Are you a hero or a villain? A semantic role labelling approach for detecting harmful memes

Shaik Fharook and **Syed Sufyan Ahmed** and **Gurram Rithika**
Sumith Sai Budde and **Sunil Saumya** and **Shankar Biradar**
Department of Computer Science and Engineering
Indian Institute of Information Technology
Dharwad, Karnatka, India
(fharookshaik.5@gmail.com)@iiitdwd.ac.in

Abstract

Identifying good and evil through representations of victimhood, heroism, and villainy (i.e., role labeling of entities) has recently caught the research community's interest. Because of the growing popularity of memes, the amount of offensive information published on the internet is expanding at an alarming rate. It generated a larger need to address this issue and analyze the memes for content moderation. Framing is used to show the entities engaged as heroes, villains, victims, or others so that readers may better anticipate and understand their attitudes and behaviors as characters. Positive phrases are used to characterize heroes, whereas negative terms depict victims and villains, and terms that tend to be neutral are mapped to others. In this paper, we propose two approaches to role label the entities of the meme as hero, villain, victim, or other through Named-Entity Recognition(NER), Sentiment Analysis, etc. With an F1-score of **23.855**, our team secured **eighth** position in the **Shared Task @ Constraint 2022**.

1 Introduction

The availability of smartphones and the internet has caught the interest of today's youth in social media. These applications provide a large platform for users to communicate with the outside world and share their thoughts and opinions. With these advantages comes a disadvantage: many people exploit the platform to spread offensive content on social media under the guise of freedom of expression (Boon, 2017). This incendiary material is usually directed towards a single person, a small group of people, a religious group, or a community. People create offensive content and aggressively spread it over social media (P. Fortuna, 2018; T. Davidson, 2017). For many purposes, including commercial and political benefit, this type of information is created (Jeff Goodwin and Polletta, 2009; Biradar et al., 2022). This type of communication can disturb societal harmony and spark riots. It also has the ability to have a negative psychological impact on readers. It has the potential to harm people's emotions and behavior (Stieglitz and Dang-Xuan, 2013; Biradar et al., 2021). As a result, identifying such content is crucial. Further, researchers, politicians, and investors are working to build a reliable method for dissecting the dangerous memes present over the internet.

Framing allows a communication source to portray and describe a problem within a "field of meaning" by employing conventional narrative patterns and cultural references (Scheufele, 1999). By connecting with readers' existing knowledge, cultural narratives, and moral standards, framing helps to construct events (Green). It can portray the characters in a story as heroes, villains, or victims, making it easier for the audience to anticipate and comprehend their attitudes, beliefs, decisions, and actions. Narrative frames can be found in various media, including memes, films, literature, and the news. Narrators use emotionality to plainly distinguish between good and evil through vivid descriptions of victimization, heroism, and villainy, which is a major feature of the popular storytelling culture (Diego Gomez-Zara, 2018). Positive adjectives are used to portray heroes, whereas negative terms depict victims and villains. In popular culture, heroes represent bravery, great accomplishments, or other noble attributes, whereas villains represent malicious intents, conspiring, and other undesirable characteristics (Diego Gomez-Zara, 2018). To summarise, narrative frames are essential for understanding new situations in terms of prior ones and therefore making sense of the causes, events, and consequences.

The standard method for detecting frames of the narrative is by examining the semantic relationships between the various elements in the meme about the events it portrays. Understanding the events in a narrative and the roles that the entities

in that meme play in those events, on the other hand, is a complex, tough, and computationally expensive task.

Thus, rather than determining all of the specific events and event types described in the meme, as well as the semantic relationships among the entities involved in those events in great detail, we propose methodologies in which the entities are analyzed at a much higher level of abstraction, specifically in terms of whether they hold the qualities of heroes, victims, villains, or none as conveyed by the terms used to characterize them. As a result, we arrive at a rather basic realization. The terms nearest to each entity are evaluated for their sentiment polarity or closeness to associated terms with heroes, villains, or victims.

