THE UNIVERSITY *of* EDINBURGH

Demographically-Aware Computational Humor

Julie-Anne Meaney

Doctor of Philosophy

THE UNIVERSITY OF EDINBURGH

2023

Abstract

Computational Humor is a subfield of humor research aimed at using computational methods to understand humor. This understanding can entail analysing large datasets to find differences in humor appreciation, training systems that can detect and rate humor, or even generating jokes. In this thesis, we draw on findings from the broader field of Humor Research to build a large dataset of humor ratings. To capture responses to offensive humor, we also collect ratings of how offensive the funny texts were perceived to be, as well as demographic characteristics about the annotators who provided the ratings. We organised a humor and offense detection competition, and 60+ research groups submitted cutting edge computational systems to detect how humorous and offensive our data was. We then analysed our dataset's ratings when grouped by age and gender, and found that men give higher humor ratings than women, and that both women and older people enjoy offensive humor less than other groups. Lastly, we built humor and offense detection systems which included the annotators' demographic data, and found that incorporating demographic information as a textual description of the annotator improved how well models learned during training, but did not help the models to generalise to unseen data.

Lay Summary

Humor is a very common part of everyday life, both in the media and in our interactions with others. Instinctively, we know that there are differences in how funny some people find various kinds of humor. We might also be aware that some people take offense to humor, particularly if it makes fun of a touchy subject, or of a group that we belong to. This project is about asking, do people really respond differently to humor if they come from a different age group or gender? And can a computer learn to understand these differences?

For this research, we collected ratings of humor and offense on 10,000 texts. We looked at differences in how funny or offensive men and women of different ages find the texts. We found that when women feel a joke is offensive, they find it less funny. This happens to men too, but only when they feel personally offended by the joke. Women also admitted that they didn't 'get the joke' more often than men. Among the different age groups, we found that younger people are the least likely to get offended, or to laugh less when they are offended, but as they get older, people do get offended more and find offensive things less funny.

We organised a humor detection competition, in which 60 research teams trained computers to predict if a text was funny and offensive, and the best systems were able to do this with over 90% accuracy! We then built our own prediction system, where we found that giving the computer information about the gender and age of the person who rated the joke helped it to predict the humor and offense ratings given by that person, but when we tested the system on jokes it had not seen before, the model that got the best results was the one that had just seen the text, and not the information about age and gender.

Acknowledgements

Thanks firstly to my parents, Tom and Rita, for their support during this period - we overcame pandemics, stress and illness together.

Thanks to my supervision team for their contribution to this work. In particular, I am so thankful to Dr Steve Wilson for his sustained support, expertise and inquisitive nature. I am also very grateful to Prof. Rada Mihalcea for acting as a panel member - her seminal work on computational humor inspired much of this thesis.

I was incredibly lucky to have Dr Björn Ross and Prof. Nikos Aletras examine my work. Not only did they facilitate an interesting and enjoyable viva, but their suggested corrections improved this work greatly, as well as inspiring future projects.

A further source of inspiration for this project came from the Spanish-language tasks organised by Dr Santiago Castro and Dr Luis Chiruzzo. Luis, in particular, has become a close collaborator, sharer of jokes and friend.

A massive thanks to my friends in Edinburgh and further afield. To Kate Fogliaresi, I owe an incredible debt of gratitude for the encouragement, companionship, laughter and love. I jixelled a lot, thanks to you. To Dr Ibrahim Abu Farha, whose friendship and encouragement over regular lunches sustained me. Also to Yousef Nassif for being the warmest, kindest host, with the most delicious food. I am so thankful for my CDT colleagues, particularly Mr and Mrs William Toner and Kasia Prus.

To the people who make Marchmont an amazing place - to Christine Simpson, for being my constant focusmate for over two years. Dr Alexander Robertson, Jenny Sanger and Cat Chisem - you were always available for chats, walks and half pints. To David Hodges, who brought fun and creativity during hazy periods. To my dear neighbour and friend Maggie Barrons, for showing what resilience and good spirits look like. To the dynamic duo of Liane Phillips and Kayleigh Gordon for providing dog treats, dog-sitting and the odd bit of hardware.

I am so grateful to Dr Nicolas Chevalier and Dr Bonnie Auyeung for hiring and inspiring me. To Maggie Hogan, for modelling what taking on a big project looks like. To Charlotte Armstrong for demonstrating toughness and patience. To Dr James Garforth and Dr Yazmin Morlet-Corti for always checking in and suggesting fun plans. To Tracy Small for lovely Sunday walks. And to Victoria, of course.

I am also thankful for the physical and emotional support provided by Chris Sharp and Simonetta Logan. Also to the teachers at Lila Yoga, in particular Lisa Shaw and Angie Lake - your work helped immensely to soothe a tired mind and body! And to Austin Shirley, who has been with me every step of the way, showing me what mastery of a subject looks like.

I am incredibly grateful to Ian Dunmayne - you won't believe the positive impact that having a quiet, homely space to work and write made. You made all the difference.

And obviously, to Phoebe, a constant companion to the work, who encouraged me to take breaks, walks, and to play.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Julie-Anne Meaney

Contents

Figures and Tables

Figures

Tables

Chapter 1

Introduction

Computational humor has the primary aims of understanding the nature of and responses to humor, as well as training systems to detect and generate humor. It is a subfield of the much larger field of Humor research, which comprises research from Psychology, Sociology, Linguistics, Education etc. A key argument of this thesis is that the broader humor research has much to offer computational humor. Among the key issues that have yet to be incorporated into how we build datasets and algorithms for computational research:

- Subjectivity: humor is subjective, and responses to humor may depend on demographic factors such as age, class and education (Kuipers, 2015). Humor ratings are often treated as objective measures, which may be averaged to provide an objective ground truth reference for computational humor systems to predict. We argue that this approach requires more nuance.
- Interpretation: what is intended to be humorous may elicit other reactions, e.g. offense. Indeed humor often masks hate speech (Sue & Golash-Boza, 2013). This is a central concern to platforms which host user-generated content, and intend to automatically moderate content, as the lines between humor and hate speech can often be blurry.

- Transience: what is funny now may not be even five years from now (Highfield, 2015). This may be because the world knowledge which makes a text funny now is not the same in the future, or because the group that is targeted in the text is no longer deemed appropriate to be the butt of a joke (Lockyer & Pickering, 2005). Computational Humor researchers need to be mindful that their datasets age fast, and what may be humorous or offensive/inoffensive to users can shift over time, and throughout the lifespan.

The scope of this project is to examine systematic differences in the ways that different demographic groups perceive written humor, and to build systems which can predict humor and offense ratings, when given information about these demographic differences.

Computational humor detection - the automatic classification and/or rating of humor in text - poses a number of unique problems to natural language processing (NLP). Where several 'core' NLP problems focus on reducing semantic, syntactic or phonological ambiguity, humorous texts often centre on increasing ambiguity, where often two or more interpretations are required to understand a text (Meaney, 2020). The ability to process such ambiguity and detect/rate humor automatically has many applications, such as content filtering, recommendation, and improved human-computer interaction in applications such as chatbots.

Based on the summary of the linguistic, psychological and computational literature found in the following chapter, we devised the following research questions:

Following from these hypotheses, this thesis addresses the following research questions (RQ):

- **(RQ1)**: How can we improve on current datasets to incorporate signals such as demographic information?
- **(RQ2)**: Which models are most effective at capturing humor and offense ratings in text data?
- **(RQ3)**: To what extent do ratings of humor and offense vary as a function of demographic characteristics?

- **(RQ4)**: Can computational humor and offense detection performance be improved by incorporating the demographic characteristics of the annotators who rate the data?

1.1 Contributions

1.1.1 Humor and Offense Dataset

This is the first dataset to combine humor and offense ratings of texts. Offense is typically modelled separately, under the umbrella of hate speech, however this dataset accommodates the idea that a joke could provoke alternative reactions, besides humor. We offer suggestions on how to capture offence, specifically on splitting the concept of offense two: general offense, which asks the reader to consider what their community broadly might find offensive, and personal offense - which impacts the reader personally. This is also the first humor dataset to allow the rater to label the text as a joke, based on genre characteristics, but also to report that they don't get the joke. Previous datasets would ask for a rating regardless, and so lack of comprehension could be confounded with low humor ratings. The data collection process ensured that we captured opinions from a wide range of ages, in order to represent a larger section of society than the typical crowdsourced datasets. Along with their ratings of the data, we also collected data about the annotators' demographic characteristics for further analysis.

1.1.2 HaHackathon Shared Task

We released an aggregated version of the above data as a shared semantic evaluation task. This was the first task to model humor and offense contemporaneously. This task attracted submissions from more than 60 teams, who used cutting-edge approaches to advance the state of the art. We report on the output of this group effort.

1.1.3 Replication of Humor Research Findings

Although findings of group differences in humor appreciation are both compelling and intuitive, they often come from experimental setups with a relatively small number of participants. We take a big data approach to replicating these previous findings, by exploring group differences between 1,800+ annotators, on 200,000+ ratings. We find that, similar to previous results, humor and offense ratings did vary as a function of gender and age. We also explore differences between humor detection and comprehension between groups, given the use of the 'I don't get it' rating option.

1.1.4 Demographically-Aware NLP Systems

We explore what auxiliary sources of information may help improve the performance of humor detection systems. We incorporate variables such as age, and gender into the prediction system. We also explore including the ordinal nature of the ratings themselves as a further auxiliary.

1.2 Thesis Outline

Overall, this thesis highlights the issue of addressing subjectivity in humor ratings. It explores how subjectivity can vary systematically, as a function of the demographic characteristics of the humor audience. It also allows for alternative responses to intended humor, e.g. perceived offense. We explore ways that this can be operationalised to improve humor and offense detection systems.

Chapter 2: Background This thesis draws on the multidisciplinary field of humor research to explore the mechanisms of humor, both in terms of linguistic features and interpersonal functions. We also highlight findings from psychology and sociology to explore group differences in humor appreciation, and the overlap of humor and offense. We then review previous approaches to computational humor detection, in terms of datasets, feature engineering, and the newest humor understanding approaches.

Chapter 3: Datasets This chapter addresses **RQ1**, and describes the data collected for the HaHackathon challenge. It details our initial pilot studies, the selection of texts, the choice of annotators, the annotation procedure and potential grouping variables for ratings. It also explores the goal of including a wide variety of humor types, while still maintaining high inter-annotator agreement between annotators.

Chapter 4: Systems This chapter focuses on **RQ2**, and describes the best-performing systems which participated in the HaHackathon challenge. Although the vast majority of systems used transfer learning approaches, we also describe domain adaptation methods and data augmentation strategies, as well as approaches which disimproved performance.

Chapter 5: Statistical Analysis Focusing on **RQ3**, this chapter looks at replicating previous findings from humor research, using the HaHackathon dataset. We look at correlations between humor and offense across gender and age groups, how the distributions of ratings differ between these groups, and whether there are differences in humor detection and comprehension.

Chapter 6: Demographic Systems To address **RQ4**, this chapter describes modelling approaches which are often used to predict rating data, e.g. classification, and describes an alternative approach - ordinal regression. We explore which metrics are best suited to this task, which allow us to compare the two approaches. We incorporate the demographic information collected about the dataset annotators to determine if this helps to improve performance.

1.3 Thesis Outcomes

The following papers have been published from this thesis, and can be found in this Appendix B.

- Meaney, J. (2020). *Crossing the Line: Where do Demographic Variables Fit into Humor Detection?* In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 176–181)/

- Meaney, J., Wilson, S., Chiruzzo, L., Lopez, A., & Magdy, W. (2021). *Semeval 2021 task 7: Hahackathon, Detecting and Rating Humor and Offense.* In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021) (pp. 105–119).
- Meaney, J., Wilson, S. R., Chiruzzo, L., & Magdy, W. (2022). *Don't take it personally: Analyzing gender and age differences in ratings of online humor.* In International Conference on Social Informatics (pp. 20–33)
- Chiruzzo, L., Castro, S., Góngora, S., Rosá, A., Meaney, J., & Mihalcea, R. (2021). *Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in Spanish*. Procesamiento del Lenguaje Natural, 67 , 257–268.

Chapter 2

Literature Review

2.1 Introduction and Scope

Humor is one of the most common elements of public discourse. It occurs across media, film, political discourse, across different genres and narratives, it permeates every area of social life and interaction (Lockyer & Pickering, 2005). Given the omnipresence of humor, Humor Research is an inter-disciplinary field with major contributions from Linguistics, Psychology, Sociology, Philosophy, Anthropology, Management Studies, as well as Computational Research.

In light of the prevalence of humor, and the number of approaches to its research, it is necessary to scale back which aspect of humor this thesis will focus on. The scope of this work is termed *verbal humor* in the literature - a somewhat misleading name which encompasses both spoken and written humor. It does not consider other types, such as visual humor or physical humor (i.e. slapstick). The focus is further narrowed by looking at reception of jokes - how does the listener respond to the text, without considering how this varies as a function of the joke teller, their delivery, their own demographic characteristics. Lastly, we focus on humor *detection*, not humor generation.

This literature review will bring together insights from the longstanding field of Humor Research together with the much younger sub-field of Computational Humor. We will first focus on the mechanisms of humor - what are the cognitive and emotional mechanisms that provoke a humorous response? This is contextualised by looking at the linguistic mechanisms that create these effects. We will look at distinct humor

styles, and what function these styles serve to the joke-teller. We will then examine where systematic differences exist in humor responses - do particular groups of people respond differently to humor than others? We will examine the case where humor provokes an alternative reaction - specifically offense.

From the Computational Humor field, we will examine the ways in which researchers have incorporated findings from the broader research into their computational systems. We will explore the datasets created for humor detection, the features which are useful for this task, the algorithmic approaches which have been applied, and the most recent attempts at understanding what cutting-edge models understand about humor.

2.2 Mechanisms of Humor

The processes which result in humor have been investigated as far back as Aristotle - whose contribution to humor theory is difficult to understate. His theory of incongruity remains one of the most prominent explanations of humor to this day. This incongruity theory is laid out in Aristotle's *Rhetorics*, and states that some forms of humor arise due to the unexpected juxtaposition of two normally disparate ideas. The setup of a joke establishes one frame of reference, and the punchline subverts that frame of reference. In other words, the humor occurs between what is expected and what actually occurs.

> "'I'm sorry' and 'I apologize' mean the same thing. Unless you're at a funeral."
>
> -Demitri Martin

In the above quote from Demitri Martin, the first frame of reference established is that 'I'm sorry' and 'I apologize' are synonyms. The subversion of this idea is that, in the context of a funeral, the first phrase expresses sympathy, while the other expresses responsibility. A second example from Groucho Marx highlights that a common device used to create incongruity is ambiguity.

> "One morning I shot an elephant in my pajamas. How he got in my pajamas, I don't know."
>
> –Groucho Marx

In the above quip, the expectation that the speaker is wearing the pajamas is subverted by the revelation that the elephant was wearing them. This centres around the two possible ways of parsing the prepositional phrase 'in my pajamas'.

Although incongruity is a widely seen to be a necessary condition of humor, it is not a sufficient one - that is to say that not everything that is incongruous will be humorous. A further contribution from Aristotle is superiority theory. In his *Poetics*, Aristotle puts forward the idea that we may elicit laughter at the expense of inferior individuals because we feel joy at our perceived superiority to them. A common manifestation of superiority theory is seen in the type of jokes that most countries make at a neighbouring country's expense, or even targeting a provincial area of their own country. The *Philogelos*, a collection of 265 jokes from 4-500 CE regularly makes the people of Kyme the butt of superiority-based jokes. The below quip highlights the perceived inferiority of the Kymean, who does not understand that it is the direction of the house that precedes the aspect of the windows.

> Shopping for windows, a Kymean asks if there are any that look South.

Although incongruity and superiority can account for many sub-genres of joke, there are certainly cases in which they do not account for the causes of humor, for example puns. Relief theory, popularised by Freud (Fedor, Gaynor, & Reik, 1950) (although like originating earlier), states that humor allows us to experience temporary relief from the rules and constraints of life, this can release pent up energy from the nervous system. Freud extended this theory to puns, which he stated temporarily free us from the rules of language, causing mirth, although others have strongly argued the key mechanism of puns is incongruity/ambiguity (Hempelmann, 2004). Although the theory of relief may account for jokes about taboo subjects, or so-called *black humor*, theorists broadly agree that a comprehensive theory of humor still does not exist.

2.2.1 Linguistic Mechanisms of Humor

Raskin (1985) proposed the first formal semantic theory of jokes, which closely aligns with incongruity theory. The Semantic-Script Theory of Humor (SSTH) is based on the notion of *scripts*, i.e. a semantic unit of information about the world. It holds that two conditions must be true to create humor in a text:

1. The text is compatible, fully or in part, with two different scripts.
2. The two scripts with which the text is compatible are opposite in a special sense.

The theory holds that the semantics in both scripts must overlap, and must be incongruous. In the setup of the joke, both scripts are evoked in a way that somewhat obscures the less obvious script. The punchline of the joke requires the audience to switch their interpretation to the less likely script. He offered the following text to illustrate:

> "Is the doctor at home?" The patient asked in his bronchial whisper. "No," the doctor's young and pretty wife whispered in reply. "Come right in."

The first script is the idea that the patient is seeking the doctor's assistance for his lung ailment, which causes him to speak with a bronchial whisker. The less likely script is that the patient is checking that the coast is clear of the doctor, so that he may engage in intimate activities with the doctor's wife - whose own whisper is conspiratorial in nature.

Some criticisms of the SSTH, are that it does not distinguish between puns, which make reference to the phonological features of the utterance, and non-puns, which are based simply on the meaning of the text. A second criticism was that some jokes appeared to be more similar to each other than to other types of jokes, for example:

1. What do you get when you cross a cow and a lawnmower? A lawnmooer!
2. What do you get when you cross a lemon and a cat? A sourpuss!

The above two jokes comprise a very similar setup and punchline structure, and rely on wordplay associated with an animal for comic effect. However, the SSTH does not distinguish between these texts and any of the other examples in this chapter. An expanded version of the SSTH was the General Theory of Verbal Humor (GTVH) (Attardo & Raskin, 1991), which was created with the specific aim of answering 'when are two jokes the same joke?' The GTVH proposed six knowledge resources which can be used to create a joke, and these can be used to determine the locus of humor, and therefore the similarity between texts. The knowledge resources are:

1. **Language**: which phonological, morphological, syntactic and lexical ways can a text be interpreted?
2. **Narrative Strategy**: what ways can this story be told? Is it a simple setup and punchline text, or a longer anecdote?
3. **Target**: in aggressive styles of humor, who (if anyone) is the target, or "butt" of the joke?
4. **Situation**: in which environment does the joke take place? The example of the funeral joke above makes use of the situation knowledge resource to humorous effect.
5. **Logical Mechanism**: how is the incongruity of the joke explained away, or justified?
6. **Script Opposition**: identical to the SSTH

Computational approaches which use these knowledge resources for feature engineering are detailed in Sections 2.4.5.

2.2.2 Humor Styles and Functions

Although humor may be enjoyed outside the company of others, it occurs more frequently in interaction with others, and psychologists have long been interested in the social and interactive functions of humor. Martin et al. (2003) developed the Humor Styles Questionnaire, to capture differences in the use of four types of humor:

1. **Self-enhancing humor**: this is characterised by an amusement at the incongruities of life, and is seen by some psychologists as a means of emotional regulation, or a coping mechanism against negative emotions.
2. **Affiliative humor**: this aims at boosting relationships with other by means of witty banter which serves to entertain, and to reduce social tension.
3. **Aggressive humor**: while this type of humor is also aimed at boosting relationships, it comes at the expense of others. This may entail sarcasm, teasing, or types of humorous takes which do not take the listener's perspective into account, e.g. sexist or racist humor.
4. **Self-defeating humor**: this attempts to enhance relationships at one's own expense. It can be seen as a way to ingratiate oneself into a group through self-deprecation.

Individuals employs each type of humor to a greater or lesser extent, depending on the social standing they hold in the group they are in, the characteristics of the other group members or even the acceptability of each humor style in their culture (Lockyer & Pickering, 2005).

2.3 Demographic Differences in Humor Use

Much research effort has gone into trying to systematically understand the subjectivity of humor, and demographic differences provide an attractive schema in which to explore this subjectivity. Jiang et al. (2019) surveyed differences between Asian and Western views towards humor, and found that Chinese people are more ambivalent about humor than their Western counterparts. Where Westerners considered humor to be as essential part of everyday life, and a desirable personality trait, Easterners felt that it was a skill best left to experts, and not a trait they tend to foster. Similarly, Kalliny et al. (2006) focused on differences between Arab and American use of humor, and found that Americans used more self-defeating and self-enhancing humor than their Arab counterparts, although they found no significant differences in the usage of

affiliative and affirmative humor between the cultures. The authors speculate that the use of self-deprecating humor may be an attempt to mitigate the power imbalances present in American society, and that the use of self-enhancing humor may be a more accepted coping mechanism than in other cultures. This theory contradicts Kazarian et al. (2004) who hypothesise that people from individualist cultures such as U.S. may use aggressive humor in order to reinforce hierarchies and inequality.

2.3.1 Gender differences in humor use

The main demographic variables of interest in this thesis are gender and age. This is because they are among the most well-studied variables in the literature, albeit often in relatively small-scale *experimental* studies. This represents an opportunity to replicate these results in a larger-scale *descriptive* study, in a more natural setting.

The Humor Styles Questionnaire mentioned in Section 2.2.2 comprised four separate 8-item scales to record how individuals perceived their own humor style. The items were selected to ensure that each scale was minimally inter-correlated, and responses were elicited from approximately 600 participants, divided into a younger and older age group and also binned by gender. The authors found that males tended to self-report significantly higher scores of affiliative humor than women. Males also reported higher scores for aggressive humor than females, and younger men used this kind of disparaging humor more than older men. Older women reported higher scores on the self-enhancing type of humor than younger women, who tended to score significantly more highly on self-defeating humor.

In a recent systematic review of 77 papers, Hoffman et al. (2020) found that men showing greater tolerance or appreciation of aggressive humor was the most often-replicated result across cultures and age groups. Similarly, Proyer and Ruch (2010) report that men tended to score higher on *kagelaticism* - the joy of laughing at others, which suggests that as long as a joke does not target men explicitly, it may be offens-

ive towards other groups, without impacting men's humor ratings. This may extend to dark humor, as a recent Italian study of covid-related humor (Bischetti, Canal, & Bambini, 2021) reported that increasing age, as well as being female was related to finding pandemic humor more aversive and less funny.

Lampert and Ervin-Tripp (1998) aim to contextualise these results, and offer that men's tolerance of aggressive humor may be due to the socialisation of males to dominate societal power structures. Although this seems like a simplistic analysis, it may be supported by Knegtmans et al. (2018), who found that participants (regardless of gender) whose social power had been manipulated to place them in a high-power state rated jokes which targeted others as less offensive, and gave higher humor ratings.

In terms of tolerance of sexual humor, studies as far back as 1937, indicate found that young men gave higher ratings to 'shady' (i.e. sexual) jokes than their female, and older counterparts did (Omwake, 1937). While this is an often-replicated result, Lampert and Ervin-Tripp again aim to contextualise this finding, as their experimental results suggest that women are more tolerant of sexual humor than previous studied reported, but only in the case that the material itself is not sexist. They also stress the importance of considering not just the content of jokes, but also the function, e.g. 'building group solidarity, establishing social norms, building self-image and controlling behaviour'.

2.3.2 Age differences in humor use

Similar surveys of humor throughout the lifetime found that appreciation of jokes shifts as people age, once again highlighting that humor is a moving target. Svebak et al. (2004) in a survey of 65,000 Norweigians found that "overall humor scores" were higher for men than they were for women. They also found that humor appreciation declined with age: the mean scores for total sense of humor on average declined across the age cohorts from highest score in the 20s to lowest score among those aged 70. Although the number of individual surveyed is impressive, this finding should

be tempered by the fact that humor appreciation was assessed using only three items. Indeed more fine-grained experiments have indicated that *humor appreciation* increases with age, even as *humor comprehension* (and the cognitive abilities which is requires) declines (Greengross, 2013; Schaier & Cicirelli, 1976).

