Multimodal and Multilingual Laughter Detection in Stand-Up Comedy Videos

Anna Kuznetsova, Carlo Strapparava
University of Trento, Fondazione Bruno Kessler
Trento, Italy
kuzannagood@gmail.com, strappa@fbk.eu

Abstract

This paper presents the development of a novel multimodal multilingual dataset in Russian and English, with a particular emphasis on the exploration of laughter detection techniques. Data was collected from YouTube stand-up comedy videos with manually annotated subtitles, and our research covers data preparation and laughter labeling. We explore two laughter detection approaches presented in the literature: peak detection using preprocessed voiceless audio with an energy-based algorithm and machine learning approach with pretrained models to identify laughter presence and duration. While the machine learning approach currently outperforms peak detection in accuracy and generalization, the latter shows promise and warrants further study. Additionally, we explore unimodal and multimodal humor detection on the new dataset, showing the effectiveness of neural models in capturing humor in both languages, even with textual data. Multimodal experiments indicate that even basic models benefit from visual data, improving detection results. However, further research is needed to enhance laughter detection labeling quality and fully understand the impact of different modalities in a multimodal and multilingual context.

Keywords: multimodal humor, stand-up, laughter detection

1. Introduction

Automatic detection of humor in natural language is a challenging task due to its complex nature. Yet, it has gained significance as interactions with robots and smart assistants increasingly rely on humor for successful engagement. A review by Kalloniatis and Adamidis (2023) mentions various datasets collected for this task - they are sourced from various platforms, including social media, news articles, and communication platforms. There are also different ways to label the data - usually it is a manual annotation by domain experts. However, this process is time-consuming and expensive, and inter-annotator agreement can vary. Distant supervision, using domain knowledge like hashtags on Twitter, is another approach, often followed by crowd-sourcing to ensure reliable labeling.

In recent years, the NLP community has shown a growing interest in multimodal aspects of humor recognition. This focus has resulted in various models and datasets. Most of the datasets focus on the sitcom genre, especially The Big Bang Theory (TBBT) (Castro et al., 2019; Kayatani et al., 2021; Patro et al., 2021). They are labelled either using laughter detection software or laughter extraction from the audio track. TED talks are another medium to source datasets (Hasan et al., 2019). They have reliable transcriptions and "laughter" markers to label humorous content. Mittal et al. (2021) is the only work focusing on stand-up, they collected show recordings and transcripts from the web and manually segmented them into clips.

It is worth noting that most papers focus on English datasets with scripted humor. Given that humor is heavily influenced by culture, it would be interesting to explore the differences and nuances that may exist when working with different language datasets and more interactive humor domain.

This study's primary focus is on advancing research in multilingual multimodal humor detection, achieved through the creation of a unique dataset sourced from Russian and English stand-up comedy. Stand-up comedy, an underrepresented domain in current datasets, was chosen due to the scarcity of suitable sources in Russian sitcoms - the more popular source of datasets. The research seeks to investigate various approaches to humor annotation using laughter, striving to minimize manual annotation requirements. Additionally, the study aims to evaluate the dataset's potential for multimodal humor detection. The code and dataset is available at `https://github.com/kuzanna2016/multimodal_humour`

2. Dataset collection

The dataset collection process began with a focus on the Russian language, as it offered greater novelty and value for the research. Experiments were primarily conducted on the Russian dataset, followed by a similar pipeline for the English dataset.

2.1. Video collection

For the Russian dataset, we manually compiled a list of YouTube channels featuring stand-up performances in Russian with subtitles. We collected 46

LREC-COLING 2024, pages 11884–11889

Figure 1: Frames from the collected videos expressing different non-verbal laughter cues.

videos from 8 channels (17 hours). Most videos (31) were from a Vladivostok-based stand-up club, primarily featuring male comedians. As for the English dataset the biggest Stand-Up YouTube channel was used resulting in 56 videos with 20 hours of content in total. English dataset had a greater diversity, including performances by female comedians (42%) and comedians of color (39%). The collected videos showcased a diverse range of non-verbal humor expressions, encompassing facial expressions and body gestures (see Figure 1).

2.2. Annotation and preprocessing

To facilitate further experiments on laughter detection, manual annotations were conducted on 5 short videos from each dataset. In the case of the Russian dataset, videos were selected from multiple channels. Utilizing ELAN software (ELAN), we annotated segments of laughter and applause. Notably, the annotation process revealed challenges in Russian subtitles, including incorrect segmentation where a single subtitle phrase encompassed multiple sentences, or sentences were divided between multiple subtitles. Furthermore, certain subtitle time spans incorporated laughter pauses, which deviated from the intended detection of laughter following each phrase.

