CHoRaL: Collecting Humor Reaction Labels from Millions of Social Media Users

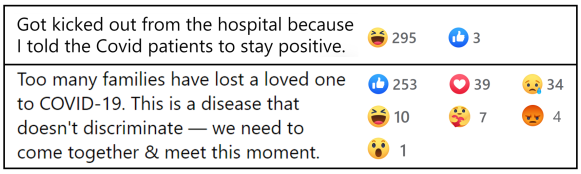
Zixiaofan Yang and **Shayan Hooshmand** and **Julia Hirschberg**
Department of Computer Science
Columbia University
New York, NY, USA
{zy2231,shayan.hooshmand}@columbia.edu, julia@cs.columbia.edu

Abstract

Humor detection has gained attention in recent years due to the desire to understand user-generated content with figurative language. However, substantial individual and cultural differences in humor perception make it very difficult to collect a large-scale humor dataset with reliable humor labels. We propose **CHoRaL**, a framework to generate perceived humor labels on Facebook posts, using the naturally available user reactions to these posts with no manual annotation needed. CHoRaL provides both binary labels and continuous scores of humor and non-humor. We present the largest dataset to date with labeled humor on 785K posts related to COVID-19. Additionally, we analyze the expression of COVID-related humor in social media by extracting lexico-semantic and affective features from the posts, and build humor detection models with performance similar to humans. CHoRaL enables the development of large-scale humor detection models on any topic and opens a new path to the study of humor on social media.

1 Introduction

Humor is ubiquitous — it forms a crucial part of people's lives both online and off. Automatically detecting humor, then, has become an important task, with applications from misinformation to advertising to philosophy. From a psychological perspective, humor represents anything people say or do that others perceive as funny and tends to make them laugh (Martin, 2010). Humor perception, though, is highly individualistic (Ruch, 2001), making it hard to reliably annotate humor.



Got kicked out from the hospital because I told the Covid patients to stay positive.	😆 295	👍 3	
Too many families have lost a loved one to COVID-19. This is a disease that doesn't discriminate — we need to come together & meet this moment.	👍 253 😆 10 😮 1	❤ 39 🤗 7	😢 34 😡 4

Figure 1: User reactions to a humorous Facebook post (top) and a non-humorous post (bottom).

Researchers have proposed various methods to collect humorous and non-humorous data with minimal annotation needed. Most attempts have focused on distinguishing between jokes and news, which both have natural labels on humor and can be scraped automatically. This large stylistic difference makes detecting humor easier — but it is far from most real-world scenarios where humorous and non-humorous texts come from the same domain. Another technique collects social media posts by humor- and non-humor-related hashtags, but this method suffers from large data noise and low labeling accuracy (Zhang and Liu, 2014). Finally, there have been studies to use the number of Reddit upvotes as humor labels (Weller and Seppi, 2019, 2020). Though this technique sources data from the same domain, that domain is too limited in scope: all the data comes from one single subreddit. This specificity means that the data represents only the humor perception of a particular group of Reddit users, dedicated to producing witty jokes.

To address these problems of specificity and domain discrepancy in humorous data collection, we propose **CHoRaL**, a framework for **C**ollecting **H**umor **R**eaction **L**abels. CHoRaL generates perceived humor scores using the naturally available reactions on Facebook posts. Our framework includes several advantages: (1) labeling humor on any Facebook post, without the need for extra human annotations; (2) providing both binary labels and continuous scores for humor and non-humor; (3) enabling the collection of large-scale social media datasets on humor.

We use CHoRaL to present the largest dataset to date on humor, containing 785K Facebook COVID-19 related posts, each assigned a humor score. We chose to focus on COVID-19 because of its universality as a phenomenon that affects all Facebook users. CHoRaL, however, can be easily adapted to other topics, making it the most extendable humor

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4429–4435
November 7–11, 2021. ©2021 Association for Computational Linguistics

data collection framework yet.

