Arjun Balasubramanian
CS 4395
Section 004

Reading ACL Papers

The title of the paper I chose is "Racist or Sexist Meme? Classifying Memes beyond Hateful", written by Haris Bin Zia, Ignacio Castro, and Gareth Tyson all from the Queen Mary University of London in the United Kingdom.

The issue that this paper talks about is offensive memes. Memes are a staple of modern day internet culture which usually consist of an image and a corresponding text which typically is used to present humor. Meme culture has somewhat evolved over time as they started as innocuous ways to make others laugh over the internet, but have now turned into a broad category with multiple sub-genres and niches. One genre of meme is the offensive meme. Offensive memes are intentionally hateful comedy at the behest of a group. The paper gives a good example of what they considered an offensive meme which I have presented on the right here. This image is considered an offensive meme as it has hate directed towards the muslim community. The goal of the paper is to find a method to successfully identify offensive memes and classify them by subcategory of offensive memes.

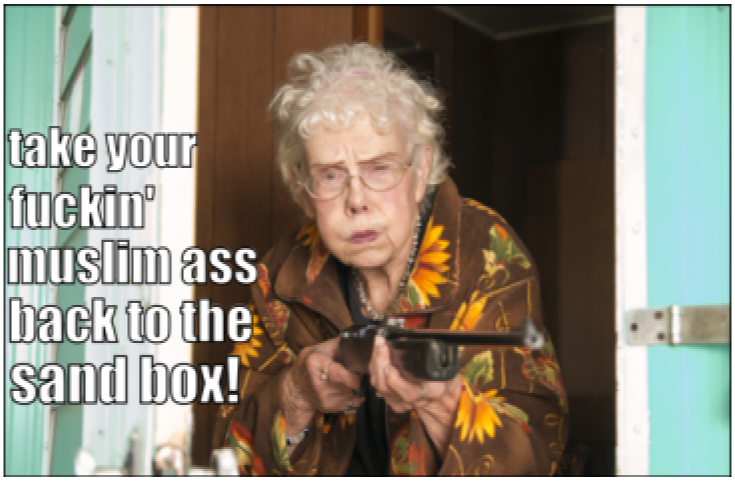

Figure 1: An example of a hateful meme. The meme is targeted towards a certain religious group.

The biggest inspiration and greatest source of prior work for this paper was the Facebook Hateful Memes Challenge. The hateful memes challenge was an effort by Facebook to find methods to reduce the amount of hateful memes on their platform. For the challenge Facebook built an annotated dataset with over ten thousand memes for researchers to use in training models with a prize of

$100,000. The big difficulty with building this problem as identified by Facebook is that the meaning of a meme is made up by a combination of the picture and the image. So the problem is finding a way to have the model take in both aspects of the meme.

The researchers' unique way to solve the problem of classifying types of offensive memes is to use a pipeline of two distinct classifications. They broke up the classification into figuring out which category of group was being attacked for example religion, race, sex, etc. Then using another classification to classify what type of attack the meme was using for example mocking, dehumanization, slurs, etc. Both of the classifications however, were also multilabel, meaning that a meme could fall under multiple different categories of group attacking and attack type. To do this and tackle the issue of having both a text and visual component the team of researchers decided to build a pipeline of specific tasks. First they extracted the text from the meme, then used visual and text embeddings and concatenated them to input it into a model for a classifier. I found it quite cool the way that they passed in the image and text data together. Rather than trying to process them together they encoded both of the parts of the meme and concatenated them together. The way they decide what model to use is they use a series of different models with different types of embedding in order to look at the accuracy to determine the best method to get the most accurate results.

The metrics that the research team used were accuracy to decide what type of model to choose. After testing on different types of models the one that they came up with was the multimodal model with CIMG, CTXT, LASER, and LaBSE embeddings. With that model they got an accuracy of .96 on the group targeted by the meme, and on the task of classifying by attack type the model got an accuracy of .97. With that best model they also used an F1 score for each of the classes in the groups targeted and attack types.

The first listed researcher on this paper is named Haris Bin Zia. Haris is a researcher at the Queen Mary, University of London. He has 8 publications listed under his name all published at the University. Most of his work is related to the decentralized web which I found odd considering that this paper is a very different category. Haris was a PhD student there and has completed his doctorate. The next listed is Ignacio Castro who is a lecturer at the Queen Mary, University of London he is the researcher with the most publications with over 20 papers written. His work is far more related to the topic of the paper as he has a lot of research related to deep learning. The last listed researcher is Gareth Tyson, he has the second most amount of papers written with around 15. He is a professor as well. I found it quite unique that the paper put the least experienced researcher first, but there typically isn't a restrictive rule.

Bibliography:

- Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or Sexist Meme? Classifying Memes beyond Hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219, Online. Association for Computational Linguistics. https://aclanthology.org/2021.woah-1.23/