Multimodal Sentiment Analysis using Transformers: Final Assignment

Abstract

Information on social media comprises of various modalities such as textual, visual and audio. NLP and Computer Vision communities often leverage only one prominent modality in isolation to study social media. However, the computational processing of Internet memes needs a hybrid approach. The growing ubiquity of Internet memes on social media platforms such as Facebook, Instagram, and Twitter further suggests that we can not ignore such multimodal content anymore. To the best of our knowledge, there is not much attention towards meme emotion analysis. The objective of this proposal is to bring the attention of the research community towards the automatic processing of Internet memes. The task Memotion analysis will release 8K annotated memes - with human-annotated tags namely sentiment, and type of humor that is, sarcastic, humorous, or offensive.

The Multimodal Social Media

In the last few years, the growing ubiquity of Internet memes on social media platforms such as Facebook, Instagram, and Twitter has become a topic of immense interest. Memes, one of the most typed English words (Sonnad, 2018) in recent times. Memes are often derived from our prior social and cultural experiences such as TV series or a popular cartoon character (think: One Does Not Simply - a now immensely popular meme taken from the movie Lord of the Rings). These digital constructs are so deeply ingrained in our Internet culture that to understand the opinion of a community, we need to understand the type of memes it shares. (Gal et al., 2016) aptly describes them as performative acts, which involve a conscious decision to either support or reject an ongoing social discourse. Online Hate - A brutal Job: The prevalence of hate speech in online social media is a nightmare and a great societal responsibility for many social media companies. However, the latest entrant Internet memes (Williams et al., 2016) has doubled the challenge. When malicious users upload something offensive to torment or disturb people, it traditionally has to be seen and flagged by at least one human, either an user or a paid worker. Even today, companies like Facebook and Twitter rely extensively on outside human contractors from start-ups like CrowdFlower, or companies in the Philippines. But with the growing volume of multimodal social media, it is becoming impossible to scale. The detection of offensive content on online social media is an ongoing struggle. OffenseEval (Zampieri et al., 2019) is a shared task which is being organized since the last two years at SemEval. But, detecting an offensive meme is more complex than detecting an offensive text – it involves visual cue and language understanding. This is one of the motivating aspects which encourages us to propose this task. Multimodal Social Media Analysis - The Necessity: Analogous to textual content on social media, memes also need to be analyzed and processed to extract the conveyed message. A few researchers have tried to automate the meme generation (Peirson et al., 2018; Oliveira et al., 2016) process, while a few others tried to extract its inherent sentiment (French, 2017) in the recent past. Nevertheless, a lot more needs to be done to distinguish their finer aspects such as type of humor or offense. We hope Memotion analysis - the task will bring research attention towards the topic and the forum will be the place to continue relevant discussions on the topic among researchers.

Objective

The goal of this project is to design and develop a multimodal sentiment analysis system using **Transformers** for the task of **Memotion Analysis**. This will involve analyzing Internet memes, which are composed of both **textual** and **visual** data, and classifying them based on their sentiment, humor type, and offense level. The project will utilize **state-of-the-art transformer models** to perform **sentiment classification**, **humor classification**, and **offense detection** across multimodal inputs.

Task Breakdown

This project involves three main tasks:

1. **Sentiment Classification**
2. **Humor Classification**
3. **Scales of Semantic Classes** (including Humor, Sarcasm, Offense, Motivation)

Each task has specific input formats, output formats, and evaluation criteria.

Task A: Sentiment Classification

- **Objective**: Classify the sentiment of a given meme as **Positive**, **Negative**, or **Neutral**.
- **Approach**:
    - Use the **text** (caption or embedded text in the meme) and **image** (visual content in the meme) to predict sentiment.
    - The transformer model (e.g., **BERT**, **RoBERTa**, **DistilBERT**) can be used for processing the textual content, while a **Vision Transformer (ViT)** or **CNN-based model** can be used for processing the image.
    - Combine the outputs of both models using a **multimodal fusion** layer.
- **Input**:
    - Text: Caption or embedded textual content of the meme.
    - Image: The image of the meme itself.
- **Output**:
    - Positive (+1)
    - Negative (-1)
    - Neutral (0)
- **Evaluation Criteria**: Macro F1 score.

Task B: Humor Classification

- **Objective**: Identify the type of humor expressed in the meme. Categories include **Sarcastic**, **Humorous**, **Offensive**, or **Other**.
- **Approach**:

    - Use the multimodal approach where both text and image contribute to humor classification.
    - **Transformer models** can be used for the textual component, while **CNN** or **Vision Transformers (ViTs)** can handle the image part.
    - Use a **classification head** after multimodal fusion to predict the humor type.
- **Input**:
    - Text: Meme caption or embedded text.
    - Image: The image of the meme itself.
- **Output**:
    - **Humor**: 0 (Not Humorous) or 1 (Humorous, Funny, Hilarious).
    - **Sarcasm**: 0 (Not Sarcastic) or 1 (Sarcastic, Twisted Meaning, Very Twisted).
    - **Offensive**: 0 (Not Offensive) or 1 (Slight, Very Offensive, Hateful Offensive).
- **Evaluation Criteria**: Macro F1 for each subtask (Humor, Sarcasm, Offense) and average.