In contrast, the transcriptions for the English dataset exhibited less noise and maintained a relatively consistent style. These annotations included informative audio descriptions such as "(crowd cheering)" and "(music playing)," along with labeled instances of laughter, such as "(audience laughing)." These annotations present a valuable resource for distant supervised annotation.

To address the wrong subtitle segmentation and be able to use words aligned with frames we forced aligned subtitle words using Montreal Forced Aligner (MFA)[1]. As for segmentation, a novel approach was developed. It combined punctuation-based segmentation with segmentation based on pauses between utterances. The algorithm used full stops followed by capitalized words, punctuation signs, and substantial pauses to segment phrases.

[1] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

3. Laughter detection

To evaluate the detection of laughter, we adopted a pragmatic approach, considering laughter as an event within specific time window after an utterance. Our analysis of annotated videos revealed that laughter frequently follows an utterance, occurring either immediately afterward or slightly within the pause between utterances. Given that, we devised an algorithm to define these laughter windows, bounded by the subsequent subtitle pause or, if the pause is less than 0.2 seconds, by the end of the subtitle plus 0.7 seconds. This method allowed us to label the utterances, marking the presence (1) or absence (0) of laughter within these windows. Subsequently, we employed these labels to evaluate laughter detection using standard binary classification metrics. Validation was performed on the manually annotated subset of videos.

3.1. Peak detection approach

Some works have employed peak detection method for automatically labeling laughter in datasets with high-quality audio from sitcoms, often recorded with two channels (left and right) featuring centered actors' voices. By subtracting the left channel from the right channel, the voice component can be removed, leaving only laughter and music in the remaining audio track. The peak detection process involves applying a simple signal energy-based detector, such as the *auditok* library[2], to identify peaks in the preprocessed audio track. Afterwards, the detected segments can be filtered either manually, using signal postprocessing techniques (Kayatani et al., 2021), or by employing clustering algorithms (Liu et al., 2022).

For the voice removal step, we subtracted the audio channels in the Russian dataset. The majority of audio exhibited unclear results, however, five successful subtractions were observed, with two being part of the annotated section of the dataset (Figure 2). The first video achieved near-perfect separation with only some laughter segments missing from the audio, making it an ideal case for voice separation. The second video left a lot of artefacts in the track, representing a non-ideal case.

Building on the work of Liu et al. (2022), we further employed the *auditok* library for separating audio segments with and without audio. With the audio-specific optimal thresholds the ideal audio had a laughter detection F-score of 81%, while non-ideal audio achieved 71%. To enhance the precision of filtering music, applause, and noises, clusterization techniques were explored, however, the best-performing clustering method did not surpass the non-clustering approach.

[2] https://github.com/amsehili/auditok

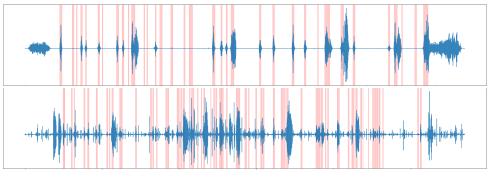

Figure 2: Audio tracks after the channels have been subtracted. The annotated laughter segments are highlighted in red spans. First row - ideal audio subtraction case, second row - non-ideal.

These observations underscore the inconsistent result of voice removal with channel subtraction and clustering. We acknowledge that we did not thoroughly investigate the specific reasons for the poor performance of the channel subtraction method in our study. Our assumption was that the overall udio quality in YouTube videos might be comparatively lower than that in sitcoms (which were initially used in previous works that employed this method). Different videos exhibited diverse behaviors, suggesting the presence of multiple contributing factors. Unfortunately, further exploration of this approach is put on hold until a dedicated study on voice removal in videos of different qualities is conducted.

3.2. Machine learning approach

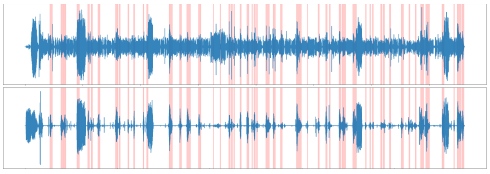

Figure 3: Comparison of an audio track after channels subtraction (first row) and vocal-remover (second row). Annotated laughter is shown in highlighted red spans.

In our review of existing literature, we identified two papers, Mittal et al. (2021) and Castro et al. (2019), that used the laughter detection library developed by Gillick et al. (2021). We will also follow that approach. We first examined the feasibility of applying the laughter detection approach to the entire audio. We conducted a hyperparameter search to determine different minimum probability threshold values for identifying laughter segments. The initial threshold of 0.1 yielded an F-score of 61%. However, this approach detected significant noise, prompting us to refine the search to the specific time window (see the window description at the beginning of 3). With a threshold of 0.2 within this context, we achieved improved performance, resulting in an F-score of 69%.

