When to Laugh and How Hard?
A Multimodal Approach to Detecting Humor and its Intensity

Khalid Alnajjar[1,2], **Mika Hämäläinen**[1,2]
Jörg Tiedemann[1], **Jorma Laaksonen**[3] and **Mikko Kurimo**[3]
[1]University of Helsinki, Finland
[2]Rootroo Oy, Finland
[3]Aalto University, Finland
firstname.lastname@{helsinki.fi[1],aalto.fi[3]}

Abstract

Prerecorded laughter accompanying dialog in comedy TV shows encourages the audience to laugh by clearly marking humorous moments in the show. We present an approach for automatically detecting humor in the *Friends* TV show using multimodal data. Our model is capable of recognizing whether an utterance is humorous or not and assess the intensity of it. We use the prerecorded laughter in the show as annotation as it marks humor and the length of the audience's laughter tells us how funny a given joke is. We evaluate the model on episodes the model has not been exposed to during the training phase. Our results show that the model is capable of correctly detecting whether an utterance is humorous 78% of the time and how long the audience's laughter reaction should last with a mean absolute error of 600 milliseconds.

1 Introduction

Humor is a topic that has piqued interest of the computational creativity research community over the years. There are numerous systems that can generate humor using a variety of different methods (Weller et al., 2020; Tyler et al., 2020; Alnajjar and Hämäläinen, 2021b). But just as important as it is to research generation from the computational creativity perspective, it is to study automatic assessment of humor.

The role of humor is an important one for us humans as it has it's own social function (Ziv, 2010). It helps us talk about difficult topics (Vivona, 2013; Monahan, 2015) and relieves tension (Shurcliff, 1968). Laughter has a role in building relationships (McCabe et al., 2017; Kurtz and Algoe, 2017) and it has a positive effect on brain chemistry (Gonot-Schoupinsky and Garip, 2018). Humor and laughter are therefore an integral part of who we are as a species.

Humor is a phenomenon that requires surprise and coherence (see Hämäläinen and Alnajjar (2019)); or incongruity and its resolution in other terms (Raskin, 1985; Attardo and Raskin, 1991). However, what is surprising or incongruous, depends heavily on the context where humour is presented. Quite indeed, something intended as a joke can be seen as a severe insult just by a change in context. Therefore, we believe that a multimodal approach needs to be researched when when assessing humor automatically; something that thus far has been researched by focusing on the textual modality alone.

Annotated multimodal datasets are scarce, but an access to such a dataset is crucial for any computational attempt on humor detection and assessment. For this reason, we embrace a clever approach: we use episodes of the beloved American sitcom *Friends* as our data source. The TV show has prerecorded audience laughter which provides us with an ultimate source for annotations. Every time the audience laughs, we know that there was something humorous immediately before the laughter. A lack of laughter indicates no humor. In addition to this, the audience can laugh for a short or a long time, which allows us to gather data on how funny a given joke was.

We propose a pipeline consisting of two neural models. One of them detects whether a sentence is humorous or not and the other rates how funny the sentence is in case it was deemed humorous. We evaluate our models on episodes of *Friends* that were not used during training or validation. Our work is a first step towards multimodal humor detection and assessment. We have also established several important data processing practices that make it possible for future research to automatically gather annotated multimodal data in a similar fashion as we did in this paper.

The main contributions of our paper are as follows:

- Methodology for automatically annotating a multimodal humor corpus based on laughter

Proceedings of the 29th International Conference on Computational Linguistics, pages 6875–6886
October 12–17, 2022.

cues.

- A multimodal humor detection model that does not rely on an explicit split in a setup and a punchline.
- A multimodal humor assessment model that can predict how funny a given joke is.

2 Related work

There is a an extensive body of literature that focuses on humor generation (Dybala et al., 2010; Alnajjar and Hämäläinen, 2018; Mishra et al., 2019; Yamane et al., 2021) and also a growing body of work that deals with multimodality in natural language understanding (Soldner et al., 2019; Wang et al., 2020b; Rodríguez Bribiesca et al., 2021). However, in this section, we focus on some of the recent papers that deal with humor detection and analysis.

In a recent work (Xie et al., 2021), the authors study humor detection in a context where there is a setup and a punchline. They use a GPT-2 model (Radford et al., 2019) to assess uncertainty and surprise to determine if a setup-punchline pair is a joke or not. Similar setup and punchline based approaches for humor detection have been widely studied in the past using different computational methodologies (Cattle and Ma, 2016, 2018; Wang et al., 2020a). However, such methods are very different from our approach as we do not expect our model to receive a setup and a punchline that are explicitly marked in the data.

