Predicting Humor by Learning from Time-aligned Comments

Zixiaofan Yang, Bingyan Hu, Julia Hirschberg

Columbia University, USA

zy2231@columbia.edu, bh2447@columbia.edu, julia@cs.columbia.edu

Abstract

In this paper, we describe a novel approach for generating unsupervised humor labels using time-aligned user comments, and predicting humor using audio information alone. We collected 241 videos of comedy movies and gameplay videos from one of the largest Chinese video-sharing websites. We generate unsupervised humor labels from laughing comments, and find high agreement between these labels and human annotations. From these unsupervised labels, we build deep learning models using speech and text features, which obtain an AUC of 0.751 in predicting humor on a manually annotated test set. To our knowledge, this is the first study predicting perceived humor in large-scale audio data.

Index Terms: humor prediction, automatic labeling, multimodal corpus

1. Introduction

Humor is one of the most interesting yet complex components in our daily communication, in which producers evoke positive emotional reactions from perceivers [1] using various strategies [2]. Identifying humor is an essential step toward fully understanding human communicative activity. Our motivation is twofold: first, we are interested to learn when the speaker is being humorous rather than serious so we can evaluate the content of what they say properly. In addition, we believe that defining a set of metrics to identify humor can lead to interesting work in speech synthesis: for example, it would permit the production of "humorous speech" that are designed to be engaging, including applications of interactive games and advertisements.

While researchers have attempted to find patterns in humorous expressions and to build models to recognize humor, most work has been done on text alone; very little has been done on multimodal humor including text and speech information. Unlike other cognitive processes such as emotions, the perception of humor is highly individualistic [3]. Thus, more effort is needed to obtain annotations of humor with high accuracy. A major difficulty in this is the lack of multimedia data annotated with humor. To address this problem, we propose a novel approach using time-aligned user comments on videos to generate unsupervised humor labels, which we validate by human annotations. We then train deep learning models for predicting humor using speech and text features and achieve high AUC on a held-out, manually annotated test set. In Section 2 we describe related work. Section 3 introduces the Bilibili corpus we collected. In Section 4, we explain our approach to humor prediction. We discuss the experimental settings in Section 5, and analyze the results in Section 6. We conclude in Section 7 and present directions of future work.

2. Related work

Most previous work on humor prediction has been done on text. Mihalcea and Strapparava [4] and Yang et al. [5] examined the

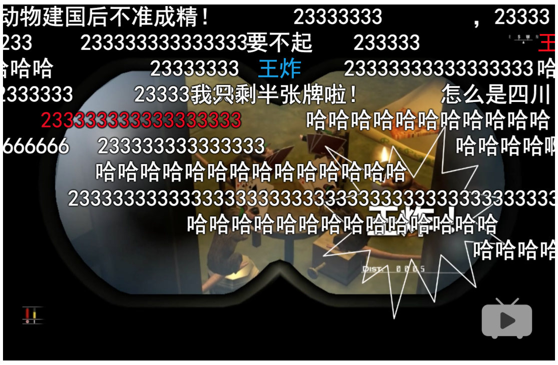

Figure 1: *Screenshot of a humorous scene with laughing comments containing "233" and "哈哈".*

expression of humor in one-liners; Mihalcea and Pulman [6] analyzed humorous features in news and blogs; Raz [7] and Zhang and Liu [8] collected and classified humorous tweets; Radev et al. [9] predicted humor ranking in The New Yorker Cartoon Caption Contest. Major findings from this research are that humor in text is associated with semantic classes relevant to human-centeredness and negative polarity [6, 9]. Research on humor in videos has focused on TV sitcoms, using canned laughter as indicators of humor. Purandare and Litman [10] examined speech features of the "FRIENDS" sitcom, while Bertero and Fung [11, 12, 13] built deep learning models with text and speech features to predict canned laughter in "The Big Bang Theory" and "Seinfeld". However, no study has shown that canned laughter represents the audience's actual perception of humor. Such information can only tell us what the sitcom producers want the audience to find humorous. Another drawback of this approach is the limitation of the genre; models trained on a particular TV show may not generalize to other shows.

