# The Persuasive Memescape: Understanding Effectiveness and Societal Implications of Internet Memes

**Gitanjali Kumari**[1] **Pranali Shinde** [1] **Asif Ekbal**[1]
[1]Department of Computer Science and Engineering,
[1]Indian Institute of Technology Patna, India
{gitanjali_2021cs03,pranali_1901cs41,asif}@iitp.ac.in

## Abstract

Persuasive meme identification is a crucial task in automatically categorizing memes based on their persuasive nature. Memes, being highly influential in online communication, have the ability to shape individuals' attitudes, behaviors, and beliefs, both positively and negatively. They can be utilized to promote positive actions, challenge social norms, and raise awareness, but they can also perpetuate harmful ideologies, spread misinformation, stereotype, and manipulate emotions. In this paper, we are addressing this challenge by empirically investigating three novel tasks, *viz.* (i) Task 1: Persuasive meme detection, (ii) Task 2: Identification of the effectiveness of persuasive memes, and (iii) Task 3: Identification of persuasion techniques used in persuasive memes. To this end, we make the very first attempt to release a high-quality, large-scale dataset, *Persuasive_meme* [1], since there is no publicly available such dataset for the Hindi-English code-mixed (Hinglish) domain.[2] We further developed several baseline unimodal and multimodal models for these tasks. Empirical evaluation with respect to both, qualitative and quantitative analysis, on the *Persuasive_meme* dataset highlight the significance of multimodality in addressing these tasks effectively. Additionally, we discuss the limitations of the current models and emphasize the need for further research to overcome these challenges.

## 1 Introduction

The rise of internet memes has revolutionized communication in the contemporary digital landscape, surpassing the boundaries of traditional textual and visual mediums. Memes, with their engaging, humorous, and relatable nature, have emerged as powerful tools for conveying messages and ideas (Kirk et al., 2021). These memes possess the potential for both positive and negative impacts, captivating audiences and shaping their perceptions. Many memes, despite being humorous, use extremism and dark humor to promote societal harm (Kiela et al., 2020a; Kumari et al., 2021; Bandyopadhyay et al., 2023). Meme analysis is, therefore, essential for detecting offensive content, analyzing psychological responses, and gaining a deeper understanding of the persuasive strategies employed in online communication, etc. (Rijhwani et al., 2017; Sharma et al., 2020; Kiela et al., 2020a; Suryawanshi et al., 2020; Hossain et al., 2022; Sharma et al., 2022). Prior studies have primarily focused on the humorous and cultural aspects of memes, overlooking their potential as persuasive tools (Seiffert-Brockmann et al., 2018; Nee and Maio, 2019). Consequently, the examination of memes' persuasive effectiveness remains an underexplored area, limiting our understanding of their true impact on individuals and society.

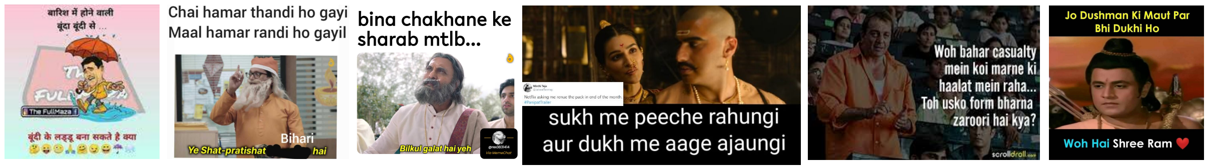
Persuasion is a fundamental aspect of communication that seeks to influence the beliefs, desires, and actions of an audience (Nee and Maio, 2019). Several research has emphasized the importance of assessing persuasive communication in the digital domain (Somasundaran and Wiebe, 2010; Tan et al., 2014; Trabelsi and Zaïane, 2014; Jaech et al., 2015). Fahmy and Omneya (2021) highlighted the need for a comprehensive analysis of persuasive strategies employed in online visual content, including memes. Similarly, (Seiffert-Brockmann and Diehl, 2016; Seiffert-Brockmann et al., 2018; Nee and Maio, 2019) have highlighted the potential of memes to influence public opinion and disrupt political decision-making, thus emphasizing the need for thorough investigations into the persuasive nature of memes.

However, despite the existing studies, there is still a gap in empirical research within the field of Natu-

[1] Code and dataset are available at this link: https://www.iitp.ac.in/ ai-nlp-ml/resources.html#Persuasive-Meme as-well-as at our GitHub repository. https://github.com/Gitanjali1801/Persuasive_meme.git

[2] WARNING: This paper contains meme samples that are offensive in nature.

*Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–959
November 1–4, 2023.



{Can we make boondi laddus from the drizzle that occurs in the rain?}

(a) [0,---,--]

{"My tea has cooled down, My girl became pro***** ". He is 100 percent Bihari.}

(b) [1,-2, Wordplay and Pun, Invective]

{Drinking alcohol without having snacks. This is absolutely wrong.}

(c) [1,-1, Hyperbole]

{I will follow you in happiness, I will come forward in sorrow.}

(d) [1,0, Wordplay and Pun]

{If a person remains in the condition of death outside the casualty section, then is it necessary for him to fill the form?}

(e) [1,1, Irony]

{The one who is sad even on the death of enemies, that is Lord Ram.}

(f) [1,2, Hyperbole

Figure 1: Examples from our *Persuasive_meme* dataset. The labels are in the format Persuasion_[Identification, Effect, Techniques]. For Identification, {0, 1} correspond to Non-persuasive and Persuasive memes, respectively. For Effect, {-2, -1, 0, 1, 2} correspond to Highly Negative Persuasion, Moderately Negative Persuasion, Neutral Persuasion, Moderately Positive Persuasion and Highly Positive Persuasion, respectively. For Techniques {Metaphors, Analogies, Hyperboles, Irony, Alliteration, Personification, Puns and wordplay, and Invective}. Texts in { } are the English translation of code-mixed Hindi-English meme texts.

ral Language Processing (NLP) that systematically evaluates the persuasive effectiveness of memes and explores their persuasive strategies. Persuasive memes possess the power to influence society both positively and negatively. A positively persuasive meme is used to convey powerful messages, challenge norms, and advocate for social change (c.f. Example (e,f) in Figure 1). Contrarily, negatively persuasive memes can perpetuate harmful ideologies, spread misinformation, and manipulate emotions (c.f. Example (b,c) in Figure 1). Understanding the impact of persuasive memes is crucial in shaping public opinion and promoting responsible communication (Nee and Maio, 2019). This research addresses this gap by analyzing the persuasive effectiveness of memes, offering valuable insights for informed and ethical digital discourse.