Sentiment analysis has been used in humor detection (Liu et al., 2018). As many of the existing approaches, their approach also operates on a setup and a punchline. The authors found that sentiment conflict and transition between the setup and the punchline are useful in humor detection. The authors use an existing discourse parser (Feng and Hirst, 2012) combined with TexBlob[1] sentiment analysis and heuristic rules to detect humor.

Apart from humor detection, there is a line of work on assessing the humor value of a joke (Weller and Seppi, 2019). The authors propose a model that does not detect humor, as it expects jokes as its input, but instead, the model rates how humorous a given input joke is. Their model also expects a setup and a punchline division in the data. The authors train a transformer (Vaswani et al., 2017) based model for the task.

[1]https://github.com/sloria/textblob

Humorous headlines have been automatically ranked based on how funny they are (Dick et al., 2020). The authors use ridge regression and an LSTM (Long short-term memory) model with manually engineered features such as whether Donald Trump and his hair have been mentioned and the length of the headline. The authors conclude that a language model is simply not enough for assessing humor, but a wider context is needed to help the model understand humor.

As we can see, most of the previous approaches on humor detection and assessment do either or. The models can either tell whether something is funny or not, or rate how funny a given joke is. In addition, many models seem to expect a clear division into a setup and a punchline, which makes it impossible to use them to detect humor in free formed speech or text. In addition, there are many types of humor, for instance sarcastic one (Hämäläinen, 2016), that does not require an explicitly uttered setup in natural language, but rather the setup of the joke can be deduced from the context itself. Our approach tries to tackle these shortcomings in the current state of humor detection.

3 Humor theories

Humor is an integral part of our social lives as humans and because of that, it has provoked the interest of many scholars in the past. Some of the early theories of humor (Hobbes, 1651) saw it as a question of superiority, where a superior person laughs at the misery of those inferior to them. While this explanation might be valid in the context of schadenfreude, more modern takes on humor theory reject it as it cannot explain humor as a whole.

For Koestler (1964), humor is a part of creativity together with other two components: discovery and art. What is seen as characteristic to humor, in his view, in comparison to the other two constituents of creativity, is that its emotional mood is of an aggressive nature. Humor comes from bisociation which is a collision of two frames of reference happening in a comic way.

Raskin (1985) presents a theory that is quite similar to the previously described one in the sense that in order for a linguistic expression to be humorous, it has to be compatible with two different scripts. The different scripts have to somehow oppose one another, for example in the sense that one script is a real situation and the other is hypothetical.

Attardo and Raskin (1991) sees humor to be consisting of six hierarchical resources of knowledge: language, narrative strategy, target, situation, logical mechanism and script opposition. Similarly to the previous theories, the incongruity of two possible interpretations is considered to be an important aspect of humor. An interesting notion that sets this theory apart from others is that of target. According to the authors, it is not uncommon for a joke to have a target, such as an important political person or an ethnic group, to be made fun of.

Two requirements have been suggested in the past as components of humor in jokes: surprise and coherence (see Brownell et al. (1983)). A joke will then consist of a surprising element that will need to be coherent in the context of the joke. This is similar to having two incongruous scripts being simultaneously possible.

Veale (2004) discusses that the theories of Raskin (1985) and Attardo and Raskin (1991) entail that people are forced into resolution of humor. He argues that humor should not be seen as resolution of incompatible scripts, but rather as a collaboration, where the listener willingly accepts the humorous interpretation of the joke. Moreover, he argues that while incongruity contributes to humor, it does not alone constitute it.

4 Data construction

We focus on the sitcom TV show *Friends*. The show is one of the most popular American sitcoms ever produced and it aired from 1994 to 2004. Our data consists of the entire show, i.e. 10 seasons and 236 episodes each of a duration around 20 minutes. All episodes had English well-aligned subtitles that correspond to what is said in the audio track of each episode. We have randomly sampled an episode from each season to assess the quality of the subtitles, and found no major errors or clear delays.

While there are some multimodal annotated data for sarcasm detection (Castro et al., 2019; Alnajjar and Hämäläinen, 2021a), multimodal annotated data of humor in more general terms is very scarce. While several textual humor datasets exist (Hossain et al., 2019; Meaney et al., 2021), they are not suitable for our need as we are interested in multimodality. This is mainly due to the great subjectivity of what humans deem to be funny, and to the high amount of work and funds needed to manually annotate a dataset. To overcome this obstacle, we embrace an automatic approach for annotating humor in the TV show by recognizing laughter in the audience as described in the following subsection.

4.1 Data annotation

After the first few seasons, *Friends* was shot entirely in front of a live audience and a great deal of the laughter in the aired version was original, which would even cause the cast to panic when no laughter is heard while it was expected (Winston, 2021). This makes this show, and other sitcom shows that are shot live a mine of humor annotations that is calling for extraction, given that the laughter is an indication of truly landing jokes rather than being something cued in or added later in the post-editing phase.