3. The Bilibili corpus

To perform our experiments in humor prediction, we collected 241 videos and their user comments from *bilibili.com* . These videos consist of two categories: comedy movies and gameplay videos, which both contain many spoken utterances and humorous scenes. Different from traditional video sharing websites where audiences post their comments in a specific comment area under the video, *bilibili.com* allows users to post comments about a specific scene while watching the video. When others watch the same video, all previous comments from other viewers are displayed on the video field, synchronized with the scenes. Figure 1 shows examples of videos displayed on the website. Based on findings that laughter is the most explicit expression of perceived humor [3, 14], we use laughter indicators to identify perceived humor in our videos. The sequence

"233" is commonly used by Chinese video viewers to indicate laughter [15], while "哈哈" ("haha" in Chinese) also strongly correlates with humor as an onomatopoeia of laughter. By calculating the number of comments that contain "233" or "哈哈" within a response window, we can estimate a video scene's degree of humorous.

3.1. Comedy movies

We selected comedy movies directed by Stephen Chow based on the large number of user views on Bilibili to ensure the quality of the comments. The 8 comedy movies have 14 hours of video and 63,582 comments, including 8,821 comments with "233" and 2,700 with "哈哈". The percentage of comments containing either laughter indicator is 18.12%. Specifically, 13.87% of the comments contain "233" in them, which is higher than the average percentage of 10.44% reported in the previous work [15]. This further indicates that our video selection strategy is able to pick out the videos with intensive humorous scenes.

3.2. Gameplay videos

We used all 233 gameplay videos uploaded by a popular video creator famous for describing the games in a humorous way. These 64 hours of video have 410 billion total views and 494,438 comments, including 39,789 comments with "233" and 10,515 comments with "哈哈". The percentage of comments containing either laughter indicator is 10.17%. The lower percentage of laughing comments in gameplay videos can be demonstrated in Figure 3. The humorous scenes in gameplay videos are sparse. Moreover, unlike the comedy movies which are all carefully crafted to express humor, not all gameplay videos are intended to be humorous. The existence of gameplay videos recorded for walkthrough and advertising purpose make the percentage of laughing comments lower than average.

4. Approach to humor prediction

To identify humor, we first processed time-aligned comments to infer humor labels, generating these labels by estimating user response time to a scene and performing contextual smoothing on the number of comments. Next, we trained classifiers on text and speech using the unsupervised labels and predicted perceived humor on a held-out, manually labeled test set.

4.1. Constructing unsupervised labels

We performed an unsupervised labeling method suggested by an initial study of user behavior on Bilibili, using the keywords "233" and "哈哈" which represent laughter in Chinese network culture [15] as indicators of perceived humor. We calculated the time delay for responding to a humorous scene by estimating the reaction time and the typing time. Most time-aligned comments are posted while users are watching the videos without any pauses for commenting. Therefore, most comments have time delays which we need to take into account. In the study reported by Schröger and Widmann [16], human reaction time to audiovisual stimulus is 0.316s. For typing time, an average keystroke takes 0.2s for a skilled typist. Therefore, typing every single key character takes 0.2s. An average Chinese character can be represented by 4.2 Roman characters in Pinyin [17], so we estimated that each Chinese character takes 0.2*4.2 = 0.84s to type. Moreover, sending the comment takes 0.2s. To study the time delays, we scraped and computed the response time

