Multimodal Indicators of Humor in Videos

Zixiaofan Yang, Lin Ai, Julia Hirschberg
Department of Computer Science
Columbia University
zy2231@columbia.edu, la2734@columbia.edu, julia@cs.columbia.edu

***Abstract*—In this paper, we propose a novel approach for generating unsupervised humor labels in videos using time-aligned user comments. We collected 100 videos and found a high agreement between our unsupervised labels and human annotations. We analyzed a set of speech, text and visual features, identifying differences between humorous and non-humorous video segments. We also conducted machine learning classification experiments to predict humor and achieved an F1-score of 0.73.**

***Keywords*-humor; multimodal analysis; multimodal dataset development;**

I. Introduction

Humor is one of the most interesting yet least-studied components in our daily communication. From a psychological perspective, humor often consists of a social context in which it occurs, a cognitive-perceptual process, an emotional response and a vocal-behavioral expression such as laughter [1]. So, to measure and predict humor, we need a context in which both the producer and the perceiver of humor are involved [2]. Computational linguists have attempted to find patterns in such humorous expressions and to build models to recognize humor. However, most work done on automatic humor prediction has focused on text data; very little research has been done on humor in multimedia contents. Our motivation is twofold: first, we are interested to learn whether speakers judged humorous share certain acoustic-prosodic and visual characteristics, and how these interact with lexical content. 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' for speech generation applications that are designed to be engaging, including advertisements.

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 quality on large amounts of data, since one major difficulty in doing research on humor is the lack of multimedia data annotated with humor. To address this problem, we propose an approach using time-aligned user comments to automatically generate unsupervised humor labels in videos. Our hypothesis is that audiences tend to respond to humor in videos with laughing comments as they find them funny, so a high volume of 'humor-related' laughing comments at a given time in a video may well indicate that the video content is humorous at that point. Using this approach, we have collected a large corpus of Chinese videos with humor labels. We then examined a set of multimodal features to identify differences between humorous and non-humorous videos scenes, and built a humor classifier using these features.

The remainder of the paper is organized as follows. We describe related work in Section II. Section III describes the Bilibili corpus we collected from a series of fast-talking satirical videos made by 'papi酱', one of the most popular Chinese online celebrities. In Section IV, we explain our approach to generate unsupervised humor labels and verify this approach using manually annotated gold labels. We analyze multimodal indicators of humor in Section I and describe our classification results in Section . Finally, we discuss our conclusions and future work in Section VII.

II. Related Work

Due to the greater ease of scraping and annotating text data, most previous work of humor prediction has been done on text. Mihalcea and Strapparava analyzed humorous and non-humorous one-liners [4]; Mihalcea and Pulman extended this work to longer news and blogs [5]. Yang et al. identified semantic structures behind humor from one-liners [6]. For humorous tweets, Raz proposed to classify funny tweets by 11 humor categories [7]. Zhang and Liu created a twitter humor dataset with hashtag and keyword search [8]. Radev et al. developed unsupervised learning methods to predict the humor ranking in The New Yorker Cartoon Caption Contest [9]. Chen and Lee built models to recognize humor in TED Talk transcripts [10]. This research found that humor in text is associated with semantic classes relevant to human-centeredness and negative polarity [4], [5], [9].

To predict humor in multimedia content such as videos, most prior research has focused on data from TV sitcoms using canned laughter as indicators of humorous scenes. Purandare and Litman examined acoustic-prosodic features of the TV sitcom 'FRIENDS' [11]. Bertero and Fung built deep learning models with text and speech features to predict humor in 'The Big Bang Theory' and 'Seinfeld' [12], [13], [14]. However, no study has shown that canned laughter actually represents the audience's perception of humor. Even