2 Literature review

The topic of entity role detection from narrative has recently piqued the interest of several corporate and academic researchers in recent times. However, there were just a few efforts to extract knowledge and present it from newspaper articles that especially utilized the newspaper article bodies to derive meaning, focusing on the headline (Boon, 2017; Dor, 2003; Diego Gomez-Zara, 2018). But there have been hardly any attempts to identify the entities that had been exalted, demonized, or victimized (Melodrama and of Communication, 2005). Instead, studies were conducted to see how satire delivered through the means of internet memes affects brand image (Christopher Kontio). However, no existing approach has been able to handle harmful content identification in multimodal data employing the role labeling notion. In this paper, the emphasis is on detecting which entities are vilified, glorified or victimized in a meme by assuming the frame of reference from the meme author's perspective (Sharma et al., 2022).

3 Task and Dataset description

3.1 Task

As noted in the competition's problem statement, the focus is on recognizing whether entities are glorified, condemned, or victimized within a meme by assuming the meme author's frame of reference[1].

Given a meme and an entity, the task is to determine the role of each entity detected in the meme as hero or villain or victim or other. The constraint here is that the meme has to be analyzed from the perspective of the author of the meme (Sharma et al., 2022).

[1] https://constraint-lcs2.github.io/

3.2 Dataset description

The dataset for this task was provided by the organizers of the competition Shared Task @ Constraint 2022. This dataset is a collection of memes and their associated entities from two domains: Covid-19 and US Politics. It is organized into three parts: train, validation, and test set, respectively. Each item of the dataset from train and validation contains an image of the meme and its pre-extracted OCR with its entities mapped to Hero, Villain, Victim, and Other Categories. A sample item of the dataset can be seen in Figure 1.

Corresponding JSONL Text input

{
"OCR": "Bernie or Elizabeth? \n Be informed. Compare them on the issues that matter.\n Issue: Who makes the dankest memes?\n",
"image": "covid_memes_18.png",
"hero": [],
"villain": [],
"victim": [],
"other": ["bernie sanders", "elizabeth warren"]
}

Figure 1: Train/Validation Dataset sample

Each item of the test dataset contains an image of meme and its corresponding pre-extracted OCR and its entities. The total dataset contains 6920 items, and a detailed domain-wise distribution of train, validation, and test sets can be seen in Table 1.

4 Methodology

This study has proposed two submissions based on two different methods. In the first method, we perform entity recognition then sentiment analysis.

	Train	Valid -ation	Test
Covid-19	2700	300	718 (Combined)
US Politics	2852	350	
Total	5552	650	718

Table 1: Data set Distribution

In the second method, we perform entity recognition and then use Wu-Palmer similarity (S. Bird, 2009) to calculate similarity scores of entities with each of the roles, i.e., hero, villain, victim, and other.

4.1 Data Processing

The following data processing steps were performed while creating an end-to-end system, i.e., given a meme image, the OCR text recognizes the entities present in that meme by performing entity recognition on the text. However, in the competition, as the entities are already recognized and given as an entity list, we can skip the entity recognition step here for the competition.

Then each entity is linked to its corresponding parts of the sentence (words surrounding the entity) present in the OCR text of that respective meme. Here a fair assumption was made that the words nearer to the entities weigh more than those farther from the entity in its role assignment. So first, we search for entity occurrence in the OCR sentences. Then using a window approach(i.e., selecting the n-words occurring before that entity and the n-words occurring after the entity), we create a sub-part of that sentence. By doing this on the whole OCR of that respective meme, we create a list of sub-sentences, one for each entity present in that particular meme as shown in Figure 2.

```
"memes_4576.png": {
    "nation": [
        "this great nation must bear the"
    ],
    "thomas paine": []
},
```

Figure 2: Entity sentence linking example

4.2 Methods and models

In this study, two different frameworks have been experimented for role detection. The description of the frameworks are discussed in the following subsections.

4.2.1 Framework-I

1. For each entity given in a particular meme, identify the words close(i.e., surrounding words) to these entities by linking the entity sentence.

2. Perform sentiment analysis to determine the polarity of these words, thus making out the sentiment attributed to the entity.

