Hate Speech Meme Classification

AIM 5011, FALL 2021

The Katz School at YU

meme /mēm/ *noun* - a virally transmitted image, usually with embedded text

In an evolving world where social media is everything and the entire economy is single handedly driven by the tech sector. In a world where 97 percent of fortune 500 companies invest heavily in social media (Porteous, 2021), the content floating around these platforms should make a difference to us. While all sorts of content makes its way across the internet, memes have become a popular way to convey sentiment and connect with like-minded people. While the concept offers many benefits, memes can also be used to spread hate, misogyny and ignorance. This has led to problems such as hateful sentiment spread by radical groups (*Pepe the Frog*, n.d.). Last year, over 3000 people from across the world participated in a competition to build an AI model which can recognize hateful memes (Kiela, 2020). As students in a program at a Jewish school studying AI and ML, we felt it was an appropriate project to work on.

In the AI space where many tasks which are challenging elsewhere are relatively straightforward, building a hate-meme detector is a complex task. Memes rely on context to get a point across, good or bad. This means that any successful AI model needs to account for both the text in the meme and the actual picture. The picture part of the meme itself needs to be broken down into different parts. For example, we will likely want to know about objects in the image, if there are people, faces and about their ethnicities, age and/or gender.

When we are done with that, we need to use all the information to somehow determine if the image is a hate meme. This is quite the undertaking.

Project Architecture and Methods Applied

1. **The definition of Hate.** For this project, we had to agree on a clear set of rules to determine if a meme is considered hateful. Some might hear “hate meme” and think “discrimination”, however the correlation between “discrimination” and “hate meme” is not high enough to simply label a discriminatory meme as a hate meme. We therefore agreed that “Hate” in our project refers to something that is meant to propagate negative actions or generalization of a social group.
2. **The determination of the type of hate speech that exists was challenging.** Originally, a multi-categories classification project was considered, as the memes do regard different aspects in nature. Also, there are evolving ways of understanding the various groups and categories of people. Generally hate speech targets a group of people who share some common appearance, belief, background, etc. Official definitions from the United States and the United Kingdom were utilized to enhance the ultimate categories of people who are generally the target of hate speech.

3. **Limitation of data.** Finding a meme dataset and hateful memes, which are actually banned online, brought us a new challenge.We wanted to get familiar with, arguably the hardest part of an AI project, sourcing data. We went to great lengths not to use the Facebook Dataset and instead found our own memes and did our own labeling. Even though we spent considerable time looking for proper memes for the project (thanks Mazal-tov), the meme files are still limited, which contain only around 1000 records, and the hate memes are even less, not necessarily an optimal size to train a high quality model. But we worked with the resources we had.

4. **Labeling the memes.** Information sensitivities are different in different countries. Hate memes seem more important in the culture of the United States due to the culture diversity. The same challenge happened in our team as well. We are a culture diversified team, so the team members' cultural background and first languages are different. For these differences, it is difficult to understand the memes thoroughly without meme context, which is easy to cause bias . Plus the limitation of tata mentioned above, to anticipate running a high quality model would be even harder.

5. **Reasonable workflow architecture.** Unlike pure languages, image meaning is a big characteristic of memes. That is the reason why sending a meme is more expressive than merely saying words when users would like to convey some ideas and attitude. The image information is definitely beyond a pure Natural Language Processing task, which we focus on the languages more. With two kinds of information, image-meaning and text, the problem of building a

reasonable architecture brought up all sorts of questions. We spent some time researching different meme-recognition concepts like multimodal AI model training (*[1909.11740] UNITER: UNiversal Image-TExt Representation Learning*, 2019) and did learn from approaches by some of the Facebook project contenders (Zhu, 2020). Relying on self-study and cross-courses, we finally separated the workflow into two pathways, the image meaning classification and pure text classification, on our self determined dataset. In the meantime, we tried to use different models such as Densenet, VGG16, Resnet, EfficientNetB7, NASNetLarge and InceptionV3 to do transfer learning and also designed a CNN model with 10 layers on image classification. We got the best prediction accuracy, 38%, on a testing dataset by using Densenet. It looks like only using CNN to detect the meme is hard to reach an ideal result, so we have to combine image classification results with text classification. (Zeyu's section) Finally, we used ensemble training to fuse our models and improve the overall accuracy.