In terms of aggressive/offensive styles of humor, a study from the Netherlands (Kuipers, 2017) found significant differences in humor preferences along the lines of gender, age, social class and it found that the older generation rated their younger counterparts' humor as more offensive than that of their own age group. This replicates studies from Ruch (1990) and Stanley et al. (2014) who demonstrated that that tolerance for aversive humor declines with age.

Despite the plentiful evidence of demographic differences in humor appreciation furnished by psychologists and sociologists, there has been very little integration of these nuances in computational treatments of humor. This may mean that humor detection systems, or joke recommendation systems which are trained to work across the board may provide results which are too aggressive for some users, or too tame for others.

2.3.3 Humor and Offense

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.' However, humor is still used as a mechanism to mask hate speech, specifically because humor situates itself as the opposite of *serious discourse*, allowing perpetrators to frame hateful comments as benign, or 'softening' the impact of prejudice (Sue & Golash-Boza, 2013).

Despite the shift away from using racist or sexist comments in public arenas such as comedy clubs or work settings, offensive language still abounds online (Davidson, Warmsley, Macy, & Weber, 2017; Nobata, Tetreault, Thomas, Mehdad, & Chang, 2016). This can reinforce racial stereotypes, or have a damaging impact on com-

munities. Furthermore, it can pose a liability issue for companies which host hateful content, and it is thought that all platforms which host user-generated content have an urgent need for content moderation, to protect against these risks (Nobata et al., 2016).

2.4 Approaches to Computational Humor

Having surveyed the mechanisms of humor, and the impact of demographic variables on humor appreciation, we will explore computational humor detection approaches. We will see that many detection systems have attempted to incorporate the linguistic mechanisms of humor into feature engineered approaches, but have paid scant attention to the subjectivity of humor, the impact of demographic differences on appreciation, or the link between humor and offense.

2.4.1 Datasets

Early datasets specifically limited themselves to a very restricted scope of humor, e.g. *Knock Knock* jokes or one-liners, noting that a deep comprehension of all types of humor was ambitious, and would likely exceed the computational capacities available at the time of writing (Mihalcea & Strapparava, 2005). These datasets were also limited by their negative, or non-humorous examples, which were either synthetic, or were drawn from domains that were so distinctive (e.g. news), that many humor detection systems trained on these datasets easily determined the news domain, without learning about the humorous one.

One of the first computational humor detection datasets was created by Taylor and Mazlack (2004) who collated approximately 200 Knock Knock jokes. This is a tightly constrained type of humor of the form seen in Table 2.1.

Table 2.1: Knock Knock Structure and Samples from Taylor and Mazlack (2004)

Joke Structure	Sample Joke	Negative Example
Line 1: "Knock, Knock" Line 2: "Who's there?" Line 3: any phrase Line 4: Line 3 followed by "who?" Line 5: Wordplay based on line 4.	Knock, Knock Who's there? Justin Justin who? Justin time for dinner.	Knock, Knock Who's there? Justin Justin who? Justin awoke in the middle of the night.

Synthetic examples were generated without the final wordplay. Their recognition system used bigram and trigram language models to identify words which were unlikely to co-occur, and were therefore likely to be jokes based on wordplay. They manually identified a key word in Line 3 of the text, and generated possible candidates for wordplay based on substituting phonemes in the key word. If their n-gram language model determined these candidates to be valid substitutions, they selected the most likely candidate and checked if it was in the punchline. Their system was trained on 66 jokes; and tested on 130 positive examples and 66 synthetic negative examples whose structure was similar to Knock Knock jokes.

Mihalcea and Strapparava's first humoor dataset (Mihalcea & Strapparava, 2005) focused on one-liners - a short-form humorous sentence which employs rhetoric devices such as alliteration and rhyme to produce humor with very few words. To collect data, they used a simple bootstrapping algorithm to find websites that contained one or more 'seed' one-liners, with the hypothesis that such websites likely contained other positive samples. They sought negative examples from Reuters news website, the British National Corpus (BNC) (Consortium et al., 2007) and a set of proverbs, with the hypothesis that these samples would be of similar length and structure to the one-liners. Their positive samples contained approximately 20,000 one-liners, with an estimated 9% noise (e.g. negative examples). They noted that Reuters sentences were easier to distinguish from the one-liners than simple conversations utterances (e.g. BNC sentences), likely due to stylistic differences between these two genres of writing.

Table 2.2: Sample one-liners, Reuters titles, BNC sentences, and proverbs (from Mihalcea and Strapparava, 2005)

Source	Examples
One-liners	Take my advice; I don't use it anyway. I get enough exercise just pushing my luck. Beauty is in the eye of the beer holder.
Reuters titles	Trocadero expects tripling of revenues. Silver fixes at two-month high, but gold lags. Oil prices slip as refiners shop for bargains.
BNC sentences	They were like spirits, and I loved them. I wonder if there is some contradiction here. The train arrives three minutes early.
Proverbs	Creativity is more important than knowledge. Beauty is in the eye of the beholder. I believe no tales from an enemy's tongue

By 2007, more joke collections were available online, and Sjoberg and Araki (2007) manually created a set of 6,100 one-liners from such a set. Their negative examples were based on the BNC, given Mihalcea et al.'s finding that the news domain was too distinct to use as a source of non-humor. Similarly, van den Beukel and Aroyo (2018) scraped data from websites such as *funnyshortjokes.com* and took negative examples from news, proverbs and wikipedia. Faruqi and Shrivastava (2018) also scraped from websites such as *jokeoftheday.com* etc, and took negative examples from Reuters, BNC and proverbs. However, with big data, there tend to be issues of how to assess the quality of the data collected, and the authors sampled only 100 texts, from their 400,000 examples to determine how many examples were actually humorous. This is arguably too small a sample to determine the quality or quantity of jokes in the dataset.

With the rise in popularity of websites such as Twitter and Reddit, sourcing texts from these domains became more common. However, curating humor datasets from such sources can be difficult, as it requires either pain-staking manual curation, or the ability to acquire negative examples with a similar enough lexical style so as not to be obviously different. Khandelwal et al. (2018), is an example of a high-effort, low-yield dataset. In compiling their corpus of Hindi-English code-switched texts, they

Table 2.3: Sample texts from Hinglish, Italian and Dutch datasets

Source	Examples
Khandelwal et al.	“For #WontGiveItBack to work, Dhoni needs to say ‘Trophy toh ghar par hi bhul aaye’ ” (For #WontGiveItBack to work, Dhoni needs to say “We forgot the trophy at home”)
Buscaldi et al.	“Lo sa che io ho perduto due figli” “Signora lei una donna piuttosto distratta” (“You see, I lost two sons” - “Madame, you are quite a scatterbrain”)
Winters et al.	Joke: “Wat is groen en plakt aan de muur? Kermit de sticker!” (“What’s green and adheres to the wall? Kermit the Sticker”, pun on “kikker” (“frog”) Non-joke “Wat is groen en telefoneert aan de muur? Kermit de spin!” (“What’s green and telephones on the wall? Kermit the Spider”)

first scraped a large number of tweets, then manually selected those which contained Hindi and English, before sourcing bilingual annotators to rate the texts. This effort resulted in 1755 humorous examples and 1698 non-humorous texts. They highlight that this required a lot of effort, and that the type of humor found on Twitter is often topical in nature, and would not make sense to a reader in a year’s time.

Similarly, Buscaldi and Rosso (2007) scraped Italian-language quotations from Wikiquote. Although their dataset may suffer from issues of validity due to being annotated by just one annotator, this effort resulted in a dataset of 1966 examples, with 471 positive examples.

Those who use weak supervision to source texts from social media platforms face the same issues of data quality checking, and sourcing negative examples seen in the other datasets mentioned above. Reyes et al., (2012), exploit the use of hashtags, such as *#humor #irony* to source humorous tweets, and *#politics* and *#technology* to find negative examples. The authors do not report on lexical or stylistic differences between the positive and negative examples, nor do they randomly sample texts to examine the noise ratio. This is a crucial omission, as large datasets are meaningless unless data quality is assured.

Ermilov et al. (2018) faced similar problems of sourcing non-humorous texts. They used a dataset of Russian one-liners scraped from VK and Twitter, and took negative examples from Russian classical novels, news and proverbs. Once again, the domain differences were so great that even a simple bag-of-words model could distinguish the humorous and non-humorous examples with high accuracy. Winters and Delobelle (2020) took a novel approach to generating negative examples from the positive ones, using dynamic templates to identify key context words and replace them with similar parts of speech.

2.4.2 Shared Task Datasets

Shared tasks deserve a special mention in the literature, as they are an important means of attracting new researchers to the field of computational humor. However, amongst the existing shared tasks in English and Spanish, there has been a consistent trend of either featuring data with a very narrow scope of humor and high agreement, which may restrict the extent to which a system built on this data can generalise to other data, e.g. Potash, Romanov, and Rumshisky (2017), or a broad scope of humor with very low agreement amongst the annotators, which may limit how much models built on this data can learn, e.g. Castro, Chiruzzo, and Rosá (2018).

Semeval 2017 Task 6, *Hashtag Wars* (Potash et al., 2017), took 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: first, the producers selected what they considered to be the ten tweets most humorous submissions, and most appropriate for the show's type of humor. Then the show's audience voted on their number one submission. The following samples were submitted for the hashtag "FastFoodBooks". A ranking of 0 means that the submission did not reach the top ten, a ranking of 1 means the text was selected by producers to be in the top 10, and a ranking of 2 means that the text was in the top 10, and was selected as the best tweet by the audience.

Table 2.4: Sample Submissions to @Midnight for the hashtag #FastFoodBooks

Text	Ranking
Orange Julius Caesar #FastFoodBooks @midnight	0
@midnight Diary of a Calorie Girl #FastFoodBooks	0
Are you there, God? I'm Ready to Place My Order #FastFoodBooks @midnight	1
The Diarrhea of Anne Frank #FastFoodBooks @midnight	1
As I Lay Dying of congestive heart failure @midnight #FastFoodBooks	2

Participants in the humor detection challenge took part in two subtasks: Subtask 1 required systems to pair the tweets, and for each pair, predict which one had achieved a higher ranking, according to the audience. Subtask 2 was to predict the labels given by this stratified annotation, e.g. 0) submitted but not top-10, 1) top-10 or 2) 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.

Similarly, 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, Krumm, Gamon, & Kautz, 2020). Instead of collecting previously existing humorous texts, the organisers generated their humorous examples by scraping news headlines from Reddit, and then hiring crowdworkers to edit the headlines to make them funny. Edits were defined as 'the insertion of a single-word noun or verb to replace an existing entity or single-word noun or verb'.

Table 2.5: Sample submissions to the Edited Headlines Task

Original text	**Substituted Word**	**Votes**	**Mean Score**
President Trump's first year <anniversary/>report card, with grades from A + to F	Kindergarten	3,3,3,3,3	3.0
Richard Spencer 's white-nationalist <think/>tank broke Virginia nonprofit law	fish	3,2,2,2,2, 1,1,1,1,0	1.5
Donald Trump should lift <sanctions/> and use aid instead of weapons on North Korea	weights	3,3,2,2,2, 2,1,0,0,0	1.5
Congo 's mining revenue ' <missing/>' - Global Witness	Stolen	1,0,0,0,0	0.2
State officials blast unprecedented DHS <move/>to secure electoral system	idea	0	0.0

A second set of annotators rated the headline's humor from 0-3. An abusive/spam option was included, but presumably to discard ineffective edits, rather than highlight a text which would cause offense. Although inter-annotator agreement between raters was relatively high, (Krippendorff's α 0.64), 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 drastically limits 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. Furthermore, thematically, the texts are limited in that they deal with topics that featured in the news in 2019. This focus on topical humor can limit the longevity of a dataset, as even a human annotator may struggle to comprehend the humor a year after the news story has hit the headlines.

In terms of broad scope and low agreement, the HAHA challenge (Humor Analysis based on Human Annotation) ran 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 was 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?

Sourcing the data from 50 different Spanish-speaking accounts means that this dataset has a much wider scope than the two others mentioned above. However, the inter-annotator agreement for the humor rating question was extremely low (Krippendorf's α of 0.1625). It's possible that sourcing the texts from fifty different sources introduced too many genres to allow a consensus about what was funny to emerge 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

Table 2.6: Sample texts from the HAHA task, 2018

Text	Is humor	Votes	Mean humor
La frase "No eres tu, soy yo" viene del latín "Me Gusta Alguien Más" *The phrase "It's not you, it's me"* *comes from the Latin for "I like someone else"*	1	1,2,2,2, 2,2,3,3	2.125
— ¿Aquí es la reunión para personas impuntuales? — Ops, fue ayer. *- Is this the meeting for tardy people?* *— Oops, it was yesterday*	1	1,1,3,4,4	2.6
Quisiera saber que hago durante la siesta de la cual me levanto más cansado que cuando me acosté a dormir. *I would like to know what I do during the siestas* *in which I wake up more tired than when I* *laid down to sleep.*	0	0,2,0,0,0	NULL

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. The low inter-annotator agreement is an obstacle in computational humor tasks, as it can impact the signal that we expect the machine to follow. If there is limited agreement on what the signal is for humans, this may make it difficult for machines to follow it too.

2.4.3 Offense

A closer examination of the HAHA task data was influential for this thesis. When filtering the texts based on those that had a bimodal distribution, e.g. those that were clustered at both 1 and 5, we found that many dealt with sensitive topics such as race and gender, and often in a disparaging way. A sample of these texts is seen in Figure 2.7. Given that the annotators were very split in terms of how they rated these texts,

and that this was not reflected in the aggregated mean humor score, we determined that another form of rating was necessary - that of offense. We were also curious about the demographics of the annotators that rated sexist and racists texts as very humorous, and whether this related to the gender and age differences mentioned in Section 2.3.

Table 2.7: Sample Offensive Texts with Bimodal distribution from HAHA

Text	Is humor	Votes	Mean humor
—Mi amor, ¿me veo gorda con este pantalón? —Júrame que si te digo la verdad no te vas a enojar. —No. —Me estoy cogiendo a tu hermana. *—Do I look fat in these pants my love?* *—Promise me if I tell you the truth,* *you're not going to get mad* *—No* *—I'm sleeping with your sister*	1	1,1,1,5,5	2.60
Que hacen 10 negros en una pared blanca? R. Un codigo de barras! *What do 10 black [people] make when standing against a white wall?* *A barcode!*	1	1,1,4,4,5	2.83
Cuando mi mujer habla ,yo escucho. . . .. la radio o la tele.. *When my wife speaks, I listen. . . ..* *to the radio or the TV*	1	1,1,2,5	2.25

2.4.4 Feature Engineering for Humor Detection

Before the widepsread use of transfer learning in NLP, handcrafted features which aimed to capture characteristics of the input domain were common. For humor, the most commonly used features were incongruity, ambiguity and stylistic features. However, given the figurative nature of humor and the wide variety of humor styles, feature engineering appraoches were not always successful.

Mihalcea et al. (2010) implemented incongruity features, in an attempt to operationalise Attardo and Raskin's work on GTVH. They used a sample of 150 one-liners, removed the punchline and replaced it with three alternative, non-humorous punchlines which had been generated by annotators. To capture incongruity, they hypothesised that the correct punchline would have the minimum semantic relatedness to the setup. They calculated this relatedness in a number of ways using WordNet (Fellbaum, 2010), e.g. counting the number of nodes between two words and calculating the overlap in definitions between two words. They found that this was not very useful for this setup. They also represented the data using a TF-IDF weighting, and calculated the cosine similarity and pointwise mutual information between the resulting vectors, and found that this setup did improve detection performance. A further implementation used latent semantic analysis, which uses singular value decomposition to reduce the dimensionality of a term-by-document matrix, but found that it did not improve performance, which contradicted their own previous findings. On a different dataset, Barbieri and Saggion (2014) calculated the frequency of words in their texts according to the American National Corpus of frequency data, hypothesising that words that appear infrequently together were more likely to be incongruous. They found that the less frequent a word was, the more useful this was as a feature.

As mentioned above, a key device of incongruity is that multiple, ambiguous interpretations may be available for a word or constituent-phrase. However, feature engineered approach to this ambiguity have had mixed success. Sjoberg and Araki's (2007) system looked up the number of senses of a word on dictionary.com and calculated the average ambiguity of the words as well as the maximum ambiguity. This was a useful feature for predicting their data. However, Reyes et al. (2012) replicated this approach on their data and found that it was not a useful feature for their dataset. In a separate implementation, Barbieri and Saggion (2014) also calculated the set of cognitive synonyms (synsets) for nouns, verbs and adjectives in their data and found it was not useful for humor detection.

Phonetic features have also proved useful to some humor detection systems, owing to the fact that many puns rely on phonetic similarities between keywords in the text. Beyond puns, Mihalcea et al. (2006) assert that alliteration and rhyme are not just stylistic features, but can be used of an incongruity device, as they create an expectation, which is then subverted. However, as is the case with many features of humor, the performance of phonetic features varies, depending on the implementation and the dataset. The Carnegie-Mellon University Pronouncing Dictionary is often recruited in phonetic feature engineering for humor. It was used by Sjoberg and Araki (2007) to calculate the number of words which share at least four letters, one of which was a vowel. However, in this dataset, the feature did not improve their prediction performance. Conversely, for their case, Mihalcea et al. (2006) saw improved performance when they used the CMU dictionary to capture alliteration chains, in texts like:

> Veni, Vidi, Visa: I came, I saw, I did a little shopping.

Finally, Van de Beukel and Aroyo (2018) used homophones from the CMU dictionary and homographs by matching to a list of common homographs from Wikipedia. They found significant performance improvements for homograph features, but not for homophones.

The use of syntactic features approach can not only lead to improved classification performance, but can also give an insight into the characteristics of humorous texts. Such features were implemented successfully by Liu et al, (2018), who counted the number of noun phrases, verb phrases and adjective phrases as a measure of complexity of the text. They also used phrase lengths and dependency relations between phrases. They found that the humorous examples used simpler words, but more complex syntax. This syntax entailed more negations and rhetorical questions, but also included more personal pronouns - similar to conversational text. Their detection system which included features that captured these syntactic differences achieved the best performance among their systems.

Many of the handcrafted features used for humor detection have fallen into disuse, given the automated feature extraction offered by PLMs. However, Chapter 4 highlights that some features which were tailored to the domain of humor were still effective at boosting performance, while data augmentation approaches that interfere with the subtler elements of humor, such as back-translation, were not successful.

2.4.5 Systems for Humor and Offense Detection

Supervised learning algorithms learn to predict a known set of labels by training on a large subset of a labelled dataset and validating performance on the unseen subset. Among the traditional supervised learning algorithms, such as support vector machines (SVMs), Naive Bayes (NB), random forests (RF) and logistic regression (LR), SVMs have consistently shown the best performance on humor detection (Castro et al., 2018; Mahajan & Zaveri, 2020), as they appear to generalise well to unseen data. In a number of cases, when comparing SVMs, NB, LR and RF, researchers have found RF to be the most effective approach to humor detection in their datasets (Hossain et al., 2020; Jaiswal, Mathur, Mattu, et al., 2019).

Deep learning approaches, which use multiple layers of neural networks have largely superseded the traditional machine learning algorithms mentioned above. The most commonly used implementations are multi-layer perceptrons, convolutional neural networks (CNNs) and long short-term memory networks (LSTMs). LSTMs are a type of recurrent neural network which allow the nodes to retain information about previous states which may be relevant to future nodes. Bi-directional LSTMs enhance this ability by retaining information about past and future states. This can be very useful for NLP, as it can help to resolve long-distance dependencies which are common in language. LSTMs have tended to be the most successful for many NLP tasks, and humor is no exception. All of the best performing submissions to the Hashtag Wars challenge (Potash et al., 2017) and the first edition of the HAHA task (Castro et al., 2018) tasks implemented this algorithm for the task. Where CNNs have been

used successfully, it has tended to be in combination with LSTMs. Diao et al., (2020) implemented a CNN to capture information about scripts within a humorous text, a Bidirectional LSTM for overall semantic representation, and an attention mechanism to capture information about homographic puns.

As quickly as deep learning models became the state of the art for NLP, they were usurped by transfer learning systems. This involves pre-training a model on vast quantities of data, using a masked language modelling objective, and then adding a small number of layers on top of this trained model to finetune it to a different task. The sheer quantity of data used to train these models is of great benefit to humor detection, because jokes normally require a lot of contextual and world knowledge to comprehend them. All of the best performing systems in the Edited Headlines shared task (Hossain et al., 2020) and in our own HaHackathon task used PLMs, and this approach will be expanded upon in Chapter 4.

2.4.6 Demographically-Aware NLP Systems

Until this project, there were no existing approaches to demographically-aware humor detection, despite the mounting evidence that humor is demographically modulated. In an example of the usefulness of demographics to NLP, Hovy (2015) incorporated demographic information into three classification tasks, but trained separate word embeddings based on texts that came from women and men, as well as those that came from people under 35 and over 45. On the three tasks - sentiment analysis, topic detection and author attribution classification - the demographically-aware models outperformed the demographically-agnostic ones. This result was replicated in five languages. Similarly, Garimela et al. (2017) found that the words we first associate with other words varies along the lines of gender and age, and they built demographically-aware word embeddings for these groups. Heng et al. investigated if transfer learning approaches which incorporate demographic information can improve performance on downstream tasks. They experimented with adapting the language representations derived from transformers for different gender and age groups, and

found substantial performance gains across four languages. However, when they controlled for confounds such as the domain of the text and the language proficiency of the PLM, they found that their performance improvements were not in fact down to demographic information. This thesis continues the abovementioned pioneering work to incorporate the demographic information of the annotator or joke recipient into humor detection and rating systems.

2.5 Conclusion

This literature review brings together theories and functions of humor, with demographic differences in humor appreciation. We explored how computational humor has incorporated much of the linguistic theories of humor, but has not accounted for demographic differences, or overlaps between humor and offense. The following chapter details our efforts to incorporate the modelling of offense into humor detection, and the collection of demographic information about the dataset annotators.

Chapter 3

Improving datasets for Humor detection with offense ratings and demographic data

3.1 Introduction

Research workshop challenges, in which annotated data and research problems are presented for analysis, are a vital means of attracting interest to subfields of NLP, such as humor detection. Aside from drawing attention to the field, they also provide an easy entry point for newcomers, by offering research questions and datasets. The number of participants in humor detection competitions has been growing year-on-year (see Figure 3.1), and this has been followed by an increase in publications on humor detection in Information and Computing Sciences.

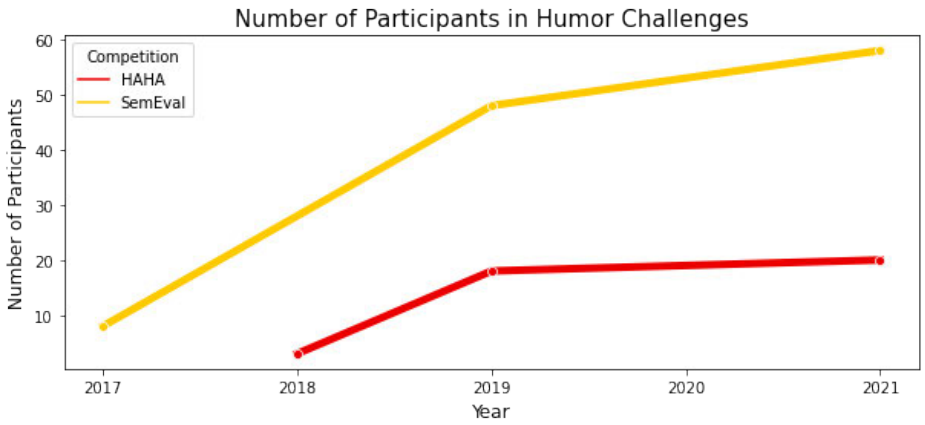


Figure 3.1: The Trend of Increasing Participation in Humor Challenges since 2017

In light of how influential humor detection challenges are to the emerging field, and given that these datasets may form the basis for how researchers think about humor, the focus of this chapter is on the following question:

RQ1: How can we improve on current humor detection datasets?

The need for improvement is introduced in Section 2.4.2, which highlighted that the datasets previously used in computational humor challenges have either been limited by a narrow scope of humor, or by low inter-annotator agreement. These issues can have a detrimental impact at various points in the research process. During training, low agreement may mean that the signal we intend the system to pick up is weak, which can impact how well the model learns. After training, a narrow scope of humor can mean that the system does not generalise well to unseen examples. Furthermore, to date, no computational humor datasets have attempted to incorporate the findings of broader Humor Research, such as the interplay of humor and offense (Lockyer & Pickering, 2005), and the variance seen in ratings from different demographic groups. As findings from Sociology and Psychology demonstrate, there may be significant differences in humor ratings that come from different genders (Aillaud & Piolat, 2012), age groups (Tsai, Chen, Hung, Chang, & Huang, 2021) and social classes (Kuipers, 2017).