3. Use sentiment polarity to role label the entities, according to the proposed semantic classes.

After performing entity sentence linking, we determine the sentiment score of the words(sub-sentences) linked with an entity; we do this for all the entities mentioned in that particular meme. To do this, we calculate the sentiment(i.e., word polarity) for each word using a standard toolkit like VADER-Sentiment[2](as it has a huge vocabulary of the word polarities), thus getting a polarity for each word, which ranges between [-1, 1] (i.e., very-negative to very-positive). These sentiment-polarities are then summed up for each sentence. Finally, the sentiment-polarities for each sentence are normalized and then averaged to get an overall sentiment ascribed for the entity.

As we know, that hero is linked with positive words with positive sentiment. Similarly, victims and villains are linked with negative words with negative sentiments. If the words(sub-sentences) have no polarity, they don't glorify or vilify or victimize any entity thus semantically similar to the class "other" as described in Figure 3.

4.2.2 Framework-II

1. For each entity given in a particular meme, identify the words close(i.e., surrounding words) to these entities by linking the entity sentence.

2. Determine the resemblance of these words with the words used to describe heroes, villains, and victims by curating word sets or dictionaries for each role.

3. Role label the entities by analyzing their similarity scores with those of hero, villain, and victim. If the scores are zero or almost the same, role label it to "other" class.

[2]https://pypi.org/project/vaderSentiment/

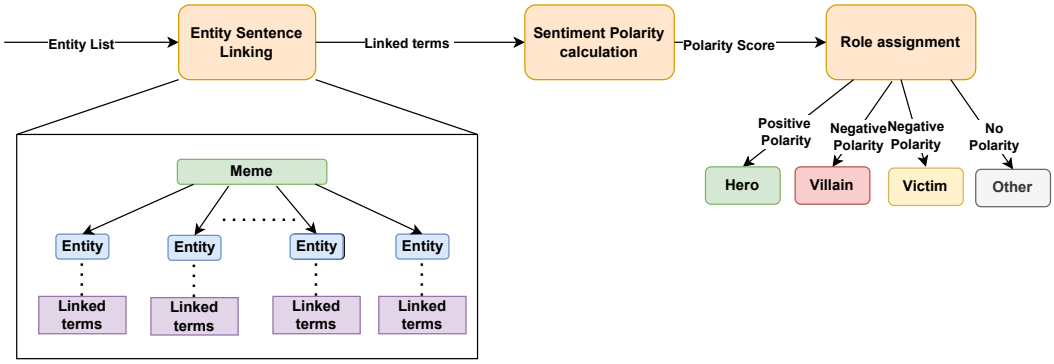

Figure 3: Framework-I architecture

After performing entity sentence linking, We create three dictionaries, one for each hero, villain, and victim containing the words or terms similar to them, respectively. Then by using a method like Wu-Palmer similarity[3] we calculate the similarity score of each word from the entity-sentence linking step with hero dictionary, villain dictionary, victim dictionary to create the similarity dictionary Figure 5. Then the similarity score for each entity is determined by summing the similarity scores of all the words found in the sub-sentences. Then it is normalized to get an overall similarity of a particular entity with the roles of hero, villain, victim, and others. We assign an entity to the role whose similarity score is the highest using these similarity scores. If the similarity scores with each of the roles are almost similar or zero, we assign it to the class "other" in the proposed role assignment approach as described in Figure 4. Implementation details of the proposed model are made publicly available [4]

5 Results

In the competition, teams were ranked based on macro F1-Score across all the classes. The suggested method and model secured the eighth position in the competition for the task of dissecting harmful memes for Semantic role-labeling of entities. Table 2 shows the rankings of various teams, and the performance of the proposed system is indicated in bold letters.

The model performs well in the role labeling task. However, in some cases, the model under performs in identifying the categories due to the difficulty in capturing some of the attributes or traits related to the roles. As a result, the overall systems' macro F1-score has been low at 23.855. In addition, the ensembling of multiple NLP sub-tasks also have contributed to the decrease of the F1-score of the system. The systems' performance can be further improved by modeling those NLP sub-tasks in the proposed methods using better parameters which could potentially increase the score.

