This thesis focuses on the multimodal and multilingual detection of humor in stand-up comedy videos. A novel multilingual dataset was collected, primarily targeting the Russian language, to address the lack of specific multimodal datasets for humor detection in this language. The dataset was obtained from stand-up comedy videos with subtitles sourced from YouTube. The thesis investigates various aspects of the data preparation process, including word-level forced alignment, segmentation, and labeling with laughter detection. Two automatic laughter detection approaches are explored: the peak detection approach, which employs preprocessed voiceless audio and an energy-based peak detection algorithm with clusterization filtering, and the machine learning approach, which utilizes a pretrained model to detect laughter presence and duration. Results indicate that for now the machine learning approach outperforms the peak detection approach in terms of accuracy and generalization, however the peak detection approach is considered promising. Additionally, thesis delves into the unimodal textual and multimodal humor detection on the new dataset. The results demonstrate the ability of neural models to capture humour in both languages even in the textual only setting. While multimodal experiments showed that even in simple models the addition of visual modality improves the results. However, further experiments and research are needed to enhance the laughter detection labeling quality and investigate the influence of different modalities in the multimodal and multilingual approach.