Task C: Scales of Semantic Classes

- **Objective**: Quantify the extent to which the meme expresses specific effects such as **Humor**, **Sarcasm**, **Offense**, and **Motivation**.
- **Approach**:
    - Use a **regression head** on top of the multimodal fusion model to predict a scale for each category.
    - The transformer model will be responsible for encoding the text, while a CNN or ViT will process the image.
    - For each category (Humor, Sarcasm, Offense, Motivation), predict values based on the scale provided.
- **Input**:
    - Text: Caption or embedded textual content of the meme.
    - Image: The image of the meme.
- **Output** (on a scale of 0-3):
    - **Humor**: 0 (Not Funny), 1 (Funny), 2 (Very Funny), 3 (Hilarious).
    - **Sarcasm**: 0 (Not Sarcastic), 1 (General), 2 (Twisted Meaning), 3 (Very Twisted).
    - **Offense**: 0 (Not Offensive), 1 (Slight), 2 (Very Offensive), 3 (Hateful Offensive).
    - **Motivation**: 0 (Not Motivational), 1 (Motivational).
- **Evaluation Criteria**: Macro F1 for each subtask and average.

Model Architecture

1. **Textual Processing (Transformer Model)**:
    - Use a **BERT-based model** (or any other transformer architecture like RoBERTa or GPT) for the **textual data**(meme caption). Fine-tune the model to understand the nuances of the meme language (sarcasm, humor, offense).

2. **Visual Processing (Vision Model)**:
    - Use a **CNN** or **Vision Transformer (ViT)** for the **visual data** (the meme image). Fine-tune the model on the meme dataset to learn the visual aspects of humor, sarcasm, and offense.
3. **Multimodal Fusion Layer**:
    - Fuse the output embeddings from the **textual** and **visual** components. This can be done using simple concatenation, cross-attention mechanisms, or other fusion techniques.
4. **Classification/Regression Head**:
    - For **Sentiment Classification**, a classification layer will predict Positive, Negative, or Neutral sentiment.
    - For **Humor Classification**, the model will output the categories for humor, sarcasm, and offense.
    - For **Scales of Semantic Classes**, a regression layer will output the scale values for Humor, Sarcasm, Offense, and Motivation.

Dataset

- **Memotion Dataset**: Use the **dataset** provided for Memotion analysis. This dataset contains meme images along with human-annotated labels for sentiment, humor type, offense level, and motivation.
- **Preprocessing**:
    - **Text Preprocessing**: Tokenize the meme captions and handle special tokens like emojis or slang.
    - **Image Preprocessing**: Resize images to a consistent size, normalize, and apply any augmentation techniques to improve generalization.
- Dataset link: Multimodal_Sentiment_Analysis_FinalAssignment

---

Training and Evaluation

- **Training**:
    - Train the model using the provided dataset. Use **cross-entropy loss** for classification tasks (Sentiment and Humor) and **mean squared error** for regression tasks (Scales of Semantic Classes).
    - Utilize **early stopping** and **checkpointing** to save the best model based on the validation set's performance.
- **Evaluation**:
    - After training, evaluate the model's performance using **Macro F1 score** for each task.
    - Report the individual F1 scores for **Sentiment Classification**, **Humor Classification**, and **Scales of Semantic Classes**.

Deliverables

1. **Code**:
    - Provide the implementation for multimodal sentiment analysis, including the transformer-based model for text and visual data processing, the multimodal fusion technique, and the classification/regression heads.
2. **Trained Models**:
    - Include the trained models (for each task) and checkpoints.
3. **Evaluation Metrics**:
    - Report the evaluation metrics (Macro F1 score) for each task.

Tools and Libraries

- **Transformers** library for pre-trained transformer models.
- **PyTorch** or **TensorFlow/Keras** for model training.
- **OpenCV** or **Pillow** for image preprocessing.
- **Hugging Face Datasets** for accessing the Memotion dataset (if publicly available).
- **Matplotlib/Seaborn** for data visualization and evaluation.

Conclusion

This project aims to push forward the research on **multimodal sentiment analysis**, particularly in the context of **Internet memes**, by leveraging state-of-the-art **transformer models** for both text and visual processing. The Memotion Analysis dataset provides a great opportunity to explore the intricacies of meme culture, sentiment, humor, and offense detection in an automated and scalable manner.

DEADLINE:

Tuesday(4th feb), EOD