In our qualitative exploration of the errors in the laughter detection model, we encountered a diverse range of false positives without a clear systematic pattern, therefore, we are not providing any qualitative or quantitative error analysis here. However, to improve precision and reduce noise and false positives, we explored noise reduction and voice separation techniques. We utilized the *vocal-remover* library[3], originally designed for extracting instrumental tracks from songs, which demonstrated effectiveness in isolating the comedian's vocals from background sounds (Fig 3 shows the quality of the separation). This modified pipeline resulted in improved detection scores, ultimately achieving an F-score of 71% for the Russian and 77% for the English dataset. Because of the greater application success we used the machine learning laughter detection approach for labeling the whole dataset (Table 1).

	funny	non-funny
RUS	9,117 (48%)	9,696 (52%)
ENG	10,660 (52%)	9,654 (48%)

Table 1: Class distribution in the datasets.

4. Multimodal humor detection experiments

To conduct the experiments, we divided the dataset into context-utterance chunks, with four sentences in the context and the fifth sentence as the target utterance. This approach was initially chosen to test the datasets, but more sophisticated splitting methods that, for example, rely on temporal information could be explored.

For the unimodal setting we explored two key approaches for humor detection in text: SVM with TF-IDF vectors and transformer-based models, particularly Conversational RuBERT (Burtsev et al., 2018) for Russian and the original BERT base cased model for English.

In the multimodal setting, we used textual, visual and facial modalities. We intentionally decided not to use audio modality. Firstly, in previous studies, audio tended to outperform other modalities, and we wanted to focus on the visual modality more. Additionally, since we used audio for dataset annotations, incorporating it into training raised concerns. Given the prominence of laughter as a cue for humor, we were cautious about unintentionally training a laughter classifier instead of a humor classifier, especially if the audio segments were not perfectly segmented.

We explored two classification heads - SVM and neural projection. This involved integrating features from diverse encoders. For textual informa-

[3] https://github.com/tsurumeso/vocal-remover

tion, [CLS] embeddings of the context and utterance from BERT for English and Conversational RuBERT for Russian were utilized. VideoMAE was used for the visual modality - 16 equally spaced frames were sampled from the entire context plus utterance window and averaging was performed along the first dimension to align with the BERT embeddings. Facial features were integrated using OpenFace, focusing on gaze direction, facial action units, and non-rigid face shape parameters. These features were processed by averaging the context and utterance frames separately and concatenating them. For the SVM head embeddings were concatenated and normalized, while for the projection head each modality's embeddings were projected into a shared space and later fused for classification task. The number of projection layers, projection dimensionality and batch size were searched with hyperparameter search, the dropout rate was set to 0.1.

To conduct the multimodal experiments, we performed cross-validation utilizing the StratifiedShuffleSplit method with 4 splits and a test size of 0.2. We did classification using each modality separately, in pairs and using all of them.

5. Experimental Results

The SVM + TF-IDF approach scored 59.5% for Russian and 65.5% for English, providing a valuable benchmark against the multimodal configurations. We observed that text-only BERT-based models delivered the most robust performance, showcasing F-scores of 62.9% for Russian and 69.2% for English, underscoring the effectiveness of BERT models for textual data. In the multimodal setting, combining modalities significantly boosted classification scores, with the Visual modality playing a pivotal role. Notably, the neural projection head, while strong, did not surpass the text-only BERT models, primarily due to non-fine-tuned embeddings. The highest performance was achieved by combining Text and Visual modalities, yielding 61% (SVM) and 63.8% (Projection head) for Russian, while for English, the strongest setups included Facial and Visual modalities (67.8% SVM) and Visual and Text modalities (68.4% Projection head). Overall, our results highlight the potential of multimodal fusion, where diverse data sources collectively enhance classification performance, with the Visual modality exerting a substantial influence. Also, the results of our initial experiments, where even a basic model showed improvement over the baseline, indicate that the dataset is suited for further analysis and model development.

	T	V	TV	VF	TVF
Russian					
SVM	56.6	59.2	**61**	59	60.9
NN	61.5	61.4	**63.8**	59.8	62
English					
SVM	60.9	67.3	67.2	**67.8**	67.5
NN	68.3	**68.4**	**68.4**	67.1	68

Table 2: Best multimodal classification results (F-score). Letters represent the modalities used. T - textual, V - visual, F - facial. Modalities with lower scores were omitted.

6. Discussion and Future Work

Using laughter as the primary humor marker for dataset labeling has limitations, especially in stand-up comedy. Stand-up comedy laughter differs due to audience expectations and the unique comedian-audience dynamic (Abrahams, 2020). Additionally, various laughter types exist in stand-up comedy, such as warming-up, main, and follow-up laughter (Bochkarev, 2022). Consequently, any conclusions drawn from stand-up comedy data should consider these factors, including the distinct nature of laughter and comedian-audience dynamics.

