3 Major Strengths of the paper -

- One of the major strengths of the paper is that, in the model they have captured aspects like - emotions and hostility. Well, this is important in the case of harmful memes, as it should be able to understand which crowd we are trying to harm or their image per se. The emotions are also captured by doing the sentiment analysis of the text available on the meme. Further, they are correlated with the contextual information, that has been annotated by 2 external people.
- The sampling distribution of the MCC is quite rich. Each sample has the following things - meme image, contextual document (of max 350 tokens - this is the descriptive part which helps to understand the meme in the intended way), OCR extracted meme's text (the actual small sentence of 5 to 10 words written in the meme itself), and set of ground truth evidence (to support the contextual document). Further, the MCC has been manually annotated to create a multimodal dataset with related context. This approach increases the meme inferential knowledge and its representation.
- The conventional transformer uses self-attention. But does not take into consideration additional contextual information. These people have proposed the usage of MHA (multi-headed). This helps to compute self-attention over contextual representation. Self-attention does allow us to weigh its relevance to other tokens, but there is no way to incorporate the contextual information. Now, see MHA solves this by focusing to multiple subspaces of input data. So, we can capture different types of relationships and context information.

3 Major Weaknesses of the paper -

- There is no mention of how the OCR process was conducted. Moving in the same direction, to strengthen my point. It is clearly stated, "lower precision, as against the higher recall scores, suggests the inherent noise being additionally modeled." Now this leads to a reduction of integration of visual information not matching correctly with the factual knowledge, which is currently lacking in the MIME.
- Memes come with audio playback like memes related to money, people buying big things, Ambani Ji's wedding memes and Jethalal's power of money will have audio like - "Moti Chain, Mota Paisa." Recently, such playbacks have been replaced by "Millionaire - Honey Singh." Sad memes like Heartbreak or something related to that sort will have audio playbacks like - "Yaar ka Sataya hua hai, etc." Memes related to news of the Army and all have songs like - "Arjan Vally." There is no mention in the paper of how these audio pieces are influencing memes (of any sort - especially harmful ones).
- The scope of the dataset is too low, only limited to political and historical memes. This might be limiting the model's applicability. Also, note that only two annotators have been used for creating the dataset, this might as well introduce bias in the ground truth labeling work. A larger and more diverse strategy like cross-annotation might avoid such issues.

3 Major Improvements of the paper -

- The first scope of improvement is we can use a wider range of meme types. This will make the model more generalized across different contexts and cultures. Also, as stated in the weakness, we need to include the audio playbacks for all the memes (if they have). We can build a dictionary of data structures, where the most common audio playbacks can be mapped to their emotions. Finally, it will be giving the memes one more layer of context to understand.
- Furthermore, we noticed that the current MIME approach has limitations in modeling the complex level of abstraction that the memes exhibit, which leads to bad predictions. We can ask the model to prepare for hate speech detection and sentiment analysis of the text written on the memes. This will increase the applicability of the MEMEX and the MIME model.
- As per the research, there are different types of errors occurring in the model currently, which do not identify the abstract concepts correctly (possibly). I think we need to refine the training data - 4 things are good for the next iteration - Add audio playbacks with their context dictionary, hate speech detection model, sentiment analysis of the written text, and better OMR process (Adobe API currently gives the best OMR in detecting text from images). This will enhance the overall ML model and even extend it to RL (implementation of a feedback mechanism to learn from its mistakes).