3.2 Hypotheses

Given the above issues from the humor literature, we formulated the following hypotheses:

H1: What may be perceived as humor to one listener may be offensive to another. Modelling offense alongside humor may be an improvement on existing datasets

H2: We can increase the scope of humor datasets without decreasing inter-annotator agreement

H3: Demographic characteristics may provide an important signal to understanding some of the subjectivity of humor. The inter-annotator agreement within certain demographic groups may be higher than the overall agreement.

While the inclusion of a broader scope of humor types and a rating for offense are relatively straightforward, increasing inter-annotator agreement is not a trivial matter, particularly due to the highly subjective nature of humor. To tackle this, we chose to group the annotators who provided the ratings based on demographic characteristics and investigate if agreement was higher *within* groups than *between* groups. We selected age and gender as two potential grouping variables, as these are long-standing targets of investigation in the wider Humor literature (Engelthaler & Hills, 2018; Kotthoff, 2006; Omwake, 1937), and were also well-represented across different groups on the annotator recruitment platform that we were using. At the time of data collection, there were a large number of annotators from ages 18-70, and they were split evenly between male and female, which made it feasible to investigate age and gender as grouping variables. We then undertook a pilot study to check which of these variables would be best for grouping.

3.3 Pilot Study

Our pilot study presented a small sample of 60 texts to 40 annotators, and had the following aims:

1. To assess if it was meaningful to model humor and offense together.
2. To select between age or gender as a grouping variable for annotators.
3. To clarify guidelines for annotation.

We used Prolific Academic [1] to crowdsource ratings from annotators who were evenly spread across four age groups: 18-24, 25-40, 41-55 and 56-70. The raters were also evenly divided into male and female annotators (these were the best represented genders among Prolific users). The pilot study and subsequent data collection were granted ethical approval from University of Edinburgh Ethics committee. All users provided informed consent and they were given the right to withdraw from the study at any time.

Data

The pilot data was sourced from the Kaggle Short Jokes Dataset [2], and the negative examples came from Twitter. The data comprised:

1. 20 non-humorous texts
2. 20 texts which were intended to be humorous, but which would likely be considered offensive
3. 20 texts which were intended as humorous and likely not offensive.

Each of the texts contained one or more keywords from Fortuna's (2017) list of common targets of offensive speech (e.g. 'black', 'woman', 'girlfriend', 'blind', 'gay', etc.). The texts which were intended to be humorous contained some genre characteristics of humor, e.g. setup and punchline, or absurd content. In the texts that were potentially offensive, the featured keyword (e.g. woman, Muslim) was the target or 'butt' of the joke. The humorous texts which were less likely to be offensive contained the keyword, but it was not the target of the joke. The non-humorous examples contained the keyword but did not feature any genre characteristics of humor, and so no joke was made at this group's expense. The average number of tokens in the potentially offensive texts was 18.4, in the unoffensive texts was 19.1 and in the non-humorous texts, was 20.2.

---

1. https://www.prolific.co/
2. https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes

Table 3.1: Examples of Text which Mention/Target Common Hate Speech Keywords

Joke	Humorous	Keyword is target?
'What is the Terminators Muslim name? Al Bi Baq'	✓	✗
'Mattel released a Muslim Barbie... It's a blow-up doll.'	✓	✓
'As an Australian-born Muslim, a city slicker. I'd never gotten to see the beauty of the outback.'	✗	✗

During the annotation procedure, we informed annotators that they were assisting with humor research, and that we wanted to elicit their personal opinions. We asked them the following questions, the first of which comes from Castro et al. (2018):

1. Is the intention of this text to be humorous? (Yes/No)
   (a) (If yes) How humorous do you find this text? (1-5)
2. Is this text generally offensive? (Yes/No)
   (a) (If yes) How generally offensive do you find this text? (1-5)
3. Is this text personally offensive? (Yes/No)
   (a) (If yes) How personally offensive do you find this text? (1-5)

Annotator Instructions

The annotator instructions outlined that the first annotation question was intended to determine the *genre* of the text, and should be distinguished from *funniness*. Annotators were instructed to look at the structure of the joke, e.g. setup and punchline, or the content of the joke, e.g. absurdity, in order to determine if the intention was to be humorous. We specified that they did not need to find the text funny in order to identify that the intention, or genre, was humorous.

We split the concept of offense into two, based on feedback from a pre-pilot trial that indicated that users were not sure what we meant by 'offensive'. We clarified that generally offensive meant that a text targeted a group of people, simply because they belonged to that group, or the annotator thought that a large number of people would

be offended by the text. Personally offensive meant that the annotator belonged to a group that was targeted by the joke, or they were hurt on someone else's behalf. We also included an open text box to ask the annotators why they selected the ratings they did.

Results

In terms of our first aim - to assess that it was meaningful to model humor and offense together - we looked at the extent to which the ratings were correlated using Spearman's ρ (), and all correlations reported were significant, with a p-value < 0.05. Table 3.2 indicates that humor ratings were negatively correlated with general offense ratings, and to a greater extent, personal offense ratings. We also saw that general and personal offense were correlated, but not perfectly, meaning that they were related, but annotators still made a meaningful distinction between the two categories. This was confirmed by responses in the open text box, in which annotators confirmed why they had selected general or personally offensive ratings, and this was in line with what we had requested in the annotation instructions.

Table 3.2: Correlation Coefficients (Spearman's ρ) for Ratings in Pilot Study

	Humor	**General Offense**	**Personal Offense**
Humor	1.0000	-0.2279	-0.3292
General Offense	-0.2279	1.0000	0.7208
Personal Offense	-0.3292	0.7208	1.0000

For the second aim, to assess if age was the most appropriate grouping variable, we calculated the inter-annotator agreement on the humor, general and personal offense ratings of all texts for all annotators, as well as the agreement on ratings grouped by age, and then by gender. We used Krippendorff's α as the measure of inter-annotator agreement (Krippendorff, 2011). Figure 3.2 demonstrates that, with the exception of the age group 18-25 whose opinions were more mixed, the inter-annotator agreement for humor and offense ratings was higher within age groups than overall. Agreement within gender groups was lower than agreement overall, so we determined that age was the better grouping variable.

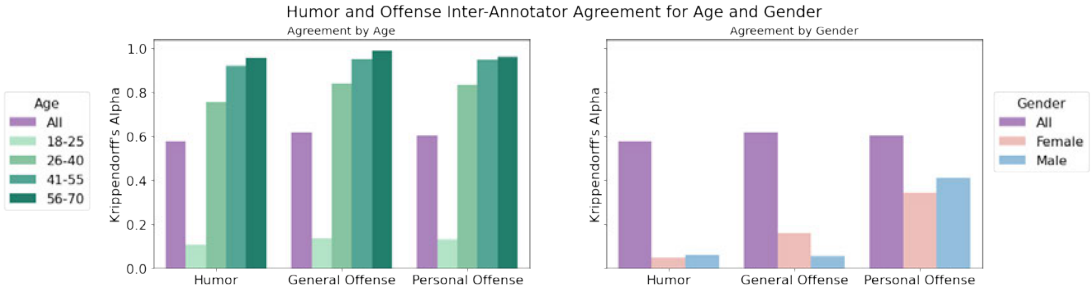


Figure 3.2: Inter-Annotator agreement in Krippendorff's alpha for Age and Gender Groups

Feedback from the free response text box indicated that we had met two of the aims for this pilot study: that the annotators were able to understand they did not need to find the text funny in order to label it as humor, and that general and personal offense were distinct concepts. Feedback from 4 annotators indicated that for one or more texts, they could identify that the genre was humor, but they did not get the joke, and therefore they rated the humor as 1. We took the position that a low rating of humor should *not* be annotated in the same way as a failure to understand the humor. We therefore decided to include an option of 'I don't get it' for the larger data collection, and in our analyses, this equated to a humor rating of 0.

3.4 HaHackathon Dataset

Given the pilot feedback we proceeded with collecting ratings for 10,000 texts, as this amount of data was within the range provided in two previous English-language humor tasks (Hashtag Wars provided 12,734 texts and Edited Headlines gave 8,248 texts). The annotation procedure used the same questions as the pilot study. In order to provide sufficient statistical power for later analysis and modelling, we sought 20 annotations per text, comprising 5 ratings from each age group – 18-25, 26-40, 41-55, and 56-70. Below we outline how we selected the texts and annotated them.

3.4.1 Data Collection

In order to examine a broad selection of naturally-occurring humorous and offensive texts in English, we sourced 80% of our data from Twitter. The remaining 20% of texts, we selected from the Kaggle Short Jokes dataset[3] in order to meet three quotas:

- **Humor Quota:** We wanted to ensure that a large sample of the texts in the dataset were intended to be humorous. Our annotation procedure asks raters if the intention of the text is to be humorous (as evidenced by the the setup/punchline structure, or absurd content). Given that the texts in the Kaggle Short Jokes Dataset were sourced from the /r/jokes and /r/cleanjokes subreddits, we were confident that the intention of the text was to be humorous.
- **"Traditional" Humor Quota:** We wanted to represent jokes which have a traditional setup and punchline structure. Twitter humor is known to use a number of unique features, such as short dialogues and textual memes which are self-referential (Zhang & Liu, 2014), and which may not be equally recognisable to all annotators. The Kaggle texts contain a selection of conventionally recognisable jokes, which allowed us to gauge the annotators' humor detection abilities and to use as a quality check for the crowdworkers (see below).
- **Offense Quota:** To ensure that a proportion of texts were likely to be considered offensive by the annotators, half of the texts selected according to the procedure below.

Potentially Offensive Texts

To select potentially offensive texts, we expanded the list of keywords associated with Silva et al.'s (2016) sub-categories of hate speech in social media, and queried the Kaggle dataset for these.

---

3. https://www.kaggle.com/abhinavmoudgil95/short-jokes

Target	Keywords
Sexism	She, woman, mother, girl, b*tch, he, man, blond, p*ssy, hooker, slut, wh*re
Body	Fat, thin, skinny, tall, short, bald, amputee, redneck
Origin	Mexico, Mexican, Ireland, Irish, Indian, Pakistan, China, Chinese, Polish, German, France, Welsh, Vietnam, Asian, American, Russia, Arab, Jamaican, homeless
Sexual Orientation	Gay, lesbian, d*ke, f*ggot, homo, aids, LGBT, trans, tr*nny
Racism	Black, Africa, African, wop, n***** white people
Ideology	Feminism, leftie/lefty
Religion	Muslim, Islam, Jew, Jewish, Catholic, Protestant, Hindu, Buddhist, ISIS, Jesus, Mohammed
Health	Wheelchair, blind, deaf, r*tard, Steven Hawking, Stevie Wonder, Helen Keller, dyslexic

Table 3.3: Targets and Sample Keywords

From these texts, we identified the target, or butt, of the joke and made the assumption that a text could be potentially offensive to our annotators if the hate speech keyword was the target of the joke. We selected 1,000 texts this way. We also assumed that the text would likely be considered not offensive if the keyword was mentioned, but was not the target and selected a further 1,000 texts like this. This was to reduce the probability that a humor/offense detection system would learn to classify texts simply based on the presence of a hate speech keyword.

Text	Keyword = Target
A **fat** woman just served me at McDonalds and said "Sorry about the wait". I replied and said, "Don't worry, you'll lose it eventually".	Yes
Don't worry if a **fat** guy comes to kidnap you... I told Santa all I want for Christmas is you.	No

Table 3.4: Sample of potentially offensive and non-offensive texts

Selection of Twitter texts

In order to avoid introducing annotation confounds such as a lack of cultural or linguistic knowledge (Meaney, 2020), we selected the texts and the annotators from the same region – the US. When sourcing the humorous Twitter data, we selected accounts according to whether they were based in the US and posted almost exclusively humorous content (e.g. @humurous1liners, @conanobrien). For the non-humorous Twitter accounts, we elected not to use news sources, e.g. CNN due to stylistic differences between news and humor (Mihalcea & Strapparava, 2006) making them easy to differentiate.

The non-humorous accounts we selected featured US celebrities (e.g. @thatonequeen, @Oprah). We also chose organisations that represent the targets of hate speech groups (e.g. @BlkMentalHealth), in order to increase the occurrences of the keywords in a non-humorous and non-offensive context). We selected trivia accounts (e.g. @UberFacts), because the question and answer structure is similar to some types of setup and punchline). We also featured accounts that post tv/movie quotation (e.g. @MovieQuotesPage), in order to resemble the dialogue-type jokes that are common on Twitter). Please see the appendix for a comprehensive list of accounts.

Using the Twitter API, we crawled up to 2,000 tweets from each account, and removed retweets and texts containing links. As topical humor can be difficult to appreciate once the event it is tied to has passed (Highfield, 2015) we also removed tweets that contained references to US Politics, the Covid-19 pandemic, or TV show characters. From an initial 76,542 texts, we were left with 8,000 tweets. From these, we removed hashtags that labelled the texts as humorous, e.g. #joke. Then using Ekphrasis (Baziotis, Pelekis, & Doulkeridis, 2017), we split up any remaining hashtags into their constituent words so as to make them less easy to differentiate from the Kaggle texts.

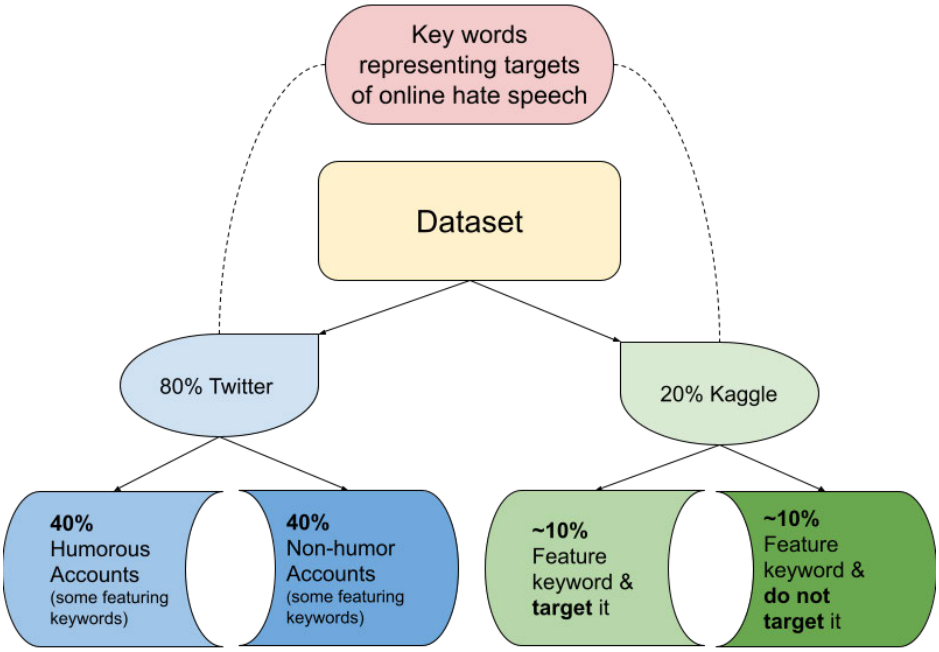


Figure 3.3: Occurrence of Keywords in Humorous and Non-humorous Texts

3.4.2 Annotation

Participants were recruited from Prolific based on their self-reported native English-speaker status, US citizenship, and membership of one of the four selected age groups: 18-25, 26-40, 41-55, 56-70. Each text was annotated by 5 members of each age group, giving a total of 20 annotations per text. Batches comprised 100 texts which were presented in a randomised order, and annotators answered the same questions as in the pilot study, using the interface presented Figure 3.4.

For the humor rating, the user was also given the option to select ‘I don’t get it’, meaning that they recognised by the structure or content that the text was intended to be humorous, but that they did not comprehend why the text was funny. This is distinct from a rating of 1, which is a comprehension of humor, with little appreciation for it.

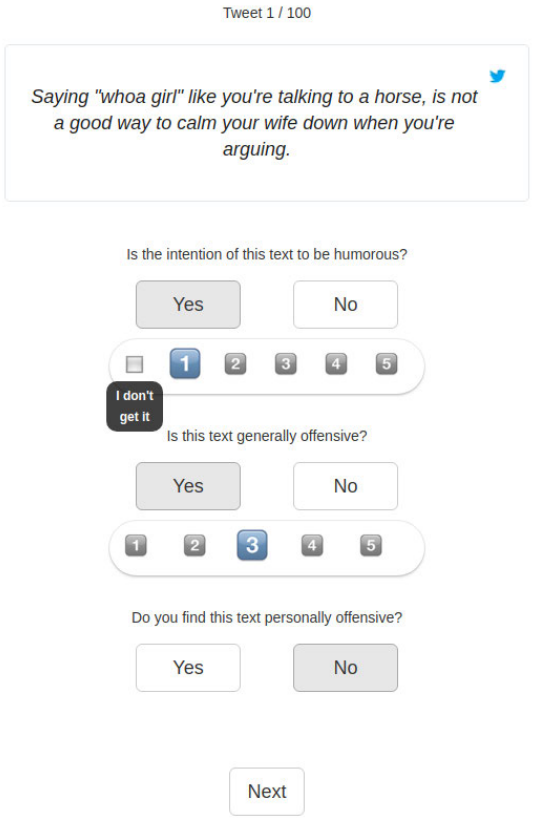


Figure 3.4: Screenshot from the tool used to annotate the texts.

Annotator Characteristics

Besides selecting annotators according to their age, first language and citizenship, we also collected the following demographic information about the annotators, where available, for further analysis.

- **Basic demographics**: gender, US state of residence
- **Language, Ethnicity and Origin**: country of birth, ethnicity, English speaking monolingual (yes/no), first language, fluent languages, mono/multi cultural, nationality.
- **Socioeconomic status**: employment status, student status, health insurance (US), highest education level completed, household income (USD).
- **Lifestyle/Health**: relationship/marital status, diet restriction (e.g. adherence to a vegetarian, vegan, halal diet etc.), smoking status, mental health/illness/condition, sexual orientation

- **Political Leaning**: political spectrum (US), pro-life/pro-choice, religious affiliation, concern about environmental issues
- **Personality traits and internet cohort**: measures of the 'Big Five' personality traits, i.e. extroversion, agreeableness, conscientiousness, emotional stability, openness Digman (1990) and what 'internet cohort' the user belongs to, i.e. the websites the annotator used when they first started using the internet, which can be an index of linguistic usage and technological ability McCulloch (2020).

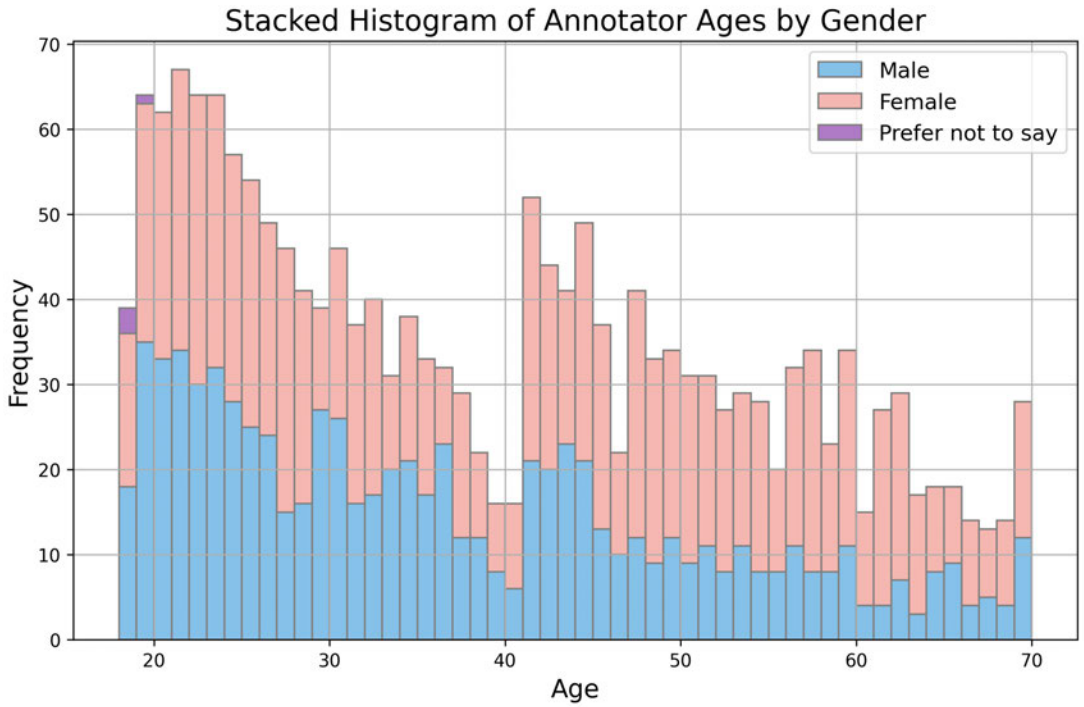


Figure 3.5: Distribution of Ages and Genders in the the Annotator Pool

Figure 3.5 demonstrates that the gender variable is mostly balanced across age groups, but from age 40 onwards, there is a slightly higher proportion of females in the dataset. Those who preferred not to disclose their gender were exclusively 18 and 19 years old. Figure 3.6 demonstrates two issues with other demographic variables we collected. Firstly, many of the responses to other demographic questions were strongly skewed towards one particular group - this means that some statistical models will not be appropriate for modelling this data, due to the imbalance in groups. Secondly, in relation to the two variables shows in Figure 3.6, we can see that the

annotator pool comprises predominantly white and liberal people. As highlighted in Section 2.3, this is likely to bias the humor responses in a certain direction, and we speculate that it may also impact what the annotators consider to be generally and personally offensive.

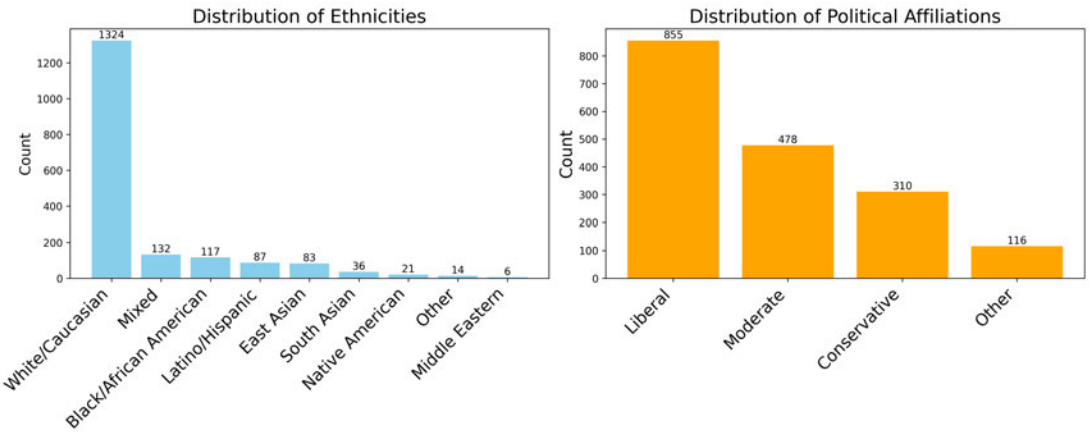


Figure 3.6: Distribution of Ethnicities and Political Affiliations in the the Annotator Pool

3.4.3 Quality Control and Data Discarded

Each batch of 100 texts comprised approximately 20% of texts from Kaggle. As the majority of these have a setup and punchline structure, or other recognisable humor traits, we used these as a quality control. If an annotator did not label at least 60% of these as humor, it was inferred that they they did not follow the instructions for the first question, and annotated based on perceived humor, as opposed to observation of humorous characteristics. We therefore discarded these submissions and replaced the annotators. Of 2,364 annotation sessions (in batches of 100), 301 submissions were discarded and replaced. The final dataset features 2,061 annotation sessions make up the dataset. Of these, 1,569 annotators rated one batch of texts with an additional 492 doing a second batch.

