Saga Hansson
Artificial Intelligence: cognitive systems 7,5 hp
HT20

Multimodal Classification: Memes and Hate Speech

Memes, i. e. social media posts consisting of both image and text, are at times used to convey hate speech. Often times, it is the combination of the image and text that make the meme hateful, making it impossible to simply filter out “bad words” on social media sites. For that reason, I am interested in examining different ways to classify memes as hateful or not. The paper that inspired this project is *Exploring Hate Speech Detection in Multimodal Publications* by Gomez et. al. (2019), in which the *MMHS150K Dataset* is presented. The MMHS150K is a “manually annotated multimodal hate speech dataset formed by150,000 tweets, each one of them containing text and an image” (Gomez et. al., 2019). Another source of inspiration is the *Hateful Memes Challenge* by Facebook[1]. In this challenge, the *Hateful Memes Dataset* is provided. A third source of data is created in *Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text* by Suryawanshi et. al. (2020), in which the *MultiOFF dataset* is constructed by annotating an already existing Kaggle dataset consisting of memes regarding the 2016 US election[2] as offensive or not. Lastly, *MemeSem: A Multi-modal Framework for Sentiment Analysis of Meme via Transfer Learning* by Pranesh and Shekhar (2020) is also an interesting paper that provides a clear explanation of the architecture of the model.

As I would like to explore the datasets further, I have not yet chosen which one I would like to work with in this project. Hence, that is one of the tasks that needs to be completed. Other tasks that I have planned are model research (reviewing what models are used in similar projects, what kind of model is doable, etc.), coding (trial and error included), evaluation (including evaluation research, i. e. surveying how similar tasks have been evaluated), reporting and presentation design. I suspect that the most challenging part will be to decide on the architecture of the model, especially since I have not worked with images before.

1 https://ai.facebook.com/tools/hatefulmemes/
2 https://www.kaggle.com/SIZZLE/2016electionmemes

References

Gómez, R., Gibert, J., Gómez, L., & Karatzas, D. (2020). Exploring Hate Speech Detection in Multimodal Publications. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 1459-1467.

Pranesh, R.R., & Shekhar, A. (2020). MemeSem:A Multi-modal Framework for Sentimental Analysis of Meme via Transfer Learning.

Suryawanshi, S., Arcan, M., & Buitelaar, P. (2020). Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. TRAC@LREC.