Furthermore, the dataset's quality may not be optimal due to automated labeling and limited validation subset, potentially affecting model performance as there's no human-annotated gold standard. While achieving reasonably good labeling results (71% for Russian and 77% for English), there's room for improvement. For future work, enhancing labeling techniques should be a priority. This can involve exploring peak detection or template-based methods, improving vocal separation, and addressing variations in audio recordings across different environments. Moreover, considering more than two classes in clustering could enhance noise separability.

Another important thing to acknowledge is the laughter windowing - in future work, we might need to consider not only the context frames before the target utterance but also those after it. This is particularly critical in the case of visual comedy, where it has been observed that there can be codas or additional comedic elements that occur after the initial utterance.

Finally, the new dataset provides several future research opportunities. First, we can explore humor across different languages and cultures to understand its universal and culture-specific aspects. Second, delving into gender-specific humor differences is a promising avenue. Finally, we can dive deeper into how different elements like facial expressions, body language or prosody impact humor. Though this questions will require advanced models probably with attention mechanisms. These fu-

ture investigations not only advance humor analysis but also have wider applications in cross-cultural communication, gender studies, and multimodal machine learning.

7. Conclusion

In conclusion, this paper has primarily focused on multimodal humor dataset collection from variable quality sources (such as YouTube videos) and automated laughter labeling techniques. The proposed machine learning laughter labeling technique achieved F-scores of 71% for Russian and 77% for English. The initial experiments with multimodal humor detection showed potential in using the dataset for further analysis. Looking ahead, we recommend further refinement of labeling methods and continued dataset expansion with other languages, alongside the pursuit of advanced multimodal models for feature analysis.

8. Ethics Statement

We collected videos from YouTube as part of our dataset acquisition process. YouTube videos are considered publicly available information, and while YouTube users retain the copyright to their videos, we argue that research constitutes "fair use" of copyrighted materials. Moreover, to adhere to copyright regulations and data use permissions, we decided not to redistribute the videos themselves. Instead, we provided links to the videos and a web scraper script designed to download videos with manually sourced subtitles from a list of channels.

Since we have not personally reviewed the entire dataset, the quality of the content is mainly dependent on the filtering done by the individuals who uploaded the stand-up routines on YouTube. Therefore, we cannot ensure that the content is not offensive, does not propagate hurtful stereotypes of marginalized groups, and avoids explicit language. The English part of the dataset was collected from a well-known source, which may provide some level of reassurance. However, the Russian part is sourced from different channels, and during the annotation process, we encountered instances of disparaging humor and a significant amount of profanity. It is essential to acknowledge that while this content falls within the realm of humor, it can also be hurtful.

Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

9. Bibliographical References

Daniel Abrahams. 2020. Winning Over the Audience: Trust and Humor in Stand-Up Comedy. *The Journal of Aesthetics and Art Criticism*, 78:491–500.

Arsentiy I. Bochkarev. 2022. Classification of Laughter in Stand-Up Comedies. *European Proceedings of Educational Sciences*, Topical Issues of Linguistics and Teaching Methods in Business and Professional Communication - TILTM 2022.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-Source Library for Dialogue Systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia. Association for Computational Linguistics.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards Multimodal Sarcasm Detection (An _obviously_ Perfect Paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

ELAN. 2023. ELAN (Version 6.5) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive.

Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Bamman. 2021. Robust Laughter Detection in Noisy Environments. In *Interspeech 2021*, pages 2481–2485. ISCA.

Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, Mohammed, and Hoque. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. 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 2046–2056. ArXiv:1904.06618 [cs, stat].

Antony Kalloniatis and Panagiotis Adamidis. 2023. Computational Humor Recognition: A Systematic Literature Review.

Yuta Kayatani, Zekun Yang, Mayu Otani, Noa Garcia, Chenhui Chu, Yuta Nakashima, and Haruo Takemura. 2021. The Laughing Machine: Predicting Humor in Video. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2072–2081, Waikoloa, HI, USA. IEEE.

Zhisong Liu, Robin Courant, and Vicky Kalogeiton. 2022. FunnyNet: Audiovisual Learning of Funny Moments in Videos. pages 3308–3325.

Anirudh Mittal, Pranav Jeevan, Prerak Gandhi, Diptesh Kanojia, and Pushpak Bhattacharyya. 2021. "So You Think You're Funny?": Rating the Humour Quotient in Standup Comedy. ArXiv:2110.12765 [cs].

Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh Sarvesh, Hunar Singh, and Vinay P. Namboodiri. 2021. Multimodal Humor Dataset: Predicting Laughter tracks for Sitcoms. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 576–585, Waikoloa, HI, USA. IEEE.