Our approach for annotating the show relies on the model proposed by Gillick et al. (2021) for automatically detecting laughter. The model is designed to be robust enough to work on real data and be capable of detecting laughter "in the wild". On a lower level, the model's implementation relies on ResNet (He et al., 2016). The model allow us to indicate the minimum length of the laughter and a cutoff threshold of how confident the model is. We set the minimum laughter length to 0.2 seconds and the threshold to 80% based on our empirical experiments. A shorter length resulted in numerous non-continuous short segments of laughter while a longer length limited the results to a few segments per episode. The case was similar for the confidence threshold. A shorter duration for laughter also lead to many non-laughter noises such as yelling to be detected falsely as laughter.

We ran the model on all the episodes of every season of the entire show and obtained 7422 laughter segments. To construct a dataset for training a supervised neural model for identifying humor, positive and negative samples are required. A crucial aspect of what makes something funny is the context it is present in (Tsakona, 2020), just like how the utterance "You guys just keep getting cooler and cooler" can indicate the opposite of what is expressed and be sarcastic based on the tone of the speaker and what the "guys" have done or said[2]. Furthermore, the context prompting the humorous interpretation can vary in nature, especially that we are dealing with multi-modality. In our case, it

[2]The utterance was said by Chandler sarcastically in the *The One Where the Stripper Cries* episode.

could vary, for instance, based on its length (i.e., how distant is the required knowledge for the act to be deemed funny? A scene, episodes or seasons?), type (i.e., what aspects in the context contribute to the humor? does the humorousness arise from what is said, how it was said or what is done, or a combination?), and familiarity with the characters' personalities and common knowledge of the topic of the discourse.

As it is unfeasible for an unsupervised automated approach to achieve an understanding of the world and, therefore, explain the humor in a given scene and link all contexts in the entire show contributing to it, we resort to defining the context as a fixed duration for the sake of simplicity. For the positive humorous samples, we consider the last 10 seconds of all the laughter segments detected by the model to be the context of the joke. From our experiments, 10 seconds seemed to include some context (e.g., two characters saying two sentences) and not collide with previous laughter segments. An example of such a humorous segment can be seen in Figure 1.

Figure 1: An example of a humorous segment that preceded laughter

To build a set of negative samples, for each positive sample segment, we consider what is prior to it until the laughter segment preceding it to be a non-humorous segment. If possible, we split the segment into 10-second clips and randomly pick 3 clips that has some context. The presence of the context is determined by inspecting the subtitles of the clip. In the case where it was empty, the clip was discarded. This is important to remove segments that have no verbal communication at all such as camera spanning across a scenery in the beginning of a scene. Sometimes, laughter segments are very close to each other that no non-humorous segment exists before a humorous segment.

Sometimes humor was expressed non-verbally in the TV show as seen in Figure 2. We filtered out the cases where the audience laughed and there was no subtitle text before the laughter in the humor segment. We do this because we are focusing on multimodality (text and audio) in humor detection. Such non-verbal humour would require the video to be considered as well.

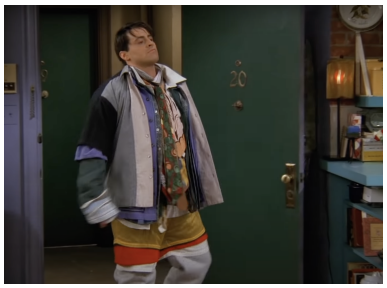

Figure 2: Joey entering the room wearing many layers of clothes provoked audience laughter non-verbally

By the end of the annotation process, our dataset consisted of a total of 16710 clips where 7422 of them are humorous and the remaining 9288 are non-humorous clips[3]. Furthermore, we indicated in the dataset the length of the laughter segment for all humorous clips. This metadata will be used to predict the intensity of the humor.

4.2 Preprocessing

In our current work, we focus on text and audio in detecting humor, and leave including visual content for future work. This is due to the fact that processing video requires a research on its own right. Video is such an information rich resource with so many potential things one could extract that may or may not be relevant for humor such as facial expressions, body poses, object recognition, action detection and so on. For our text and audio modalities, two types of preprocessing are applied: 1) cleaning and reformatting the subtitles[4] and 2) resampling the frequency of the audio clips.

Dashes, -, at the beginning of subtitle lines were commonly used to imply that the speaker has not changed. We treat such dashes as noise and prune them out. Furthermore, new lines are usually added in subtitles to facilitate reading them and/or to separate talks by different speakers. We substitute new lines with spaces in order to convert the subtitle into complete sentences. Italic tags, "" and "", were stripped out. When inspecting the subtitles, we noticed that a frequent typo of having a capital I instead of an l existed. We addressed

[3]Due copyright we are unable to publicly release the dataset.

