Crossing the Line: Where do Demographic Variables Fit into Humor Detection?

J. A. Meaney
School of Informatics
University of Edinburgh
Edinburgh, UK
jameaney@ed.ac.uk

Abstract

Recent shared tasks in humor classification have struggled with two issues: scope and subjectivity. Regarding scope, many task datasets either comprise a highly constrained genre of humor which does not broadly represent the genre, or the data collection is so indiscriminate that the inter-annotator agreement on its comic content is drastically low. In terms of subjectivity, these tasks typically average over all annotators' judgments, in spite of the fact that humor is highly subjective and varies both between and within cultures. We propose a dataset which maintains a broad scope but which addresses subjectivity. We will collect demographic information about the data's humor annotators in order to bin ratings more sensibly. We also suggest the addition of an 'offensive' label to reflect the fact a text may be humorous to one group, but offensive to another. This would allow for more meaningful shared tasks and could lead to better performance on downstream applications, such as content moderation.

1 Introduction

Interest in computational humor (CH) is flourishing, and since 2017, the proliferation of shared humor detection tasks in NLP has attracted new researchers to the field. However, leading researchers in CH have bemoaned the fact that NLP's contribution is not always informed by the long and interdisciplinary history of humor research (Taylor and Attardo, 2016) (Davies, 2008). This may result in the creation of humor detection systems which produce excellent evaluation results, but which may not scale to other humor datasets, improve downstream tasks like content moderation, or contribute to our understanding of humor.

A central issue is the conception of humor classification tasks as humor-or-not, similar to image classification's view of an image as dog-or-not. However, while one can be an expert in whether or not an image depicts a dog, and this is stable within and between cultures, humor is more nuanced than that. Unlike image classification:

- Humor differs *between* cultures. Even within the same language, different nationalities perceive jokes differently. This is particularly relevant to stereotyped humor, which may be perceived as funny to one culture, but offensive to another. (Rosenthal and Bindman, 2015)
- Humor differs *within* cultures. Age, gender and socio-economic status are known to impact what is perceived as humorous. (Kuipers, 2017)
- Humor differs within the same person. Mood is thought to impact what is considered to be humorous or not. (Wagner and Ruch, 2020)

Currently in NLP shared tasks, there is scant admission of these issues. Humor is treated as a stable target, and humorous texts are subjected to binary classification and humor score prediction, with little recognition that gold standard labels for these constructs simply do not exist.

1.1 Proposal

To the extent that humor is multi-faceted, and subject to multiple interpretations, incremental improvements to shared tasks can be made by:

- Acknowledging that texts may not be perceived as humorous by all readers, and allowing for a different interpretation, e.g. offensive.
- Collecting demographic information about the annotators of humor datasets to learn more about which sectors of society find a text humorous versus offensive.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 176–181
July 5 - July 10, 2020. ©2020 Association for Computational Linguistics

1.2 Why Offensive as an Alternative Label?

Cultural shifts in many parts of the world have seen a decline in racist and sexist jokes, and the growth of humor that acknowledges marginalized people. Lockyer and Pickering (2005) argue that this is not just a recent phenomenon, but that all pluralist societies navigate the space between humor and offensiveness, between 'free speech and cultural respect'

Despite the shift away from using racist or sexist comments as humor, offensive language is still plentiful on the internet (Davidson et al., 2017), (Nobata et al., 2016). This can reinforce racial stereotypes, or have a damaging impact on communities. In light of the fact that many shared tasks source their data online, either by scraping Twitter, Reddit, or crowdsourcing, we believe it is worth capturing the impact of these texts on users.

1.3 Why Demographic Factors?

Studies as far back as 1937 demonstrate gender and age differences in the appreciation of jokes, where young men gave higher ratings to 'shady' (e.g. sexual) jokes than their female, and older counterparts did (Omwake, 1937).

More recently, in the Netherlands, Kuipers (2017) found significant differences in humor preferences along the lines of gender, age, and in particular, social class or education level. An interesting finding was that the older generation rated their younger counterparts' humor as offensive. This contradicts the popular opinion that the millennial generation is perpetually offended (Fisher, 2019).

In terms of gender-specific offensive humor, a US study found that males tended to give higher ratings to female-hostile jokes, and females did the same with male-hostile jokes. Both genders found female-hostile jokes more offensive overall (Abel and Flick, 2012).

The body of work from CH on demographic differences in humor perception is absent in current work, but can be incorporated into shared tasks with some simple adjustments.