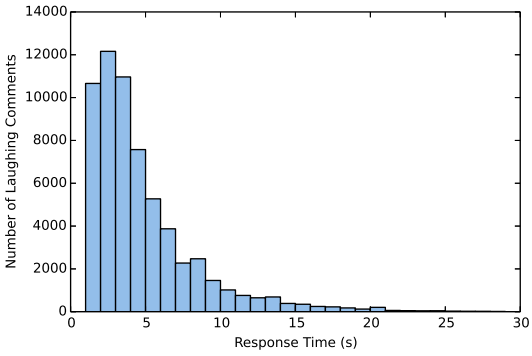

Figure 2: *Histogram of response times.*

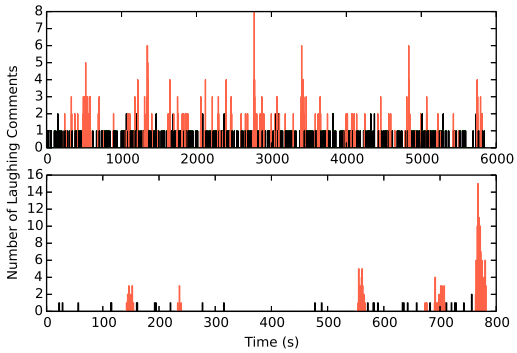

Figure 3: *Result of smoothing and labeling on a comedy movie (upper) and a gameplay video (lower)* .

for 61k laughing comments. The histogram of response time distribution is shown in Figure 2. We can see that the median response time is 3.716s and 54.65% of the comments have a reaction time within 4s. To take into account users who *do* pause the video to type comments, we estimated that the 4s response window will cover even more comments towards a particular scene. Therefore, we used 4s as the response window size.

Comments are posted and aligned with videos with millisecond precision; however, humor typically occurs in a broader context. To further reduce the influence of response time delay on our labels, we applied contextual smoothing over the number of laughing comments. We smoothed peaks of comments using a sliding window with a window size of 4s and a stride of 1s over the whole video. To estimate the "humorousness" of each window, we calculated the total number of laughing comments posted within that window. The "humorousness" threshold was set at 3 in order to filter out scenes with low agreement among the audience. For every sliding window with 3 or more laughing comments, we labeled the 4 one-second units in the window and the previous 4 one-second units as humorous. Figure 3 shows the result of smoothing and labeling on a comedy movie and a gameplay video, each with millions of views and thousands of comments. In this figure, red bars represent 1s units with humor labels, and black bars represent 1s units with non-humor labels. The bar height indicates the number of laughing comments within the corresponding 1s unit. In this way, we capture peaks of laughing comments in the video,

while ignoring portions with low agreement among the audience. After smoothing laughing comments, we generated unsupervised labels on each 1s unit as described above. We see that continuous non-humorous chunks have significantly longer duration than continuous humorous chunks, making it possible for a classifier to "cheat" by looking at chunk length. To make the prediction task fair, we cut all chunks into 8s segments. The segment size is twice the response window size to cover one window and most of its responses. For comedy movies, there are 1159 segments with unsupervised humor labels and 4502 segments with non-humor labels. For gameplay videos, there are 4018 segments with unsupervised humor labels and 23421 segments with non-humor labels. We chose fixed size segmentation method since we wanted to build models on a large-scale automatically-annotated dataset. On such a corpus, it would be very time-consuming to segment utterances manually; moreover, the Chinese Automatic Speech Recognition (ASR) transcription is not accurate enough to segment utterances reliably. Table 1 shows the statistics of our corpus, including the number of videos, comments and labeled segments.

4.2. Human annotation

To verify our unsupervised labels, we asked human annotators to annotate the test sets of both comedy movies and gameplay videos. For the 8 comedy movies, 1 movie with 775 segments was randomly chosen as the test set; for the 233 gameplay videos, we randomly chose 10% of the videos with 2188 segments as the test set. Three native Chinese annotators were asked to watch the videos *without* seeing the time-aligned comments and to label each segment with humor/non-humor labels. Fleiss' Kappa indicates moderate agreement (0.532) between annotators for comedy movies, and substantial agreement (0.683) for gameplay videos. One possible explanation of the difference in the inter-annotator agreement is that the comedy movie has more small punchlines crafted to make people laugh throughout the whole movie, while the humorous scenes are more concentrated in gameplay videos, making them easier to identify. We obtained gold labels on the test set by taking the majority vote of all 3 annotators' labels. The accuracy between unsupervised labels and gold labels on comedy movie test set is 0.881, and the accuracy on gameplay videos is 0.942. These numbers we believe are high enough to validate our unsupervised labeling method for humor prediction.