6. **Technical challenges** (best ways of image processing an text extraction)

Considering that the project was based on the information we manually created from scratch, the raw data cleanliness and human labeling accuracy cannot compare with the data reached from real world products. How to separate the image information and text information needs a lot of attempts to decide. For example, for OCR, a number of tools were tested and finally Google Vision was selected due to its remarkable versatility and relatively high accuracy. For facial feature detection we used an AI model called FairFace built on PyTorch which uses a deliberately sourced set ethnically, agewise and gender diverse images

(*FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation*, 2021). While we were successful in running the model with incredible results, we have not yet merged the results with the rest of the model. Some other challenges included finding an optimal CNN architecture, training time, model fine tuning and more which were all time consuming to the process.

7. **Google Vision API.** As stated before, we decided to use Google Cloud Vision service through its API so that we could create a custom code that fitted our needs. The Google Cloud Vision API service was set up using a gmail account, and the keys were generated for the service. From the API we could extract multiple information such as text, faces, different objects and their position, colour distribution, logo detection, landmarks and so on. We decided to extract the text, objects, and faces positions at first. The results from the API were saved as json files, from where we could extract the needed data. At the end we only used the text and objects to be further processed and feeded into our model.
8. **Image Classification.** For memes in hate detection, image plays a key role in this architecture because memes not only are shown on text but also on images. Before building a model or doing transfer learning, cleaning data is important for successfully classifying images. The image data comes from Facebook Dataset having different dimensions like 334*876*3, 986*232*1, 763*876*3. Here 3 and 1 represent 3 channels and 1 channel. Image with 1 channel means it is a gray

image. To better input data into the model, all images should be resized to the same size and channel. After considering the effect and memory pressure, all images were resized to 200*200*3. Even though choosing this dimension, the Densenet model with more than 10000 images occupies 40GB of memory. For enlarging our data and enhancing the robustness of the model, an image data generator was also used in the model. The final model we chose is Densenet with a pretraining weight and 1 neural output before softmax activation. And for letting the network be trained, we let all the layers' trainable to be True. After a long time of transfer learning, the accuracy reached 0.3825 on the testing dataset. This accuracy can be accepted because a single image sentiment classification without an NLP analysis can't reach an ideal result.

9. **Large team management.** There were a total of eight on this team, which is relatively large to have each person's individual opinion being considered. Due to the regular work schedule and location differences, it brought us a huge challenge to set up a time for group meetings and progress discussion to clarify the general plan. Combining the aforementioned technical and workflow difficulties, it took a long time to have the preliminary preparation and hesitate to divide tasks from scratch.

For all the challenges we encountered listed above, we can briefly conclude what should be planned ahead in the future.

1. If starting a project from scratch, do not start by trying to build the final "version 5" product. Trying to focus on a specific problem first, just a minimum viable

product. If we have an idea of a project, the idea doesn’t need to be “big”. This can offer the opportunity for more testing on a wider range for a specific idea.

2. As a data practitioner, the sensitivity to data is supposed to be absorbed into blood. In daily life, think of more ways of obtaining data, for instance, improving data obtaining and data manipulation skills. It might be a good start to have a new topic or even a new product.

3. For a larger group, individual talents need to be taken into account. However small a project is, the team is supposed to be divided into different roles. Grouping team members into different roles in mini groups might increase efficiency, as we finally did.

References

[1909.11740] UNITER: UNiversal Image-TExt Representation Learning. (2019, September 25). arXiv. Retrieved December 23, 2021, from https://arxiv.org/abs/1909.11740

FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. (2021). CVF Open Access. Retrieved December 23, 2021, from https://openaccess.thecvf.com/content/WACV2021/papers/Karkkainen_FairFace_Face_Attribute_Dataset_for_Balanced_Race_Gender_and_Age_WACV_2021_paper.pdf

Kiela, D. (2020, December 11). *Hateful Memes Challenge winners*. Facebook AI Research. Retrieved December 23, 2021, from https://ai.facebook.com/blog/hateful-memes-challenge-winners/

Pepe the Frog. (n.d.). ADL. Retrieved December 23, 2021, from https://www.adl.org/education/references/hate-symbols/pepe-the-frog

Porteous, C. (2021, March 18). *97% of Fortune 500 Companies Rely on Social Media. Here's How You Should Use It for Maximum Impact.* Entrepreneur. Retrieved December 23, 2021, from https://www.entrepreneur.com/article/366240

Zhu, R. (2020, December 15). *[2012.08290] Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning*

Solution. arXiv. Retrieved December 23, 2021, from https://arxiv.org/abs/2012.08290