[4]This step improved the accuracy of our models by 3%.

this issue by replacing all *I*s with *l*s if they were happened to be in the middle of the word. In case a clip contained multiple subtitle segments across the time-span of the clip, they were all joined together.

In terms of audio processing, we only apply frequency resampling to adjust the frequency to 16kHz. This step is performed to ensure consistency when feeding the data to a neural model.

5 Assessing humor

We present a pipeline of two neural models for assessing humor. The goal of the first model is to identify whether an utterance at the end of the segment is humorous or not given the rest of the segment as context, whereas the second model predicts the intensity of the humor once it has been detected.

5.1 Humor detection

Here, we describe two different models for detecting humor. The first model relies solely on the textual data, and the second is a multimodal model that accepts both textual data and audio signals as input.The task for these models is a binary classification downstream task which is to determine whether the input contains humor or not. We experiment with the two models to gain a better understanding on what the effect of the audio is for humor detection.

We group our dataset by episodes and, then, randomly decide which episodes will be used for training, validation and testing. This division is conducted to prevent the model from getting exposed to shared contexts during the training and testing phases, which would introduce undesired bias. When test data is sampled from completely different episodes than what the training and validation contained, we can ensure that the model learns to detect humor in completely novel contexts rather than detecting merely episode specific recurring jokes. The test dataset is constituted of 25 full episodes, which are 1x09, 2x06, 2x22, 3x13, 3x20, 4x09, 5x04, 5x07, 6x09, 6x11, 6x15, 6x16, 7x09, 7x16, 7x19, 7x20, 7x22, 8x03, 8x14, 8x21, 9x05, 9x11, 9x21, 10x17 and 10x18. In total, the training, validation and testing splits contained 13506, 1477 and 1708 samples, respectively. Both of the models used the same splits to ensure comparable results.

5.1.1 Text only model

We build our text only model by applying transfer-learning and fine-tuning a BERT model (Devlin et al., 2019) using the transformers Python library (Wolf et al., 2020). The pretrained model we used is the uncased English BERT model[5]. For a given input, it is first tokenized using BERT tokenization. If the input contained subtitles from different scenes, they we combined together and separated using the special token "[SEP]".

The architecture of the neural network is composed of the BERT model, a BERT pooler layer, a dropout layer (Srivastava et al., 2014) and a fully connected dense layer that has two outputs. Once the input has passed through the BERT model, the pooler layer returns the last layer hidden-state of the first token of the input sequence. Dropout is applied on the pooler output with a probability of 20% to reduce overfitting. The linear layer is introduced so that the network would learn a way of interpreting the features produced by the past layers and assign a probability score for each of the two labels. In total, the model has 109 million trainable parameters. We utilize Adam optimizer (Kingma and Ba, 2014) with a learning rate of $1e{-}4$, along with the cross-entropy loss function to optimize the neural network. The fine-tuning process was run for 3 complete epochs.

5.1.2 Text and audio model

The multimodal model we propose utilizes the textual and audio input by combining BERT (Devlin et al., 2019) with Facebook's HuBERT[6] (Hsu et al., 2021) neural models. We use the same uncased English BERT model in our multimodal model as we did in our text only model to examine the effect of incorporating audio features for detecting humor. The choice of HuBERT, in contrast to the popular XLSR-Wav2Vec2 (Conneau et al., 2021), is due to its superior or, in worse case scenario, neck-to-neck performance.

Our multimodal model architecture is similar to a siamese neural network architecture in the sense that the output of two models are considered collectively. In our model, one side of the network is dedicated to text and the other to audio. We ensure that both sides produce an equal size of features by 1) setting a fixed input length to BERT where padding and truncating is applied where necessary and 2) having two average pooling layers following the output of each side. For the textual output,

[5] https://huggingface.co/bert-base-uncased

[6] https://huggingface.co/facebook/hubert-xlarge-ls960-ft

Length	0.2-0.5	0.5-1.5	1.5-2.5	2.5-3.5	3.5-4.5	4.5-5.5	5.5-15.5
N	459	2895	2328	948	374	184	234

Table 1: The duration of laughter (in seconds) for different ranges in our dataset

a global average pooling is applied, whereas an adaptive average pooling is applied to the audio output. Afterwards, the pooled output is concatenated and followed by a dropout layer with a probability of 20%. Lastly, a fully connected dense layer is employed as the classification layer. The network has 424 million trainable parameters. We use the same hyperparameters for optimizing the multimodal model as the text only model; in other words, we fine-tuned it for 3 full epochs with a learning rate of $1e{-}4$.