**Code-mixing** The widespread use of code-mixed memes on social media platforms presents a significant challenge for meme analysis and understanding (Edwards, 1995; Bali et al., 2014; Rijhwani et al., 2017; Kamble and Joshi, 2018; Ghanghor et al., 2021; Hossain et al., 2022). To the best of our knowledge, there is no publicly available dataset for persuasion identification for English-Hindi (Hinglish) code-mixing. In order to address this gap and facilitate research in this area, we have created a dataset of Hinglish memes across four domains, namely political, religious, racist, and sexist. This dataset enables the analysis and exploration of persuasive techniques in code-mixed memes and contributes to the understanding of the persuasive nature of Hinglish memes in various contexts.

**Our Contributions:** In this paper, we study persuasive memes, and formulate three tasks. **Task 1 (Persuasive meme detection):** Given a meme, detect whether it is persuasive or not. **Task 2 (Identification of the effectiveness of persuasive memes):** Given a persuasive meme, analyze the various categories of persuasive impact, ranging from highly negative to highly positive persuasion. **Task 3 (Use of persuasion techniques):** Given a persuasive meme, identify which techniques are used to enhance the persuasiveness and impact. To this end, we develop a novel code-mixed Hinglish dataset, named *Persuasive_meme*, containing 6k real memes in the Indian scenario, which we collected from social media and carefully annotated. In our study, we meticulously develop comprehensive annotation guidelines for all three tasks. We extensively evaluate multiple state-of-the-art unimodal and multimodal models to establish benchmark performance for these tasks.

## 2 Related work

**Persuasion detection in textual data:** In the field of persuasion detection, researchers have shown significant interest in text-based analysis. For instance, Tan et al. (2014) investigated the influence of wording in predicting the popularity of social media content. Guerini et al. (2015) explored the impact of sounds on persuasiveness by examining euphony and focused on the phonetic aspect rather than language usage. Park et al. (2016) developed an interactive system to assist human moderators in selecting high-quality news. Additionally, Reddit has become an important platform for research on social news analysis and recommendation, as demonstrated by previous studies exploring language use, community reactions, and comment analysis (Buntain and Golbeck, 2014; Jaech et al., 2015; Tan et al., 2016). Wei et al. (2016) investigated the identification of persuasive comments in online forums using several feature identification techniques.

**Persuasion detection in visual data:** The advent of social media has driven researchers to incorporate images into their analyses to better understand persuasion and its intentions. Political science and mass media scholars have explored audiences' emotional and cognitive responses to televised images of political leaders (Bucy and Bradley, 2004; Masters et al., 1986) and investigated the selective use of images by the media for persuasive purposes (Barnhurst and Steele, 1997; Grabe and Bucy, 2010). Joo et al. (2014) have demonstrated the value of systematically examining communicative intents in uncovering deeper insights into the meaning and persuasive impact of images, transcending surface-level feature classification.

**Other studies on memes:** The widespread proliferation of memes and their increasing impact on online communication have recently sparked research interest in meme analysis in the NLP community. However, the existing efforts in meme analysis have primarily centered around identifying hateful or offensive memes (Rijhwani et al., 2017; Sharma et al., 2020; Kiela et al., 2020a; Suryawanshi et al., 2020; Hossain et al., 2022; Sharma et al., 2022), or detection of propaganda techniques (Dimitrov et al., 2021) with limited attention given to the identification of persuasive memes.

**Code-mixing:** Furthermore, most of the existing works for memes in the code-mixed settings have been performed on textual data (Kamble and Joshi, 2018; Bali et al., 2014; Mathur et al., 2018; Tang et al., 2020; Bohra et al., 2018). Persuasiveness identification in multimodal, especially in Hinglish scenarios, is primarily unexplored due to inadequate resources and tools. As a result, there is a significant gap in the research landscape when it comes to identifying persuasive memes. This paper aims to address this gap by focusing specifically on the task of persuasive meme identification and exploring their impact on social media platforms, mainly in Hinglish.

# 3 Corpus Creation

We create a new dataset due to the lack of an existing Hinglish dataset for persuasion identification in memes. During the preparation of this corpus (henceforth referred to as *Persuasive_Meme*), we take the following steps: (i) Data Collection, (ii) Annotation process, (iii) Annotation guidelines, and (iv) Data statistics. We discuss these steps below in more detail:

## 3.1 Data Collection

We collected memes covering various domains, such as politics, religion, and social issues like terrorism, racism, sexism, etc., using a list of 126 keywords (c.f. Table 1). To keep a strategic distance from copyright issues, we only retrieved the freely available memes in the public domain with the help of a browser extension called Download All Images of Google's image search engine[3]. We finally retain only around 6K unique memes after removing the duplicates. (c.f. Figure 2 for the data collection process.)

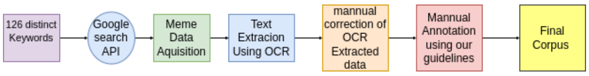


Figure 2: Data collection procedure

## 3.2 Annotation process

For the annotation purpose, we require experienced annotators with an expert-level understanding of the code-mixed Hindi-English language. The annotation team comprises of three highly qualified members, both male and female, 20–25 years old, who possessed undergraduate degrees and extensive experience in code-mixed Hindi-English and NLP research. Their expertise and fluency in the language ensured the quality of the annotation process. It is important to note that no incentives were provided to the annotators to maintain objectivity. Furthermore, we only included annotators familiar with the Indian scenario. To address this, we divided the process into three distinct phases: (i) Dry Run, (ii) Final Annotation, and (iii) Consolidation. Details of each step are described in the following sections.

| Keywords |
|---|
| Demonetization, Odd-even rule, GST, Liquor ban in Bihar, Fatwa, Beef ban, Love jihad, Hindu-Muslim, JNU incident 2016, Article370, Intolerance, Islamophobia, Citizenship Bill, Anti Hindu, darkisbeautiful, fake Feminism, No Acid, article377, me too, Aurat Azadi March, LGBTQ, Dowry, Parental Expectations, Indian Festivals, Cricket Rivalries, Swachh Bharat Abhiyan, Make in India, Beti Bachao, Beti Padhao, Digital India, Jan Dhan Yojana, Atmanirbhar Bharat, |

Table 1: Examples of lexicons used to collect memes

[3] https://download-all-images.mobilefirst.me/

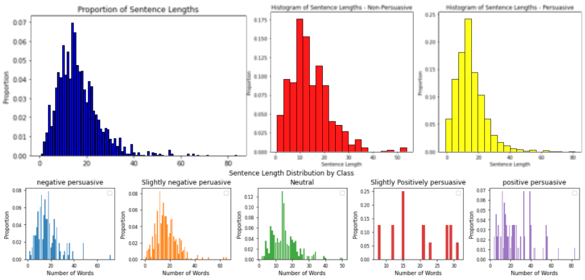
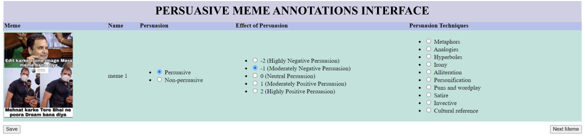


Figure 4: Histogram of the length of the meme' text for each class: On the top for Persuasion Identification, and for the effectiveness of persuasive memes on the bottom.