2 Previous Work

SemEval 2017 posed two humor detection tasks. Task 7 (Miller et al., 2017) covered puns, which we do not include here as the identification/interpretation of puns is less ambiguous than other forms of humor, except in the case that the audience does not possess the tacit linguistic knowledge required to understand them (Aarons, 2017).

2.1 Limited Scope

Task 6, Hashtag Wars (Potash et al., 2017), sourced its name and data from a segment in the Comedy Central Show @Midnight with Chris Hardwick, which solicited humorous responses to a given hashtag from its viewers, submitted on Twitter. These submissions were effectively annotated twice: the producers selected ten tweets as most humorous, and most appropriate for the show's type of humor. The show's audience then voted on their number one submission. Task 1 was to pair the tweets, and for each pair, predict which one had achieved a higher ranking, according to the audience. Task 2 was to predict the labels given by this stratified annotation: submitted but not top-10, top-10, number one in top-10.

The task's organisers highlighted the data's limited scope, and were keen to point out that this task does not aim to build an all-purpose, cross-cultural humor classifier, but rather to characterise the humor from one source - the show @Midnight. This task's dual annotation and ecologically valid task make it arguably one of the most effective humor challenges in recent years. However, it remains to be seen how well a system built on this data would generalize to another humor detection task.

Semeval 2020 featured another humor challenge with two subtasks: predicting the mean funniness rating of each humorous text, and given two humorous texts, predicting which was rated as funnier (Hossain et al., 2019). Instead of collecting previously existing humorous texts, the organisers generated them by scraping news headlines from Reddit, and then paying crowdworkers to edit the headlines to make them funny, and annotators to rate the funniness of the new headlines.

Edits were defined as 'the insertion of a single-word noun or verb to replace an existing entity or single-word noun or verb'. The annotators rated the headline as funny from 0-4. An abusive/spam option was included, but presumably to discard ineffective edits, rather than highlight a text which would cause offense. Nonetheless, inter-annotator agreement between raters was moderately high, (Krippendorff's α 0.64)

Of interest to CH research is that the authors' analysis of the generated humor finds support for established humor theories, such as incongruity, superiority and setup and punchline being central to the this task. However, the editing rules enforced such tight linguistic constraints that many common features of language were not permitted, e.g. the use of named entities with two words, phrasal verbs, even apostrophes. This scales down the humor that can be generated, not in terms of genre, as was the case with the 2017 SemEval task, but rather in terms of arbitrary linguistic constraints.

Finally we must consider that, given that the humorous texts were presented alongside the original headline, it's possible that affirmative humor ratings do not mean that the text is humorous in and of itself, only that it is funnier than the contemporary news — arguably a low bar in the current climate.

2.2 Unlimited Scope

The HAHA challenge (Humor Analysis based on Human Annotation) has run in 2018 (Castro et al., 2018) and 2019 (Chiruzzo et al., 2019) with two subtasks: binary classification of humor, and prediction of the average humor score assigned to each text.

The data were collected from fifty Spanish-speaking Twitter accounts which typically post humorous content, representing a range of different dialects of Spanish. These were then uploaded to an online platform, which was open to the public who were asked the following questions to annotate the data:

1. Does this tweet intend to be humorous? (Yes, or No)
2. [If yes] How humorous do you find it, from 1 to 5?

A strength of this annotation process is that the first question allows the user to objectively identify the genre of the text by identifying its intention, before giving their subjective opinion of it. However, the inter-annotator agreement for the second question was extremely low (Krippendorf's α of 0.1625). It's possible that sourcing the texts from fifty different accounts introduced too many genres to gain a consensus about what was funny amongst annotators. Similarly, the organizers targeted as many different Spanish dialects as possible in their data collection, which could lead to cultural and linguistic differences in humor appreciation. Finally, the annotations were sourced on an open platform, with only three test tweets to assess whether an annotator provided usable ratings or not. There were no questions as to whether the user was a Spanish speaker, and as the task was unpaid, there may have been little incentive to do it accurately.

3 Methodology

The datasets featured in both SemEval tasks had tight constraints on the genre of humor involved. This led to high inter-annotator reliability, but may not generalize well to other forms of humor. The Spanish tasks featured no such constraints, however, there was extremely low inter-annotator agreement, suggesting that the dataset is noisy, and that a system which is built on this may also fail to generalize.