5.2 Predicting laughter intensity

We define the intensity of the laughter based on its duration. Thus, a strong laughter for multiple seconds indicates a great joke. This is intuitive because the funnier the humor gets, the longer the audience laughs. As the duration of a laughter is a continuous value, we treat the task a regression problem and adopt an artificial intelligence neural network for addressing it.

Our dataset for this part is only the humorous clips, a total of 7422 clips. Table 1 shows the number of humorous clips grouped by various durations. We cap all the durations to 3 seconds given that the majority of laughter segments are within this limit. The data is then split for training, validation and testing with 80%, 10%, and 10% ratios, respectively.

In our laughter intensity prediction model, we only train a multimodal model given that *what* is said in the joke and *how* it is said or performed have great influence on the reaction from the audience. Textual features are extracted using BERT like the aforementioned models. For audio features, we extract them from the entire humorous clips using Google's VGGish model[7] (Hershey et al., 2017). VGGish is a Convolutional Neural Network inspired by the VGG network (Simonyan and Zisserman, 2015) that is trained for image classification. However, in VGGish, the input image is the log mel spectrogram of frames derived from the audio. The network has achieved the state-of-the-art results for audio classification given its ability to capture acoustic features, tones, volume and so on. Unlike the earlier models, the pretrained models, i.e. BERT and VGGish, are frozen and not fine-tuned during the training step.

The model we present here is a sequence of layers where the first one is a 2D average pooling layer that converts the extracted features by BERT and VGGish models into a fixed-size set of features by averaging neighbouring features until the desired size is reached. We set the size here to 128 as this is the size of features that the VGGish model returns per frame. As a result, the output of this layer is a vector of 256 features. The layer is then followed by a dense fully-connected linear layer that takes in the averaged features and learns a new representation of 64 features. ReLU activation (Agarap, 2018) and a dropout with a probability of 10% are then subsequently applied to the 64 features. The network architecture ends with a fully-connected dense layer that returns one output representing the intensity score.

As the problem here is regression, we make use of the mean squared error (squared L2 norm) as the loss function. This model has 16.5 thousand trainable parameters and we optimize them for 100 epochs using Adam optimizer, however we use use early stopping to stop the training of the model before 100 epoch in the event of the model converging early.

6 Results and evaluation

We run both, the text only and multimodally trained models, for humor detection on the test split. Their performance is assessed using precision, recall, F1 and accuracy scores, which are given in Table 2. Both of our models outperform the baselines of choosing a label at random or the most frequent label, their accuracies were 51% and 56%, in the order given.

The results indicate that the multimodal model clearly outperformed the text only model, by a 16% increase in detection accuracy. This suggests that audio cues were helpful in recognizing humor. For instance, sarcasm and irony are sometimes marked with clear intonations and tones. Both of these phenomena are frequently used in sitcoms for humorous effect, which would aid the model in distinguishing when "Yeah, right" is meant as a sincere confirmation or as a sarcastic remark for humorous effect.

Recognizing humor is a challenging task, even

[7] https://github.com/harritaylor/torchvggish

	Precision	Recall	F1	N
Text only model				
Funny	0.58	0.52	0.55	758
Not funny	0.65	0.70	0.68	950
Accuracy	62%			
Text + Audio model				
Funny	0.69	0.90	0.78	758
Not funny	0.90	0.68	0.78	950
Accuracy	78%			

Table 2: Accuracy, Precision, recall, and F1 scores of the two models for detecting humor

for humans and it is no surprise that computational models would struggle. As these models are tested on entirely novel contexts (i.e., new discourses that are not covered during the training phase), the performance achieved by them is impressive.

To test the model for predicting laughter intensity, we compute the mean absolute error between the predicted intensity and the intensity of the laughter in the dataset. The average mean absolute error was only around 600 milliseconds. This means that the model can predict how long the audience will laugh after a given joke rather accurately, given that there is some flexibility in the duration of the laughter as it is not an absolute measurement of the humor of a joke. With these results, we can say that the model has learned to predict the intensity of humor in a joke well, given that for jokes provoking less laughter, the model predicts a short laughter, and for jokes provoking a lot of laughter, the model predicts a long laughter. Even though the model is not quite accurate in knowing the exact duration of the laughter.

6.1 Error analysis

When we look at the results of the models, we can see some cases where both of the models failed at predicting the humor accurately. For example, the following dialog provoked a laughter in the audience:

- **Phoebe:** It's amazing. My headache is completely gone. What were those pills called?
- **Monica:** Hexadr*i*n.
- **Phoebe:** I love you Hexadr*i*n.