5. Experiments

Our overall goal is to build models for predicting humor by learning from our unsupervised labels. Only 24.52% of the comedy movie segments and 17.50% of the gameplay video segments are annotated as humorous in the manual labels. Given this imbalance, we used the area under the receiver operating characteristic curve (AUC) to evaluate the model's performance on the test sets. The AUC baseline is 0.5 for a binary classification problem. For comedy movies, there are 4886 segments in the training set and 775 segments in the test set. For gameplay videos, there are 25251 segments in the training set and 2188 segments in the test set.

We converted all corpus videos to audio files sampled at 44.1kHz and segmented into 8s segments. We used the Chinese Speech-to-Text API of Google Cloud Platform for transcription, and the Jieba package for word segmentation. For machine learning models, we used a Random Forest (RF) classifier with 500 estimators as a baseline. RF is good at preventing overfitting, and it suits our task well since the training and test sets are very different. Text features are TF-IDF transformed unigrams and speech features are the openSMILE toolkit's baseline set [18, 19]. We also added speaking rate feature and keywords with human-centeredness and negative polarity — shown in prior work to be related to humor expression [6, 9].

For the neural network (CNN) model, we used 13 mel-frequency cepstral coefficients (MFCC), pitch, energy and zero-crossing rate as speech features, computed with 25ms frame length and 10ms stride. These features and their first and second order delta coefficients were stacked along the time axis. We used Chinese word vectors pre-trained on Wikipedia dumps as text features. Two sets of convolutional and max-pooling layers were used to extract patterns from the frame-level speech features and word vectors separately. For speech, we used a kernel size (10,10), a filter number of 50, and a pooling size of (2,5); for text, we used a kernel size of 4, filter number of 256, and max-over-time pooling. The convolutional outputs were then concatenated and fed into a fully connected layer with size 512 to generate the final predictions. All convolutional layers are followed by Dropout layers with 0.5 probability. We used 32 as batch size, cross-entropy as loss function, and Adam with a learning rate of 10^{-4} as optimization algorithm. The hyperparameters were tuned in a cross-validation manner.

6. Results and analysis

6.1. Classification results

Acoustic-prosodic features such as pitch and energy have been shown to be relevant to humor expression in TV sitcoms [10]. So, we first conducted experiments using only speech features; results are shown in the first two columns of Table 2. In both video categories, the CNN performs better than the RF model, since it can capture more complex patterns in the data. We notice that the AUC for gameplay videos is higher than the AUC for comedy movies. From manual analysis, this appears to be primarily because the humor expressed in the gameplay videos is more straightforward than in the comedy movies. To attract as many viewers as possible, the gameplay video creator shows emotion explicitly in his voice and sometimes exaggerates his reactions. However, in the comedy movies, the expression of humor in the actor's voice is more subtle, and a humorous scene often results from the joint effect of speech prosody, speech content, facial expression, and body gestures.

To test whether the speech content itself provides useful indicators of humor, we conducted experiments using features from the text transcription only. However, using text features alone, results were not significantly better than chance. For comedy movies, the main problem appears to be the insufficiency of training data. The themes of all 7 training movies are quite different, so it was not possible to learn lexical correlates of humor from transcripts. For gameplay videos, the theme is unified but the speaker has a strong accent, partly as an element of humor production, making the standard Chinese ASR system quite inaccurate; so these transcripts are barely usable. However, by combining both speech features and text-based features, our results were slightly better than using speech features alone for both video categories.