2 Related Work

Most corpora for textual humor detection use online joke compilations as humor data and more serious sources, like news or proverbs, as non-humor data. Mihalcea and Strapparava (2005) built a model to distinguish one-liners from short sentences such as news titles, and Mihalcea and Pulman (2007) extended the work to longer humorous articles and news articles. Yang et al. (2015) identified the semantic structures of humor by studying the differences between puns and news. Chen and Soo (2018) built deep learning humor detection models on four datasets with jokes as humor data and news as non-humor data. Blinov et al. (2019) collected jokes in Russian, combining with forum posts that have low similarity to the jokes as non-humorous samples. More recently, Annamoradnejad and Zoghi (2020) combined Reddit jokes with news headlines and used a BERT-based model to classify these two sets of data.

For other forms of naturally labeled humorous texts, Reyes et al. (2012) obtained humorous tweets with the hashtag "humor" and non-humorous tweets from other hashtags. Radev et al. (2016) obtained humor scores from a cartoon caption contest, and, similarly, Potash et al. (2017) obtained humorous tweets from the official website of a TV show. Chen and Lee (2017); Hasan et al. (2019) generated humor labels using the audience laughter marker in the transcripts of TED talks. Hossain et al. (2019, 2020) asked annotators to edit news headlines to make them funny. There are also some hand-annotated humor datasets (Chiruzzo et al., 2020; Zhang and Liu, 2014). However, these methods either need extensive human annotation or suffer from low label accuracy.

The line of work most relevant to our paper is the rJokes dataset (Weller and Seppi, 2019, 2020), where humor scores are obtained from the number of upvotes toward each post in the r/Jokes subreddit. However, all the posts in the subreddit are intended to be jokes, making the dataset include only successful jokes and failed jokes, which is far from the natural distribution of posts in social media.

For multimodal humor detection, researchers used canned laughter in TV sitcoms (Purandare and Litman, 2006; Bertero and Fung, 2016a,b,c), and time-aligned comments in online videos (Yang et al., 2019a,b). Multimodal humor is also examined in internet memes (Chauhan et al., 2020; Sharma et al., 2020).

3 CHoRaL Framework and Dataset

In this section, we introduce our Facebook post collection process, as well as our algorithm to assign humor and non-humor scores to the posts. Although CHoRaL can be applied to any topic, we chose COVID-19 as the topic for our dataset. There has been extensive discussion on the pandemic with a wide range of audiences, so this topic prevents us from biasing our posts and labels toward a specific demographic group.

3.1 Data Collection and Cleaning

We collected our Facebook posts from CrowdTangle by searching COVID-related keywords ("covid-19, coronavirus, corona, covid 19, sars-cov-2, covid, sars cov 2"), and downloading posts from January 20th, 2020 until March 18th, 2021. We set the language as English and post type as Status on CrowdTangle, in order to ensure that we retrieve text-only posts without images or videos attached. This initial retrieval surfaced 2 million posts.

We further cleaned these 2 million downloaded posts locally. We removed posts with duplicate text fields and some remaining non-English posts. We also removed posts with rendered links to minimize the influence of non-text elements on the viewers' perception of humor. For posts with non-rendered links, we replaced the links with a special token. This replacement allowed more posts to pass our final filter, which was to cap post length at 500 characters to suit the max token length of BERT-based models. About 785K posts remained in our corpus after this local filtering round.

3.2 Defining the Humor Score (HS)

We used Facebook's built-in reactions feature to determine how funny a post is in the perception of users. Our assumption is that the higher the Haha percentage among all reactions, the more humorous the post. An example of a post with a high percentage of Haha reactions (laughing face) is shown at the top of Figure 1.