In this example the humor comes from Phoebe's lack of knowledge of how medicines work as she continues by calling the instructions booklet a story. This is an example of humor that requires some world knowledge and also some understanding of Phoebe's care-free character. Just relying on text in this case or even including the audio does not give the model a context wide enough to reach to a correct interpretation.

In some of these cases where neither of the models predicted the humor right, it is evident that with an access to the video, the model could situate the humor better in the context. The following dialog is an example of such humor:

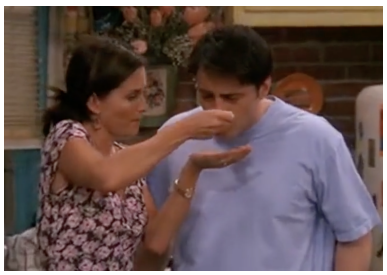

Figure 3: Joey inquiring whether Monica had cooked a person after tasting her food.

- **Monica:** Remember the guy that gave me a bad review? Well... I'm getting my revenge.
- **Joey:** You cooked him?

In this dialog, Monica is preparing food and letting Joey taste some of it after her line in the dialog as seen in Figure 3. A great part of the joke is in the visual action of Joey tasting Monica's food before asking whether she had cooked the guy who had given a bad review. A model that can take the visual modality into account as well could potentially benefit from the humor intensifying action seen on the video.

When we look at the results where the multimodal model predicted humor right and the text only model predicted it falsely as non-humorous, we can see cases where the speaker's use of their voice gave additional context for the humor. An example of such is in the following dialog:

- **Katie:** You have selected a lot of nice things. So do you uh, want these things delivered Mr. and Mrs. Geller?
- **Rachel:** Oh
- **Ross:** Oh
- **Rachel:** Oh, no, no. No, no.

In this example, the tone of Rachel's voice makes it more evident that Katie was wrong in her assumption that Rachel and Ross were married. The model had learned to capture such a tone of voice as an indication of humor. Another example where the audio is beneficial is the following dialog

- **Waiter:** It's just that we do have some large parties waiting.
- **Phoebe:** Oh, one really does have a stick up one's ass, doesn't one?

In this case Phoebe's line was delivered with a mean and fed up tone, which was helpful for the model in determining humor. Of course, such a tone is not related to humor in every day speech, so this is an indication of a potential bias in the TV show where such a tone is probably used more often to deliver a punchline of a joke than to actually upset the interlocutor. Even though the multimodal model predicted it correctly, including video modality could strengthen the signal to the model because Phoebe had an uncharacteristically nasty facial expression when uttering her line as seen in Figure 4.

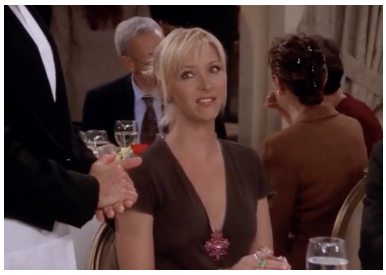

Figure 4: Phoebe having a grumpy facial expression while delivering a laughter provoking line.

The both of the models also produce false positives. The following is an example of a dialog that resulted in being falsely labeled as humorous:

- **Monica:** Chandler, don't joke with me. Okay? I'm very, very upset right now.
- **Chandler:** Is this the most upset you could be?
- **Monica:** I think so.

When this dialog is presented without a wider context, it becomes difficult even for a person to know whether it is supposed to be a joke or not. Monica might very well be talking sarcastically, which is a typical type of humor in the corpus, but in this particular case, she is being sincerely upset. Giving the model more context could alleviate this issue, but of course more context might result in more noise because not all contextual information is relevant. The following is another example where both of the models predicted a false positive:

- **Casting agent:** In your love scene with Sarah... she talks about how she's never seen a naked man who wasn't Jewish.

The audience did not laugh after this statement, although in the right context, it might be funny. Here the context was serious. The audience only laughed later on when it became evident that Joey, who was being cast to the movie, did not understand what the agent meant by this utterance. It is clear that what is humorous and not is not always that clear cut especially in a narrow context.

7 Discussion

In this section, we discuss our work and the results obtained. As mentioned earlier, one crucial aspect for understanding humor is the context. We have defined the context in our work as a fixed 10 seconds but, in reality, the context might be wider than that. As another attempt to define context, we have split the TV show based on changes in the visual scene. This approach however was not very practical for our needs as it would mistakenly cut scenes based on changes in the camera angel. A potential solution to overcome the problem of fixed contexts is to observe semantic changes for discovering consecutive scenes that share the same topic. We keep this for future work.