PERSUASIVE MEME ANNOTATIONS INTERFACE

| Meme | Name | Persuasion | Effect of Persuasion | Persuasion Techniques |
|---|---|---|---|---|
| | meme 1 | ● Persuasive<br>○ Non-persuasive | ○ -2 (Highly Negative Persuasion)<br>● -1 (Moderately Negative Persuasion)<br>○ 0 (Neutral Persuasion)<br>○ 1 (Moderately Positive Persuasion)<br>○ 2 (Highly Positive Persuasion) | ○ Metaphors<br>○ Analogies<br>○ Hyperboles<br>○ Irony<br>○ Alliteration<br>○ Personification<br>○ Puns and wordplay<br>○ Satire<br>○ Invective<br>○ Cultural reference |

Edit karke toone image Mera meme bana diya
Mehnat karke Tere Bhai ne poora Dream bana diya
Save
Next Meme

Figure 3: Our Annotation interface. The right part shows labels for each task and the left part shows a meme for which annotation has to be done.

### 3.2.1 Phase 1: Pre-processing and Text editing

The collected raw memes are (i). noisy such as background pictures are not clear, (ii) non-code-mixed Hindi-English, i.e., meme texts are written in other languages except code-mixed Hindi-English, and (iii) non-multi-modal, i.e., memes contain either text or visual content. We manually discarded these memes to reduce manual data annotation effort. Next, we extracted the textual part of each meme using an open-source Optical Character Recognition (OCR) tool: Tesseract[4]. The OCR errors are manually post-corrected by the annotators. Finally, we consider 6,000 memes for data annotation. The average meme text length for the memes samples in our dataset is between 10-20 words. (Refer to Figure 4 to see the plot. )

### 3.2.2 Phase 2: Dry run

This stage is the pilot annotations to train the annotators to understand our annotation guidelines. In Figure 3, we have shown our annotation interface. We annotated 200 samples on our own to use as a quality checker while evaluating the annotators' ability. We conducted a dry run on the same 200 memes, which helped the annotators understand well the definitions of all the labels, as well as eliminate the uncertainties/challenges about the annotation guidelines. For the preliminary data, one meme is annotated by three annotators. For Tasks 1 and task 2, we computed the inter-annotator agreement(IAA) in terms of Cohen's Kappa coefficient (Bernadt and Emmanuel, 1993), and for Task 3 in a multilabel scenario, we reported Krippendorff's Alpha Coefficient (krippendorff, 2011). We average the Cohen's Kappa/Krippendorff's Alpha Coefficient score of all three annotators $a_i$ for i= 1 to 3 for each meme for all three tasks. It was observed that the initial scores for all three tasks were low (0.6529, 0.8097, and 0.5874), which is typical for a first pilot test.

| Dataset | Domain | Language | Modality | | Label | Statistics |
|---|---|---|---|---|---|---|
| | | | T | V | | |
| Change My View (CMV) (Tan et al., 2016) | Open | English | ✓ | — | Persuasion | 1,785 |
| Echoes of Persuasion (Guerini et al., 2015) | Twitter/Movie/CORPS | English | ✓ | — | Persuasion | 11k/2k/6k |
| Joo et al. (2014) | Political | English | — | ✓ | Persuasion intents | 1124 |
| MultiOff Dataset | 2016 U.S.Pre. Ele. | English | ✓ | ✓ | Offensive | 743 |
| Hateful meme (Kiela et al., 2020b) | Open | English | ✓ | ✓ | Offensive | 10K |
| Harmful meme (Sharma et al., 2022) | Open | English | ✓ | ✓ | Offensive | 3.5K |
| Memotion (Sharma et al., 2020) | Open | English | ✓ | ✓ | Offensive | 7K |
| MAMI (Fersini et al., 2022) | Misogynous | English | ✓ | ✓ | Offensive | 10K |
| MUTE (Hossain et al., 2022) | Open | CM Eng-Ben | ✓ | ✓ | Offensive | 4K |
| Persuasive_memes (Ours) | Multi-domain | Hinglish | ✓ | ✓ | Persuasion/Effect/Technique | 6K |

Table 2: Comparison of our dataset with some existing dataset. Here, *2016 U.S. Pre. Ele.*: U.S.Presidential Election, *CM Eng-Ben:* Code-Mixed English-Bengali, *Hinglish:* Code-Mixed Hindi-English

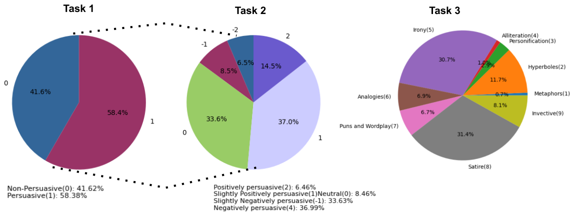


Figure 5: Class distribution of each label of *Persuasive_meme* data.

### 3.2.3 Phase 3: Final Annotation

After phase 3.2.2, the training process was completed. Now, in this phase, we started the final annotation process. We asked the annotators to annotate a given meme with the correct label of each layer as given in the annotation guidelines. After confirming the validity of the meme, we proceed toward the consolidation phase of the annotation.

### 3.2.4 Phase 4: Consolidation

Phase 3.2.4 is the process in which the annotations from phase 3.2.3 are consolidated. This step was critical for maintaining quality, as well as providing additional training for the entire team, which we found really beneficial. In the case of disagreements, we solved them by agreeing on a common point after a lot of discussions. At the end of this phase, we finally obtained the IAA score of 0.7197, 0.89380, and 0.69721, respectively, for all three tasks, which is interpreted as substantial agreement.

[4] github.com/tesseract-ocr/tesseract

| Task 1 | | Task 2 | | | | |
|---|---|---|---|---|---|---|
| **Persuasive** | **Non-persuasive** | **-2: Strongly Disagree** | **-1: Disagree** | **0: Neutral** | **1: Agree** | **2: Strongly Agree** |
| Mitro | Friend | Chutiya | Ungli | Shadi | School | Bhai |
| Pappu | Shadi | Pappu | Corona | Samaj | achha | khush |
| Corona | Mobile | Chaukidar | Modi | ghar | raat | ghar |
| Chaukidar | Party | Maar | election | Whatsapp | Soch | bachpan |
| Allah | School | Nichi | ladkiyaan | log | Aao | cricket |

Table 3: Top-5 most frequent words per class using Tf-Idf

### 3.3 Annotation guidelines

Based on the context of memes, annotators have annotated each meme with three labels: (i) Level 1: Persuasive/non-persuasive,(ii) Level 2: (a) -2: Strongly Disagree (Highly Negative Persuasion), (b) -1: Disagree (Moderately Negative Persuasion), (c) 0: Neutral (Neutral Persuasion), (d) 1: Agree (Moderately Positive Persuasion), and (e) 2: Strongly Agree (Highly Positive Persuasion), (iii) Level 3: Persuasion Techniques/ Rhetorical Devices (multi-label), i.e., (a) Metaphors, (b) Analogies, (c) Hyperboles, (d) Irony, (e) Alliteration, (f) Personification, (g) Puns and wordplay, (i) Invective, and (j) Satire for each meme.