Of course, the fewer the total reactions in a post, the less confidence we had in conclusions drawn from its reaction distribution. So, we also discounted unpopular posts with a tanh multiplier proportional to the total number of reactions. The multiplier is stretched by 50, so that posts with about

100 total reactions or more are similarly weighted, while there is a steep decline in weighting as total reactions approach zero. The following formula summarizes our Humor Score (HS):

$$\text{HS} = \frac{h}{t} * \tanh(\frac{t}{50}) \quad (1)$$

where h = number of haha reactions, t = total number of reactions, and 50 is used as our popularity stretcher.

3.3 Defining Non-Humor Score (NS)

Besides finding humorous posts using HS, we also want to retrieve non-humorous negative samples for building a binary humor detection model.

Intuitively, it makes sense to use those posts with the lowest HS as non-humorous data. But these posts that have an extremely low Haha percentage also represent too extreme of an opposite to humor — for COVID-related posts, this opposite turns out to be almost exclusively sad posts about people's deaths and illness. Though sad posts are certainly non-humorous, they don't represent the full scope of non-humorous expression. Thus, we need a new technique to retrieve a broader range of non-humorous posts, which should include neutral posts, sad posts, as well as other emotional posts that do not evoke a humorous reaction.

We instead define our Non-Humor Score (NS) as posts whose reaction distributions have the lowest divergence from the standard Facebook post distribution. Given the fact that the vast majority of posts have a very low HS, we assume that standard Facebook posts are non-humorous, as the example shown at the bottom of Figure 1. To use our NS, we first average the distribution of reactions over our 785K cleaned posts. Then, for a new post, its NS is defined as the negative log of the mean-squared error between its reaction distribution and the averaged distribution. Thus, a higher NS indicates a lower divergence. We also include a tanh popularity multiplier for the same reasons as above. The following formula summarizes our NS:

$$\text{NS} = -\log(\tanh(\frac{t}{50}) * \sum_{r \in R} \frac{(S(r) - O(r))^2}{|R|}) \quad (2)$$

where t = total number of reacts, R = the set of Facebook reactions, S maps a reaction to its percentage in the standard distribution, and O does the same with respect to the observed post.

# of Posts	784,965
# of Poster Accounts	264,685
# of User Reactions	126,839,984
# of Haha Reactions	6,525,247

Table 1: Statistics of the dataset.

4 Humor Analysis

Table 1 shows the summary of our dataset with 785K posts posted by 265K accounts. There are a total of 149M user reactions and 6M of them are Haha reactions, which we use as indicators of humor. To better understand the expression of humor, we performed lexico-semantic and affective analysis by extracting lexicon-based features from the posts, aiming for explainable results. We used Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) for lexico-semantic analysis; for affective content, we used the Revised Dictionary of Affect in Language (DAL) (Whissell, 2009) and the Vader sentiment tool (Hutto and Gilbert, 2014); we also analyzed the complexity of posts, and the use of emojis as a social media specific feature. All word-level features were normalized by the total number of words after using the Twitter-aware tokenizer of the NLTK Toolkit (Bird, 2006). We calculated Pearson's correlation between the features and the HS of posts, and all reported results are significant with a $p < 0.05$.

LIWC The top categories that positively correlate with HS include singular first-person pronouns, total pronouns, anger words, negative emotional words, and negations. This agrees with previous findings that humorous texts have more negative polarity and human-centeredness (Mihalcea and Strapparava, 2005; Radev et al., 2016). Also among the top 10 categories are informal words, swear words, and sexual words, which correspond to the characteristics of humorous posts on social media. On the other hand, there are fewer word categories that negatively correlate with HS, indicating that serious posts share less lexical similarity. Some negatively correlated categories are relativity words related to space and time, possibly suggesting that humorous posts have a less detailed writing style.

Affect and sentiment To further investigate the affective component found to be related to humor in previous work (Reyes, 2013; Mahajan and Zaveri, 2020), we computed average activation, imagery, and pleasantness scores for each post using the DAL lexicon and sentiment scores using the Vader tool. Both imagery and pleasantness scores in DAL, as well as the sentiment score in Vader, are negatively correlated with humor, indicating a more abstract and negative style in humorous posts, which agree with the LIWC findings.