Sitcom TV shows are a great resource for computational humor as there is a multitude of humor forms that they present. For this reason, it is challenging for a neural model to capture all humor forms (e.g., irony, sarcasm, satire, exaggeration, personification, silliness, pun and parody), whether they are expressed verbally or visually. Our multimodal model develops its own understanding of what is funny based on the textual and audial features embedded by BERT and HuBERT. Thus, it is incapable of explaining the humorousness it perceives. In other words, the model can say whether something is funny or not, but it cannot say *why* something is funny. A future direction would be to break down these types of humor and feed them collectively to the model, which would enable it

to recognize humor from different aesthetics and pinpoint the humor type.

From the error analysis, it is evident that including video modality can help the model in understanding humor by situating what is said to the context presented in the video. However, it is not that clear to know which features would be needed. In Figure 3, we could see that it is the action that makes the humor more interpretable, whereas in Figure 4 we saw that the facial expression was revealing of humor. If the humor occurs only in the visual modality as seen early on in Figure 2, it is in the silliness of how Joey looks while wearing all those clothes. Needless to say, a simple automatically extracted vector representation of the video such as video2vec (Hu et al., 2016) is not capable of capturing all these different nuances expressed in the video. It might very well be that an entire TV show does not have enough data for the model to learn to use the video in a meaningful way.

We did try to include video features in our models by obtaining textual descriptions of what is happening in a scene by using an existing state-of-the-art video captioning model (Luo, 2020). The idea was that a captioning model could extract relevant information from the video into a textual format that could then be understood by a language model such as BERT. In practice, this turned out to be an impossible task with the current models and the image datasets they were trained on. Figure 5 shows the poor performance the model had on images from *Friends*. In our experiments, we did not see a single correctly captioned image from the TV show while test data from the dataset the model was trained on produced decent results.

We have also seen that the 10 second contextual window is not always enough to resolve whether something is humorous or not. The issue of increasing the context is that more context will also increase the amount of irrelevant context. We do not believe that simply increasing the context from 10 to 20 seconds, for example, is the most optimal way to go about it because some jokes require more context whereas others do not. Perhaps the best way would be to introduce a third model to the pipeline that is trained to determine how much context is needed by identifying how far back in the dialog we can go and still stay in the same topic. A change in topic would indicate that the context goes too far away from what is needed to interpret the joke.

(a) a man and a woman sitting on a bench

(b) a man and a woman sitting at a table

(c) a man and a woman is sitting on a couch with a dog

(d) a man and a woman playing a video game

Figure 5: Descriptions produced by the image captioning model

8 Conclusions

This work has shown the first steps towards humor interpretation in a multimodal data. Unlike the existing methods, our method does not rely on implicitly marked setup and punchline but can rather detect humor even in cases where the setup of the joke was not made explicit in the text. We have also trained a model that can rank the intensity of humor based on how long the audience laughed. The results are promising and our current research has a lot of potential for future research especially in studying how to deal with video and how much context one should include.

The trained models can be incorporated in other computational creativity models for generating humor. For instance, a system for generating humorous transcripts could utilize our models for determining whether the plot is funny and which version of it would make the audience laugh the most.

An interesting application of our approach would be to pipeline it with a laughter generator. The models presented in this paper could be used to identify where laughter should be inserted and with what intensity in a comedy show that does not have prerecorded laughter. This could save time in post-editing if used in a professional setting.

Because *Friends* has been translated into multiple languages, this makes it possible to rerun our experiments in different languages with a minimal effort. It also creates an ultimate test-bed for multilingual models where we can test whether a model learning humor from the data in all languages can learn a better representation.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Khalid Alnajjar and Mika Hämäläinen. 2018. A master-apprentice approach to automatic creation of culturally satirical movie titles. In *The 11th International Conference on Natural Language Generation*, page 274–283.

Khalid Alnajjar and Mika Hämäläinen. 2021a. ¡Qué maravilla! multimodal sarcasm detection in Spanish: a dataset and a baseline. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 63–68.

Khalid Alnajjar and Mika Hämäläinen. 2021b. When a computer cracks a joke: Automated generation of humorous headlines. In *Proceedings of the 12th International Conference on Computational Creativity (ICCC 2021)*, pages 292–299.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor-International Journal of Humor Research*, 4(3-4):293–348.

Hiram H Brownell, Dee Michel, John Powelson, and Howard Gardner. 1983. Surprise but not coherence: Sensitivity to verbal humor in right-hemisphere patients. *Brain and language*, 18(1):20–27.

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.

Andrew Cattle and Xiaojuan Ma. 2016. Effects of semantic relatedness between setups and punchlines in Twitter hashtag games. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 70–79.

Andrew Cattle and Xiaojuan Ma. 2018. Recognizing humour using word associations and humour anchor extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1849–1858.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Anna-Katharina Dick, Charlotte Weirich, and Alla Kutkina. 2020. HumorAAC at SemEval-2020 task 7: Assessing the funniness of edited news headlines through regression and trump mentions. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1019–1025, Barcelona (online).