3.4.4 Data Statistics

Post-annotation, we classed a text as humorous if the majority of its twenty votes labelled it as such. In a small number of cases where votes were tied, we assigned the label humorous. For the texts labelled humorous, we calculated the average humor score, which was the average of the numerical votes. “No" ratings did not count towards this value, and votes of “I don’t know" were counted as 0, because this was deemed to be a recognizable humor structure, but one in which the humor was not successful.

Label	Affirmative	Negative	Average Rating
Humorous	6179	3821	2.24
Controversial	3052	3017	N/A
Offensive	5754	4246	1.02

Table 3.5: Data Statistics

Figure 3.7 indicates that the most frequent humor rating for Twitter texts was ’not funny’. This is to be expected, as all of the negative examples were drawn from that source. The Kaggle texts marked as ’not funny’ are due to annotator error, as these texts were all drawn from a joke dataset, and displayed a setup and punchline structure or absurd content.

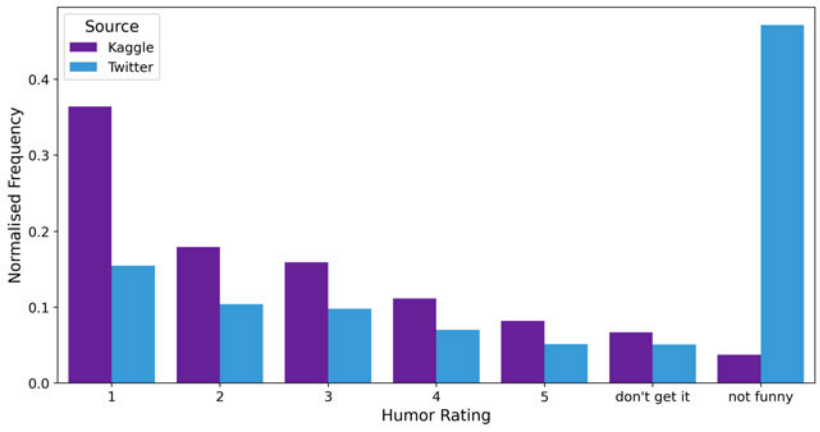


Figure 3.7: Humor Ratings by Source (Normalised

In terms of offense ratings, Figure 3.8 indicates that the majority label for both sources of data was 'not offensive'. Kaggle jokes tended to elicit more ratings of offense than Twitter, and this is expected given the selection of keywords and targets outlined in Section 3.4.1. This figure also highlights that there are far fewer personally offensive ratings than generally offensive ones, and this limitation will be discussed further in the next chapter.

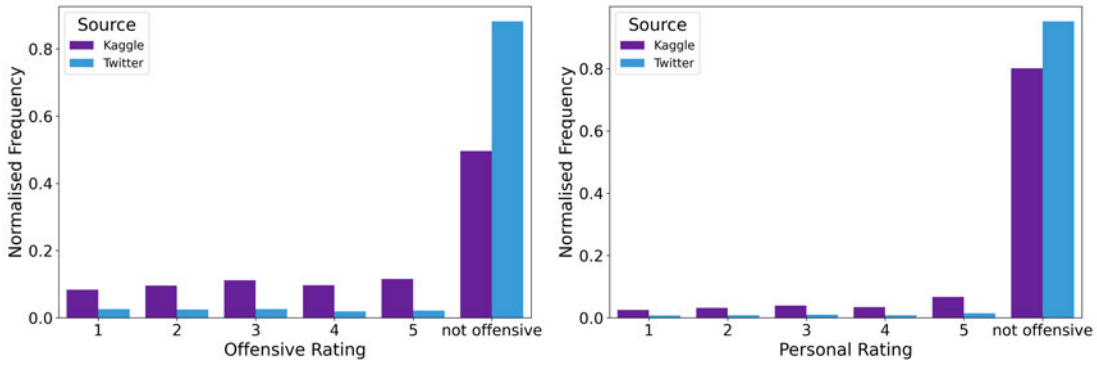


Figure 3.8: General and Personal Offense Ratings by Source (Normalised)

There are a number of common themes that arise in the humorous and offensive texts. Figures 3.9 and 3.10 show the most commonly seen words in the texts that were rated 5 for humor or offense. Women feature commonly in both, but in the humorous texts, they are referred to more often as 'wife' and 'girlfriend'. Similarly, men appear in both lists, but they are referred to as 'guy' more in the humorous texts. The term 'black' occurs much more frequently in the offensive texts than in the humorous ones, and the offensive texts also contain more references to religion and sexuality, like 'Muslim', 'Jew' and 'gay'. Perhaps surprisingly, the word 'fat' appears on the most frequent words in humorous texts list, but not on the offensive, perhaps hinting at the acceptability of fat jokes to the American audience who annotated the dataset.

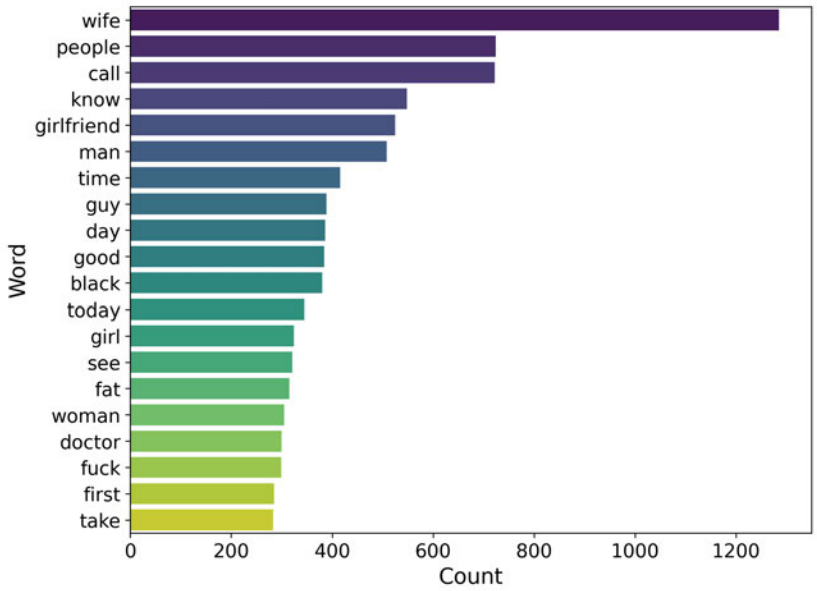


Figure 3.9: Top 20 most Frequent Words in Humorous Texts rated 5

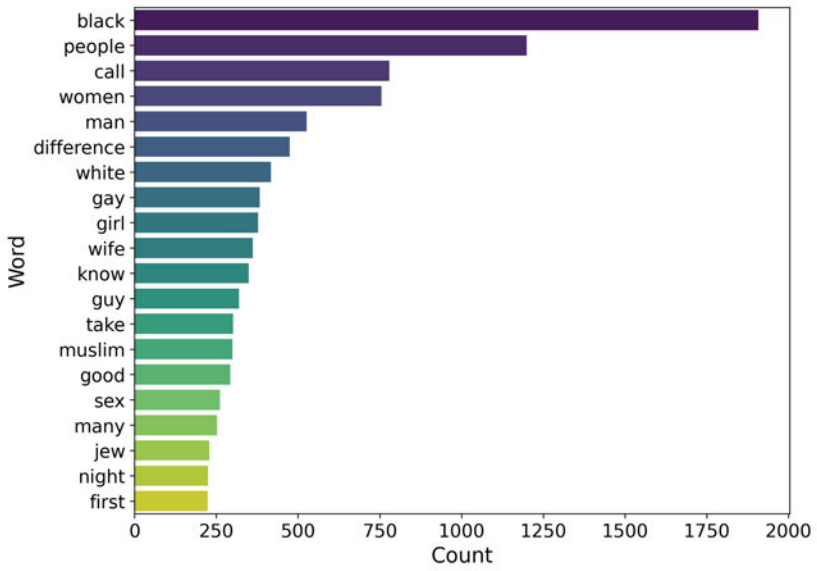


Figure 3.10: Top 20 most Frequent Words in Generally Offensive Texts rated 5

3.5 Discussion and Limitations

We proposed to improve humor detection datasets based on these hypotheses:

H1: What may be perceived as humor to one listener may be offensive to another. Modelling offense alongside humor may be an improvement on existing datasets

H2: We can increase the scope of humor datasets without decreasing inter-annotator agreement

H3: Demographic characteristics may provide an important signal to understanding some of the subjectivity of humor. The inter-annotator agreement within certain demographic groups may be higher than the overall agreement.

In terms of **H1**, the correlation between humor and offense in the pilot study suggested that these two concepts were meaningfully linked. This was replicated in the HaHackathon dataset. Table 6.1 illustrates that ratings for humor and offense are moderately correlated. Although the directionality of the connection between these two variables is not assumed, it seems plausible that the more offensive a text was, the less humorous the annotator found it, rather than the opposite interpretation - that less humorous texts were more offensive.

Furthermore, ratings of general and personal offense are strongly correlated, but not perfectly so, indicating that annotators still made a meaningful distinction between them.

Table 3.6: Correlation Coefficients (Spearman's ρ) for Ratings in HaHackathon Dataset

	Humor	**General Offense**	**Personal Offense**
Humor	1.000	-0.1241	-0.1711
General Offense	-0.1241	1.000	0.6696
Personal Offense	-0.1711	0.6696	1.000

To explore **H2**, we increased the scope of humor in our dataset by sourcing data from a more traditional, setup and punchline dataset, such as Kaggle, as well as a more modern, and less stylistically constrained source of humor, such as Twitter. Furthermore, we reduced the risk of cultural confounds biasing the data, by selecting texts and annotators from the same background (i.e. native speakers of US English). Arguably, this decision does reduce the scope of the dataset, and therefore the gen-

eralisability of the humor detection systems built on top of it. However, given that a secondary aim of the dataset was to explore the relationship between demographics and humor/offense ratings, we deemed it important to reduce the potential impact of this known cultural confound.

This is also the first Humor Detection challenge dataset to provide an 'I don't get it' option to annotators. This was an important option to offer, as it separates humor detection from humor comprehension, and may prevent texts from being incorrectly labelled as 'not humorous' simply because the annotator did not possess the world knowledge to understand the humor presented. Of a total of 202,369 ratings, 'I don't get it' was used 10,924 times.

One hypothesis that was not borne out, was **H3**, that inter-annotator agreement would increase by binning ratings based on demographic groups. Our pilot study indicated that agreement within age groups was higher than agreement within gender groups, and overall agreement. Figure 3.11 demonstrates that this was not the case for the HaHackathon dataset. We can see that agreement overall is low, and when split by gender and age groups, it hovers around the same level, or is lower. This means that the question of increasing inter-annotator agreement while also maintaining a broad scope of humor is still an open one. Future research could investigate the use of different demographic variables to bin ratings. Alternatively, it could experiment with using a pre-annotation sense of humor survey, to determine if the annotator favours affiliative or aggressive types of humor.

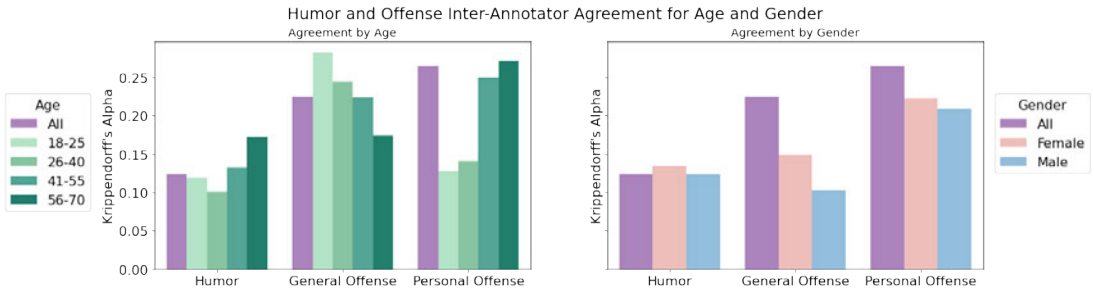


Figure 3.11: Inter-Annotator agreement in Krippendorff's alpha for Age and Gender Groups in HaHackathon Dataset

3.6 Conclusion

This chapter suggests areas for improvement on previous humor detection datasets. We created the first humor dataset to collect offense ratings, and allow raters to admit that they do not understand the joke. We also collected demographic information about the annotators, for social science analysis, and to increase the inter-rater reliability. Our initial pilot study indicated that binning annotations by age would yield higher reliability within groups than overall, but this was not replicated in the HaHackathon dataset. Suggestions were given for future research. The next chapter describes the best-performing systems on the data we collected. We then give an analysis of the demographic ratings, in order to validate theories from the Psychology of Humor, before suggesting ways to model the unaggregated data.

Chapter 4

Systems for Detecting Humor and Offense

4.1 Introduction

As the previous chapter highlighted, humor detection challenges are an important means of attracting researchers, and stimulating progress in computational humor detection, and our challenge was no exception. SemEval Task 7: HaHackathon, Detecting and Rating Humor and Offense attracted submissions from 63 teams, who built cutting-edge and innovative systems to achieve high performance on all of the tasks presented. This chapter focuses on the research question:

RQ2: Which models are most effective at capturing humor and offense ratings in text?

We describe the tasks presented in the challenge, as well as the best systems for approaching them. We also summarise the findings from the teams who submitted experiments with their system descriptions. As the use of pre-trained language models (PLMs) predominated in this competition, we discuss which PLMs performed best, which architectures and training strategies improved performance, and which domain adaptations had a positive or negative impact on performance. We also investigate why one of the sub-tasks was difficult for teams to achieve good results on, and why a rule-based system was one of the best approaches to this task.

4.2 Tasks and Evaluation

The data is an aggregated version of the dataset presented in the previous chapter. Although our aim was to increase inter-annotator agreement by binning ratings, we found that overall inter-annotator agreement was at the same level, or higher, than in the binned annotations. For this reason, to assign the 'humorous' and 'offensive' labels we selected the majority label given by the annotators. If the majority label was humorous or offensive, we took the average of the ratings, where "I don't get it" had a value of 0. This means that the challenge participants worked on a dataset of 10,000 texts and ratings. Personal offense ratings were not included in this task in order to streamline participants' workflow. The detection and rating tasks are the same as those seen in Castro et al. (2018) and Chiruzzo et al. (2019), and the humor controversy task is novel.

Task 1a: Humor Detection

This was a binary classification task to detect, given a text, if the majority label assigned to it was humorous or not. This was evaluated using F-score for the humorous class and overall accuracy.

$$Accuracy = \frac{C}{N}$$

$$F_1 = 2 * \frac{Precision \times Recall}{Precision + Recall}$$

Task 1b: Humor Rating Prediction

This was a humor rating regression task. Participants predicted the average rating given to texts from 0-5. Texts which had not been labelled as humorous by our annotators did not have a humor rating, and predictions for these texts were not counted towards the final score by our scoring system. The metric for this task was root mean squared error (RMSE).

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{N}\right)^2}$$

Task 1c: Humor Controversy Detection

This task was also a binary classification task to predict whether the humor ratings given to the text showed it to be controversial or not, that is to say that the annotators gave a wide variety of ratings. We calculated the median variance of the humor ratings in the training set, and assigned a label of 'controversial' if the variance in a text's ratings was higher than the median. This was also evaluated using F-score and accuracy.

Task 2: Offense Detection

This was an offense rating regression task. Unlike the humorous task, this rating was not dependent on the text having been labelled as humorous. All annotator ratings were considered, and each text had a rating from 0-5. The metric was RMSE.

4.3 Systems

4.3.1 A Primer on Large Language Models

The previous state of the art representation for words in NLP was achieved by learning word representations or embeddings, using a shallow neural network model such as Skipgram or CBOW (Mikolov, Chen, Corrado, & Dean, 2013). These could be trained on a large amount of data and subsequently downloaded for use by researchers. This feature extraction method transforms raw text into a meaningful alignment of word vectors in the embedding space. However, these embeddings did not capture context, and therefore were unable to represent higher-level concepts, such as word sense disambiguation, syntactic or semantic roles (Qiu et al., 2020).

Models built on transformer architectures (Vaswani et al., 2017) brought about a sea change in word representation because they capture both semantics and context. This created much more powerful and flexible word representations. Arguably the most famous example in research settings is the Bi-directional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2018), which uses self-attention to compute representations of the input, without recurrent connections or convolutions. It uses 12-24 encoder layers, each with feed-forward networks and multiple attention heads to train on a masked language modelling objective. This allows it to pick up patterns in language. The resulting PLM can be used for feature extraction, i.e. it can be used to output embeddings like Word2Vec embeddings. More commonly, they are used for finetuning, e.g. building a task-specific classifier on top of the pre-trained model, while also updating the parameters of the pre-trained model during training. The success of BERT has spawned a lot of similar models, which may differ in terms of the model's architecture (e.g. decoding in an autoregressive manner or using more layers) and they also may be trained with different data and pre-training tasks. Each pre-training task introduces its own bias, which can impact downstream tasks (Ladhak et al., 2023; Navigli, Conia, & Ross, 2023).

4.3.2 Baselines

We created simple, linear baselines using Scikit-Learn (Pedregosa et al., 2011) for the classification tasks which consists of a Naive Bayes classifier with bag of words features. For the regression tasks, we used a support vector regressor with term-frequency inverse document frequency features.

We also built a BERT-base classification/regression model which was run for one epoch, with a batch size of 16 and a learning rate of 5e-5, for all sub-tasks. As this system out-performed the linear benchmarks on all sub-tasks, we refer to this as the baseline in the rest of the chapter.

Table 4.1: Performance of top 3 teams and baseline models on each task

Task	1a		1b	1c		2
	Acc	**F1**	**RMSE**	**Acc**	**F1**	**RMSE**
1st	0.9820	0.985	0.4959	0.4943	0.6302	0.4120
Team	PALI		abcbpc	PALI		DeepBlueAI
2nd	0.9750	0.9797	0.4977	0.4699	0.6279	0.4190
Team	stce		mmmm	mmmm		mmmm
3rd	0.9600	0.9676	0.5210	0.4699	0.6270	0.4230
Team	DeepBlueAI		Humor@IITK	SarcasmDet		HumorHunter
baseline	0.911	0.9283	0.8000	0.4731	0.6232	0.5769
	BERT		BERT	BERT		BERT
baseline	0.8570	0.8840	0.8609	0.4374	0.4624	0.6415
	Linear		Linear	Linear		Linear

4.3.3 Winning Systems

The best-performing systems all used PLMs to great effect. Enhancing the attention mechanism was a novel strategy used by PALI (Zhou, Ma, Yang, Jiang, & Mo, 2021), while ensembling the predictions of several models also proved useful. Similarly, strategies such as adversarial training and domain adaptations were used by the top-scoring teams.

Tasks 1a and 1c

PALI achieved the highest scores on the humor classification task. Their approach focused on improving the token representations obtained from PLMs They used RoBERTa (Y. Liu et al., 2019) as a backbone model and then implemented a sequential attention module (SAM) on the embeddings obtained from this language model. As existing attention mechanisms tend to focus on the token level, this team's attention module first calculated the importance of the representation's features at each dimension, and then re-weighted the features at the token level. This SAM module performed better than a simple Bi-LSTM layer and an attention-based Bi-LSTM and achieved the best scores in the humor detection and controversy detection tasks.

Task 1b

Team abcbpc (Pang et al., 2021) compared two PLMs - DeBERTa (He, Liu, Gao, & Chen, 2020) and ERNIE 2.0 (Sun et al., 2020). On a single task setup, they found ERNIE 2.0 performed better by a small margin, which was unexpected, given that this model outperformed RoBERTa in most tasks. They then used this model in a multi-task setup with different loss functions for the classification and regression tasks, and found that this achieved better results than single-task learning. Finally, they ensembled the models by taking the most popular label for the classification tasks and the average rating for the regression tasks, and found that this also improved performance and made the predictions more robust.

Task 2

DeepBlueAI (Song, Pan, Wang, & Luo, 2021) achieved high performance in subtasks 1a and 2 using PLMs with a variety of domain adaptations and training methods. Their system stacked RoBERTa and ALBERT (Lan et al., 2019) models and used task-adaptive pre-training (Gururangan et al., 2020) to adapt their PLMs to the dataset. This is an auxiliary task in which a masked language model is trained on the humor and offense dataset.

They then augmented the dataset by using pseudo-labelling to generate labels for the test set, and added these to the training data. Then, after encoding the input, they used adversarial training (Miyato, Dai, & Goodfellow, 2016), i.e. the addition of perturbations to the embedding layer, to obtain more stable word representations and to improve generalization. They ensembled all models produced and took the majority vote, or average vote to produce predictions. This team conducted an interesting ablation study in which they demonstrated that a combination of the above techniques delivered the best performance. However, using a knowledge distillation mechanism which trains a student model to predict the dataset targets and then uses this student to predict the probability of each sample being humorous and offensive, before minimising the loss between these hard and soft targets disimproved performance.

4.3.4 Trends in Experimental Results

Many of the teams carried out experiments to test which approaches are best suited for this problem. From their experimental setups, the main questions they focused on were:

1. Which PLMs are best suited for this task?
2. Does ensembling predictions from multiple models make predictions more robust?
3. Does multi-task learning help?
4. Which strategies for data augmentation, domain adaptation and training strategies improve performance?

Comparison of Pre-trained Language Models

A popular experiment amongst teams was to compare the performance of different PLMs. The most common finding was that when comparing single-task setups, RoBERTa gave the best performance on all tasks. DeepBlueAI found that RoBERTa large outperformed ALBERT. SarcasmDet (Faraj & Abdullah, 2021) found RoBERTa best for tasks 1a and 2, but BERT large performed best in task 1b. MagicPai (Ma et al., 2021) compared BERT, RoBERTa, XLNET (Yang et al., 2019) and ERNIE 2.0, and found that the models performed similarly, but RoBERTa was marginally better. Similarly, CSUMP6 (Essefar, Mekki, Mahdaouy, Mamoun, & Berrada, 2021) found that RoBERTa was better than BERT. Among teams who compared other models, ABCBPC and Humor@IIITK (Gupta, Pal, Khurana, Tyagi, & Modi, 2021) found that ERNIE performed better than DeBERTa. RedwoodNLP (Chi & Chi, 2021) found that ELECTRA (Clark, Luong, Le, & Manning, 2020) slightly outperformed RoBERTa. ES-JUST (Bashabsheh & Alasal, 2021) found RoBERTa better than XLM and BERT.

Table 4.2: Teams which compared PLMs, in order of performance on Task 1a

Team	Tasks	Models Compared	Best Model
DeepBlueAI	1a, 2	ALBERT (base) RoBERTa (large)	All: **RoBERTa** (large)
SarcasmDet	1a, 1b, 1c, 2	ALBERT BERT RoBERTa XLNET	1a: **RoBERTa** (large) 1b: BERT (large) 1c: rule-based 2: **RoBERTa** (large)
MagicPai	1a, 1b, 2	BERT (large) ERNIE (large) RoBERTa (large) XLNET (large)	All: **RoBERTa** (large)
DLJUST	1a, 1b, 2	BERT (base, large) BERTweet RoBERTa (base, large, irony) XLM-RoBERTa (large) XLNet	1a BERTweet 1b BERT (large) 2 BERT (large)
CS-UMP6	1a, 1b, 1c, 2	BERT (base, large) RoBERTa (base, large)	**RoBERTa** (large)
Abcbpc	1a, 1b, 1c, 2	DeBERTa (large) ERNIE 2.0 (large)	ERNIE 2.0 (large)
Humor@IITK	1a, 1b, 1c, 2	BERT DeBERTa ERNIE 2.0 RoBERTa XLNet	1a ERNIE 2.0 1b DeBERTa 1c **RoBERTa** 2 **RoBERTa**
HumorHunter	1a, 1b, 1c, 2	BERT (base, large) RoBERTa (base, large) DeBERTa (base, large)	DeBERTa (large)
RedwoodNLP	1a	BERT (large) ELECTRA (large) RoBERTa (large)	ELECTRA (large)
ES-JUST	1a, 1b, 1c, 2	BERT (large) RoBERTa (large) XLM-RoBERTa (large)	All: **RoBERTa** (large)
YoungSheldon	1a, 1b, 1c, 2	ALBERT (large) BERT (base) ELECTRA (base) MPNET (base) RoBERTa (base) XLNET (base)	1a BERT (base) 1b ALBERT (large) 1c MP-NET (base) 2 ALBERT (large)
hub	1a, 1b, 1c, 2	ALBERT BERT	1a ALBERT 1b ALBERT 1c BERT 2 ALBERT

4.3.5 Ensemble Methods

Ensemble methods, i.e. combining multiple base models to increase the robustness of the overall predictions, and improve generalisability, were a popular choice among competitors. Ensembles can be different models trained on the same dataset, or the same model trained on slightly different datasets. HaHackathon participants experimented with three types of ensemble methods:

1. Hard/soft voting: Hard voting takes the majority from all the base models and selects that as the prediction. In the case of regression, the predictions are averaged. Soft voting entails summing the predicted probabilities for each class and selecting the class with the highest sum probability. Other voting rules, such as a weighted sum, the median or the maximum are also possible, though they were not implemented by participants.
2. Cross-validation: a resampling method in which the data are split into k folds. An equal number of models are trained, and a different fold of the data acts as the test set each time. A voting method is used after.
3. Bagging/bootstrap aggregation: this entails subsampling datapoints with replacement, such that several models are trained on slightly different datasets. A voting method is used after.

A number of teams successfully used hard voting to improve performance. MagicPai and Humor@IIITK used a weighted average of models, depending on how well they performed in cross-validation. SarcasmDet used a hard-voting ensemble based on RoBERTa and BERT. Amherst685 (Gugnani, Zylich, Brookman, & Samoray, 2021) used a combination of hard and soft voting, and also selected the combination of models that performed best on the development set, and used that combination to predict the test set. Redwood, UPB (Smădu, Cercel, & Dascalu, 2021) and CSECU (Sultana, Ayman, & Chy, 2021) also used hard voting, and mostly found that it improved performance, although their failure to see improvements in all tasks should be seen in context of their low placement in the competition overall.

Table 4.3: Teams experimented with ensemble approaches, in order of performance on Task 1a

Team	Ensemble Method	PLMs Used	Improvement?
DeepBlueAI	Cross Validation Hard Voting	ALBERT RoBERTa	Yes
SarcasmDet	Hard Voting	BERT RoBERT	Yes
EndTimes	Bagging Soft and Hard Voting	BERT	Yes
MagicPai	Hard Voting (weighted)	BERT ERNIE RoBERTa XLNET	Yes
Humor@IITK	Hard Voting (weighted)	BERT DEBERTA ERNIE RoBERTa XLNET	Yes
Amherst685	Soft and Hard Voting	DistilBERT	Only for 1b and 1c
Abcbpc	Cross Validation	ERNIE 2.0	Yes
RedwoodNLP	Hard Voting	BERT ELECTRA RoBERTa	Yes
UPB	Hard Voting	BERTweet	Only for 1b and 2
CSECU	Hard Voting	ALBERT, BERT, ELECTRA RoBERTa MPNET XLNET	Yes

DeepBlueAI used 7-fold cross-validation, as well as training models with different hyperparameters, before using hard voting for the final predictions. Their experiments revealed that ensemble models performed better than single models. However, after examining correlations between predictions, they concluded that combining the least correlated outputs was more efficient than combining all of them. Similarly, Abcpbc's best solutions used 8-fold cross-validation with majority voting.

EndTimes (Pandey, Singh, & Mangla, 2021) used bagging before selecting the best models from each bag, and then used hard voting to finalise the predictions. They also tried summing the softmax score of each epoch for all of the bags, before using the argmax to generate the predictions for each bag. After this they used majority voting. This resulted in their best performing system.

4.3.6 Multi-Task Learning

Multi-task learning is often pitched in terms of human learning, in that humans often use knowledge from one task to aid in completing another task. From a machine learning perspective, it can be seen to introduce an inductive bias into the model, causing it to prefer one hypothesis over another. Along with ensemble methods, several teams also experimented with multi-task learning. There were two main approaches:

1. Hard parameter sharing: using one model for all tasks, which shares the hidden layers, but which has a specific layer/head for each separate task.
2. Soft parameter sharing: building a separate model for each task, and using regularisation between the parameters to encourage them to be more similar to each other.

Table 4.4: Teams who experimented with MTL approaches, in order of performance on Task 1a

Team	Tasks	MTL Method	Improvement?
MagicPai	1a, 1b	Interactive multi-task learning	MTL best for all
Humor@IITK	1a, 1b, 1c, 2	Hard parameter sharing	1c and 2 ✓ 1a STL 1c ensemble.
CS-UM6P	1a, 1b, 1c, 2	Hard parameter sharing	MTL best for all
Amherst685	1a, 1b, 1c, 2	Hard parameter sharing	1a and 1b ✓ 1c and 2 ensemble
	1a, 1b, 1c, 2	Hard parameter sharing	1a, 1b, 2 ✓ 1c STL
UPB	1a, 1b, 1c, 2	Hard parameter sharing	1a ✓ 1b, 1c, 2 ensembles

CS-UM6P (Essefar et al., 2021) used hard parameter sharing and found that multi-task learning gave their best results for all tasks. However Humor@IITK, Amherst685, abcbpc and UPB, using similar hard parameter sharing setups found that, in contrast to ensemble methods, multi-task learning gave better results in about half of the tasks, while ensembles prevailed in others. MagicPai's winning system implemented an interactive multi-task setup, in which the output from task 1 was concatenated to a copy of the BERT embedding and passed to a head to make predictions for task 1b.

4.3.7 Training Strategies/Data Augmentation

Beyond selecting the best model, using ensemble methods and multi-task setups, several teams experimented with training strategies or data augmentation approaches, in order to boost performance. DeepBlueAI used adversarial training, e.g. adding a perturbation to the embedding layer in order to may the model more robust to noise and found that it improved performance. MagicPai used a similar approach, and found it gave improvements, when combined with loss correction. However, DeepBlueAI found that knowledge distillation disimproved their scores. Finally pseudo-labelling, i.e. predicting the labels of the dataset and feeding them into the training data im-

proved, DeepBlueAi's score. Another data augmentation approach from Amherst685, back translation from English to other language and back again, in order to increase the amount of data available to the model did not improve performance - and the authors furnish interesting examples where the humor was not preserved on translation.

4.3.8 Domain Adaptations

Several teams experimented with domain adaptation, i.e. training a model on a source domain so that this knowledge may improve performance on a target domain. In terms of the HaHackathon data, the domain could be considered to be the genre of humor/offense, or it could be regarded as Twitter, which is the source of the majority of the data.

Adapting to the domain of humor/offense gave mixed results. For example, Amherst685 continued pre-training their PLMs using two large humor and offense datasets, but this did not result in performance improvements for them. However, IIITH created incorporated HurtLex features (Bassignana et al, 2018), a multilingual lexicon of offensive, aggressive and hateful words and found that it did improve their performance on both humor and offense. Amherst685 also experimented with ColBERT (Annamoradnejad & Zoghi, 2020), a model that is specifically finetuned on humor data, and again, did not note improvements. Taking a different approach, DeepBlueAI and IIITH experimented with task-adaptive pre-training (TAPT), in which finetuning is continued using the unlabelled HaHackathon dataset, and both teams found that this did help.

Adapting to the domain of Twitter gave more consistent results. For example, DLJUST found that BERTweet, a BERT-based model trained on 850 million tweets performed better than larger language models, such as RoBERTa base. Lending further credence to the value of adapting to the domain of Twitter, DLJUST (Al-Omari, Abedul-

Nabi, & Duwairi, 2021) and YoungSheldon (Sharma, Kandasamy, & B, 2021) used Ekphrasis (Baziotis et al., 2017) - a Twitter-specific preprocessing tool which normalises spelling correction such as ('cooooooooool' to 'cool') found that it improved their scores.

4.3.9 The Rule-Based System for Controversy

Many of the teams found it difficult to achieve good results for task 1c - humor controversy. Interestingly, SarcasmDet implemented a rule-based system: if their system predicted a humor score of 3 or more for task 1b, they automatically labelled the text as controversial. They placed an impressive 3rd with this strategy.

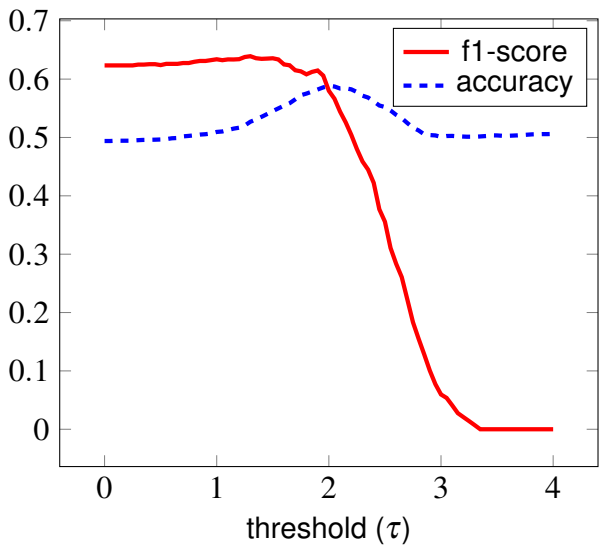


Figure 4.1: For varied values of a threshold, τ, accuracy and f1-score achieved by a hypothetical model predicting the label *controversial* for all texts in the test set with ground-truth humor score $> \tau$. Note that participants did not have access to these ground-truth scores for the test set, making these results an upper-bound for this type of threshold-based approach.

As we were interested in seeing if that rule could have been improved, we investigated the upper-bound of success for any threshold-based heuristic which determines whether a text was controversial given the humor score alone. Figure 4.1 shows the hypothetical F1-score and accuracy that could be achieved by such a system. Assuming a perfect score on humor rating prediction, if teams assigned a controversial label

for any text with a humor rating of over 2, they could achieve first place in this task in terms of accuracy with a score of 0.580. A threshold of 1.45 given perfect knowledge of the humor labels would result in a leaderboard-topping F1-score of 0.635. However, the teams that took part did not obtain the perfect humor rating scores required for this simple rule to work so effectively, yet were still able to achieve similar scores on the task. This suggests that their systems were learning something, but that ultimately the task is a difficult one.

Although we aimed to increase inter-annotator agreement in this task's annotation procedure, by matching the origin of the texts and annotators, the agreement on humor ratings was low, and indeed the task which aimed to capture this controversy proved difficult.

4.4 Conclusion

This chapter addresses the question of which models are most effective at capturing humor and offense ratings in text data. Our humor and offense detection challenge presented two classification tasks (humor and controversy detection) and two regression tasks (humor and offense rating) on the HaHackathon dataset, presented in Chapter 3. A total of 63 teams submitted systems for this challenge, making it the most well-attended SemEval humor detection challenge yet. Teams overwhelmingly used large language models in their systems. Many participants contributed interesting experimental findings with their system descriptions. For example, RoBERTa outperformed other large language models. Ensembling systems and multi-task learning were also successful approaches for most teams. Other useful approaches were adversarial training and pseudo-labelling, however knowledge distillation and back-translation were not helpful to teams who used them. Domain adaptations were more successful when adapting to the domain of Twitter than that of humor.

Chapter 5

Demographic Differences in Humor and Offense Ratings

5.1 Introduction

One of the driving motivations behind this thesis has been a pair of complementary goals: to incorporate findings from broader humor research into computational approaches to humor, and also to contribute an analysis of a large-N dataset to the field of Humor. As mentioned in Chapter 2, factors such as age (Kuipers, 2015), gender (Hofmann et al., 2020), personality (Ruch, 2010) and other demographic variables all modulate responses to humor, and so this chapter focuses on using the HaHackathon dataset to explore the following question:

RQ3: To what extent do ratings of humor and offense vary as a function of demographic characteristics?

Although the annotators of this dataset provided demographic data about their age and gender, this was not released as part of the humor detection task, and this is the first analysis of the impact of these age and gender on the humor and offense ratings in this large dataset. The analysis aims to uncover if humor and offense are as meaningfully linked in big datasets as they are in small-N studies, while validating evidence that there are gendered differences in the distribution of humor ratings (Svebak et al., 2004), as well as tolerance of aggressive humor.

5.1.1 Research Questions

Bearing in mind the annotation procedure, and demographic data collection described in Section 3.4 We break down the above RQ into the following sub-questions:

RQ3.1: Is there a correlation between annotators' perceptions of humor and offense? Does this vary by age and gender?

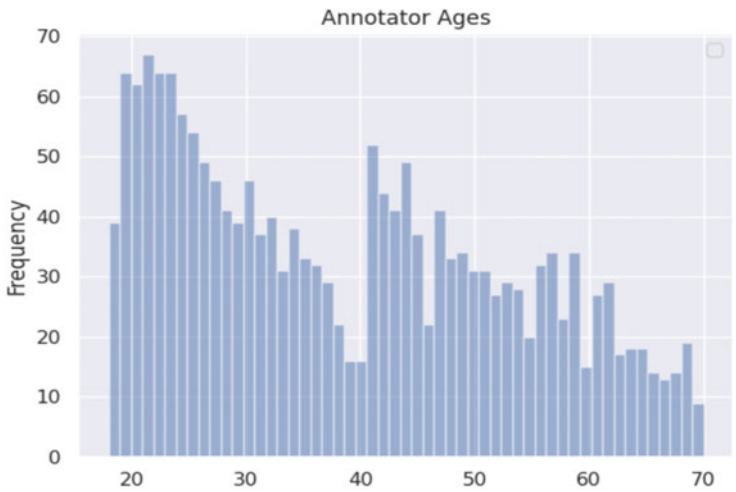
RQ3.2: Are there differences in humor *detection* and *comprehension* between groups?

RQ3.3: Are there differences in the distributions of humor and offense ratings between groups?

5.1.2 Data

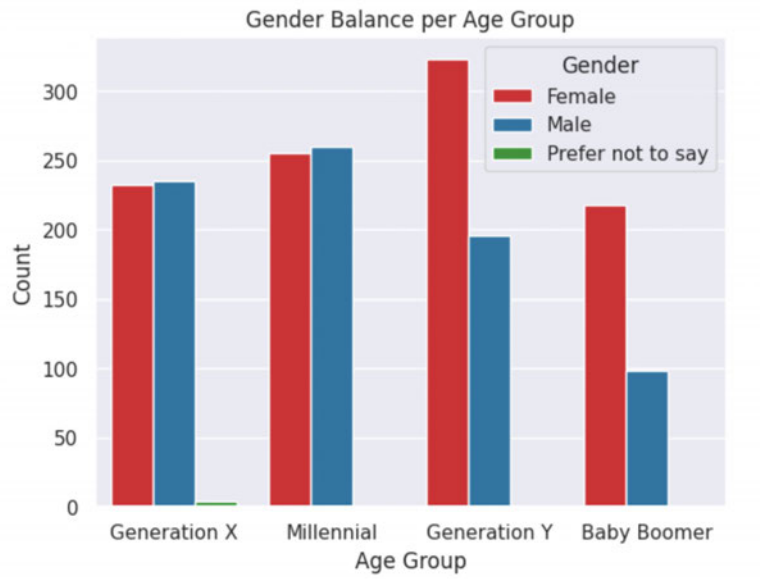
For this analysis, we excluded texts which had been labelled as 'not humorous' by our annotators, and removed outliers (e.g. texts that had fewer than 3 humor ratings). This left 121,622 ratings of 6,918 texts from 1,821 unique users. As described in the annotator selection procedure (Section 3.4.2), we aimed to represent age groups from 18-70 and Figure 5.1 indicates the distribution of ages in that range. We also binned these age groups roughly into Generation Z (18-25), Millennials (26-40), Generation Y (41-55) and Baby Boomers (56-70).

Figure 5.1: Annotator Ages



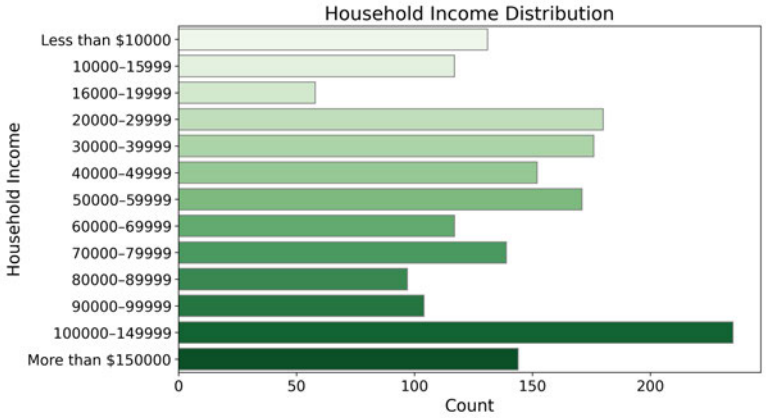
In terms of gender, 56% self-identified as female, 43% as male, and 1% declined to disclose their gender.

Figure 5.2: Annotator Gender



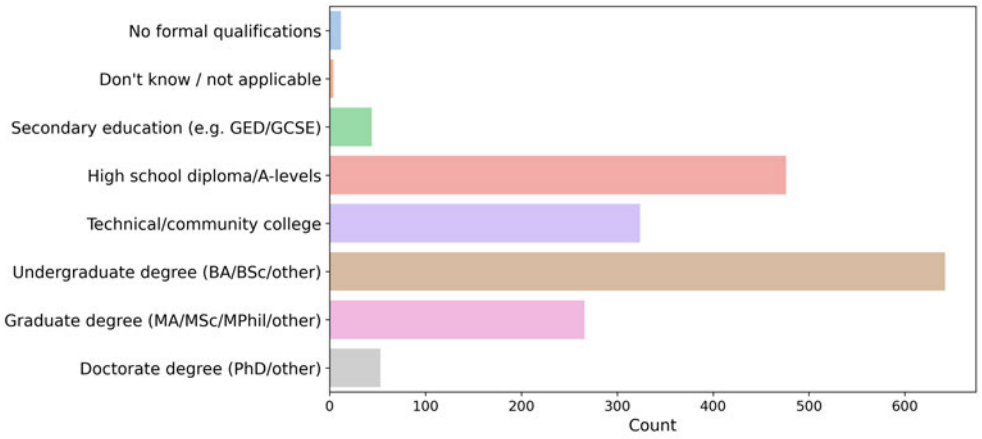
We saw in Section 3.4.2, that the annotators in the dataset are in the majority, White and Liberal. As an indicator of socio-economic status (SES), we looked at their self-reported household income in Figure 5.3, which appears to indicate an even spread of values. However, Kuipers et al. (2015) offer that self-reported income is not always the most trustworthy index of SES, owing to disparities in reporting of gross or net income and lack of knowledge about a household's total earnings.

Figure 5.3: Annotators' Household Income (USD)



In addition to household income, Figure 5.4 illustrates the highest level of educational award achieved by the annotators. In countries where education is not subsidised by the state, this can provide further insights into SES. The majority of the annotators have an Undergraduate degree, which in 2021, cost on average US$108,000 for a four-year degree Hanson (2021). This suggests that these annotators come from majority middle-class families who can support such expenditure, or take out a sufficient line of credit.

Figure 5.4: Annotators' Highest Level of Educational Attainment



5.2 Methodology

Given that the humor and offense responses were measured using an ordinal scale, for **RQ3.1**, we used the Spearman rank correlation (Spearman, 1904) to report the correlations between these variables. The Spearman rank correlation is a generalisation of the Pearson correlation (Freedman, Pisani, & Purves, 2007) which is used for discrete and ordinal data which captures the strength and direction of the relationship between two variables by ranking the values of each variable, summing the square differences and calculating the covariance of the ranks. This returns a correlation coefficient, ρ, ranging from -1 to +1, the magnitude of which indicates the strength of the relationship and the sign signifies the direction. It also returns a p-value - the probability that the value of the coefficient could occur under the null hypothesis.

To answer **RQ3.2**, we calculated the proportion of annotators from each group (i.e., gender or age group) that mislabeled (failed to *detect*) and who misunderstood (failed to *comprehend*) each text. For this analysis, we examined only the Kaggle texts, which came from a dataset of jokes, and in accordance with our annotation procedure, should all have been labelled as 'humor'. For misunderstanding, we counted the number of *'I don't get it'* ratings per group. The resulting distributions were non-normal, so we chose non-parametric tests, which do not assume an underlying distribution. As we have only two values for gender in the dataset, we used a Wilcoxon Signed Rank test (Wilcoxon, 1945) to examine the null hypothesis that the samples from male and female annotators came from the same distribution. This is similar to a paired t-test, and it ranks the absolute value of the pairs of differences to calculate the test statistic, *w*. With this test, we report the Common Language Effect Size (CLES) - the proportion of pairs where the values for one group are higher than the other.

For more than two groups, i.e., our age variable, which had four bins, we use the Friedman test (Friedman, 1937), which is similar to a repeated measures ANOVA. Again, the values are ranked and the test compares the mean rank of each group for statistical significance. In the case of a significant result, we ran post hoc pairwise Wilcoxon tests. We used the Bonferroni correction to adjust the p-values for multiple comparisons, reducing the risk of false positive results.

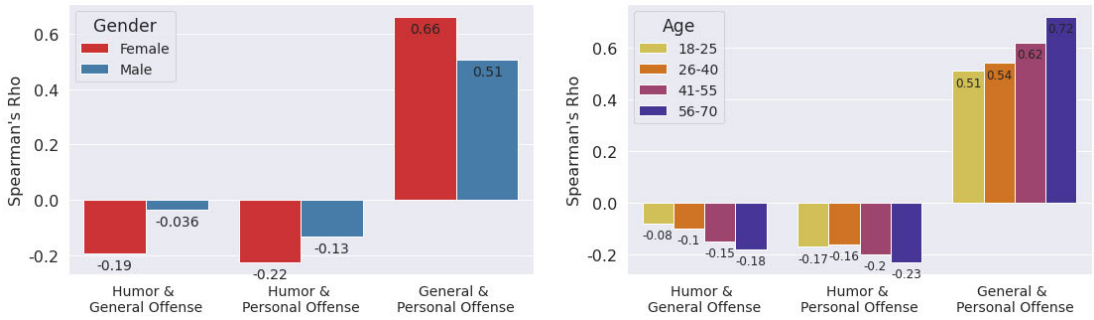
For **RQ3.3**, we first used the Wilcoxon and Friedman tests to determine if one group tended to give higher or lower ratings than another. We then used a chi-square test of homogeneity to examine how the distributions differed from each other. This test determines if the frequencies of each possible value of the dependent variable are distributed in the same way across the different groups. The test calculates the expected frequencies of each rating by each group by multiplying the number of annotators in each group by the true probability that any annotator would pick each answer. This expected frequency is then compared to the observed frequency.

5.3 Results

5.3.1 RQ3.1: Is There a Correlation between Humor and Offense?

Overall, without binning the ratings by demographic group, there was a small negative correlation between humor and general offense ($\rho = -0.13$, $p < 0.05$), and this grew stronger for humor and personal offense (ρ =-0.19, $p < 0.05$), which suggests that offensive content is negatively related to humor appreciation. There was a moderate-strong correlation between general and personal offense (ρ=0.60, $p < 0.05$), indicating that these concepts are linked, but are not identical.

Figure 5.5: Correlations between Humor and Offense by Gender and Age



Correlations between ratings by Gender

When examining the correlations between ratings split by gender, an interesting trend emerged (Figure 5.5). There was almost no relationship between humor and *general* offense for men, however *personal* offense ratings were negatively correlated with humor ratings. Conversely, for female annotators, both types of offense were more strongly correlated with a reduced humor rating for female annotators.

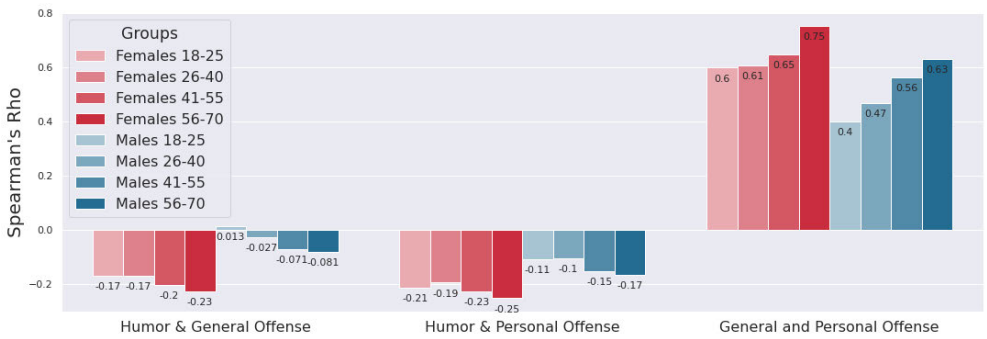
Correlations by Age

A second interesting trend emerged in terms of age: the older the annotators were, the stronger the negative link between general *and* personal offense on humor ratings was (Figure 5.5). The oldest group had the most prominent negative correlation between humor and both types of offense, as well as the strongest correlation between the two offense metrics.

Correlations by Age and Gender

Although splitting 20 ratings per text into 8 groups (for four age groups by two gender groups) would cause issues of data sparsity and statistical power, we noted that the trend of an increasingly negative correlation between humor and offense continues when this is broken down by age and gender (Figure 5.6). Female annotators relate higher offense scores to lower humor scores increasingly with age, and this trend is much less pronounced in male annotators.

Figure 5.6: Correlations between Humor and Offense by Age and Gender



5.3.2 RQ2: Are There Differences in Humor Detection and Comprehension Between Groups?

Humor Detection

To investigate differences in annotators' humor *detection*, we looked at the proportion of male and female annotators who labelled each text from the Kaggle data as 'not humorous'. We confined this analysis to the Kaggle data because all texts in this dataset were intended to be humorous, and should have been labeled as such. A paired Wilcoxon signed rank test showed that there was no significant difference between groups (z = 134201.0, p=0.29).

Table 5.1: Mislabeling and Misunderstanding in the Kaggle Jokes

	Male	**Female**
Proportion of annotations from each group	42.92%	57.08%
'Not Humor' ratings from each group	3.79%	3.70%
Unique texts with 1+ label of not-humorous	22.91%	25.42%
'I don't get it' ratings from each group	5.77%	7.35%
Unique texts with 1+ rating of 'I don't get it"	33.04%	45.81%

We used a similar procedure to test if there were significant differences between age groups in terms of humor detection. A Friedman test showed that there were no significant differences between groups (χ^2 = 6.976, p=0.07).

Humor Comprehension

After labeling a text as humorous, one of the options for humor rating was 'I don't get it'. This indicated that the annotator had recognized that the text was intended to be humorous, but that they lacked the knowledge to fully understand the joke. We first looked at the Kaggle dataset, and calculated the number of 'I don't get it' votes from men and women, as a proportion of the total votes per text from each group. A paired Wilcoxon signed rank test showed that there was a significant difference between groups (z = 214403.0, $p < 0.05$). We used Pingouin (Vallat, 2018) to calculate the Common Language Effect Size (CLES), i.e. the proportion of pairs where the

proportion of 'I don't get it' ratings provided by female annotators is greater than the proportion of male annotators who gave that rating. The resulting CLES of 0.5540 indicates that a larger proportion of female annotators indicated that they did not get the joke in 55.45% of pairs. When looking at the data from Twitter, women still admit to not getting the joke more than men (z = 2298680.0, p <0.05), but the effect is less pronounced, CLES = 0.5223.

We binned the ratings by age group to examine if any groups comprehended jokes less than another. A Friedman test showed that there were no significant differences between groups (χ^2= 0.0012, p=0.06).

5.3.3 RQ3.3: Are There Differences between Groups in Distributions Humor and Offense Ratings?

When looking at the distribution of ratings across the 6 possible values (1-5 and *'I don't get it'*) for the entire dataset (both Kaggle and Twitter), a χ^2 test of homogeneity demonstrated that there were significant differences between the distributions of humor ratings between men and women (χ^2= 202.25, $p < 0.05$) and showed that women were more likely to select 'I don't get it', while men were more likely to use higher ratings. We also explored if this difference translated into different average humor ratings per text and a Wilcoxon signed rank showed that men gave significantly higher ratings than women on humor (z = 9684516.5, $p < 0.05$) and the CLES score of 0.5333 indicated that men gave higher humor ratings in 53.33% of pairs.

For general offense, a χ^2 test of homogeneity showed significant differences between groups (χ^2 = 430.85, $p < 0.05$), and examining the expected versus observed counts showed that the trend seen in the humor ratings was reversed: men were more likely to choose low offense ratings and women were more likely to select higher values. In terms of averaged general offense ratings, group differences were significant (z = 4260050.5, $p < 0.05$, CLES = 0.4704), with men giving higher offense ratings in 47.04% of pairs.

Similarly, for personal offense, a χ^2 test of was significant (χ^2 = 1195.94, $p < 0.05$) with a more pronounced trend showing that women were more likely to select a high personal offense rating, and men systematically under-selected high ratings. This led to significant differences in the average personal offense ratings per text, where men gave higher personal offense scores in only 41.5% of pairs (z = 1234096.5, $p <$0.05, CLES = 0.4146).

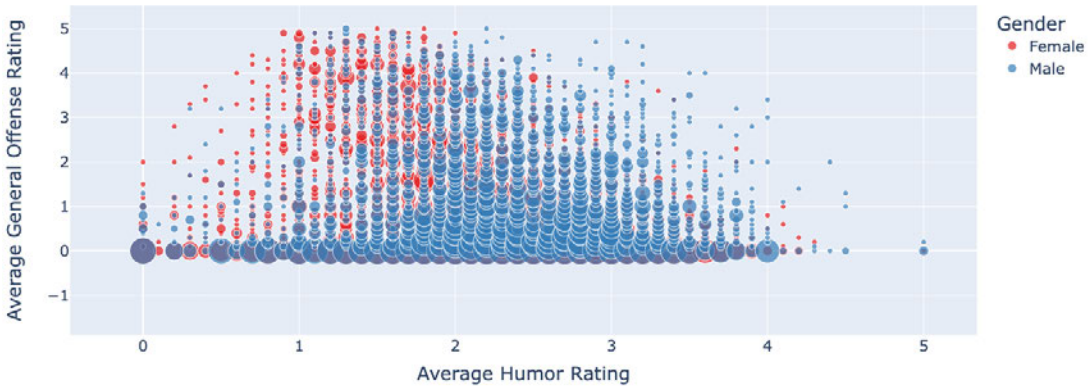
When looking at age groups, a χ^2 test showed significant differences in humor ratings between age groups (χ^2 = 239.98, $p < 0.05$). The oldest group, 56-70, were most likely to report 'I don't get it', while annotators aged 26-40 were least likely to use this, and most likely to give high ratings. In terms of general offense, there were significant group differences (χ^2 = 540.936 $p < 0.05$), and annotators 18-40 were more likely to give lower general offense ratings, while those aged 41-70 used fewer low ratings than expected, and the group ages 56-70 was most likely to give the highest possible offense rating of 5. Group differences were more pronounced in personal offense ratings (χ^2 = 1387.43, $p < 0.05$) where the two youngest groups gave consistently lower than expected ratings of personal offense, while the older group gave consistently higher ratings. This resulted in significant differences in the average personal offense scores between groups (χ^2 = 38.223, $p <$0.05).

5.4 Qualitative Analysis

The negative correlation for female annotators between humor and general offense, which was uncovered in the above analysis, is succinctly illustrated in Figure 5.7. Texts which are offensive for women tend to earn a lower humor rating, while general offense is more tolerated by men.

To examine what type of texts male and female annotators differed on with regard to general offense ratings, we selected the top 40 texts where there was at least a 1.5 point difference between the mean general offense score given by male and female annotators. We labeled the topic or target of the texts and five annotators

Figure 5.7: Relationship Between Humor and Offense by Gender



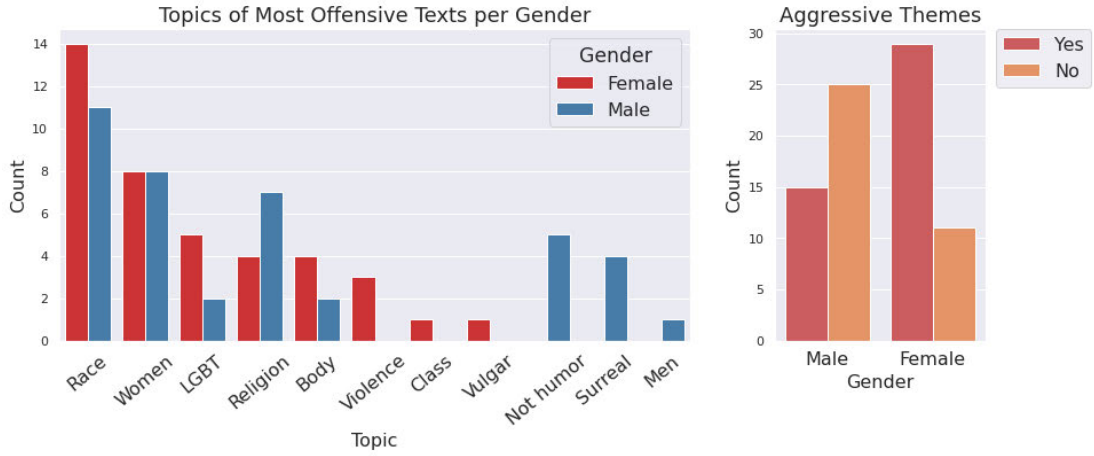
rated whether the content was aggressive or not. Annotators were instructed that a text should be deemed aggressive if it contained violent content or used racial slurs, and inter-annotator agreement was relatively high (Fleiss's Kappa, an extension of Cohen's Kappa for use with three or more raters was used, $\kappa = 0.3815$).

Table 5.2: Sample Texts Where Annotators Differed on General Offense

Text	Humor		G. Offense	
	Female	**Male**	**Female**	**Male**
Why are the labia on Japanese women oriented sideways instead of vertically? Goes better with their eyes.	1.0	2.2	4.2	1.3
In my spare time, I help blind kids I mean the verb, not the adjective	1.3	2.0	2.2	0.17
Two condoms walk by a gay bar... One says to the other, "Wanna go inside and get shitfaced?"	2.6	1.6	0.85	2.4
What did the Jewish pirate say when he heard his wife died? Argh, shiva me timbers	1.6	1.6	1.0	2.1

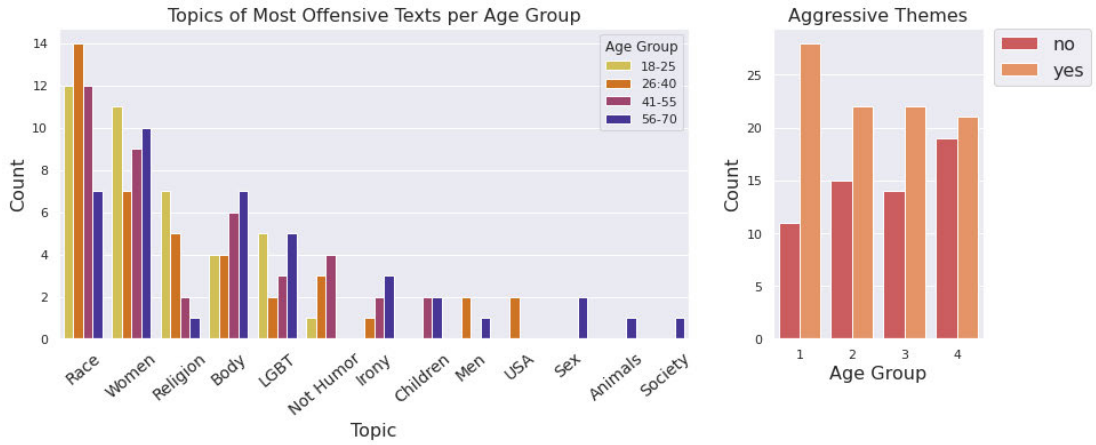
There was a sizeable overlap of topics, with women finding texts about the LGBT community more offensive than men, while male annotators found texts about religion more generally offensive. The texts that were offensive to women tended to be aggressive, while men were more tolerant of this. Interestingly, men selected several texts which were not intended to be jokes (e.g. were drawn from accounts supporting targets of hate speech) as both humorous and offensive.

Figure 5.8: Topics and Aggression where Gender Groups Disagreed on General Offense Ratings



We followed a similar procedure to examine the texts where offense ratings from different age groups differed from each other. We compared the mean general offense rating from each group to the average general offense rating from the other 3 groups combined, and looked at the top 40 texts where there was at least a 1.5 point difference. Several topics predominate, namely race, women, body (e.g. physical disability, high body weight). The texts rated as more generally offensive by younger groups focused on these topics, but as age increased, so did the variety of topics featured. The texts selected by group 1 (the youngest group) featured more which were aggressive in nature, but as age increased, aggression was less linked to offense.

Figure 5.9: Analysis of Topics and Aggression where Age Groups Disagreed on General Offense Ratings



5.5 Discussion

We used a large dataset of texts rated for humor and offense, along with some demographic information about the annotators to explore differences between age and gender groups. We looked at how the groups link humor and offense, differences in humor detection and comprehension, as well as differences in the distributions of ratings.

RQ3.1: We found that female annotators negatively link humor and offense more strongly than men. Male annotators do not relate general offense with lower humor ratings. In fact, they link humor and offense to a lesser extent, and only when personally offended.

As regards age groups, the correlation between humor and offense was weakest in the youngest group, and grew steadily with age - as did the link between general and personal offense.

RQ3.2: There were no differences in gender or age groups in terms of humor detection. However, when it came to humor comprehension, women selected 'I don't get it' more often than men.

RQ3.3: In terms of the distributions of ratings, women gave lower humor ratings and higher offense ratings, while men showed the opposite trend. Amongst the age groups, annotators 26-40 gave the highest ratings and the fewest reports of 'I don't get it'. In line with findings from **RQ3.1**, younger groups gave lower offense ratings and older groups reported higher offense.

Some of the findings above are well attested in the humor literature, albeit in smaller-N studies. Hofmann et al (2020) report that men's tolerance of aggressive humor is one of the most consistent findings in the humor field, with seven out of eight studies mentioned replicating this result. Our qualitative work shows that aggressive texts featured more prominently in the texts for which men and women's general offense ratings differer most. Perhaps relatedly, Proyer and Ruch (2010) report that men score higher on katagelasticism - the joy of laughing at others. This may be reflected in the fact that general offense does not diminish male annotators' humor ratings, only personal offense does.

A more surprising result is the increasingly strong negative correlation between humor and offense as age progressed. This contradicts the oft-touted idea of *Generation Snowflake*, which contends that those born after 1995 tend to be the most overly reactive to offensive material (Haidt & Lukianoff, 2018). The older age groups - 40-55 and 56-70 - gave higher ratings of offense than their younger counterparts, and our qualitative analysis indicated that the older groups gave higher offense ratings to a wider variety of topics. This replicates Kuipers (2017) findings that older Dutch adults found that humor which appealed younger generation was offensive.

The finding that women used the 'I don't get it' label more than men is a result that may benefit from some contextualisation from the humor literature. Bell (2013) found that when shown jokes that were *designed* to be incomprehensible, women tended to explicitly state that they did not get it, while men implicitly signaled it by asking concept-

checking questions, masking their incomprehension. It is not possible to determine if this was the case here, but it is true that the qualitative results uncovered that men more often selected negative examples (i.e. those selected from non-humorous accounts) as both humorous and offensive.

5.5.1 Implications

Given the gender and age group differences in ratings of humor and offense, it is evident that humor detection systems which average over all annotators' ratings fail to model the subjectivity that is inherent to this task. These systems may not generalise well on downstream tasks, such as content moderation, and may not be effective at moderating aggressive content if they are tuned to men's preferences, or alternatively may be more restrictive if tuned to women's preferences. Furthermore, as sociologists have pointed out (Lockyer & Pickering, 2005), the line between humor and offense is continually under revision in most societies, therefore not only are these responses subjective, but they are a moving target. We should focus on incorporating frameworks to include demographic knowledge in our systems, which can constantly be updated to reflect society's changing definitions of humor and offense.

5.5.2 Limitations

It is a limitation that the dataset did not afford the opportunity to explore the interaction between age and gender. As each text has approximately 20 annotations per text, splitting these into 8 groups to model age and gender would not have provided sufficient statistical power. Similarly, it is a limitation that there were insufficient annotations from gender non-conforming annotators, as there is a dearth of literature on their reactions to humor and offense. The lack of annotators that self-identify with genders other than female and male has been noticed in the past in different tasks as well (Excell & Moubayed, 2021; Prabhakaran, Davani, & Diaz, 2021). Indeed the

annotator pool is relatively homogeneous and it is crucial to highlight the findings of this chapter represent the views of a majority White, Liberal, educated and middle-class annotator pool. A more nuanced treatment of humor must elicit the views of a broader section of society.

A final constraint is that we are modelling only one half of the humorous interaction - the recipient of the joke. Excluding the teller of the joke can deny the recipient some important context needed to enjoy the joke. For example, a sexist joke told by a woman may be less offensive to women by virtue of the fact that it is being told by someone from within their group (Veatch, 1998). Future work should include this dimension.

5.6 Conclusion

We present the first analysis of the demographic data provided with the HaHackathon data - a large dataset used to train systems for computational humor detection. Our findings indicate that women negatively link humor to offense, while men only do so if they are personally offended. Links between humor and offense grew with age. There were no differences in humor detection by gender or age groups, but women and older annotators indicated that they did not understand jokes more than men. Distributions of humor and offense ratings replicated findings from humor research, namely that men gave higher humor ratings and lower offense ratings. We hope that these findings will inform future frameworks for computational humor detection and dataset creation.

Chapter 6

Predicting Humor and Offense Ratings

6.1 Introduction

Rating prediction tasks are ubiquitous in NLP, thanks to the proliferation of online reviews, which tend to comprise a textual comment, accompanied by a star-ratings. The dataset described in Chapter 3 is similar in nature - a text and a discrete, ordinal rating. This chapter explores the challenges and opportunities presented by predicting such ratings, in particular, what auxiliary sources of information can improve prediction performance.

As indicated in Section 2.3 of the literature review, demographic factors can help to distinguish between groups in terms of their humor preferences. This is supported by the results from the Chatper 5, which indicate that there are significant differences between genders and age groups in terms of how strongly their perception of offense impacts their perception of humor. Female and older annotators showed stronger negative links between humor and offense than did men and younger annotators. Demographic group differences were also visible in the distribution of ratings - women gave lower humor ratings and higher offense ratings than the expected distribution, and were most likely to select 'I don't get it' as a rating. Conversely, younger people gave higher humor and lower offense ratings, and were least likely to choose the 'I don't get it' option.

Given that demographic characteristics help to differentiate people in terms of humor in a real world setting, we ask if knowledge of the demographic characteristics of the annotator is a useful auxiliary source of information to improve computational rating prediction performance.

The second auxiliary source of interest is the ordinality of the rating themselves. As highlighted in the literature review, rating prediction tasks are often modelled in terms of classification, or regression. We explore whether classification models which do not attend to the ordinal nature of the data can achieve similar performance to systems which do consider the order of the labels.

6.1.1 Research Questions

RQ4.1: Does demographic information improve prediction performance for humor and offense?

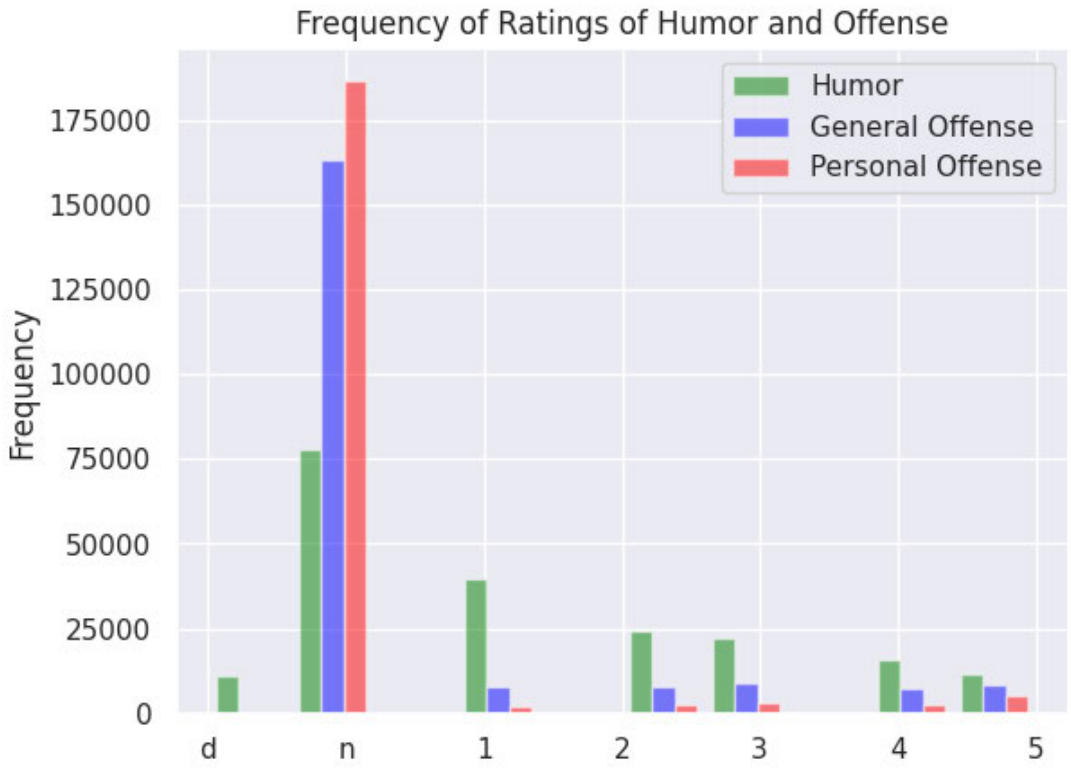
RQ4.2: Is the ordinal nature of ratings an important factor in the modelling approach taken?

6.2 The Dataset

The dataset contains the same texts and ratings as described in section 3.4, i.e. 10,000 texts with a minimum of 20 annotations per text, for a total of 202,370 ratings. Annotators were asked if the text was intended to be humorous (yes or no), and if they selected yes, they were asked to give a Likert scale rating of how humorous they found it. In the case of humor, annotators were allowed to select a rating of 'I don't get it' (indicated with 'd' in the below plot). They were also asked if the text was generally or personally offensive, and to rate the offensiveness if they selected yes. In the case that they rated the text as not humorous, they were not asked to rate its offensiveness.

Unlike the HaHackathon dataset, in which we aggregated the data into average ratings, we worked with the individual ratings in this task. Figure 6.1 shows the distribution of ratings, and indicates that for all tasks, the majority label was not humorous/offensive. This plot also demonstrates how few positive examples there are in the personally offensive data, and for this reason, we omit this from our analysis and focus on humor and general offense.

Figure 6.1: Histogram of Humor, General Offense and Personal Offense Ratings



Following from the annotator selection procedure outlined in section 3.4.2, this dataset contains ratings from 1,821 unique users. 56% self-identified as female, 43% as male, and ~1% declined to disclose their gender. We also collected information about the annotators' income, employment status, sexual orientation, political views, relationship status and a measure of the big 5 personality traits or extroversion, agreeableness, conscientiousness and emotional stability.

6.3 Modelling Options for Ordinal Data

As mentioned above, many rating prediction tasks model the task as classification. For **RQ4.2**, we aim to contrast this approach to other methods which take into account the ordinal nature of the data.

The exploration of different modelling approaches is motivated by the idea that there are flaws with both classification and regression. Classification algorithms treat the labels as nominal data, a strong assumption which discards the ordinal nature of Likert scale data. At a system level, deep classification models trained with cross entropy loss (e.g. Frangidis, Georgiou, and Papadopoulos (2020); Venkata Raju and Sridhar (2020)), all incorrect predictions are considered to be equally wrong by the model, e.g. given a ground truth label of 1, a system that predicts 2 is considered the same as a system that predicts 5. Intuitively, this is problematic, and particularly for the application of offense detection, a system that sees a text with a ground-truth label of 5 for offense, but which mislabels it as a 1 would be unsuitable for deployment in a real-world scenario.

Regression treats the rating as an interval rating, which is similar to an ordinal rating, with the exception that it assumes that there is a fixed and equal distance between a rating of 1 and 2, which equals the distance between a rating of 4 and 5, and that it is valid to subtract these differences. This is a strong assumption to make of ordinal data, which considers only that one rating is higher or lower than another, but does not presuppose the distance between them. This is supported by several psychological studies, which report that participants do not assume equidistance between Likert scale items, and tend only to interpret the endpoints of the scale consistently (Bishop & Herron, 2015; Dawes, 2008; Leung, 2011).

An alternative approach, *ordinal regression* (also known as ordinal classification and as ranking learning) is posed as an intermediate approach between classification and regression. It is motivated by teaching the model to get better at predictions that are closer to the ground truth label.

Ordinal regression is commonly implemented as an extended binary classification problem. Given a dataset $D = \left\{x^{[i]}, y^{[i]}\right\}_{i=1}^{N}$ where $x^i \in X$ indicates the inputs to a supervised model, and y^i refers to the label or rank of the $x^{[i]th}$ example. The labels belong to the set $\{r_1, r_2 \ldots r_k\}$, and the ranks are ordered $r_k > r_{k-1} > \ldots > r_1$. Ordinal regression splits the prediction task into multiple binary classification subtasks, the output of which $y_i^k \in \{0, 1\}$ represents the prediction of whether y_i exceeds rank r^k. In practice, given a set of K ranks, the output layer of the model consists of $K-1$ nodes. The first node outputs a prediction of whether the label exceeds rank 1 , the second node gives a prediction that the label exceeds rank 2, up to the $K-1^{th}$ node, which outputs a probability that the label exceeded the $K-1^{th}$ label, which would assign it to the K^{th} label. To output a final label prediction, the number of nodes where the predicted rank exceeds the rank relating to the current node are summed, and one is added.

Niu et al. (2016) used ordinal regression to predict age as an ordered variable, achieving better predictive performance over a standard regression system, which did not account for ordinality. However, despite their encouraging results, the authors noticed a problem of rank inconsistency in their predicted labels. This is the case where a system predicts that a label is greater than 4, but *not greater than* 3. Although this issue did not appear to impact performance on their data, the authors suggested that such inconsistency could interfere with the final label prediction output, which is the sum of nodes where the predicted rank is greater than the rank related to the current node.

The rank-consistent ordinal regression system, CORAL (Cao, Mirjalili, & Raschka, 2020) attempted to address this rank inconsistency by imposing weight sharing across all output nodes, such that they only differ in their bias terms. While this system achieved better performance on age prediction than a classifier trained with cross-entropy, the authors admit that the weight-sharing constraint likely limited how much their network could learn. It appears contradictory to hamstring the model by discarding the output layer weights simply in order to maintaining the ordinal information,

and for this reason, we did not implement this type of ordinal regression. Section 6.4.5 describes the CORN model (Shi, Cao, & Raschka, 2023), an alternative rank-consistent ordinal regression network, which does not impose weight-sharing, and is implemented in this chapter.

6.3.1 Metrics for this Ordinal Data

Although mean absolute error, i.e. the absolute difference between system predictions and ground truth (Equation 6.1) is a commonly used metric for ordinal regression, and is the one used in the original CORN paper to assess the model, it has a number of limitations. Several papers have pointed out that this metric assumes a predefined interval between the ordinal ranks, which as discussed in section 6.3, is a strong assumption to make. It is also unsuitable for unbalanced datasets, because all errors are treated alike. That is to say, if the majority class is 1, and the model consistently makes predictions that are close to 1, while never correctly predicting scores at the other end of the scale, these more serious errors will be averaged out. In the case of offense detection, particularly in a production setting, it is important to penalise a system that fails to detect extremely offensive content.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6.1}$$

Baccianella et al. (2009) proposed macro-averaged MAE (henceforth MAE^{M}, Equation 6.2) to tackle the issue of imbalance. As opposed to computing MAE across documents, MAE^{M} computes MAE per class (where Cej is the set of texts whose ground truth label is y_j) before averaging the results across classes. For datasets with one or more ranks that appear very frequently, MAE^{M} gives an equal weight to each class, instead of allowing a class's contribution to the error calculation to be weighted proportionally to its frequency. However, it does still suffer from the assumption of equidistance between ranks, and is recommended only as a secondary measure of performance (Sakai, 2021).

$$MAE^M = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{|C_j|}\sum_{y_i \in C_j}|y_i - \hat{y}_i| \tag{6.2}$$

After a series of empirical comparisons, Sakai (2021) proposed Cohen's Kappa (Cohen, 1968) (Equation 6.3) as the most robust metric to use for ordinal regression systems. This is a comparison of the expected agreement between the system predictions and the gold labels when these are independent. Equation 6.3 defines Kappa, where p_o is the probability of agreement between labels and ground truth, and p_e is the expected agreement when the labels are assigned at random. This will be the primary metric, implemented with TorchMetrics (Detlefsen et al., 2022) with MAE^M as the secondary metric, implemented with Imblearn (Lemaître, Nogueira, & Aridas, 2017).

$$\kappa = (p_o - p_e) / (1 - p_e) \tag{6.3}$$

6.4 Methodology

This section describes two methods of incorporating demographic information in humor and offense-rating detection models. We also describe the dataset split and the cross-validation approach used to improve model generalisation. Finally, we describe the baseline system, and the transformer-based classification and an ordinal regression models.

6.4.1 Text and Demographic Information

To incorporate the demographic information, we experimented with two different approaches:

- **Textual Descriptions:** Concatenating a description of the annotator to the text. In the input, the text and description are separated by a `[SEP]` token, in order for the model to understand that they are two separate inputs. See Table 6.1 for examples.
- **Multi-layer Perceptron on Categorical Features:** Passing each demographic characteristic (captured as categorical and numerical features) to its own multi-layer perceptron (MLP) before concatenating it to the textual embedding output from the transformer and passing it to a final classification or ordinal regression layer. This approach was implemented with the Multimodal Toolkit (Gu & Budhkar, 2021)

We also used two different quantities of information - sparse and full. In the sparse setting, the information used is the annotator's gender, age, and the generation they belong to (e.g. Generation Z, Millennial, Generation X or Baby Boomer). These variables were selected based on the statistical differences seen between gender and age groups discussed in the previous chapter, which we hypothesised could serve as a useful input to the predictive model. The full demographic information is a textual description of all the variables collected. Examples are seen in Table 6.1.

6.4.2 Cross Validation

In order to avoid overfitting to annotators seen in the training data, we implemented 5-fold cross-validation. The data was split into 80% training data and 20% test set. Although some texts were shared between train and test, there were no shared annotators between data splits, meaning there were 1456 unique annotators in the training set, and 366 in the test set.

Table 6.1: Text and Demographic Information Examples

Input	Example
Text Only	The hardest decision to make at midnight on New Year's Eve is what room of the house you want to hide and cry in.
Sparse Demographic	The hardest decision to make at midnight on New Year's Eve is what room of the house you want to hide and cry in. **[SEP]** I am an East Asian Female who is 22.0 years old. I am in generation X
Full Demographic	"The hardest decision to make at midnight on New Year's Eve is what room of the house you want to hide and cry in. **[SEP]** I am an East Asian Female. I am 22.0 years old, so I am in generation X. I was born in Taiwan and I currently live in California (CA), United States. My sexual orientation is heterosexual and romatically, I am in a romantic relationship with someone from the same cultural background as me . My nationality is United States. Am I living abroad? No. Culturally, I identify as a multicultural individual. My first language is English, I know one other language in addition to English. My highest level of education completed is Undergraduate degree (BA/BSc/other) Am I a student currently? No. My employment status is Other and my household income is $100000–$149999. Politically, I am Democrat and Pro-choice and my religion is Non Religious . My concern about environmental issues is rated as 4. In terms of personality,I score a 1.5 for extroversion, 3.5 for agreeableness, 6.0 for conscientiousness, 2.0 for emotional stability, and 3.5 for openness.

6.4.3 Baseline Model

As Sakai (2021) points out in his paper, trivial baseline systems, e.g. those which select the majority label for each text, are not appropriate for use when Cohen's Kappa is the evaluation metric, as they will result in $\kappa = 0$ for such systems. For this reason, our baseline model is a dummy classifier from Scikit-learn (Pedregosa et al., 2011), which ignores the input text and, using a stratified approach, generates a label at random, respecting the distribution of labels in the training set.

For both humor and offense, although the distributions of predicted labels are very similar to those seen in the training sets, the κ and MAE^M values show poor performance (Tables 6.3 and **??**).

Table 6.2: Distributions of ground truth labels and dummy classifier predictions

Rating	Ground Truth Humor	Dummy Humor	Ground Truth Offense	Dummy Offense
Don't get it	7892	8300	N/A	N/A
No	1121	1097	16977	17129
1	4135	4138	933	773
2	2823	2462	930	790
3	2410	2309	931	967
4	1704	1759	736	731
5	1208	1228	786	903

6.4.4 Classification Model

Both the classification and ordinal regression models are based on DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2019), a BERT-based language model. BERT (Bi-directional encoder representations from transformers) (Devlin et al., 2018) is a pre-trained language model (PLM) based on the transformer architecture (Vaswani et al., 2017) which comprises a stack of encoder-decoder networks and an attention mechanism, and is designed to focus on contextual relationships between words. BERT learns bidirectional language representations from large corpora of unlabelled data during pre-training. DistilBERT is trained by using a large *teacher* model to train a smaller *student* model to reproduce the behaviour of the teacher using cosine embedding loss. This knowledge distillation reduces the number of BERT's model parameters by 40%, while still maintaining 95% performance. Its lightweight size and speed makes it ideal for model prototyping.

During finetuning, we initialised the model using Huggingface (Wolf et al., 2020), added two fully-connected layers and an output layer, which predicts the most probable label for a text, given the training data. All models were implemented in PyTorch Lightning (Falcon & The PyTorch Lightning team, 2019), trained for 3 epochs with five folds per epoch. We used a batch size of 32 and a learning rate of 0.0002.

The classification model is trained with multi-class cross-entropy (Equation 6.4) which calculates the the loss for each class label per observation and sums them.

$$L(\hat{y}, y) = -\sum_{k}^{K} y^{(k)} \log \hat{y}^{(k)} \tag{6.4}$$

6.4.5 Ordinal Regression Model

The ordinal regression approach implemented in this chapter is CORN, which tackles the rank inconsistency problem without imposing the weight-sharing constraints of the CORAL model. Similarly to other approaches, it splits the problem into K-1 binary classification problems, however the output of each node represents the conditional probability that the label exceeds a particular rank, given that it exceeds the previous rank.

$$f_k\left(\mathbf{x}^{[i]}\right) = \hat{P}\left(y^{[i]} > r_k \mid y^{[i]} > r_{k-1}\right) \tag{6.5}$$

The conditional probabilities are calculated by splitting the data into conditional training subsets during training, i.e. datapoints where $y^{[i]} > r_1$. The model then multiplies the conditional probabilities to obtain the unconditional probabilities.

$$\hat{P}\left(y^{[i]} > r_k\right) = \prod_{j=1}^{k} f_j\left(\mathbf{x}^{[i]}\right) \tag{6.6}$$

The loss function is seen in Equation 6.7. Instead of minimising $K-1$ loss functions for each conditional subset, the losses are summed, in order to work on all binary tasks simultaneously.

$$\begin{aligned} L(\mathbf{Z}, \mathbf{y}) = -\frac{1}{\sum_{j=1}^{K-1} |S_j|} \sum_{j=1}^{K-1} \sum_{i=1}^{|S_j|} [&\log\left(\sigma\left(\mathbf{z}^{[i]}\right)\right) \cdot \mathbb{1}\left\{y^{[i]} > r_j\right\} \\ &+ \left(\log\left(\sigma\left(\mathbf{z}^{[i]}\right)\right) - \mathbf{z}^{[i]}\right) \cdot \mathbb{1}\left\{y^{[i]} \leq r_j\right\}] \end{aligned} \tag{6.7}$$

The output layer of the model consists of $K-1$ nodes, each of which outputs a probability of whether the rating exceeds the rank associated with the current node (If the rank is predicted to exceed the $K-1^{\text{th}}$ node, it is the highest rank). The final decision is given by the sum of the yeses, plus one.

$$q^{[i]} = 1 + \sum_{j=1}^{K-1} 1\left(\hat{P}\left(y^{[i]} > r_j\right) > 0.5\right) \tag{6.8}$$

6.5 Results

On the humor data, the models which include demographic information concatenated as text achieved a higher average validation kappa when the results of the five folds of the final epoch were averaged. Incorporating the data as text was a much more successful approach than passing the information to MLPs to be embedded as categorical data. However, the success of the demographics-as-text models is tempered by the results on the test set. The highest performing models seemed to overfit to the annotators in the training data, and struggled to generalise to the new data in the test set. The ordinal models generalised marginally better than the classification models, but the model which generalised best to the test set was one without any demographic information.

Model	Concatenation	Demographics	Val κ	Test κ
Classification	Text	Full	**0.84**	0.50
Ordinal	Text	Full	0.83	0.52
Ordinal	Text	Sparse	0.74	0.51
Classification	Text	Sparse	0.72	0.48
Ordinal	None	None	0.58	**0.55**
Ordinal	MLP	Full	0.56	0.49
Classification	None	None	0.54	0.5
Classification	MLP	Full	0.52	0.42
Ordinal	MLP	Sparse	0.49	0.48
Classification	MLP	Sparse	0.48	0.46
Dummy	-	-	0.022	0.18

Table 6.3: Humor Rating Prediction Results

Model	Concatenation	Demographics	Val κ	Test κ
Classification	Text	Full	0.86	0.5115
Classification	Text	Full	0.84	0.4986
Ordinal	Text	Sparse	0.74	0.4951
Classification	Text	Sparse	0.71	0.4769
Ordinal	None	None	0.56	**0.5336**
Ordinal	MLP	Full	0.52	0.49
Ordinal	MLP	Sparse	0.51	0.49
Classification	None	None	0.49	0.4692
Classification	MLP	Full	0.48	0.45
Classification	MLP	Sparse	0.45	0.46
Dummy			0.0037	0.0018

Table 6.4: Offense Rating Prediction Results

For the offense prediction task, we see a somewhat similar pattern - demographic information presented as text massively improves performance on the validation set during training, but perhaps as a result, the models overfit to the training annotators and fail to generalise well to the test set. The best performing model on the text set was once again a text-only ordinal regression model. Again, we saw that embedding the demographic characteristics as categorical feature did not improve performance, and indeed, slightly disimproved it.

6.6 Discussion

For **RQ4.1**, we found that in humor rating prediction, the method of incorporating demographic information has a big impact on the results. Embedding the information by passing each categorical and numerical feature to an individual feed-forward network was not a successful approach for this dataset, and disimproved performance when compared to text-only models. By contrast, concatenating a textual description of the annotator to the original text yielded much improved performance on the validation set, but for both humor and offense, these results failed to generalise to unseen data. Ultimately the best performance on the test set came from the text-only model.

Although we implemented a cross-validation approach in order to reduce the likelihood that the models would overfit, this measure has not been sufficient to prevent that. The setup we used also implemented a 20% of dropout in the attention layers, which also should make the models more robust to overfitting. Future implementations may try increasing this level of dropout.

For **RQ4.2**, we found that there were only minimal differences between the results of the classification and ordinal regression models, particularly in terms of test set performance. However, metrics aside, the ordinal regression model is arguably a more ecologically valid model for this type of data, and as different approaches to ordinal regression develop, for example differential decision trees (Zhu et al., 2021), or chain maximizing ordinal metric learning (Suárez, García, & Herrera, 2021), this will be a key alternative to classification and regression approaches.

6.6.1 Error Analysis

Both the classification and the ordinal regression performed comparably well at predicting the non-humorous labels, the ordinal model performs better at predicting labels 1, 2, 3 and 4, but completely fails to predict the funniest label.

Label	Precision		Recall		F1 Score	
	Ord	Cls	Ord	Cls	Ord	Cls
Not humorous	0.90	0.89	0.89	0.95	0.90	0.92
Don't get it	0.27	0.13	0.06	0.06	0.10	0.08
1	0.38	0.17	0.28	0.03	0.32	0.06
2	0.20	0.16	0.45	0.12	0.28	0.14
3	0.22	0.17	0.39	0.13	0.28	0.15
4	0.00	0.36	0.00	0.46	0.00	0.40
5	0.00	0.00	0.00	0.00	0.00	0.00

Table 6.5: Precision, Recall, and F1 for Humor labels from Ordinal and Classification Models

In terms of the offense models, once again, both classification and ordinal regression models predicted negative examples equally well. Similar to the humor models, the ordinal regression model predicted labels 1, 2, and 3 better than the classification models, but failed to perform well at predicting the highest ratings.

Label	Precision		Recall		F1 Score	
	Ord	Cls	Ord	Cls	Ord	Cls
Not Offensive	0.91	0.89	0.95	0.95	0.93	0.92
1	0.14	0.13	0.13	0.06	0.13	0.08
2	0.13	0.17	0.11	0.03	0.12	0.06
3	0.17	0.16	0.19	0.12	0.18	0.14
4	0.23	0.17	0.22	0.13	0.22	0.15
5	0.56	0.36	0.23	0.46	0.33	0.40

Table 6.6: Precision, Recall, and F1 Scores for Offense labels from Ordinal and Classification Models (Offense Ratings)

6.6.2 Annotators who were hard to classify

We examined if there were any groups of annotators whose demographic variables featured more prominently in the errors made by both models. We used a chi-square test of homogeneity to compare the distribution of gender and generation in the test set annotators, to the distributions genders and generations of the annotators whose ratings were misclassified. For humor, there were no significant differences between the distribution of ratings from different genders (χ^2=129.5, p=0.228) or generations (χ^2=198.1, p=0.41) in the errors, relative to the distribution of sex in the test set. We found the same results for general offense, the distribution of sex in the errors was not significantly different to that in the errors (χ^2=98.12, p=0.78), this was also true of generation (χ^2=100.2, p=0.38).

6.6.3 Limitations and Future Work

An exciting extension of Chapter 5's finding that there are systematic differences between demographic groups in terms of humor and offense ratings, is that incorporating information about the demographics of the annotator can improve how the model learns. However, it is a major limitation, that in our implementation, the models overfit to the annotators seen during training, and fail to generalise to the test data. Future work could look at different validation approaches, implementing a higher probability of dropout, or at extending the dataset to incorporate a larger number of annotators. Similarly, using a weighted loss function, to account for the skewness of the offense data could improve performance. Finally, it would be interesting to apply an explainable AI system to the models to capture what information it is attending to when making predictions, and if the demographic information features, or if predictions are made mainly based on the text.

6.7 Conclusion

We present a system to incorporate demographic information into humor and offense prediction systems. We find that, for humor and offense detection, incorporating demographic information as a textual description of the annotator improve performance during training, but may not generalise well to unseen data. Our most robust models between validation and training were those which used text only to make predictions. We discussion overfitting and alternative validation approaches to remedy this. Both ordinal and classification models performed comparably well on both tasks, although we strongly advocate for using ordinal models for this task. For both types of model, there was no demographic group that was more difficult to make predictions about than another. The ordinal regression models made fewer errors overal, however, for offense, it made a slightly higher proportion of errors on the most offensive examples.

Chapter 7

Conclusion

This thesis focuses of demographically-aware computational humor and had two overarching aims: to incorporate findings from the broader humor literature into a computational to humor, and to validate some of those findings using a larger data sample.

The literature review identifies several gaps in the research. There has been much work from Psychology and Sociology on demographic difference in humor appreciation, but with the exception of a large but flawed Norwegian study, these results tend to come from a relatively small number of participants. We proposed to validate some of the findings of demographic differences in a descriptive, big data setting. We then aimed to use demographic information as an auxiliary feature to improve NLP performance on humor and offense detection.

Similarly, the overlap between humor and offense has long been known in the broader field - specifically the use of humor to obscure hate speech, but in NLP, humor and hate speech have always been modelled separately.

For **RQ1**, on improving datasets, we presented HaHackathon, which was the first humor detection shared task to simultaneously model offense. We offer guidelines on how to annotate offense - specifically that it can be subdivided into what is generally offensive to the community, and what is personally offensive to the annotator. Our annotation procedure was also the first to allow annotators to label a text as funny, but to admit that they did not 'get the joke'. We took more care than previous approaches to select texts and annotators from the same region, to avoid cultural confounds,

and we also selected annotators based on age, in order to represent a broader range of age groups than previously represented. The result is a dataset of 200k ratings of humor and offense, from 1800+ annotators, along with their demographic characteristics.

We attracted cutting-edge submissions from 60+ research teams, and for **RQ2**, we found that PLMs, particularly RoBERTa give state of the art performance for both tasks. Task participants also convincingly demonstrated that multi-task learning was more effective than single-task learning on this data, and that task-adaptive pre-training on humor and offense data improve performance. Domain adaptations to the domain of Twitter were more successful than those which adapted to the domain of humor, which again speaks to the ambiguous nature of humor as a concept. Back translation was a data augmentation approach that was not suitable for humor, and we suggest that the linguistic mechanism of humor - be it syntactic, semantic or phonetic can be undermined by translation approaches.

For **RQ3**, we found that there were systematic differences in how groups approached humor. Female annotators linked humor and offense more strongly than males, who only linked the two concepts when they were personally offended. The negative correlation between age and offense was weakest in the younger group and increased linearly with age. Replicating previous findings, women tended to give lower humor ratings and higher offense ratings overall, and used the 'I don't get it' rating more often. Also replicating previous studies, men showed more tolerance for aggressive humor.

In addition to the convincing results regarding age and gender difference in humor and offense ratings, for **RQ4**, we found that incorporating demographic information improved performance for humor and offense detection during training. However, the models which incorporated demographic information struggled to generalise to unseen data. We present a strong argument for modelling ratings as ordinal data,

and our results show that ordinal systems perform as well or better than classification systems for predicting both humor and offense. We discuss appropriate metrics for this task, and speculate about different validation procedures which could improve the robustness of demographic models.

7.0.1 Implications

Given the correlations between humor and offense demonstrated by annotators, we maintain that the separate fields of humor and hate speech detection should be more closely linked. This argument is strengthened by the superior performance of multi-task systems which model both types of ratings. Our selection of data and annotators from the same domain reduces cultural confounds, and our annotation guidelines for offense, as well as the 'I don't get it' option for humor will enrich future datasets.

Our replication of humor results from Psychology and Linguistics consolidates our understanding of demographic differences in humor appreciation. Although these demographic differences improved the models ability to predict humor and offense labels during training, this caused some overfitting, and failure to generalise. Finally, we argue that modelling the ordinality of ratings is a more ecologically valid and better performing approach to this task.

7.0.2 Limitations and Future Work

This work in this thesis, as in so much of NLP, focused exclusively on English language, and US annotators, and our findings about humor need to be considered limited to that culture. Although the literature review does present results from Norway, the Netherlands, Italy, India and Spanish-speaking countries, there is still a dearth of work on humor in other languages. This is partly due to the difficulty in sourcing a sufficient number of annotators who speak these languages on crowd-sourcing platforms. When co-organising the third edition of the HAHA task, in order to select

annotators and data from the same country, we let the availability of annotators guide our choice. The largest number of Spanish-speaking crowdworkers were from Mexico, but even so, we were forced to give bigger batches of data to fewer annotators, due to the scarcity of crowdworkers from this country.

A key aim of creating our dataset was to look at a wide range of humorous texts, while achieving relatively high inter-annotator agreement on how humorous the text is. Thanks to our stringent annotation procedure, there was good agreement on what was a joke versus what was not, but the agreement on the funniness and offensiveness ratings was still very low, even when binning the annotators by age or gender. Future research may look at alternative variables with which to group the ratings, and must explore other approaches to increasing agreement.

A further limitation is that, in many areas, we reduced the scope of our research to gender and age differences. These were well-supported in the literature, but future work could consider other demographic variables. Future work could also explore how women respond to jokes about their gender, and how black women respond to jokes about their ethnicity and gender, as opposed to other groups, such as white women and white men.

Another key point is that the silent partner in this work has been the joke-teller. Although there is evidence that responses to humor - in particular offensive humor which targets groups from different genders and ethnicities - can vary depending on whether the joke-teller is a member of this group or not, experimenting with this was out of scope for this project. Future work could present fewer texts to annotators, who see these texts as coming from within or outside the group targeted by the joke.

Finally, we end with a reminder that humor is a moving target. Not only does topical humor not age well, but our own responses to humor change throughout the lifetime. Creating humor datasets is both costly and time-consuming, and they really only capture a snapshot of humor responses at a particular moment in history, during a particular moment in the annotator's life. Rather than being daunted at this what a nebulous task humor detection is, we are mindful to take a nuanced approach, and at every step, to be inter-disciplinary.

Appendix A

Pilot Study

A.0.1 Annotator Instructions

Thank you for signing up to help with our research on humor!