Complexity We computed the percentage of longer words (more than 6 characters), percentage of complex words defined by the Dale–Chall readability formula (Chall and Dale, 1995), and the Flesch reading ease test (Flesch and Gould, 1949) for a readability measurement. All features show that humorous posts have lower complexity.

Emoji We found the number of emojis in a post to be a humor indicator. Specifically, 363 of the 1,621 unique emojis in our dataset are significantly correlated with HS (320 positive, 43 negative), with the "Face with Tears of Joy" emoji having the highest humor correlation. Interestingly, humorous posts have generally fewer heart emojis, but more broken heart emoji, echoing our results above that negative sentiment is related to humor.

5 Humor Detection Experiments

Due to the naturally imbalanced distribution of humorous posts in social media, our full dataset skews towards posts with low HS and high NS. To address this imbalance and build humor detection models, we used the 20K posts with the highest HS as positive samples and the 20K posts with the highest NS as the negative samples on humor. We randomly split the 40K posts into training and test sets, respectively consisting of 80% and 20% of the data, and balanced by binary humor labels.

Pretrained language models such as BERT have shown great success when fine-tuned for text classification tasks (Devlin et al., 2019; Sun et al., 2019), including the task of humor detection (Wang et al., 2020; Annamoradnejad and Zoghi, 2020). In our experiments, we fine-tuned 3 pre-trained language models on our CHoRaL dataset: RoBERTa-base (Liu et al., 2019), a BERT-style model pre-trained on 160GB of text data including Wikipedia, news, and other web texts; BERTweet (Nguyen et al., 2020), a model with BERT-base architecture, pre-trained using the RoBERTa procedure but on 845M English Tweets; BERTweet-covid, based on BERTweet but further pre-trained on 23M COVID-related Tweets. We trained the models in two settings: continuous regression, where continuous HS is used as ground truth of humor; and binary classification, where high HS posts have a positive label, and the high NS posts have a negative label. All models were fine-tuned for 3 epochs on the training set with a learning rate of 2e-5. To compare the model performance with humans, we asked 3 native English speakers to label 100 random and balanced posts from the test set. The inter-annotator agreement in Fleiss' kappa is 0.782. Note that due to the potential differences of humor perception between our annotators and general Facebook users, the labels provided by annotators were used not as gold labels, but as a baseline for our models. To compare the continuous models with humans directly, we used an empirical threshold of 0.18 HS to convert the predictions into binary labels.

	Continuous		Binary	
	F1	AUC	F1	AUC
Human	-	-	0.867	-
RoBERTa	0.869	0.939	0.868	0.937
BERTweet	0.879	0.947	0.881	0.950
BERTweet-covid	0.880	0.948	**0.883**	**0.951**

Table 2: Humor detection results.

Table 2 shows the humor detection results on the test set, measured by binary F1-score and Area Under Curve (AUC). First, all models have comparable F1 with human annotators, validating our idea of automatically learning crowd-sourced humor from millions of users. Comparing the different models, we found that both models pre-trained on Tweets outperform RoBERTa, and that BERTweet-covid, with further adaption to the COVID-19 topic, is slightly better than the original BERTweet. This finding suggests that the pre-training domain is quite important in detecting figurative language. Moreover, training on binary labels given by both HS and NS is generally better than training on HS exclusively, indicating the effectiveness of NS to provide additional information on non-humor.

6 Conclusions and Future Work

In this paper we present the CHoRaL framework for automatically collecting humor reaction labels, and the dataset including 785K posts with humor and non-humor scores. We also perform analysis on humor expressions in our dataset and build models to detect humor with performance comparable to human labelers. CHoRaL enables the development of humor detection models on any topic, and our dataset has the potential to help broader applications, such as distinguishing malicious misinformation posts and non-malicious humorous posts. Furthermore, CHoRaL can also be used to label other human reactions such as anger and sadness.