### 3.4 Annotation for identification of persuasive memes

Any meme will be annotated as *Persuasive memes* if it aims to influence or persuade individuals' attitudes, behaviors, or beliefs (Wei et al., 2016; Nee and Maio, 2019). These memes typically employ various persuasive techniques such as metaphors, analogies, humor, satire, or manipulation of information to convey a particular message or agenda. (Refer to samples (b-f) in Figure 1).
On the other hand, *Non-persuasive* memes refer to memes that do not have a specific persuasive intent. They are often created for entertainment purposes, to share jokes, and memes related to popular culture, or to simply evoke laughter without attempting to change opinions or promote a specific viewpoint. These memes may focus on humor, irony, or relatability without a deliberate persuasive agenda (Refer to sample (a) in Figure 1).

#### 3.4.1 Annotation for the effectiveness of persuasive meme

The measurement of effectiveness in persuasive interventions is crucial, yet it can be challenging to directly measure actual persuasiveness (Thomas et al., 2019). To measure the persuasiveness of memes, we inlined our work with previous research (Kaptein et al., 2009; Busch et al., 2016; Anagnostopoulou et al., 2017; Oduor and Oinas-Kukkonen, 2017) and assess the perceived persuasiveness rather than directly measuring the persuasive impact. To achieve this, we employed three scaling items: *Motivational*, *Appropriateness*, and *Effectiveness* (Thomas et al., 2017), to evaluate the impact of memes on society. To quantify the persuasive nature of memes based on these items, we employed the Likert scale (Likert, 1932), a known technique in the field of persuasion. This scale allowed annotators to rate their agreement or disagreement with statements, enabling us to quantify the level of persuasion perceived by the memes.

- **-2: Strongly Disagree (Highly Negative Persuasion):** Based on the evaluation of the above scaling items of persuasive memes, the annotators strongly disagree that these memes have a positive influence. They perceive such memes as highly negative and firmly believe that memes exert a strong negative impact on the target audience due to their use of aggressive tactics or manipulative strategies. (c.f. sample (a) in Figure 1).
- **-1: Disagree (Moderately Negative Persuasion):** Evaluating such persuasive memes, the annotators disagree that these memes have a positive impact. They find the memes moderately negative in their approach, as they employ fear appeals, problem identification, or social disapproval to discourage certain behaviors or beliefs. However, the annotators do not perceive these memes as excessively harmful or manipulative in nature (c.f. sample (b) in Figure 1).
- **0: Neutral (Neutral Persuasion):** Annotators hold a neutral perception of persuasive memes, neither strongly positive nor negative. They perceive the memes as neither highly impactful nor devoid of influence. They view them as presenting information or arguments without a distinct positive or negative leaning (c.f. sample (c) in Figure 1).
- **1: Agree (Moderately Positive Persuasion):** Evaluating the above scaling items aspects of persuasive memes, the annotators agree that these memes have a moderately positive influence. They find the memes to be encouraging, educational, or empowering. However, it is important to acknowledge that while these moderately positive memes hold value and impact, they may not reach the same level of overwhelming positivity as those in the highly

positive category (2) (c.f. sample (d) in Figure 1).

- **2: Strongly Agree (Highly Positive Persuasion):** The annotators strongly agree that persuasive memes are remarkably positive, serving as a source of inspiration and effectively motivating individuals towards beneficial actions, attitudes, or beliefs. They firmly believe that these memes have a substantial positive impact on the target audience. (c.f. sample (e) and (f) in Figure 1).

### 3.4.2 Annotation for the use of persuasion techniques

For this task, the annotation process requires annotators to thoroughly analyze the rhetorical strategies employed in the creation of persuasive memes and annotate the memes accordingly. It involves identifying and labeling all relevant rhetorical devices in a multi-label scenario, as they are instrumental in enhancing the persuasive nature of the meme. By annotating the rhetorical devices, we gain a deeper understanding of the persuasive techniques utilized (Burgers et al., 2016) and their contribution to the overall persuasive impact of the meme.

- **Metaphors (1):** Comparisons that highlight similarities between two things are used to make complex ideas more accessible, relatable, and persuasive in memes.
- **Analogies (2):** Comparisons between two different things to explain a concept or make a point, clarifying abstract ideas and enhancing persuasiveness in memes.
- **Hyperboles (3):** Exaggerated statements or claims that emphasize a point create humor, or evoke strong emotions in memes.
- **Irony (4):** The use of words to convey a meaning opposite to their literal interpretation, often employed in memes to create humor or emphasize a point.
- **Alliteration (5):** Repetition of consonant sounds at the beginning of words in close proximity, creating a memorable or catchy effect, making the meme more persuasive.
- **Personification (6):** Attribution of human qualities to non-human objects or abstract concepts in memes, making ideas more relatable and engaging.
- **Puns and wordplay (7):** Clever manipulation of language to create humor, surprise, or multiple meanings in memes, making them more memorable and persuasive.
- **Invective (8):** The use of abusive or strongly critical language to attack or condemn a person, group, or idea, aiming to provoke strong negative emotions and express intense dislike in memes.
- **Satire (9):** It is form of literary or artistic expression that seeks to bring attention to flaws, vices, or follies with the aim of provoking change or encouraging reflection.