6.2. Feature analysis

To identify features most related to humor, we first calculated feature importance from the RF model, since its output has a clearer relation with its input features. The most important fea-

Table 1: *Statistics of the Bilibili corpus.*

	Number of Videos	Hours of Videos	Comments	Laughing Comments	Unsupervised Humor Labels	Unsupervised Non-humor Labels
Comedy Movies	8	13.43	63582	11521	1159	4502
Gameplay Videos	233	63.85	494438	50304	4018	23421
Total	241	77.28	558020	61825	5177	27923

Table 2: *AUC for predicted humor on test sets.*

	Speech		+ Text
	RF	CNN	CNN
Comedy Movie	0.687	0.693	**0.706**
Gameplay Video	0.719	0.742	**0.751**

Table 3: *T-test of acoustic-prosodic and transcript-based features on unsupervised humor and non-humor labels.*

Feature	Comedy Movie		Gameplay Video	
	t	p	t	p
Energy max	6.10	<0.001	27.26	<0.001
Energy mean	4.19	<0.001	30.08	<0.001
Energy stddev	8.15	<0.001	29.41	<0.001
F0 max	5.79	<0.001	17.94	<0.001
F0 mean	3.85	<0.001	19.37	<0.001
F0 stddev	4.65	<0.001	23.48	<0.001
Speaking rate	3.16	0.0015	-0.33	0.744
Human centeredness	4.17	<0.001	9.69	<0.001
Negation	4.07	<0.001	8.34	<0.001

tures in comedy movies are related to MFCCs, indicating that patterns in specific cepstral components contribute to the humor expression in the movies. In gameplay videos, this is more straightforward: the arithmetic mean, standard deviation and range of root-mean-square frame energy are among the 10 most important features, as are the range of voicing probability and standard deviation of F0.

We also performed a series of t-tests between features of segments with humor and those with non-humor unsupervised labels to better examine their relation with humor. For acoustic-prosodic features, we used the maximum, arithmetic mean, and standard deviation of the RMS frame energy and F0 in the significant tests. From Table 3, we observe an increase in value and standard deviation in both energy and F0 in humorous speech. In both video categories, energy is generally more significantly related to humor than F0. This corresponds to the humor techniques of exaggeration and bombast [20, 21, 2], where the humor producer reacts in an exaggerated way or talks in a high-flown, grandiloquent, or rhetorical manner. Moreover, in humorous gameplay video segments, all energy and F0 features show higher significance than in humorous comedy movie segments, indicating that the gameplay video creator is using humor techniques of exaggeration and bombast more frequently than the comedy movie actors. According to Buijzen and Valkenburg [2], this mode of humor production is observed most frequently in commercials aimed at adolescents, which are also the target audiences for gameplay videos.

We also calculated speaking rate from the transcripts. The t-value between the speaking rate of humor and non-humor in comedy movies is 3.16 with a p-value of 0.0015. This suggests that the speaker tends to speak quicker when expressing humor, which corresponds to humor techniques of changing speed [2]. However, the t-value of the speaking rate in gameplay videos is not significant, indicating that the video creator is not changing the speaking rate in humorous expressions although the energy and F0 are changing significantly.

For textual features, we extracted manually-chosen keywords that convey human-centeredness and negation, proven in previous works to be positively related with humor expression in one-liners and cartoon captions [4, 6, 9]. In both videos categories, we observe similar trends of using human-centeredness and negation in humorous expressions, although humor is expressed with a larger context and with more modalities.

6.3. Cross-category experiments

To verify that the classifiers trained on one category of video can be used to predict humor in different categories, we performed cross-category experiments on comedy movies and gameplay videos. When training on the comedy movies training set and testing on the gameplay videos test set, the CNN model obtained an AUC of 0.648. When training on the training set of gameplay videos and testing on the test set of comedy movies, the CNN model obtained 0.658 AUC. Both results are significantly higher than the 0.5 AUC baseline, indicating that the classifiers can capture clues for humor that generalize well to different video genres.

