Humor Detection in Bengali: Creating and Analyzing Dataset through Social Media in the Bengali Language

Objective:

The proliferation of online social media has led to a vast amount of user generated content. Social media platforms like FB,reddit,x with their character limits, have spawned creative and condensed language to convey messages. This has made the automatic detection of humor a challenging task. Humor detection has been better studied for resource-rich languages given the availability of substantial annotated data. We aim to address this gap by working with Bangla, a low-resource language. We will collect and manually annotate a dataset of Bangla posts, tweets then leveraged this new resource to train and evaluate humor classification models for Bangla social media text.

Methodology:

Dataset:

To construct our dataset, we will collect posts and tweets from platforms such as Facebook, Reddit, and X by searching for relevant tags. Specifically, humorous content will be gathered by querying tags including "humor", "funny", and "bangla jokes." This tagging functionality allows us to filter for social media items that are more likely to contain comedic intent or content. By scraping tagged posts using this methodology, we aim to efficiently build a corpus of humor-related vernacular text. Additional data filtering, processing, and manual annotation steps will further refine this collection into a categorized dataset suited for humor detection research.

Working process:

We have primarily thought of using BERT and its variation, Word2Vec for detecting humor.

Analysis :

To quantify model performance on humor detection, we will report several standard classification evaluation metrics including accuracy, precision, recall, and F1-score. These will allow comprehensive comparison between approaches.