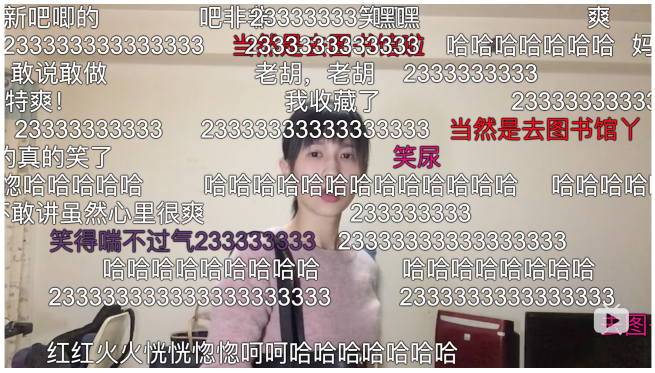

Figure 1. Screenshot of a humorous scene with multiple laughing comments containing '233' and '哈哈'. The time-aligned user comments are displayed on the video field synchronized with the scenes.

for TV sitcoms that have live audience with real laughter, there is the case where the audience is just following the instructions from the staff on when they should laugh; moreover, the producer has the full control of editing and putting the laughter at any given time during the post-production stage. Therefore, there is no guarantee that real laughter from live audience represents the audience's perception of humor neither. Trained on humorous scenes as labeled by artificial or edited laughter, the model can only learn to predict humor from the TV producer's point of view, and the model's quality depends heavily on the producer's choice of when to add the laughter. Another drawback of this approach is the limitation of genre. Models trained on the scenarios and characters of a particular TV sitcom may not generalize to other sitcoms.

III. Bilibili Corpus

For our experiments in humor prediction, we collected videos along with corresponding user comments from *bilibili.com* , one of the largest Chinese video-sharing websites. Unlike traditional video sharing websites where audiences post their comments in a given comment area under the video, *bilibili.com* allows users to make *bullet screens*, posting instant time-aligned comments about a specific scene while watching the video. When others watch the same video, previous comments from other viewers are displayed on the video as commentary subtitles, synchronized with the scenes. Figure 1 shows a screenshot of a video on *bilibili.com* . Based on findings that laughter is the most obvious expression of perceived humor [3] [15], we use laughter indicators to identify humor in our videos. Studies show that the sequence '233' is an Internet meme [1] commonly used by Chinese internet users to indicate laughter [17]. In addition, '哈哈' ('haha' in Chinese) and 'hh' are onomatopoeias of laughter and are also strongly related to perceived humor. So, by calculating the number of comments that contain '233', '哈哈' or 'hh', we can estimate the humorousness of a video scene.

We used all the videos uploaded by 'papi酱', one of the most popular Chinese online celebrities, who has millions of followers across platforms and 4 million subscribers with 296 million views on *bilibili.com* . 'papi酱' is famous for discussing trending topics in a humorous way. In most of her videos, she speaks Mandarin Chinese without any regional accent and usually faces the camera, which makes it easier for us to extract both transcript-based features and visual-based features. After filtering out videos containing dialects and advertisements, we obtained 100 videos with 93593 comments in total, including 5064 comments with '233', 7255 comments with '哈哈' and 730 with 'hh'. We randomly chose 30% of the videos as the test set, and the remaining 70% as the training set.

[1] An Internet meme is an activity, concept, catchphrase, or piece of media that spreads, often as mimicry or for humorous purposes, from person to person via the Internet.[16]

IV. Humor Labels and Annotations

We generate unsupervised humor labels by estimating user response time to a humorous scene, counting the number of laughing comments posted at each second, and performing contextual smoothing on the number of laughing comments. We then obtain human annotations on the test set to evaluate our unsupervised labels.

A. Constructing Unsupervised Labels

The unsupervised labeling method in our framework was carried out following a thorough study of user behavior when posting time-aligned comments. As discussed in Section III, we use the keywords '233', '哈哈' and 'hh' as laughter indicator. Most users do not pause the video to post time-aligned comments while watching, so most comments have time delays that we need to take into account. We divide the response time into reaction time in which users recognize humor in the videos, and typing time in which users type and post the laughing comments. In a study reported by Schröger and Widmann [18], human reaction time to audiovisual stimulus is 0.316s. For typing time, an average keystroke takes 0.2s for a skilled typist. Therefore, typing each single key character such as Roman character, number, and punctuation should take ˜0.2s. An average Chinese character can be represented by 4.2 Roman characters in pinyin [19]. However, among the 13k laughing comments we collected, 68% of the Chinese characters are '哈' which is usually represented by only 2 Roman characters. Therefore, it is reasonable to estimate that '哈' takes 0.4s to type, while other Chinese characters take 0.2*4.2 = 0.84s. Also, pressing 'enter' to post the comment takes 0.2s. Using these estimates, We calculated response time for the 13k laughing comments; the distribution histogram is plotted in Figure 2. We see that 90% of laughing comments have a reaction time within 10s. Also considering users who do pause the