7. Conclusions and future research

We have presented a framework for generating unsupervised humor labels and predicting humor by learning from time-aligned comments. We collected a corpus of comedy movies and gameplay videos and obtained human annotations as gold labels for the test set. We validated our unsupervised labels on our gold labels and found a high correlation between them. We trained classifiers on speech and text-based features to obtain an AUC as high as 0.751 on the gold test set. In future, we will collect more videos from Bilibili with clear facial expressions and use visual features as well as text and speech to predict humor. Since the comments are posted toward multimedia stimuli, using all possible features should give us more insight into humor. Moreover, our framework can be applied to other live streaming websites with real-time user comments, which also may allow us to collect unsupervised labels.

8. Acknowledgment

This work was funded by DARPA LORELEI grant HR0011-15-2-0041. The views expressed in this paper however are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S government.

9. References

[1] W. H. Martineau, "A model of the social functions of humor," *The psychology of humor: Theoretical perspectives and empirical issues*, pp. 101–125, 1972.

[2] M. Buijzen and P. M. Valkenburg, "Developing a typology of humor in audiovisual media," *Media psychology*, vol. 6, no. 2, pp. 147–167, 2004.

[3] W. Ruch, "The perception of humor," in *Emotions, qualia, and consciousness*. World Scientific, 2001, pp. 410–425.

[4] R. Mihalcea and C. Strapparava, "Making computers laugh: Investigations in automatic humor recognition," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 531–538.

[5] D. Yang, A. Lavie, C. Dyer, and E. Hovy, "Humor recognition and humor anchor extraction," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2367–2376.

[6] R. Mihalcea and S. Pulman, "Characterizing humour: An exploration of features in humorous texts," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2007, pp. 337–347.

[7] Y. Raz, "Automatic humor classification on Twitter," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Association for Computational Linguistics, 2012, pp. 66–70.

[8] R. Zhang and N. Liu, "Recognizing humor on Twitter," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 889–898.

[9] D. Radev, A. Stent, J. Tetreault, A. Pappu, A. Iliakopoulou, A. Chanfreau, P. de Juan, J. Vallmitjana, A. Jaimes, R. Jha *et al.*, "Humor in collective discourse: Unsupervised funniness detection in the New Yorker Cartoon Caption Contest," *arXiv preprint arXiv:1506.08126*, 2015.

[10] A. Purandare and D. Litman, "Humor: Prosody analysis and automatic recognition for f* r* i* e* n* d* s," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 208–215.

[11] D. Bertero and P. Fung, "Deep learning of audio and language features for humor prediction," in *LREC*, 2016.

[12] ——, "A long short-term memory framework for predicting humor in dialogues," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 130–135.

[13] ——, "Predicting humor response in dialogues from TV sitcoms," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5780–5784.

[14] R. A. Martin, *The psychology of humor: An integrative approach*. Academic press, 2010.

[15] Z. Wu and E. Ito, "Correlation analysis between user's emotional comments and popularity measures," in *Advanced Applied Informatics (IIAIAAI), 2014 IIAI 3rd International Conference on*. IEEE, 2014, pp. 280–283.

[16] E. Schröger and A. Widmann, "Speeded responses to audiovisual signal changes result from bimodal integration," *Psychophysiology*, vol. 35, no. 6, pp. 755–759, 1998.

[17] J. Wang, S. Zhai, and H. Su, "Chinese input with keyboard and eye-tracking: an anatomical study," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2001, pp. 349–356.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[19] B. W. Schuller, S. Steidl, A. Batliner *et al.*, "The INTERSPEECH 2009 emotion challenge." in *Interspeech*, vol. 2009, 2009, pp. 312–315.

[20] A. A. Berger, "Anatomy of the joke," *Journal of Communication*, vol. 26, no. 3, pp. 113–115, 1976.

[21] ——, *An anatomy of humor*. Routledge, 2017.