| | | Model | Modality | | Task 1 | | Task 2 | | Task 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | T | I | Acc ↑ | F1 ↑ | Acc ↑ | F1 ↑ | Acc ↑ | F1 ↑ | H-loss ↓ |
| *U. Baselines* | | L_FT | ✓ | | 40.42 | 38.62 | 56.94 | 23.72 | 48.63 | 25.07 | 28.04 |
| | | L_Char | ✓ | | 50.63 | 49.05 | 63.37 | 22.54 | 44.59 | 23.56 | 27.93 |
| | *Pre-trained* | m-BERT | ✓ | | 65.92 | 64.56 | 59.11 | 23.65 | 46.78 | 24.69 | 20.89 |
| | | LaBSE | ✓ | | 62.34 | 59.72 | 64.91 | 28.68 | 45.88 | 29.37 | 25.73 |
| | | Muril | ✓ | | 67.12 | 65.91 | 67.11 | 27.96 | 47.42 | 29.89 | 24.13 |
| | | Indic BERT | ✓ | | 63.75 | 60.79 | 66.30 | 29.86 | 45.73 | 31.8 | 20.64 |
| | | VGG-19 | | ✓ | 58.16 | 49.43 | 68.72 | 25.96 | 43.31 | 15.06 | 25.63 |
| | | ResNet | | ✓ | 57.49 | 50.71 | 61.02 | 28.73 | 41.32 | 18.66 | 24.53 |
| | | ViT | | ✓ | 56.02 | 51.03 | 63.37 | 32.54 | 40.81 | 23.53 | 21.72 |
| *M. Baselines* | *Early Fusion* | L_Char+VGG | ✓ | ✓ | 53.41 | 49.81 | 65.75 | 27.38 | 42.62 | 29.82 | 18.15 |
| | | L_FT+VGG | ✓ | ✓ | 42.15 | 48.71 | 64.76 | 31.38 | 44.59 | 23.56 | 19.23 |
| | | mBERT+ViT | ✓ | ✓ | 70.33 | 68.83 | 61.03 | 31.81 | 51.63 | 30.93 | 15.23 |
| | | LaBSE++ViT | ✓ | ✓ | 63.51 | 60.06 | 64.52 | 35.75 | 41.74 | 21.50 | 12.84 |
| | | Muril+ViT | ✓ | ✓ | 67.12 | 65.91 | 60.04 | 39.93 | 44.75 | 31.83 | 13.55 |
| | | Indic BERT+ViT | ✓ | ✓ | 65.22 | 61.46 | 59.76 | 38.76 | 42.95 | 32.02 | 11.52 |
| | *Pre-trained* | LXMERT | ✓ | ✓ | 68.45 | 59.19 | 59.43 | 42.03 | 50.28 | 31.52 | 11.17 |
| | | VisualBERT | ✓ | ✓ | 67.32 | 67.03 | **58.75** | 43.05 | **60.63** | **33.04** | 10.11 |
| | | mCLIP | ✓ | ✓ | 72.12 | 66.34 | 41.39 | 42.95 | 58.31 | 34.73 | 11.98 |
| | | BLIP | ✓ | ✓ | 70.26 | 64.04 | 49.42 | 41.06 | 55.38 | 33.38 | 12.74 |
| | | ALBEF | ✓ | ✓ | 68.58 | 62.72 | 52.79 | 44.72 | 51.75 | 30.08 | 12.36 |
| | | M3P | ✓ | ✓ | 75.63 | **72.98** | 57.23 | **45.67** | 58.62 | 32.58 | 10.81 |

Table 4: Results for Task 1, Task 2, and Task 3 of *baseline* models. *T*: Text, *I*: Image, *Task 1*: Persuasive/Non-Persuasive, *Task 2*: Effect of Persuasiveness, *Task 3:* Use of persuasion techniques, *Acc*: Accuracy, *F1*: macro-F1 score, and *H-loss*: Hamming Loss

## 3.5 Dataset statistics and comparison with existing datasets

Our dataset, *Persuasive_memes*, has a total of 6,000 annotated memes, which provides a substantial resource for studying persuasive memes (c.f. Figure 5). It provides several unique advantages and distinct features compared to existing datasets (c.f Table 2). It covers a broader range of domains, including political, religious, racist, and sexist themes, enabling a comprehensive analysis of persuasive communication. Unlike other datasets, Persuasive_memes is code-mixed, incorporating both Hindi and English languages and capturing cultural nuances. It is also multimodal, featuring both textual and visual components for a comprehensive examination of persuasive techniques.

## 3.6 Lexical Analysis of the dataset

Table 3 shows the most frequent words for Task 1 (Persuasive/Non-persuasive) and Task 2 (-2: Strongly Disagree to 2: Strongly Agree). In Task 1, the persuasive class includes words related to politics, social issues, and religion, while the non-persuasive class focuses on everyday topics, relationships, and school. For Task 2, strongly dis-

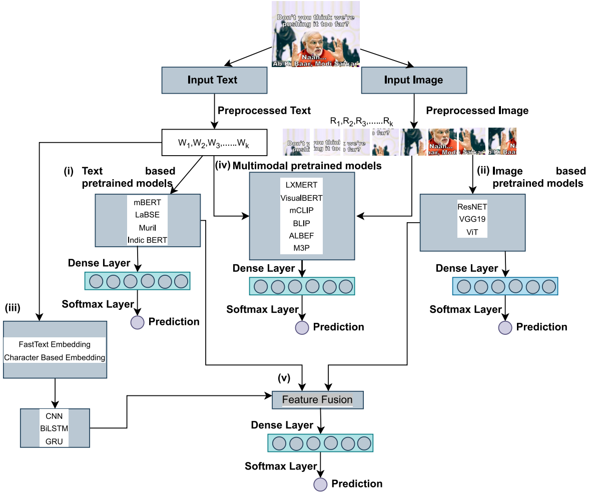


Figure 6: A comprehensive framework for our proposed tasks, incorporating five different approaches: (i) utilizing transformer-based pre-trained models for extracting textual features, (ii) employing visual pre-trained models for extracting visual features, (iii) utilizing pre-trained embeddings in conjunction with RNNs and CNNs for sentence representation and feature extraction, (iv) leveraging pre-trained multimodal models for extracting multimodal features, and (v) employing early fusion techniques to obtain a multimodal representation. These diverse representations are then passed through a dense layer and a softmax layer for the final class prediction.

agree memes exhibit negative sentiment and criticism, disagree memes involve political and societal disagreement, neutral memes revolve around daily life and social norms, agree memes align with popular opinions and thoughts, and strongly agree memes evoke positive emotions and shared interests. Overall, Task 1 involves persuasive/non-persuasive memes, and Task 2 explores the range of agreement and disagreement on various topics.

# 4 Methodology

## 4.1 Task Definition

In this paper, we aim at solving three tasks individually using a deep-learning framework: (i) **t1:** Persuasive meme detection, (ii) **t2:** Identification of the effectiveness of persuasive meme, and (iii) **t3:** Identification of persuasion techniques. Let every meme $S_i \in$ {T, V} is a set with text $T^i = (t_{i_1}, t_{i_2}, ...., t_{i_k})$, and image $I^i$ with the shape (224,224,3) in RGB pattern. In the $i_{th}$ meme, k refers to the total number of words in the textual part of the meme. Our goal is to predict the correct label of each task, i.e., $\hat{y}_{t1} \subseteq${persuasive, non-persuasive}, $\hat{y}_{t2} \subseteq${-2, -1, 0, 1, 2} and $\hat{y}_{t3} \subseteq${1, 2, 3, 4, 5, 6, 7, 8, 9} for each $S_i$. The purpose of this task is to maximize the value of the following function:

$$\underset{\theta_{tk}}{\operatorname{argmax}} \left( \Pi_{i=0}^{n} \Pi_{j=0}^{k} P\left( \hat{y}_{tk}^{i} \mid S^{i}; \theta_{tk} \right) \right) \tag{1}$$

where $tk$ is the current task in {t1, t2, t3}, $S^i$ is the current meme, P is the log-likelihood, and $\theta$ is the model parameter which is needed to be optimized.

# 5 Baseline Models

## 5.1 Unimodal Systems

For the text-based baseline model, we implement *LSTM with FastText embedding* (Bojanowski et al., 2016)(L_FT), *LSTM with Character level Encoding* (L_Char), m-BERT(Pires et al., 2019), LaBSE (Feng et al., 2020), Muril(Khanuja et al., 2021), Indic_BERT (Kakwani et al., 2020), VGG-19 (Simonyan and Zisserman, 2015), ResNet (He et al., 2015), ViT (Dosovitskiy et al., 2020).

## 5.2 Multimodal Systems

**Early Fusion:** For this category, we extract textual and visual features from different pre-trained models and then applied early fusion to get a multimodal representation. By doing so, we develop the following baseline models: L_Char+VGG, L_FT+VGG, mBERT+ViT, LaBSE++ViT, Muril+ViT and Indic BERT+ViT
**Pre-trained Models:** For the pre-trained multimodal system, we used the following pre-trained models to extract the multimodal features: *LXMERT* (Tan and Bansal, 2019), *VisualBERT* (Li et al., 2019). Further, we used another three multimodal feature extractors (*mCLIP*(Radford et al., 2021), *BLIP*(Li et al., 2022), and *ALBEF*(Li et al., 2021), M3P(Ni et al., 2020)). Each of their features is passed through a projection layer to make the final predictions for Task 1, Task 2, and Task 3.

# 6 Experimental setups

We evaluate our proposed architecture on our curated dataset. Details of each task's train and test set distribution are given in the Appendix Table 5 and 6. The optimal hyperparameters for our model are found using grid search. we have used Pytorch Lightning[5] framework for the implementation. We use Adam optimizer with weight decay (Kingma and Ba, 2015) with a learning rate of 3e-5 for all the models. We train the model for 60 epochs with 64 batch sizes and early-stopping callback. We

[5] https://www.pytorchlightning.ai/

consistently use a 32 batch size while training with a fixed random seen of 123. All the models are trained for 20 epochs, and we take the last checkpoint to evaluate the baselines. A single NVIDIA Tesla GPU is used to conduct the experiments.

## 7 Results and Analysis

### 7.1 Model results and comparison

In this section, we show the results that outline the comparison between several baseline models. For evaluation of our tasks in Table 4, we use the F1 score (F1-score), accuracy (A), and hamming loss (H-loss) as the preferred metrics. The results from Table 4 provide valuable insights into the performance of different models in the three persuasion tasks. In Task 1, the baseline models L_FT and L_Char have low accuracy scores, while pre-trained models like m-BERT, LaBSE, Muril, and Indic BERT achieved higher accuracy scores, indicating their superior performance in classifying memes as persuasive or non-persuasive due to their ability to capture language patterns and contextual information. In Task 2, baseline models had limited performance in identifying the effect of persuasiveness. However, the pre-trained models VisualBERT and M3P achieved higher accuracy scores of 58.75% and 60.63%, respectively. These results indicate that these models were effective in capturing the nuanced effects conveyed by persuasive memes, outperforming the baseline models in understanding the impact of persuasiveness. For Task 3, which focused on identifying persuasion techniques, the baseline models showed modest performance. Among the pre-trained models, M3P achieved the highest accuracy score of 58.62%. This suggests that M3P successfully captured the distinctive persuasion techniques used in the memes, highlighting the importance of leveraging pre-trained models for identifying specific persuasive strategies. Overall, the M3P model, with multimodal features, outperforms the baseline models in all three persuasion tasks, emphasizing the importance of analyzing both textual and visual aspects of persuasive memes, and its success can be attributed to capturing contextual information, linguistic nuances and visual cues essential for understanding persuasive impact.

### 7.2 Modality Importance

In this section, we highlight the significance of incorporating both textual and visual modalities in the identification of persuasive memes. By leveraging both modalities, we can access complementary information that enhances our ability to discern the intent and meaning of a meme. In Figure 7(i) (a) in the Appendix, we illustrate how relying solely on the textual modality falls short of understanding the persuasiveness of the meme. However, when we incorporate multimodal information, the M3P model gains access to hidden cues of persuasiveness, leading to accurate labeling for Task 1. Similarly, in the case of sample (b) in Appendix Figure 7 (i), relying solely on the visual modality proves insufficient in capturing the hidden cues of the meme. But once again, with the inclusion of multimodal information, the model successfully identifies the correct label for Task 1.

### 7.3 Explainability and Diagnostics

After training, our M3P model utilizes contextual information within the memes to justify its predictions. Using the LIME (Dieber and Kirrane, 2020) technique; we provide locally interpretable and human-understandable explanations for our model's predictions. In the first meme (In Appendix Figure 8), the M3P model correctly predicts the persuasive and negatively persuasive labels based on specific super-pixels corresponding to the person's face and words like "Mitro{Friends}" and "Fek{Throw}" in the text. These visual and textual elements convey persuasive intent and negative connotations. Similarly, in the second meme (In Appendix Figure 8), the M3P model accurately predicts the meme as both persuasive and negatively persuasive, with certain super-pixels representing the person's face and phrases, such as "Dilo me apni betabiyaan" (restlessness in your hearts) and "4-5 backlogs lekar chal rhe ho to Engineer ho tum" (if you are moving forward with 4-5 backlogs, then you are an Engineer) playing a significant role. These findings demonstrate how the M3P model effectively incorporates visual and textual cues to make precise predictions based on the persuasive and negative characteristics of the memes.

## 8 Error Analysis

In our analysis of the error prediction of our top-performing model (M3P), we identify categories contributing to misclassification and analyze the reasons for these errors:

**(i) Lack of contextual knowledge:** The first category relates to cases where the textual and visual components do not provide enough background or domain-related information. As a result, the model considers these memes as non-persuasive, despite their superficial persuasive nature, simply because of the absence of contextual/domain-related knowledge (c.f. sample (a) and (b) in the Appendix Figure 7).

**(ii) Lack of common sense knowledge:** The second category involves a lack of common sense knowledge, where our model struggles to intuitively reason about everyday situations and events, resulting in incorrect predictions of the persuasive class. (c.f. sample (c) in the Appendix Figure 7).

**(iii) Due to the longer sentence length:** In the third category, misclassifications occur when the sentence length is longer than the average, as illustrated in sample (d) in the Appendix Figure 7.

## 9 Conclusion

In conclusion, while internet memes have become ubiquitous in the digital landscape, their persuasive effectiveness remains a significant, yet underexplored area of research. This study aims to bridge this gap by conducting a comprehensive evaluation of persuasive memes, offering valuable insights into their impact on society and their potential to shape public opinion. By filling this void in the current literature, the research presented here seeks to advance our understanding of persuasive communication in the digital age, fostering informed discussions and facilitating responsible meme usage.

## Limitation

In Section 8, we discussed the limitations of our proposed work. Our baseline models struggle with the detection of subtle or implicit persuasive elements in memes. Some persuasive techniques may be context-dependent or rely on cultural references that are difficult for the models to capture accurately. By analyzing these errors, we gain valuable insights into the limitations and challenges faced by our model, which can guide future improvements in persuasive meme identification.

## Ethics and Broader Impact

**Individual Privacy** The resources we created for this study were derived from publicly available memes, and we strictly adhered to the restrictions on data usage to avoid any infringement of copyright laws. Furthermore, our study was evaluated and approved by our Institutional Review Board (IRB). We plan to make our code and data accessible for research purposes, subject to appropriate data agreement procedures, upon acceptance of our study.

To maintain the anonymity of any individual, we replaced the actual name with Person-XYZ throughout the paper. In addition, we also tried to anonymize the known faces presented in the visual part of the meme by masking them. We have masked these faces only to maintain the anonymity issues in the paper. During the implementation, we used the original image.

**Biases** Detecting and removing political and religious biases is an extensive research area. However, previous annotation studies show that we cannot correctly remove bias and subjectivity from the annotation process despite having some form of annotation scheme. However, any biases detected in our dataset are unintentional, and we have no intention of harming any individual or group. We ensure that our data collection is generated equally and comparably in order to answer any political and religious bias queries. Furthermore, we ensure that the topic includes various issues relevant to the Indian context over the last seven years by using a keyword-based data-gathering technique. Moreover, we made sure that the terms included were inclusive of all the conceivable politicians, political organizations, young politicians, extreme groups, and religions and were not prejudiced against any one group. Based on previous work done by (Davidson et al., 2019) to remove biases from the dataset during annotation, in our dataset, annotators were strictly instructed not to make decisions based on what they believe but on what the social media user wants to transmit through that meme.

**Misuse Potential** We suggest that researchers be aware that our dataset might be abused to filter the memes based on prejudices that may or may not be connected to demographics or other textual information. To prevent this from happening, human intervention with moderation would be essential.

**Intended Use** Our dataset is presented to encourage research into studying persuasive memes on the internet. We believe that it represents a valuable resource when used appropriately.

## Acknowledgements

The research reported in this paper is an outcome of the project “**HELIOS: Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System,**" sponsored by Wipro AI Labs, India.

## References

Evangelia Anagnostopoulou, Babis Magoutas, Efthimios Bothos, Johann Schrammel, Rita Orji, and Gregoris Mentzas. 2017. Exploring the links between persuasion, personality and mobility types in personalized mobility applications. pages 107–118.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, Doha, Qatar. Association for Computational Linguistics.

Dibyanayan Bandyopadhyay, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and Vinutha BN. 2023. A knowledge infusion based multitasking system for sarcasm detection in meme. In *Advances in Information Retrieval*, pages 101–117, Cham. Springer Nature Switzerland.

Kevin Barnhurst and Catherine Steele. 1997. Image-bite newsthe visual coverage of elections on u.s. television, 1968-1992. *Harvard International Journal of Press-politics - HARV INT J PRESS-POLIT*, 2:40–58.

Morris Bernadt and J Emmanuel. 1993. Diagnostic agreement in psychiatry. *The British journal of psychiatry : the journal of mental science*, 163:549–50.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Erik P. Bucy and Samuel D. Bradley. 2004. Presidential expressions and viewer emotion: Counterempathic responses to televised leader displays. *Social Science Information*, 43(1):59–94.

Cody Buntain and Jennifer Golbeck. 2014. Identifying social roles in reddit using network structure. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, page 615–620, New York, NY, USA. Association for Computing Machinery.

Christian Burgers, Elly A. Konijn, and Gerard J. Steen. 2016. Figurative Framing: Shaping Public Discourse Through Metaphor, Hyperbole, and Irony. *Communication Theory*, 26(4):410–430.

Marc Busch, Sameer Patil, Georg Regal, Christina Hochleitner, and Manfred Tscheligi. 2016. Persuasive information security: Techniques to help employees protect organizational information security. In *Proceedings of the 11th International Conference on Persuasive Technology - Volume 9638*, PERSUASIVE 2016, page 339–351, Berlin, Heidelberg. Springer-Verlag.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Jurgen Dieber and S. Kirrane. 2020. Why model why? assessing the strengths and limitations of lime. *ArXiv*, abs/2012.00093.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Walter F. Edwards. 1995. *Language in Society*, 24(2):302–305.

Shahira Fahmy and Ibrahim Omneya. 2021. No memes no! digital persuasion in the metoo era. *International Journal of Communication*, 15:2942–2967.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian*

*Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

Maria Grabe and Erik Bucy. 2010. Image bite politics: News and the visual framing of elections. *Image Bite Politics: News and the Visual Framing of Elections*, pages 1–332.

Marco Guerini, Gözde Özbal, and Carlo Strapparava. 2015. Echoes of persuasion: The effect of euphony in persuasive communication. *CoRR*, abs/1508.05817.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2026–2031, Lisbon, Portugal. Association for Computational Linguistics.

Jungseock Joo, Weixin Li, Francis F. Steen, and Song-Chun Zhu. 2014. Visual persuasion: Inferring communicative intents of images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 216–223.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Satyajit Kamble and Aditya Joshi. 2018. Hate speech detection from code-mixed hindi-english tweets using deep learning models. *CoRR*, abs/1811.05145.

Maurits Kaptein, Panos Markopoulos, Boris de Ruyter, and Emile Aarts. 2009. Can you be persuaded? individual differences in susceptibility to persuasion. In *Human-Computer Interaction – INTERACT 2009*, pages 115–118, Berlin, Heidelberg. Springer Berlin Heidelberg.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020a. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020b. The hateful memes challenge: Detecting hate speech in multimodal memes. *CoRR*, abs/2005.04790.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 26–35, Online. Association for Computational Linguistics.

klaus krippendorff. 2011. Computing krippendorff's alpha-reliability.

Gitanjali Kumari, Amitava Das, and Asif Ekbal. 2021. Co-attention based multimodal factorized bilinear pooling for Internet memes analysis. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 261–270, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. In *Arxiv*.

Rensis Likert. 1932. *A technique for the measurement of attitudes / by Rensis Likert.* Archives of psychology ; no. 140. [s.n.], New York.

Roger D. Masters, Denis G. Sullivan, John T. Lanzetta, Gregory J McHugo, and Basil G. Englis. 1986. The facial displays of leaders: Toward an ethology of human politics. *Journal of Social and Biological Structures*, 9:319–343.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in

Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Rebecca Coates Nee and Mariana De Maio. 2019. A 'presidential look'? an analysis of gender framing in 2016 persuasive memes of hillary clinton. *Journal of Broadcasting & Electronic Media*, 63(2):304–321.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, and Nan Duan. 2020. M3p: Learning universal representations via multitask multilingual multimodal pre-training.

Michael Oduor and Harri Oinas-Kukkonen. 2017. Commitment devices as behavior change support systems: A study of users' perceived competence and continuance intention. In *Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors*, pages 201–213, Cham. Springer International Publishing.

Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1114–1125, New York, NY, USA. Association for Computing Machinery.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

Jens Seiffert-Brockmann and Trevor Diehl. 2016. The power of memes: The digital discourse of the obama hope meme.