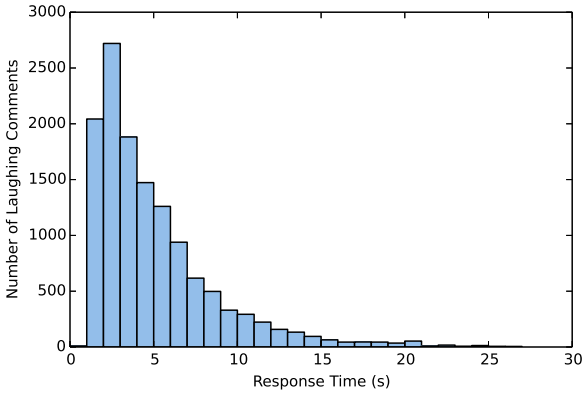

Figure 2. Histogram of response time distribution.

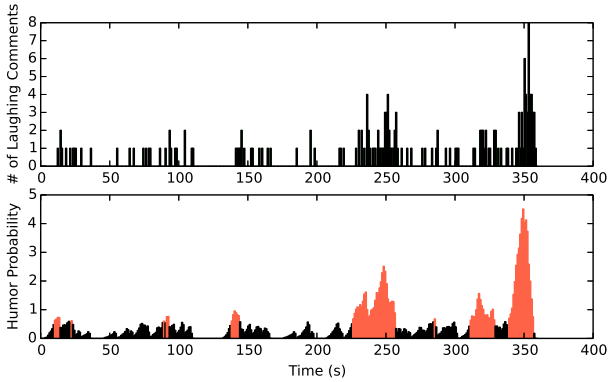

Figure 3. An example of smoothing and labeling. The upper figure shows the number of laughing comments in each one-second unit before smoothing, and the lower one shows the humor probability of each one-second unit after smoothing. Red areas are one-second units labeled as humorous and black areas are non-humorous.

video to type longer comments, we conclude that nearly all the laughing comments towards a humorous scene are posted within 10s of the scene itself. We thus normalize the number of laughing comments with response time within 10s, and treat them as the probability distribution of the time delay between the humorous scene and the laughing post. For each one-second unit in the videos, we calculate the number of laughing comments posted at each second and smooth this number out to the previous 10s according to the probability distribution of the time delay. After adding all distributed probabilities, we can obtain the probability of humor for each second. By observing the videos on the website, we found that 3 laughing comments at the same time is usually a good indicator of humor. The highest one-second bin in the probability distribution is 2-second delay with about 20%, so we set the threshold of probability of humor as 3*20% = 0.6. Using this threshold, 6508sec are labeled as humorous and 17847sec as non-humorous. Figure 3 shows an example of smoothing and labeling. The upper figure shows the number of laughing comments in each one-second unit. The lower shows the humor probability after taking into account reaction time delay, with red areas representing humorous scenes and black areas representing non-humorous. By comparing the two, we can see that the sparse comment spikes around 200s are smoothed to a lower humor probability, while the dense peak around 250s still has a high probability. In this way we can capture all peaks of laughing comments in the videos, while ignoring the portions with low agreement on humor among users. Moreover, all the peaks move forward after smoothing, indicating that humorous scenes always happen before the laughing reactions.