This proposal aims to include a wide range of genres, and to increase the reliability of the annotations by collecting information on well-known latent variables in humor appreciation — the demographic characteristics of the humor audience/annotators. This will allow for more nuanced tasks, as an alternative to simple humor-or-not definitions.

3.1 Data Collection

We plan to follow a similar data collection protocol to (Castro et al., 2018) and collect tweets from a wide variety of humorous Twitter accounts. However, unlike Castro et al., we plan to limit the dialect of the jokes collected to US English, and use a crowdsourcing platform which allows us to select annotators who use this dialect. This will help us to avoid introducing confounds such as lack of cultural knowledge, or divergent language usage. Furthermore, we will hand select the Twitter accounts which typically post humorous content, in order to ensure that the data features a wide variety of genres of humor, e.g. observational humor, wordplay, humorous vignettes, etc.

3.2 Annotation

As mentioned above, averaging over the opinions of the audience, similar to approaches in image detection is not ecologically valid for humor detection. For this reason, we plan to collect demographic information about the annotators, in order to bin the ratings into groups that may perceive humor in a similar way. In this way, we hope to increase inter-annotator reliability. We also plan to include a second label for each text — offensive.

Following Castro et al., annotators will be asked the following questions for each text:

1. Is the intention of this text to be humorous?
2. [If so] How humorous do you perceive this text to be?
3. Is this text offensive?
4. [If so] How offensive do you perceive this text to be?

The annotator guidelines will reflect that offensiveness can encompass an insult to the audience itself, or to others who are likely to find the text distasteful.

All annotators will be paid for their work, to incentivize quality ratings. They will be selected to undertake the task by virtue of fitting into the following demographic bins:

- Age: 18-25, 26-40, 41-55, 56-70 the bins are broadly designed to capture Generation Z, Millenials, Generation X and Baby Boomers respectively (Dimock, 2019).
- Gender: Male, Female, Non-binary
- Level of Education: High School, Undergraduate, Postgraduate. This will be used as an index of socioeconomic status (Mirowsky and Ross, 2003).

Subsequent to data annotation, we will select the demographic factor that gives the highest interrater reliability for this dataset. Annotations will be averaged by bin, rather than averaging over all of a text's ratings, as was the case in previous shared tasks.

3.3 Pilot Study

To examine the integrity of our assumptions, we ran a short pilot task in which we used the Prolific Academic platform to crowdsource annotations from users in the youngest and oldest age groups.

We searched for texts which related to race/origin, religion, gender, sexuality and body type. We used keywords from Fortuna's (2017) subcategories of offensive speech to source texts which could be offensive jokes, such as 'black', 'woman', 'girlfriend', 'blind', 'gay', 'Muslim', 'Jew', etc. From a readily available dataset (The Short Jokes dataset from Kaggle), we sourced 40 jokes, 20 in which the keyword also referred to the butt of the joke (average number of tokens per text = 18.4), and 20 in which it did not (average number of tokens = 19.1). Twenty neutral texts were selected from Twitter, ensuring that the semantic meaning of the keyword stayed they same, e.g. 'black' referred to race, and not to Black Friday, and that the texts were not intended to be humorous. The average number of tokens per text in this group was 20.2.

- **Keyword is not target of joke:** 'What is the Terminators Muslim name? Al Bi Baq'
- **Keyword is target of joke:** 'Mattel released a Muslim Barbie... It's a blow-up doll.'
- **Random tweet with keyword:** 'The Mosque will close this weekend due to the pandemic'.

We asked 2 groups of annotators, aged 18-25 (n=10) or aged 55-70 (n=10) to imagine they were social media moderators. Their task was to identify the genre of the texts as label them as 'humorous', 'offensive', 'humorous and offensive' and 'other'. We highlighted that they did not need to find the text humorous, or personally offensive to label them as such. If they identified the intent as humorous, or the text as possibly offensive to others, they should use the corresponding label. We omitted the numerical rating task for reasons of brevity.

In terms of results, the clearest trends emerge when the groups were split by age. Both age groups of users made use of the 'humorous and offensive' label, suggesting that annotators could identify the genre of the text as humorous, but found it in bad taste. However, there was a trend for the younger group using this label more frequently than the older group.

Examining where differences in annotation occurred, Table 1 demonstrates the disparity in labelling on the following gender-related text:

> We should really use the blackjack scale to rate women. For example: "Every girl here is ugly" "Well, what about her?" "Eh, she's like a 15 or 16. Not sure if I'd hit it"

Table 1: Variation in labelling between age groups

Age	Humorous	Offensive	Humorous & Offensive	Other
18-25	3	3	3	1
56-70	2	7	0	1

As we did not have balanced groups based on level of education, or a critical mass of non-binary users so we omit analysis for these. Similarly, regarding gender differences, there were no clear trends in terms of labelling between females and males, and there were no statistically significant differences between groups.

