Deceptive Humor Annotation Guidelines

Task Summary and Motivation

Deceptive Humor refers to humorous content that intentionally embeds fabricated or misleading claims in a satirical or playful manner. While this content may appear light-hearted, it subtly spreads misinformation by reframing false claims as jokes. This unique blend of humor and deception makes detection challenging and socially impactful.

Because of the difficulty in identifying large-scale real-world deceptive humor, we leverage the pre-trained knowledge of state-of-the-art language models to generate humorous comments grounded in verified fake claims. Although these models possess vast linguistic and cultural knowledge, they are not infallible; relying solely on generative outputs can introduce errors, biases, or inconsistencies.

To address this, we conducted a Human-in-the-Loop process to verify the linguistic and cultural quality of the generated comments. However, for the final dataset, it is essential to assess how closely human annotators' judgments align with the model-generated labels. In this human evaluation phase, annotators are asked to label the given comments independently, without knowing the labels originally assigned by the model. This enables us to measure alignment, validate label reliability, and ensure the dataset meets high annotation standards.

Annotating deceptive humor is inherently challenging because such comments often blend subtle satirical cues, cultural context, and fabricated claims. Unlike ordinary humor, deceptive humor requires understanding both surface-level humor and the underlying deceptive intent. Therefore, annotation must be conducted carefully, consistently, and contextually, ensuring that the resulting labels are accurate, interpretable, and reproducible.

If at any point the comment seems ambiguous, annotators are encouraged to look up the background of the fake claim on the internet to gain a better understanding of its context before assigning labels.

Annotation Dimensions

Each comment must be annotated along two dimensions:

1. **Satire Intensity (Ordinal Classification)**

This task involves classifying the intensity of satire, which acts as a proxy for how subtly deception is embedded in the comment. Annotators should assign one of three ordinal levels:

- **Low Satire:**
  Subtle and lightly satirical; resembles real-world statements with a mild humorous twist.
- **Moderate Satire:**
  Humor is more evident, using exaggeration or sarcasm while maintaining a balance between reality and absurdity.

- **High Satire:**
  Strongly exaggerated or overtly satirical; uses extreme irony, absurdity, or sharp contrasts with reality.

2. Humor Attribute (Categorical Classification)

Annotators should identify the dominant humor type used in the comment. While some comments may exhibit overlapping traits, choose the single most prominent attribute based on the intent and structure of the comment:

- **Irony**: Intended meaning contrasts sharply with the literal meaning, exposing contradictions or unexpected outcomes.
- **Absurdity**: Relies on exaggeration, illogical scenarios, or unrealistic premises for amusement.
- **Social Commentary**: Critiques, mocks, or highlights societal or cultural issues through satire or reflection.
- **Dark Humor**: Engages with morbid, taboo, or controversial topics in a way that is unsettling yet humorous.
- **Wordplay**: Employs clever linguistic constructions, puns, double meanings, or phonetic playfulness.

Note: If a comment blends multiple humor types (e.g., Irony + Social Commentary), select the dominant attribute that best captures the overall intent.

Annotation Procedure

1. **Read the Comment Carefully**
   - Understand the surface humor and check for cues indicating underlying fabricated claims.
   - If unclear, research the fake claim background briefly.

2. **Assign Satire Intensity**
   - Choose Low, Moderate, or High based on how explicitly humor is used to mask deception.

3. **Assign Humor Attribute**
   - Identify the dominant humor type based on the definitions provided.

4. **Mark Unclear or Invalid Samples**
   - If the comment is incomplete, nonsensical, or does not fit any category, mark it for review and leave a short note explaining the issue.

5. **Save Annotations**
   - Follow the provided annotation spreadsheet instructions to ensure labels are stored correctly.

Quality Assurance

- **Consistency Checks**: Annotators should periodically review a subset of their labels to maintain internal consistency.

- **Calibration Rounds**: Early annotation rounds may include mock annotations where multiple annotators label the same subset to resolve discrepancies and calibrate judgments.

- **Inter-Annotator Agreement**: Metrics like Cohen's/Fleiss' Kappa will be used to measure alignment across annotators and with machine labels.

- **Feedback Loop**: If recurring confusions arise even after exploring the background context of the fake claim through internet sources please discuss these issues with the annotation leads assigned to the project. If the difficulty persists, escalate the case to the faculty supervisor for resolution.

Selection of Final Labels:

Final labels are determined by majority voting. If there is disagreement, the assigned PhD students and the professor will make the final decision on labeling the comment.