B. Human annotation

Our unsupervised labels are generated from user comments and thus represent users' perception of the content of the video. However, due to the varying response times, it is difficult to precisely determine which scene of the video each comment is actually reacting to. To verify our unsupervised labels, we asked human annotators to annotate the test set with 30 randomly selected videos. Three native Chinese annotators were asked to watch the videos without seeing the time-aligned comments and to label each second as humor/non-humor. Both the average Cohen's kappa and Fleiss' Kappa for these annotators' annotations are 0.65, indicating a substantial inter-annotator agreement. We then obtained gold labels on the test set by taking the majority vote of all three annotators. The accuracy when we compare unsupervised labels to gold labels is 0.78, which is high enough to conclude that our unsupervised labeling method can generate humor labels on multimedia content with an accuracy comparable to human annotation.

V. Multimodal Indicators of Humor

In this section we present results of our analysis of acoustic-prosodic, transcript-based, and visual features. We performed a series of t-tests between features of scenes with humor and those with non-humor unsupervised labels (Table I).

A. Acoustic-Prosodic Features

We first converted all videos in the corpus to audio files sampled at 44.1kHz and 16 bits per sample. Some acoustic-prosodic features such as pitch and energy have already proven to be relevant to the expression of humor in TV sitcom [11]. Therefore, we computed the root-mean-square (RMS) frame energy and fundamental frequency (F0) with 25ms frame length and 10ms stride using the openSMILE toolkit[20]. For each one-second unit, we used a context window of five seconds and extracted the maximum, standard deviation and arithmetic mean value of the RMS frame energy and F0. These features and t-values are shown

Feature	t	p
Energy stddev	24.19	p<0.001
Energy mean	23.02	p<0.001
F0 mean	22.11	p<0.001
Energy max	21.46	p<0.001
F0 stddev	19.59	p<0.001
F0 max	12.00	p<0.001
Speaking rate	-13.94	p<0.001
Human centeredness	-3.74	p<0.001
Negation	-6.72	p<0.001
SSIM max	-6.79	p<0.001
SSIM min	3.72	p<0.001
SSIM mean	-2.76	p=0.006

Table I
T-TEST OF ACOUSTIC-PROSODIC, TRANSCRIPT-BASED, AND VISUAL FEATURES ON UNSUPERVISED HUMOR AND NON-HUMOR LABELS.

in Table I. All listed features are significant with $p<0.001$. From Table I, we observed an increase value and an increase in standard deviation in both energy and F0 in humorous speech. However, energy is generally more significantly related to humor than F0. The standard deviation of energy is the most significant indicator for humor, representing sudden and large changes of speech energy. This corresponds to the humor techniques of exaggeration and bombast [21] [22] [23], where the humor producer reacts in an exaggerated way or talks in a high-flown, grandiloquent, or rhetorical manner.

B. Transcript-based Features

To obtain speech transcripts aligned at character-level with the audios, we used the automatic speech recognition (ASR) function for Mandarin Chinese in the Google Speech API. Most videos have been speeded up by the video creator, so to improve the ASR performance, we slowed the audios down to 0.75 times their original speed before passing them to Google Speech API. We also normalized the audios on energy and F0 to reduce the effect of exaggerated emotions.

Using the automatic transcripts, we first computed speaking rate, another acoustic-prosodic feature, by calculating the number of Chinese characters in each second. Since the videos are speeded up, there can be as many as twelve characters in one second. The t-value between the speaking rate of humor and non-humor is -13.94 ($p<0.001$), indicating an increase in speaking rate in non-humorous segments. This suggests that the speaker tended to speak slower when expressing humor, which corresponds to humor techniques of exaggeration [21] [22] and changing speed [23]. The videos are speeded up from normal speech in post-processing, also indicating the humor technique of changing speed [23].

We also extracted keywords with human-centeredness and negation, which have proven in previous work to be positively related with humor expression in one-liners and cartoon captions [4] [5] [9]. However, in our corpus the t-values are -3.74 and -6.72 (both $p<0.001$), which suggests that the humor expression in multimedia is substantially different with humor in one-liners and captions. One possible reason is that in both one-liners and cartoon captions, the writers have to express humor in one short sentence; while in multimedia videos, humor happens in a larger context. The creator can set up for a punchline more slowly using several sentences, so that the punchline itself does not need to contain those keywords shown in one-liners and captions.

