. acc.*	*Test acc.*
1	50	100	25	N	Y	Y	70.4	70.1
1b	50	100	25	Y	Y	Y	65.6	62.4
2	50	50	50	N	Y	Y	69.5	67.6
2b	50	50	50	Y	Y	Y	65.3	62.3
3	20	40	50	N	Y	Y	69.4	68.6
3b	20	40	50	Y	Y	Y	65.1	61.9
3b10	20	40	50	Y	Y	N	57.7	59.5
3b01	20	40	50	Y	N	Y	65.9	65.4

DISCUSSION

Hate speech in Pixels [2]

- Not providing data makes it impossible to reproduce results
- Google queries will not always produce what is expected

Hateful Memes [1]

- Small number of annotators != representative
- Noting representativeness
- Does contain some slurs
- Multimodal?

- Balanced data effect – decrease in accuracy not proportional to decrease of label
- 10 h for 70.1% versus 2 h for 68.6%
- Easier to classify hateful memes
- Text-only model success perhaps due to most frequent words + slurs

CONCLUSION

- Is Hate Speech in Pixels [2] reproducible?
  - No
- Is the Hateful Memes Dataset [1] multimodal?
  - No

FUTURE WORK

- Further exploration of the data might explain why the model relies on the textual data
- Examining how a larger group of annotators would label Hateful Memes Dataset

THANKS :)

REFERENCES

[1] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes.

[2] Benet Oriol, Cristian Canton-Ferrer, and Xavier Giró-i-Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation.