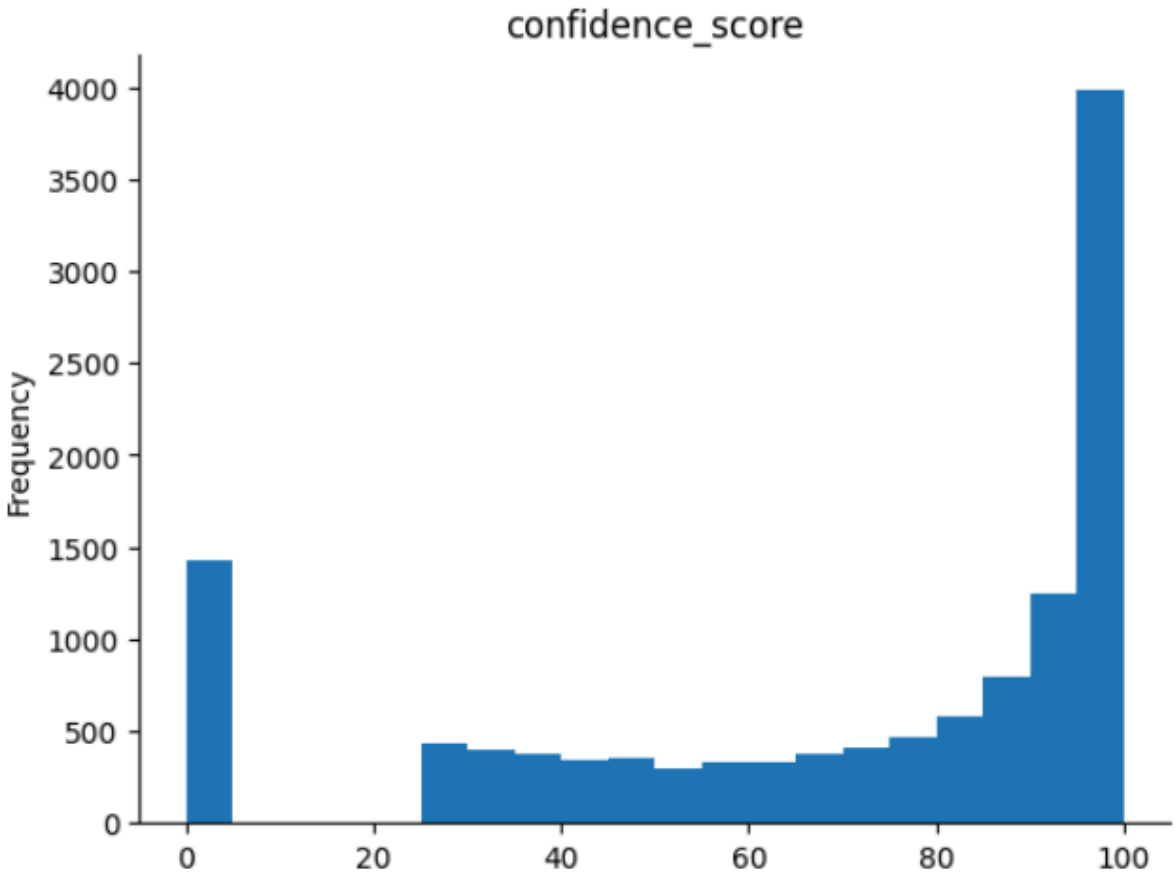
Analyzing hateful memes

⇒ Object Detection:

A. Goal: Utilize computer vision techniques to detect and identify objects within the images of the memes.
B. Tasks:
    a. Apply object detection algorithms to identify various elements within the meme images.
    b. Catalog the types of objects detected and analyze their frequency and distribution across the dataset.

Model used : Yolov4
Dataset trained on ; COCO (80 labels)
Time taken to detect 12140 images : 15.30 hrs



Note :

- The images with confidence_score 0 indicate that they are not part of the coco dataset (1426 images in total)
- Out of remaining 10894 images
- Main feature : in an image the object with largest bounding box are) is considered main feature

Main_Feature Split :

```
person         8129
               1426
dog             350
car             175
bed             147
sheep           147
cat             135
diningtable     127
tie             124
bird            120
sofa             92
```

For Complete split up please refer the code

S.no	Grouping/Class	Count (total = 10894) Excluded confidence score 0 images
1)	Images with human as main feature	8129
2)	Images with main feature as Dogs	350
3)	Images with Cars as main feature	175

Conclusion:-

Out of 12140 images 1246 are unidentified by the model and out of remaining 10894 images 5484 have confidence score greater than 0 but less than 90 percent which means in order to see whether the text in the images effect the object detection capability of the model we need to target these 5484 images mainly to test without the text

⇒ Caption Impact Assessment:

A. Goal: Assess the effect of overlaid captions on the accuracy and effectiveness of object detection.
B. Tasks:
    a. Determine how text overlays influence the object detection process.

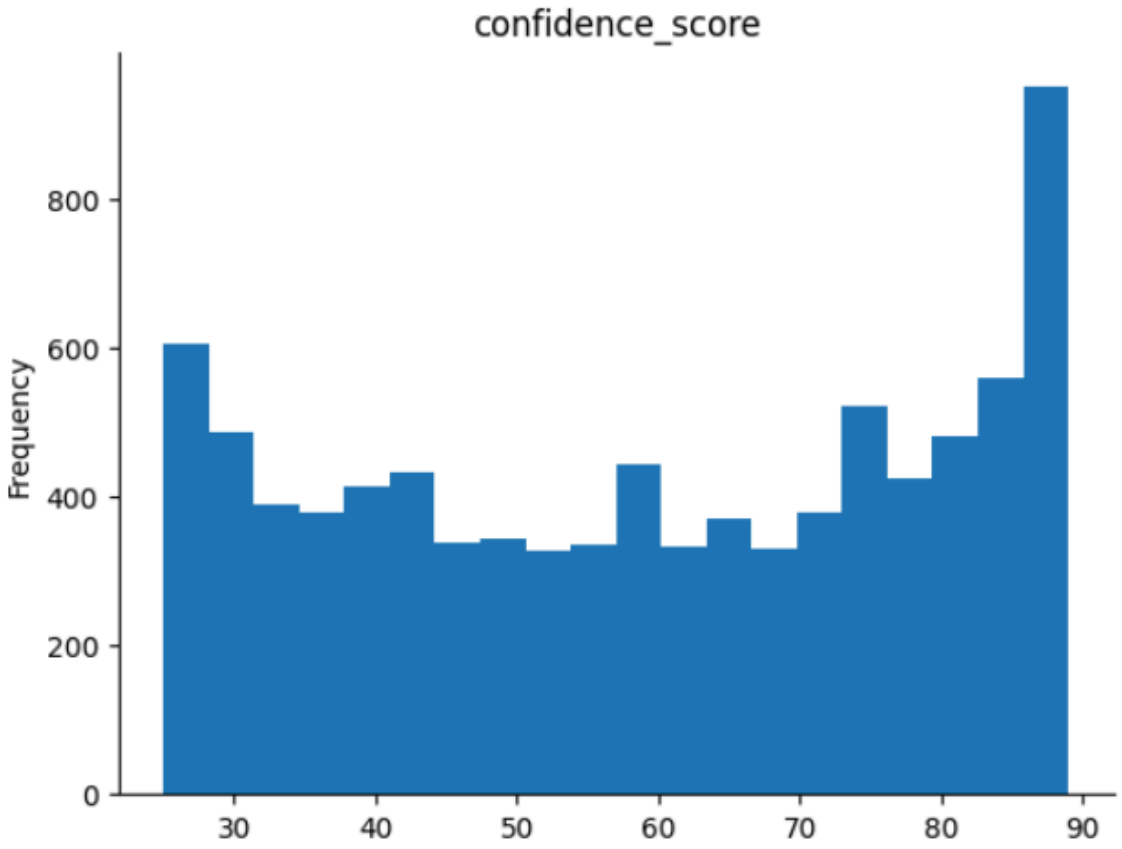
b. If necessary, develop and implement methods to minimize the impact of captions, such as using image processing techniques to filter out text. (You are not expected to make the model for this, try to find models that can do this for you)