C. Visual Features

We extracted the frame difference as a visual feature, since it may capture features such as change of scenes and large body movements of the speaker in the videos. Camera positions and scenes do not change frequently in our corpus, so the frame-by-frame differences are too small to be significant. Therefore, we calculated differences every 5 frames using the structural similarity (SSIM) index and computed the *extremum* and arithmetic mean of the SSIM scores in each second. Since SSIM measures the perceptual similarity between images, the higher the SSIM scores are, the less different the frames are. Extremely low SSIM usually indicates change of scenes and extremely high SSIM indicates that the speaker is relatively still. As shown in Table I, the t-value of maximum SSIM is -6.79 and t-value of minimum SSIM is 3.72 (both $p<0.001$). This suggests that the SSIM tends to be not too high and also not too low in humorous segments, so that the speaker can have some movement while the background scene is kept still. This finding correlates with the humor techniques of clownish behavior (making vigorous arm and leg movements or demonstrating exaggerated irregular physical behavior) and peculiar facial expression (making a funny face or grimace) [23].

VI. CLASSIFICATION RESULTS

Motivated by our analysis showing significant differences between humorous and non-humorous scenes, we trained machine learning classifiers to predict humor. As described in Section IV, 30% of the videos with 7398s in total were randomly selected as the test set and manually annotated for humor and the rest 70% of the videos with 16957s were used as the training set which was annotated in an unsupervised manner as described above..

For speech features, we extracted acoustic-prosodic features using the openSMILE toolkit's baseline set introduced in the INTERSPEECH 2009 Emotion Challenge [20] [24]. This feature set includes 384 features, such as the extremum, standard deviation and mean value of the root-mean-square frame energy, the fundamental frequency (F0), the zero-crossing rate of frame-based time signal, the voicing probabilty and the mel-frequency cepstral coefficients (MFCCs). For text features, 'Jieba' Chinese text segmentation package [2] was used to segment the transcripts into words. We applied

[2]https://github.com/fxsjy/jieba

a TF-IDF transformation on the words to generate n-gram features. We also added speaking rate and SSIM scores, since they are analyzed in Section V to be related with humor. The machine learning classifier we use is a random forest (RF) classifier with 500 estimators and the best F1-score that we have obtained on our hand-annotated test set is 0.73.

VII. Conclusions and Future Work

We have presented a framework for generating unsupervised humor labels in videos and analyzed a set of multimodal indicators of humor. We collected a corpus with 100 videos and obtained human annotations as gold labels for the test set. The high correlation between the unsupervised labels and the gold labels verified our labeling method. Our analysis of acoustic-prosodic features, transcript-based features and visual features provides insight into understanding the expression of humor in multimedia contents. Finally, we trained classifiers to automatically predict humor which achieved an F1-score of 0.73.

In future, we plan to collect more videos from different video creators, so that we can explore a larger variety of characteristics in humor expression which might generalize better to other genres of humor expression. We also plan to add more visual features, for example, facial landmarks and gestures. Since the user comments are posted toward multimedia stimuli, utilizing all possible features should give us more insight into the many dimensions of humor expression. In addition, a question that needs further investigation is whether, since users can see previous comments from other viewers, some users might be responding to each other by posting comments containing '233', '哈哈' or 'hh'. Therefore, it is possible that not all laughing comments actually address the humorousness of scenes in the videos. In this case, the annotation process might be affected, and this is an issue for future research to explore. Moreover, our framework can be applied to data collected from other live streaming websites. Using keywords related to other kinds of user reaction, we may be able to obtain unsupervised labels on even larger amounts of data in the same way.

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.

References

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

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

[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] 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.

[6] 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.

[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] L. Chen and C. M. Lee, "Convolutional neural network for humor recognition," *CoRR*, 2017.

[11] 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.

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

[13] ——, "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.

[14] ——, "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.

[15] G. E. Weisfeld, "The adaptive value of humor and laughter," *Evolution and Human Behavior*, vol. 14, no. 2, pp. 141–169, 1993.

[16] K. Schubert, "Bazaar goes bizarre," *USA Today*, 2003.

[17] 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.

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

[19] 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.

[20] 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.

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

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

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

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