Pawel Dybala, Michal Ptaszynski, Jacek Maciejewski, Mizuki Takahashi, Rafal Rzepka, and Kenji Araki. 2010. Multiagent system for joke generation: Humor and emotions combined in human-agent conversation. *Journal of Ambient Intelligence and Smart Environments*, 2(1):31–48.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 60–68.

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

Freda N Gonot-Schoupinsky and Gulcan Garip. 2018. Laughter and humour interventions for well-being in older adults: A systematic review and intervention classification. *Complementary therapies in medicine*, 38:85–91.

Mika Hämäläinen. 2016. Reconocimiento automático del sarcasmo: ¡esto va a funcionar bien! *Universidad de Helsinki, Departamento de Lenguas Modernas*.

Mika Hämäläinen and Khalid Alnajjar. 2019. Modelling the socialization of creative agents in a master-apprentice setting: The case of movie title puns. In *Proceedings of the 10th International Conference on Computational Creativity*, pages 266–273.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. 2017. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Thomas Hobbes. 1651. *Leviathan: or, the matter, forme & power of a commonwealth, ecclesiasticall and civill*. University Press.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 133–142.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Sheng-Hung Hu, Yikang Li, and Baoxin Li. 2016. Video2vec: Learning semantic spatio-temporal embeddings for video representation. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 811–816. IEEE.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980*.

Arthur Koestler. 1964. *The act of creation*. London Hutchinson.

Laura E Kurtz and Sara B Algoe. 2017. When sharing a laugh means sharing more: Testing the role of shared laughter on short-term interpersonal consequences. *Journal of Nonverbal Behavior*, 41(1):45–65.

Lizhen Liu, Donghai Zhang, and Wei Song. 2018. Modeling sentiment association in discourse for humor recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 586–591.

Ruotian Luo. 2020. A better variant of self-critical sequence training. *arXiv preprint arXiv:2003.09971*.

Crystal McCabe, Katie Sprute, and Kimber Underdown. 2017. Laughter to learning: How humor can build relationships and increase learning in the online classroom. *Journal of Instructional Research*, 6:4–7.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119.

Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. 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 6144–6154.

Kathleen Monahan. 2015. The use of humor, jesting, and playfulness with traumatized elderly. *Social Work in Mental Health*, 13(1):17–29.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Victor Raskin. 1985. *Semantic Mechanisms of Humor*. Springer Science & Business Media.

Isaac Rodríguez Bribiesca, Adrián Pastor López Monroy, and Manuel Montes-y Gómez. 2021. Multimodal weighted fusion of transformers for movie genre classification. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 1–5.

Arthur Shurcliff. 1968. Judged humor, arousal, and the relief theory. *Journal of personality and social psychology*, 8(4p1):360.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Villy Tsakona. 2020. *Context in humor research*, pages 7–18. De Gruyter Mouton.

Bradley Tyler, Katherine Wilsdon, and Paul Bodily. 2020. Computational humor: Automated pun generation. In *Eleventh International Conference on Computational Creativity: ICCC'20*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tony Veale. 2004. Incongruity in humor: Root cause or epiphenomenon? *Humor: International Journal of Humor Research*, 17(4):419–428.

Brian D Vivona. 2013. Investigating humor within a context of death and tragedy: The narratives of contrasting realities. *The Qualitative Report*, 18(50):1.

Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. 2020a. Unified humor detection based on sentence-pair augmentation and transfer learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 53–59.

Zilong Wang, Mingjie Zhan, Xuebo Liu, and Ding Liang. 2020b. DocStruct: A multimodal method to extract hierarchy structure in document for general form understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 898–908.

Orion Weller, Nancy Fulda, and Kevin Seppi. 2020. Can humor prediction datasets be used for humor generation? humorous headline generation via style transfer. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 186–191.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. 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 3621–3625, Hong Kong, China.

Ben Winston. 2021. Friends: The reunion. *HBO Max*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Yubo Xie, Junze Li, and Pearl Pu. 2021. Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 33–39.

Hiroaki Yamane, Yusuke Mori, and Tatsuya Harada. 2021. Humor meets morality: Joke generation based on moral judgement. *Information Processing & Management*, 58(3):102520.

Avner Ziv. 2010. The social function of humor in interpersonal relationships. *Society*, 47(1):11–18.

Acknowledgments

This work was partially financed by the Society of Swedish Literature in Finland with funding from Enhancing Conversational AI with Computational Creativity, and by the Ella and Georg Ehrnrooth Foundation for Modelling Conversational Artificial Intelligence with Intent and Creativity.