# **Dissecting Harmful Memes for Semantic Role Labelling**

## Code

Entire code is present here [sidtandon2014/NLP-Proj (github.com)](https://github.com/sidtandon2014/NLP-Proj) with dataset. While submitting the report I have included Jupyter notebooks for reference

## Problem Statement

Given a meme and an entity, determine the role of the entity in the meme: hero vs. villain vs. victim vs. other. The meme is to be analysed from the perspective of the author of the meme.

**Role labelling for memes:** This task emphasizes detecting which entities are glorified, vilified or victimized, within a meme. Assuming the frame of reference as the meme author's perspective, the objective is to classify for a given pair of a meme and an entity, whether the entity is being referenced as Hero vs. Villain vs. Victim vs. Other, within that meme.

* **Entities:**
  + **Hero:**The entity is presented in a positive light. Glorified for their actions conveyed via the meme or gathered from background context
  + **Villain:**The entity is portrayed negatively, e.g., in an association with adverse traits like wickedness, cruelty, hypocrisy, etc.
  + **Victim:**The entity is portrayed as suffering the negative impact of someone else’s actions or conveyed implicitly within the meme.
  + **Other:**The entity is not a hero, a villain, or a victim.

## Data Analysis

* 82% of the entities are from the OCR text and the rest 18% are from images.

|  |  |  |
| --- | --- | --- |
| **Total # of Entities in Train Data** | **# Of Entities from Text** | **# Of Entities from Images** |
| 17514 | 14314 | 3200 |
| 2069 | 1729 | 340 |

## Data Preparation

Clean OCR Text

Input OCR Text

* Special char, punctuation etc removal
* CamelCase conversion
* Fuzzy Matching

Clean Entities

Input Entities

## NER Model

Extracted entities

**HuggingFace bert-large-uncased NER**

Clean OCR Text

* The **HuggingFace bert-large-uncased NER** off-the-shelf model was able to successfully recognise *59% of the entities* in the train and text OCR with an *F1 score of 0.8*
* The BERT model trained on the input data was able to recognise only *56% percent of the entities.*

## Model Architecture

There are three main components of the model:

1. Extracting & identifying the entities present in image ([Code file](https://github.com/sidtandon2014/NLP-Proj/tree/main/Notebooks/ImageClassification))

* **Input**: Dataset Images & External data
* **Output**:
  + Entity Names (like Barack Obama, joe Biden, etc..)
  + Resnet18 layer4 output

1. Running NER pre-trained model on OCR text ([Code file](https://github.com/sidtandon2014/NLP-Proj/tree/main/Notebooks/NER))

* **Input**: OCR text
* **Output**: Named entities

1. Predicting entities as hero, villain, victim, other ([Code file](https://github.com/sidtandon2014/NLP-Proj/tree/main/Notebooks/OCR%20Classification))

* **Input**:
  + Output of step 1
  + Output of step 2
* **Output**: Classification of entities into 4 categories (Hero, Villain, Victim, Other)

Diagram

Description automatically generated

### Code Snippets

|  |
| --- |
| **Model Architecture**    Entire model is divided into 5 parts  **Step 1**: Reduce sentence embedding  **Step 2**: Reduce entity embedding  **Step 3**: Reduce image channels  **Step 4**: Concatenate Step 1,2,3 along the channels axis  **Step 5**: Run Conv layer and linear layer on top of it |

### **Quantitative Results**:

A picture containing calendar

Description automatically generated

## Future Work

* Enhance the image classification model(fine-tuning resnet18) by adding more training samples/ leveraging face identification/ object detection models
* Improve NER model by leveraging pre-trained object detection models and fine tuning existing BERT NER model
* Try out more advanced architecture with cross attention mechanism to handle the fusion of image & text features.

## References

1. Detecting Harmful Memes and Their Targets [2110.00413.pdf (arxiv.org)](https://arxiv.org/pdf/2110.00413.pdf)
2. Can images help recognize entities? A study of the role of images for Multimodal NER [2010.12712.pdf (arxiv.org)](https://arxiv.org/pdf/2010.12712.pdf)
3. Object-Aware Multimodal Named Entity Recognition in Social Media Posts With Adversarial Learning [IEEE Xplore Full-Text PDF:](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9154571)
4. A MULTIMODAL DEEP LEARNING APPROACH FOR NAMED ENTITY RECOGNITION FROM SOCIAL MEDIA [2001.06888.pdf (arxiv.org)](https://arxiv.org/pdf/2001.06888.pdf)
5. ConcatBert: Supervised Multimodal Bitransformers for Classifying Images and Text (<https://arxiv.org/pdf/1909.02950.pdf> )