Jens Seiffert-Brockmann, Trevor Diehl, and Leonhard Dobusch. 2018. Memes as games: The evolution of a digital discourse online. *New Media Society*, 20:2862– 2879.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. DISARM: Detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *CoRR*, abs/1602.01103.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

5100–5111, Hong Kong, China. Association for Computational Linguistics.

Tiancheng Tang, Xinhuai Tang, and Tianyi Yuan. 2020. Fine-tuning bert for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, 8:193248–193256.

Rosemary Josekutty Thomas, Judith Masthoff, and Nir Oren. 2017. Adapting healthy eating messages to personality. In *Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 119–132. Springer-Verlag. 12th International Conference on Persuasive Technology, PERSUASIVE 2017 ; Conference date: 04-04-2017 Through 06-04-2017.

Rosemary Josekutty Thomas, Judith Masthoff, and Nir Oren. 2019. Can i influence you? development of a scale to measure perceived persuasiveness and two studies showing the use of the scale. *Frontiers in Artificial Intelligence*, 2.

Amine Trabelsi and Osmar R. Zaïane. 2014. Finding arguing expressions of divergent viewpoints in online debates. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 35–43, Gothenburg, Sweden. Association for Computational Linguistics.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany. Association for Computational Linguistics.

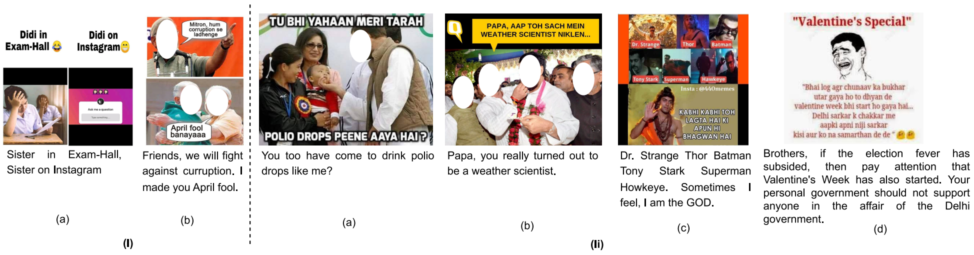
# A Appendix

| Split | #Memes | Task 1 |  | Task 2 |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  |  | Persuasive | Non-persuasive | Positively persuasive | Slightly Positively persuasive | Neutral | Slightly Negatively persuasive | Negatively persuasive |
| **Train** | 4800 | 2818 | 1982 | 99 | 75 | 978 | 1261 | 405 |
| **Test** | 1200 | 717 | 483 | 31 | 45 | 226 | 340 | 106 |

Table 5: Class wise data distribution of *Persuasive_meme* dataset for Task 1 and Task 2

| **Split** | **#Memes** | **Task 3** |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  |  | **Metaphors** | **Hyperboles** | **Personification** | **Alliteration** | **Irony** | **Analogies** | **Puns_and_wordplay** | **Satire** | **Invective** |
| **Train** | 4800 | 29 | 466 | 122 | 42 | 1238 | 275 | 263 | 1268 | 324 |
| **Test** | 1200 | 5 | 118 | 21 | 7 | 297 | 72 | 73 | 306 | 83 |

Table 6: Class wise data distribution of *Persuasive_meme* dataset for Task 3


Didi in Exam-Hall 😂
Didi on Instagram😁
Ask me a question
Mitron, hum corruption se ladhenge
April fool banayaaa
TU BHI YAHAAN MERI TARAH
POLIO DROPS PEENE AAYA HAI ?
PAPA, AAP TOH SACH MEIN WEATHER SCIENTIST NIKLEN...
Dr. Strange
Thor
Batman
Tony Stark
Superman
Hawkeye
Insta : @440memes
KABHI KABHI TOH LAGTA HAI KI APUN HI BHAGWAN HAI
"Valentine's Special"
"Bhai log agr chunaav ka bukhar utar gaya ho to dhyan de valentine week bhi start ho gaya hai... Delhi sarkar k chakkar me aapki apni niji sarkar kisi aur ko na samarthan de de"

Sister in Exam-Hall, Sister on Instagram

(a)

Friends, we will fight against curruption. I made you April fool.

(b)

**(I)**

You too have come to drink polio drops like me?

(a)

Papa, you really turned out to be a weather scientist.

(b)

Dr. Strange Thor Batman Tony Stark Superman Howkeye. Sometimes I feel, I am the GOD.

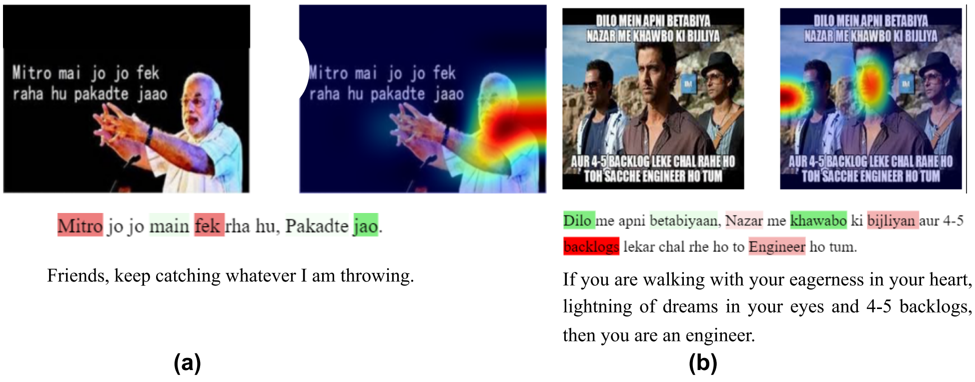
(c)

Brothers, if the election fever has subsided, then pay attention that Valentine's Week has also started. Your personal government should not support anyone in the affair of the Delhi government.

(d)

**(Ii)**

Figure 7: **(i) Modality Importance:** Test cases where unimodal systems (either text-only model or image-only model) fail to correctly predict the persuasion class whereas the multimodal system effectively predicted the persuasion class. **(ii) Error Analysis:** Mis-classification by the best performing M3P model


Mitro mai jo jo fek raha hu pakadte jaao
DILO MEIN APNI BETABIYA NAZAR ME KHAWBO KI BIJLIYA
AUR 4-5 BACKLOG LEKE CHAL RAHE HO TOH SACCHE ENGINEER HO TUM

Mitro jo jo main fek rha hu, Pakadte jao.

Friends, keep catching whatever I am throwing.

**(a)**

Dilo me apni betabiyaan, Nazar me khawabo ki bijliyan aur 4-5 backlogs lekar chal rhe ho to Engineer ho tum.

If you are walking with your eagerness in your heart, lightning of dreams in your eyes and 4-5 backlogs, then you are an engineer.

**(b)**

Figure 8: Visualization by LIME (Ribeiro et al., 2016) for best performing M3P model.