SL. no	Username / Team Name	F1 Score
1	Shiroe	58.671
2	jayeshbanukoti	56.005
3	c1pher	55.240
4	zhouziming	54.707
5	smontariol	48.483
6	zjl123001	46.177
7	amanpriyanshu	31.943
8	**Team IIITDWD (fharookshaik)**	**23.855**
9	rabindra.nath	23.717

Table 2: Top performing teams in the Competition

6 Conclusion and future enhancement

The current system implementations use NLP techniques such as entity recognition, sentiment analysis, and word sets and dictionaries, all of which have shown promising results in the role labeling task. Across all classes, the existing system implementation produced a good F1 score. However, as the model is based on simple proximity measures, it has issues when dealing with OCR text that contains composite grammatical structures such as indirect speech, passive voice etc. In this experiment, the n-words window size used for data processing is n=3. As a result, there is potential

[3] https://arxiv.org/ftp/arxiv/papers/1310/1310.8059.pdf

[4] The source code for reproducing our work can be found at https://github.com/fharookshaik/shared-task_constraint-2022

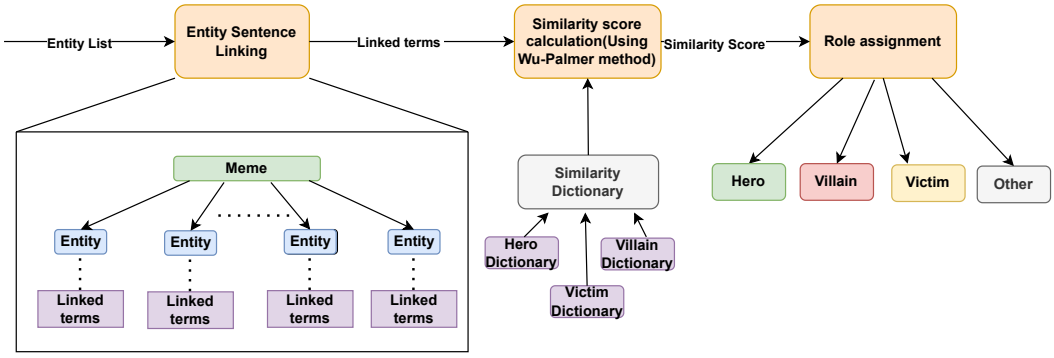

Figure 4: Framework-II architecture


```
"sentaient": [
    0,
    0,
    0
],
"inquire": [
    0.25236209659286585,
    0.2891268554312031,
    0.2690599133637109
],
```


Figure 5: Similarity Dictionary

for various future changes to increase the system's performance.

Further, in future experiments and add-ons, we plan to leverage some of the SOTA(State Of The Art) machine learning models such as SVM to discover distinct sentiment polarity boundaries for various sub-tasks to enhance the working of sub-tasks and thereby improving the system's role labeling performance.

References

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2021. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2470–2475. IEEE.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2022. Combating the infodemic: Covid-19 induced fake news recognition in social media networks. *Complex & Intelligent Systems*, pages 1–13.

Miriam L. Boon. 2017. Augmenting media literacy with automatic characterization of news along pragmatic dimensions. *ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Melker Pripp Viktor Magnusson Christopher Kontio, Klara Gradin. An exploration of satirical internet memes effect on brand image. *Linnaeus University*.

Larry Birnbaum Diego Gomez-Zara, Miriam Boon. 2018. Detection of roles in news articles using natural language techniques. *23rd International Conference on Intelligent User Interfaces*.

Daniel Dor. 2003. On newspaper headlines as relevance optimizers. *Journal of Pragmatics*.

Melanie C. Green. Transportation into narrative worlds: The role of prior knowledge and perceived realism. *Discourse processes*.

James M. Jasper Jeff Goodwin and Francesca Polletta. 2009. Passionate politics: Emotions and social movements. *University of Chicago Press*.

Melodrama and September 11. Journal of Communication. 2005. *Villains, victims and heroes: Melodrama, media, and September 11. Journal of Communication 55*.

S. Nunes P. Fortuna. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*.

E. Loper S. Bird, E. Klein. 2009. Natural language processing with python: analyzing text with the natural language toolkit. *O'Reilly Media, Inc*.

Dietram A. Scheufele. 1999. Framing as a theory of media effects. *Journal of communication*.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Findings of the constraint 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations - CONSTRAINT 2022, Collocated with ACL 2022*.

Stefan Stieglitz and Linh Dang-Xuan. 2013. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*.

M. Macy I. Weber T. Davidson, D. Warmsley. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*.