Model used

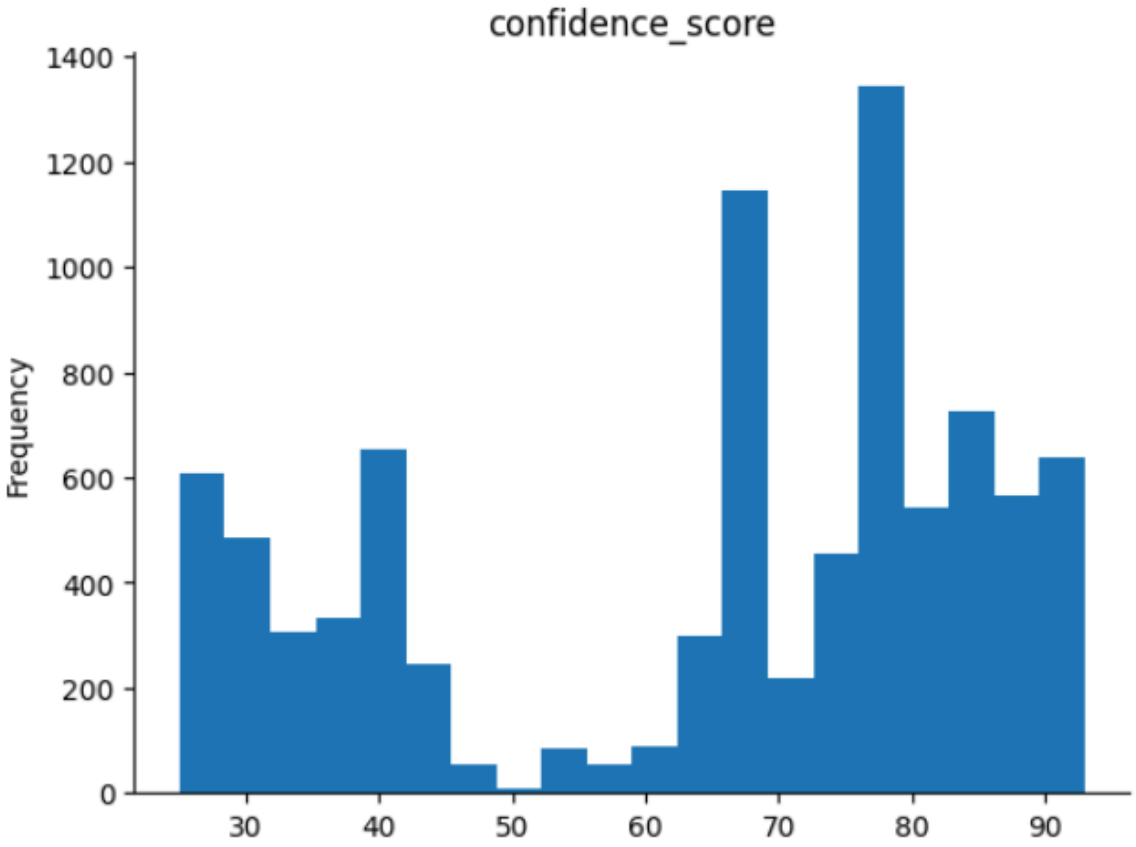
- easyOCR for text localization
- Inpainting algorithm :
    - Navier-Stokes, Fluid Dynamics, and Image and Video Inpainting (chose this because of minimum computation power with reasonable output)
    - Gan based models like (DeepFillV2, PartialConv . etc) requires heavy computation power)

Computation time taken 4.47 hrs

Before the text removal :



After cleaning the text :



Conclusion :

The confidence score score of the objects being detected has risen mainly the the number of object in whose confidence score was between 30 and 60 improved their confidence score . and the objects whose confidence score decreased is due to the draw-backs of image inpainting method chosen as it's results were not perfect as compared to deep learning based models

⇒ Classification System Development:

Goal: Develop a system to classify the images based on something non-trivial.
Suggestion: You could try classifying whether the image is a meme or not. Dataset for this is readily available as the positive class set is the dataset given, and you can easily source non-memes from other sources. You may freely choose any other classification task as well, but keep in mind that sourcing labeled data for the same might not be as easy. It is imperative that your classification task involves the provided dataset in part or as a whole. Properly report your methodologies, findings and performance of the model.

Conclusion

Here the most of memes are depicted with using human gesture and it's corresponding texts as medium and this can be proven by the section1 analysis where the main feature of 8129 images were a person/human and rest were some animals whose nature/habitat/traits they show were conveying the visual meaning and finally rest are object.

So in order to understand the meaning of the meem we need to correlate the visual and textual meanings for we need a way of representing these visual and textual features.

Images should be converted into embeddings based on the type of entity/objects being detected in the image for example , for humans we need their gesture or face expression or body posture and whereas other thing like animals or objects we need their shape as objects are detected or classified based on their shape/color/size etc but majorly shape and size so for this i want to create object representation model which will get the coordinates for body pose using pose estimation and for non humans its shape which can be obtained with edge or corner detection

I have chosen the text representation to be done using fastext embeddings because it's efficient way to represent and also to preserve the meaning of the sentence. Now the question is how to combine these both the features so that we can classify the meem efficiently with the help of understanding both visual and textual meaning . so we have made the model understand how exactly it should correlate the meaning conveyed visually and textually and then classify it .

The basic way of combining these can be making the embedding size of both image and text to be same and then multiplying them element wise to represent them as combined embedding or just concatenating them and then use MFH(multi factor high order pooling) for dimensionality reduction without/ minimal loss of semantics of the data .

Now we can use autoencoders with attention mechanisms to properly train the model and classify the memes . Therefore this is the architecture I came up with .