Ethical Considerations

All posts and reactions used in this work are from publicly available Facebook pages, and we also gained permission from CrowdTangle, a public insights tool owned and operated by Facebook, to exhibit the post examples in the paper. We did not collect or use any personal information from the Facebook users, and our 3 annotators were voluntary participants who were aware of any risks of harm associated with their participation. Since our data were collected from Facebook with a popularity stretcher, our humor analysis results and humor detection models may be biased towards English-speaking populations that are more active on social media. We tried our best to retrieve posts with as broad population coverage as possible, while maintaining the effectiveness of our humor and non-humor scores. By our inspection, we have not noticed any trend of malicious or discriminatory posts in our dataset. Because of the sheer size of our dataset, however, we cannot guarantee that no such posts exist. We will share the data and labels freely with academia; we do not, however, endorse the views expressed in the posts and the scores automatically generated according to the user reactions.

References

Issa Annamoradnejad and Gohar Zoghi. 2020. Colbert: Using bert sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*.

Dario Bertero and Pascale Fung. 2016a. Deep learning of audio and language features for humor prediction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 496–501, Portorož, Slovenia. European Language Resources Association (ELRA).

Dario Bertero and Pascale Fung. 2016b. 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*, pages 130–135, San Diego, California. Association for Computational Linguistics.

Dario Bertero and Pascale Fung. 2016c. Predicting humor response in dialogues from tv sitcoms. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5780–5784. IEEE.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large dataset and language model fun-tuning for humor recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4027–4032, Florence, Italy. Association for Computational Linguistics.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. All-in-one: A deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 281–290, Suzhou, China. Association for Computational Linguistics.

Lei Chen and Chong MIn Lee. 2017. Convolutional neural network for humor recognition. *arXiv preprint arXiv:1702.02584*.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Luis Chiruzzo, Santiago Castro, and Aiala Rosá. 2020. HAHA 2019 dataset: A corpus for humor analysis in Spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5106–5112, Marseille, France. European Language Resources Association.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Rudolf Flesch and Alan J Gould. 1949. *The art of readable writing*, volume 8. Harper New York.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rutal Mahajan and Mukesh Zaveri. 2020. Humor identification using affect based content in target text. *Journal of Intelligent & Fuzzy Systems*, (Preprint):1–12.

Rod A Martin. 2010. *The psychology of humor: An integrative approach*. Academic press.

Rada Mihalcea and Stephen Pulman. 2007. Characterizing humour: An exploration of features in humorous texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 337–347. Springer.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report, University of Texas at Austin.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57, Vancouver, Canada. Association for Computational Linguistics.

Amruta Purandare and Diane Litman. 2006. 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*, pages 208–215, Sydney, Australia. Association for Computational Linguistics.

Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. 2016. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 475–479, Portorož, Slovenia. European Language Resources Association (ELRA).

Antonio Reyes. 2013. Linguistic-based patterns for figurative language processing: the case of humor recognition and irony detection. *Procesamiento del Lenguaje Natural*, 50:107–109.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Willibald Ruch. 2001. The perception of humor. In *Emotions, qualia, and consciousness*, pages 410–425. World Scientific.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. 2020. 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, Lisboa, Portugal. European Association for Machine Translation.

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. Association for Computational Linguistics.

Orion Weller and Kevin Seppi. 2020. The rJokes dataset: a large scale humor collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6136–6141, Marseille, France. European Language Resources Association.

Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological reports*, 105(2):509–521.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Zixiaofan Yang, Lin Ai, and Julia Hirschberg. 2019a. Multimodal indicators of humor in videos. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 538–543. IEEE.

Zixiaofan Yang, Bingyan Hu, and Julia Hirschberg. 2019b. Predicting humor by learning from time-aligned comments. *Proc. Interspeech 2019*, pages 496–500.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 889–898. ACM.