The results of our pilot study suggest that pursuing demographic differentiation in humor annotation/classification is worthwhile. Specifically, we can see that age group may be relevant as the demographic factor which most distinguishes annotators' response to humor.

3.4 Tasks

We will ask systems to predict, given a group with a specific set of user demographics:

- Is this text humorous to the group, and if so, how humorous?
- Is this text offensive to the group, and if so, how offensive?

Our data will comprise texts which are either humorous and not offensive, humorous and offensive, not humorous and offensive, and not humorous and not-offensive.

In the case that there are no clear distinctions between the groups in terms of labels and ratings, we will average over these annotations, as typical tasks have done and proceed with classification and regression, as above.

The evaluation metrics for the classification task will be precision, recall and F1. The metric for predicting the humor and offensiveness scores will be root mean squared error.

4 Contribution to Computational Humor

In line with CH research, we affirm that humor is a moving target in terms of differing interpretations between demographic groups and across the lifetime. Our dataset will be the first to model the reception of a wide variety of humor genres from Twitter, presented to users of different demographics. It will also be, to the best of our knowledge, the first CH dataset to take into account the ratings of non-binary annotators.

In line with Hossain (2019), we aim to use clustering methods on the humor and/or offensive texts to determine themes that evoke these classes for different groups. We also aim to explore whether theories of humor, such as surprisal, superiority and incongruity are equally appreciated among different groups.

5 Conclusion

Humor detection and rating is a multi-faceted problem. We hope that the inclusion of demographic information will shift the state of the art away from objective classification, towards a more subjective approach. Future qualitative work could also suggest further variables whose inclusion would enhance our knowledge of humor perception. This could set a new standard for shared tasks which aim to model humor in future, and could outline a methodology that can be replicated with other cultures and languages.

6 Acknowledgements

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

Debra Aarons. 2017. Puns and tacit linguistic knowledge. In *The Routledge handbook of language and humor*, pages 80–94. Routledge.

Millicent H Abel and Jason Flick. 2012. Mediation and moderation in ratings of hostile jokes by men and women.

Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi. 2018. A crowd-annotated spanish corpus for humor analysis. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 7–11.

Luis Chiruzzo, S Castro, Mathias Etcheverry, Diego Garat, Juan José Prada, and Aiala Rosá. 2019. Overview of haha at iberlef 2019: Humor analysis based on human annotation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Christie Davies. 2008. Undertaking the comparative study of humor. *The primer of humor research, Berlin: Mouton de Gruyter*, pages 157–182.

Michael Dimock. 2019. Defining generations: Where millennials end and generation z begins. *Pew Research Center*, 17:1–7.

Caitlin Fisher. 2019. *The Gaslighting of the Millennial Generation: How to Succeed in a Society that Blames You for Everything Gone Wrong*. Mango Media Inc.

Paula Cristina Teixeira Fortuna. 2017. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. " president vows to cut¡ taxes¿ hair": Dataset and analysis of creative text editing for humorous headlines. *arXiv preprint arXiv:1906.00274*.

Giselinde Kuipers. 2017. Humour styles and class cultures: Highbrow humour and lowbrow humour in the netherlands. In *The Anatomy of Laughter*, pages 58–69. Routledge.

Sharon Lockyer and Michael Pickering. 2005. Introduction: The ethics and aesthetics of humour and comedy. In *Beyond a Joke*, pages 1–24. Springer.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of english puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Mirowsky and Catherine E Ross. 2003. *Education, social status, and health*. Transaction Publishers.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Louise Omwake. 1937. A study of sense of humor: its relation to sex, age, and personal characteristics. *Journal of Applied Psychology*, 21(6):688.

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.

Angela Rosenthal and David Bindman. 2015. *No laughing matter: Visual humor in ideas of race, nationality, and ethnicity*. Dartmouth College Press.

Julia M Taylor and S Attardo. 2016. Computational treatments of humor. In *Routledge Handbook of Language and Humor*.

Lisa Wagner and Willibald Ruch. 2020. Trait cheerfulness, seriousness, and bad mood outperform personality traits of the five-factor model in explaining variance in humor behaviors and well-being among adolescents. *Current Psychology*, pages 1–12.