Your opinions will help us to learn about how different people react to humor in different ways. We're interested in hearing your opinions today, so feel free to express your opinions without censoring yourself.

You'll see 25 texts. For each text, you'll be asked 3 yes/no questions, and then you'll give a rating 1-5.

Genre: First, we want to know if the text is supposed to be a joke or not. So the first question is 'is the intention of this text to be humorous?'. Even if you don't find the joke funny, you can normally tell if it was intended to be funny. We call this the 'genre' of the text. If you say 'yes, this text was probably intended to be humorous', after that we'll ask you if you found it funny, here you can give your personal response to the text.

Let's look at an example: Anna looks at a text, decides it's supposed to be a joke and actually she finds it really funny, so for the question 'is the intention of this text to be funny?' she answers yes. And for the question 'how funny do you find this text?', she answers 5. John looks at the same text, think 'oof that's a bad joke.' He can tell it's

supposed to be a joke, so the answer yes to question 1, but he doesn't find it funny at all, so for the question 'how funny do you find this joke?' he gives a 1. This shows that although John didn't like the joke, he still can recognise that it was intended to be humorous.

Sometimes you can tell it's supposed to be a joke, but you just don't get it. Don't worry, you can select 'yes, it's intended to be a joke' and 'I don't get it'. Sometimes this happens because it talks about a person you don't recognise, or uses language that you are not familiar with. So, what's the difference between a rating of 1, and a rating of 'I don't get it'? Rating the text as'1' means you just don't think it's funny, where 'I don't get it' means that you think it should be funny, but you're not sure why. We're just interested in your honest opinion here.

Generally Offensive: After this question, you'll be asked 'is this text generally offensive?' By this we mean, does it target a person or group of people just because they belong to that group? You can also think of it as 'would a lot of people be offended by this joke?'. For example, John sees a joke that he thinks a lot of people would be hurt by, so he selects 'yes' but he thinks that also most people would take it as a joke, and this little bit of humor wouldn't hurt, so he selects '1' for the second question. Anna also thinks the joke makes fun of a group of people and thinks it's kind of mean-spirited, so she selects '4'. For some jokes, you might think 'I find this hilarious, but some people might be offended by it.' Don't worry! Some jokes which are offensive might also make you laugh. Again, there are no right or wrong answers here, we're just interested in your opinion. If you think it's just a joke, and not really offensive, select no. Maybe it talks about a group of people, but not in a negative way. Maybe it's just innocent humor. What if you're not sure if it's offensive? Best to just say no.

Personally Offensive The last question will be ‘do you find this text personally offensive?’ If you belong to the group targeted by the joke, but you don’t really care, select no. If you belong to the group targeted and you do find it hurtful, select yes, and then tell us how hurtful. Or maybe you don’t belong to the group targeted, but you feel hurt on behalf of someone else, you can select ‘yes’ and tell us how much. Again, there are no wrong answers.

A.1 Texts Used

1. It's been confirmed by People Magazine that Bruce Jenner is transitioning into a woman I say good on him, at least he's got the balls to do it.
2. How does a Jew make his tea? Hebrews it!
3. From online museum resources on Asian art to East Asia in geographic perspective, Asia for Educators provides a rich content for educators
4. Ignorance is bliss but i'd rather be stressed, restless, and shaking in anger than blind and apathetic.
5. Muslim minority doctors first to die on front line of UK pandemic- Families & patients remember 4 doctors who contracted coronavirus
6. 5 Years ago today I asked a beautiful girl out on a date. Today at 3pm I asked that same girl to marry me. Both times she said no
7. Sometimes I run alongside trains, tearfully waving, just so people will think I have a girlfriend.
8. What is the Terminators Muslim name? Al Bi Baq
9. What do you call an overweight homosexual? Jigglypuff.
10. Woman: "Can you please call me a taxi?" Man: "You're a taxi"
11. What did the two tampons have in common? They were both stuck up bitches
12. Every religion has violent people... ... The Christians have The Westboro Baptist Church, the Muslims have the jihadists, and the Jews have the IRS
13. The church apparently thinks that gay and lesbian people laying with one another is unnatural. As opposed to walking on water.
14. Apparently I have a black person in my family tree... If you look in my backyard he's still hanging there!
15. Mattel released a Muslim Barbie... It's a blow-up doll.
16. I was walking through town yesterday, when I came across a large black man with a mohawk and jewellery. He said, "I piy the fool." I said, "Hey, you missed a T."

17. Making a documentary about homosexuality in America, if anyone that identities as homosexual would like to be in it please HMU.
18. Why should you never play Uno with a Mexican? Because they'll steal all the green cards.
19. What do you call a gay cowboy? A jolly rancher.
20. Due to COVID-19, Jewish Family Service has cancelled their upcoming Heart & Soul Gala.
21. On his first day, my gay friend lost his job at the sperm bank. He was caught drinking on the job.
22. Stumbled across Samanos Taqueria Mexican Food truck on my route. There was no way I was passing up this blessing. Such good food oh my word.
23. I was walking down the street when I saw black guy with a laptop. Is that mine? I thought. It looks exactly like mine... but that can't be... Mine should be at home cleaning my shoes.
24. Ola has done so much for Black women, girls, and Black women's literature.
25. Q: How did a blind woman pierce her ear? A: Answering the stapler.
26. Back in my day, we didn't have computers or the internet. Guys would have to walk uphill for days to tell me I'm gay.
27. I was getting a handjob from this blind hooker... She told me: WOW! This is the biggest dick I have ever felt! I said "No way! You're pulling my leg!"
28. Have more faith girlfriend. You are the most positive people with such great influence to others don't be a Debbie downer now.
29. PSA: If you've got an Islamic dog... Muslim
30. Do you know how to make a Venetian Blind? Poke them in the eye
31. Ashley Lawrence, a college senior from Versailles put her stir-craziness to good use designing face masks for the deaf and hard of hearing.
32. Wife: Honey, i think you don't love our kids equally. It seems you have one that gets less love from you. Husband: Really? Which kid do you mean? Karl, Tina or the fat one

33. I don't see why so many Americans are anti-gay rights. For a country with 4 dads, you would think they'd be more accepting.
34. What do you call an Asian woman on the moon? An Astronaut
35. I was asked how I view lesbian relationships. Apparently in HD wasn't the right answer.
36. They've thrown the same bile at you that they've always thrown at gay and bisexual people. But despite this obsessive, vicious campaign, your allies will never ever stop showing love, support and solidarity.
37. Iron Man is a super hero. Iron Woman is a command.
38. Italian nurse allegedly killed medic girlfriend, falsely claimed she gave him coronavirus: reports
39. How did a sexy but unqualified Japanese woman get voted into public office? Erections
40. Why didn't Peter Jackson cast any black actors in The Lord Of The Rings? He was afraid they wouldn't be taken on their merits, that they'd just be seen as Tolkien blacks.
41. Did you hear about the fish that went deaf? It had to buy a herring aid
42. Fat Cat Beer Hatch. Open Daily. 3pm-7pm. Mega ales and ciders. Only $3 a pint.
43. The gay blood ban ("one year with no gay sex" iteration) that emerged from the AIDS crisis is now keeping gay men from helping to address the #COVID_19 crisis.
44. Kindness is the language which the deaf can hear and the blind can see. - Mark Twain
45. Women say they like a man in uniform but I've been wearing this naughty nurse outfit all day and not a single woman has approached me yet
46. Diversity in the creative industry is dear to my heart, & for the past 3 years we've been working effortlessly to see Black talent thrive.
47. What is black and stuck to the ceiling? A crappy electrician

48. It’s funny how liberals think Donald Trump oppresses women and hates gays Like he’s a Muslim or something.
49. As an Australian-born Muslim, a city slicker..I’d never gotten to see the beauty of what the Australian outback offers, and this trip really opened my eyes
50. Vintage Oversized Sunglasses, Italian Style from 70s, Sunglasses for Woman, Lady Sunglasses by Elgratex Why wait?
51. What do you call a fruit that is a vegetable? A crippled homo.
52. I went to the doctors with hearing problems... He said "Can you describe the symptoms?" So I replied "Homers fat, and Marge has blue hair"
53. Why did the Muslim cross the road? To get to the bigger crowd
54. What did the bottle of milk say to the Mexican? Soy milk.
55. Whats better than winning gold at the paralympics? Having legs.....
56. #AlanTuring’s trial for gross indecency commences for having had a then illegal homosexual relationship with a man. Today we remember the vital contributions that he made to the work of Bletchley Park
57. 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"
58. Such a pretty girl! I hope she enjoys her cake and has a great day!
59. Oh my god, you’ve gotten so fat! Want me to make you something to eat? - my mother
60. Israelis raise half a million dollars in 24 hours for orphaned 4-year-old twins of #coronavirus victim

A.2 Twitter Accounts

These Twitter accounts were used to collect humorous and non-humorous data.

Username	Count	Username	Count
humurous1liners	924	BlkMentalHealth	37
joeljeffrey	692	mikewickett	35
UberFacts	632	BlackLoveAdvice	35
Dadsaysjokes	541	JNFUSA	35
GreysAnatomyMsg	402	JokesMemesFacts	34
ConanOBrien	340	MissyDuckWife	32
boonaamohammed	337	blackbodyhealth	32
Demented_Jokes	325	RobBenedict	31
thenatewolf	284	Boyfriend_Tips	30
DailyHealthFact	284	TheJimMichaels	29
Kasandd	219	realGpad	29
songs_lyrics	203	EverBestFilms	27
Shen_the_Bird	187	NicoleB_MD	23
BadJokeCat	130	iGirlfriendTip	23
OURSELVES_BLACK	129	Grindr	23
SupereeeGO	124	MNateShyamalan	23
Mr_Truth_Hurts	112	kecia_ali	20
GayAdvicer	112	RobbyActually	19
Wizdomstweets	103	hardwick	19
TrippAdvice	102	RabbiHarvey	19
JensenAckles	97	taylorswift13	18
BunAndLeggings	93	PGATOURWives	17
MovieQuotesPage	90	tomhanks	15
annehelen	87	BlackGirlsSmile	15
YaGayAunties	83	curtisisbooger	11
mindykaling	74	evanmarckatz	11
RyanSeacrest	70	bosshogswife	11
murrman5	59	PenguinBooks	10
TheOkraProject	59	GuyStuffAdvice	10
benyahr	57	gaystarnews	10
thatonequeen	55	DrakeGatsby	9
ZaraRahim	52	offensivefcker	9
Oprah	52	outmagazine	9
michaelstrahan	43	therapy4bgirls	8
youknowwhenshe	42	ProBonoASL	4
Blackkidsswim	40	TheAdvocateMag	3
andreavsmoak	40		

Appendix B

Published Papers

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

J. A. Meaney
School of Informatics
University of Edinburgh
Edinburgh, 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

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 inter-rater 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.

SemEval-2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense

J.A. Meaney[1], Steven R. Wilson[1], Luis Chiruzzo[2], Adam Lopez[1, 3], Walid Magdy[1,4]
[1] School of Informatics, The University of Edinburgh, Edinburgh, UK
[2] Universidad de la República, Uruguay
[3] Rasa
[4] The Alan Turing Institute, London, UK

Abstract

SemEval 2021 Task 7, HaHackathon, was the first shared task to combine the previously separate domains of humor detection and offense detection. We collected 10,000 texts from Twitter and the Kaggle Short Jokes dataset, and had each annotated for humor and offense by 20 annotators aged 18-70. Our subtasks were binary humor detection, prediction of humor and offense ratings, and a novel controversy task: to predict if the variance in the humor ratings was higher than a specific threshold. The subtasks attracted 36-58 submissions, with most of the participants choosing to use pre-trained language models. Many of the highest performing teams also implemented additional optimization techniques, including task-adaptive training and adversarial training. The results suggest that the participating systems are well suited to humor detection, but that humor controversy is a more challenging task. We discuss which models excel in this task, which auxiliary techniques boost their performance, and analyze the errors which were not captured by the best systems.

1 Introduction

Humor is a key component of many forms of communication, and so it is commanding an increasing amount of attention in the natural language processing (NLP) community (Attardo, 2008; Taylor and Attardo, 2017; Amin and Burghardt, 2020). However, like much of figurative language processing, humor detection requires a different perspective on several traditional NLP tasks. For example, the problem of reducing lexical or syntactic ambiguity differs when ambiguity is key to some humor mechanisms. Tackling these challenges has the potential to improve many downstream applications, such as content moderation and human-computer interaction (Rayz, 2017).

However, humor is a subjective phenomenon, which evokes varying degrees of funniness in its audience, while also provoking other reactions such as offense, in certain listeners. The perception of humor is known to vary along the lines of age, gender, personality and other factors (Ruch, 2010; Kuipers, 2015; Hofmann et al., 2020). That humor can also evoke offense may be partly due to differences in acceptability judgements across demographic groups, and may also be in part due the use of humor to mask hateful or offensive content (Sue and Golash-Boza, 2013). Lockyer and Pickering (2005) expand on this by highlighting that it is common for societies to explore the link between humor and offense, free speech and respect.

HaHackathon is the first shared task to combine humor and offense detection, based on ratings from a wide variety of demographic groups. Task participants were asked to detect if a text was humorous and to predict its average ratings for both humor and offense. We also introduce a novel humor controversy detection task, which represents the extent to which annotators agreed/disagreed with each other over the humor rating of a joke. A humorous text was labelled as controversial if the variance in the humor ratings was higher than the median humor rating variance in the training set.

2 Related Work

Computational humor detection is a relatively established area of research. Taylor and Mazlack (2004) were one of the first to explore recognising wordplay with ngrams. Mihalcea and Strapparava (2005; 2006) experimented with 16,000 one-liners and 16,000 non-humorous texts, using a feature-driven approach. More recently, Zhang and Liu (2014) turned to online domains, by detecting humor on Twitter with a view to improving downstream tasks such as sentiment analysis and opinion

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 105–119
Bangkok, Thailand (online), August 5–6, 2021.

mining.

Workshops on humor detection have become more prominent with each shared task, and have attracted many new researchers to the field. SemEval 2017 (Potash et al., 2017) featured Hashtag Wars, a humor task with a unique data annotation procedure. This task featured tweets that had been submitted in response to a number of comedic hashtags released by a Comedy Central program. The top-10 response tweets were selected by the show's producers and the winning tweet was selected by the show's audience. Based on these labels, (top-10, winning tweet, and other) the sub-tasks required competitors to predict the labels, and to predict which text was funnier, given a pair tweets. The winning systems were split between feature-driven support vector machines (SVMs) and recurrent neural networks (RNNs).

The first Spanish-language humor detection challenges were the HAHA tasks in 2018 (Castro et al., 2018) and 2019 (Chiruzzo et al., 2019). These collected data from more than fifty different humorous Twitter accounts, representing a wide variety of humor genres. The sub-tasks asked competitors to predict if a text was humorous, and to predict the average funniness score given to the humorous texts. In the first year, the top teams used evolutionary algorithms to optimize linear models like Naive Bayes, as well as bi-directional RNNs. In the second year, the top teams started to use pre-trained language models (PLMs) like BERT (Devlin et al., 2018) and ULMFit (Howard and Ruder, 2018).

Most recently, Hossain et al. (2020) generated data for their task by collecting news headlines, and asking annotators to make a micro-edit to the headline to render it funny. These edited headlines were rated for funniness by other annotators. The sub-tasks were to rank the funnier of two edits, and to predict the average funniness score given by the annotators. The winning teams used ensembles of various PLMs, and RNNs.

3 Data

3.1 Data Collection

In order to examine naturally-occurring humorous and offensive content in English, we sourced 80% of our data from Twitter. The remaining 20% of texts, we selected from the Kaggle Short Jokes dataset[1] for the following reasons:

[1] https://www.kaggle.com/abhinavmoudgil95/short-jokes

Target	Keywords
Sexism	She, woman, mother, girl, b*tch, he, man, blond, p*ssy, hooker, slut, wh*re
Body	Fat, thin, skinny, tall, short, bald, amputee, redneck
Origin	Mexico, Mexican, Ireland, Irish, Indian, Pakistan, China, Chinese, Polish, German, France, Welsh, Vietnam, Asian, American, Russia, Arab, Jamaican, homeless
Sexual Orientation	Gay, lesbian, d*ke, f*ggot, homo, aids, LGBT, trans, tr*nny
Racism	Black, Africa, African, wop, n***** white people,
Ideology	Feminism, leftie/lefty
Religion	Muslim, Islam, Jew, Jewish, Catholic, Protestant, Hindu, Buddhist, ISIS, Jesus, Mohammed
Health	Wheelchair, blind, deaf, r*tard, Steven Hawking, Stevie Wonder, Helen Keller, dyslexic

Table 1: Targets and Sample Keywords

- **Humor Quota:** To ensure that a sample of texts in the dataset were intended to be humorous. Our annotation procedure asks raters if the intention of the text is to be humorous (as evidenced by the the setup/punchline structure, or absurd content). As the texts were sourced from the /r/jokes and /r/cleanjokes subreddits, we were confident that the intention of the text was to be humorous.

- **Traditional Humor Quota:** We wanted to represent jokes which have a traditional setup and punchline structure. Twitter humor is known to use a number of unique features (Zhang and Liu, 2014), which may not be equally recognisable to all annotators and so we wanted to have a selection of conventionally recognisable texts in order to gauge what the audience response was, and to use as a quality check for annotators (see below).

- **Offense Quota:** To ensure that a proportion of texts were likely to be considered offensive by the annotators, half of the texts selected according to the procedure below.

To select potentially offensive texts, we used some of the keywords associated with Silva et al.'s (2016) sub-categories of hate speech in social media, and queried the Kaggle dataset for these.

Text	Keyword = Target
A **fat** woman just served me at McDonalds and said "Sorry about the wait". I replied and said, "Don't worry, you'll lose it eventually".	Yes
Don't worry if a **fat** guy comes to kidnap you... I told Santa all I want for Christmas is you.	No

Table 2: Sample of potentially offensive and non-offensive texts

From these texts, we identified the target, or butt, of the joke and made the assumption that a text could be potentially offensive to our annotators if the hate speech keyword was the target of the joke. We selected 1,000 texts this way. We also assumed that the text would likely be considered not offensive if the keyword was mentioned, but was not the target and selected a further 1,000 texts like this. This was to reduce the probability that a humor/offense detection system would learn to classify texts simply based on the presence of a hate speech keyword.

3.1.1 Selection of Twitter texts

In order to avoid introducing annotation confounds such as a lack of cultural or linguistic knowledge (Meaney, 2020), we selected the texts and the annotators from the same region – the US. When sourcing the humorous Twitter data, we selected accounts according to whether they were based in the US and posted almost exclusively humorous content (e.g. @humurous1liners, @conanobrien). For the non-humorous Twitter accounts, we elected not to use news sources, e.g. CNN due to stylistic differences between news and humor (Mihalcea and Strapparava, 2006) making them easy to differentiate. The non-humorous accounts we selected centred on US celebrities (e.g. @thatonequeen, @Oprah), organisations that represent the targets of hate speech groups (e.g. @BlkMentalHealth, in order to increase the occurrences of the keywords in a non-humorous and non-offensive context), trivia accounts (e.g. @UberFacts, as the question and answer structure is similar to some types of setup and punchline) and tv/movie quotation accounts (e.g. @MovieQuotesPage, in order to resemble the dialogue-type jokes that are common on Twitter). Please see the appendix for a comprehensive list of accounts.

Using the Twitter API, we crawled up to 2,000 tweets from each account, and removed retweets and texts containing links. We also removed tweets that contained references to US Politics, the pandemic, or TV show characters as topical humor can be difficult to understand once the event it is tied to has passed (Highfield, 2015). From an initial 76,542 texts, we were left with 8,000 tweets. From these, we removed hashtags that labelled the texts as humorous, e.g. #joke, and using Ekphrasis (Baziotis et al., 2017) we split up any remaining hashtags into their constituent words so as to make them less easy to differentiate from the Kaggle texts.

3.2 Annotation

We recruited annotators from the Prolific[2] platform. Participants were recruited based on their self-reported native English-speaker status, US citizenship, and membership of one of the following age groups: 18-25, 26-40, 41-55, 56-70. Each text was annotated by 5 members of each age group, giving a total of 20 annotations per text. Batches comprised 100 texts, and annotators answered the following questions:

1. Is the intention of this text to be humorous?
2. Is this text generally offensive?
3. Is this text personally offensive?

In the case that a user answered 'yes' to any of these questions, they were asked to rate the humor or offense from 1-5 (see figure 1). For the humor rating, the user was also given the option to select 'I don't get it', meaning that they recognised by the structure or content that the text was intended to be humorous, but that they were unsure of why the text was funny. This is distinct from a rating of 1, which is a recognition of humor, with little appreciation for it.

The annotator instructions outlined that the first annotation question was intended to determine the *genre* of the text, and should be distinguished from *funniness*. Annotators were instructed to look at the structure of the joke, e.g. setup and punchline, or the content of the joke, e.g. absurdity, in order to determine if the intention was to be humorous.

[2] https://www.prolific.co/

In terms of offense, we posed two annotation questions in order to avoid ambiguity about which type of offense was meant. We instructed annotators to consider as generally offensive, a text which targets a person or group of people, simply for belonging to a certain group. Alternatively, they could select yes for generally offensive if they thought that a large number of people were likely to be offended by the joke. The last question asked annotators if they felt personally offended by the text, or if they felt offended on another person's behalf. We used only the generally offensive ratings in this task.

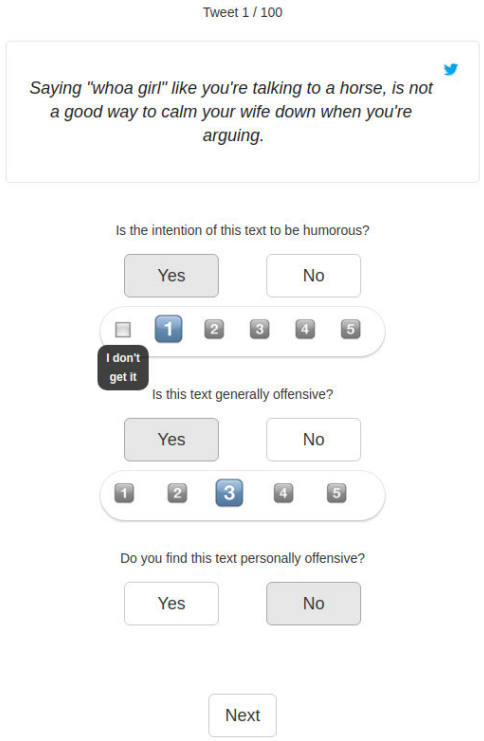


Figure 1: Screenshot from the tool used to annotate the texts.

3.3 Quality Control and Data Discarded

Each batch of 100 texts comprised approximately 20% of texts from Kaggle. As the majority of these have a setup and punchline structure, or other recognisable humor traits, we used these as a quality control. If an annotator did not label at least 60% of these as humor, it was clear that they they did not follow the instructions for the first question, and annotated based on perceived humor, as opposed to observation of humorous characteristics. We therefore discarded these submissions and replaced the annotators. Of 2,364 annotation sessions (e.g. batches of 100), 301 submissions were discarded and replaced, and the ratings of the remaining 2,062 annotation sessions make up the dataset. Of these, 1,569 annotators rated one batch of texts with an additional 492 doing a second batch.

3.4 Data Statistics

Post-annotation, we classed a text as humorous if the majority of its twenty votes labelled it as such. In a small number of cases where votes were tied, we assigned the label humorous. For the texts labelled humorous, we calculated the average humor score, which was the average of the numerical votes. "No" ratings did not count towards this value, and votes of "I don't know" were counted as 0, because this was deemed to be a recognizable humor structure, but one in which the humor was not successful.

Label	Affirmative	Negative	Average Rating
Humorous	6179	3821	2.24
Controversial	3052	3017	N/A
Offensive	5754	4246	1.02

Table 3: Data Statistics

The humor controversy label was based on whether the variance between the humor ratings was higher or lower than the median variance in the training set (median $s^2 = 1.79$). The offense rating was the average of all ratings given, including 'no' as 0. Table 3 summarises the labels in the dataset, and in the case of offense, affirmative indicates that the rating is higher than 0.

Ratings	Krippendorff's α
Class label	0.736
Humor rating	0.124
Offense rating	0.518

Table 4: Inter-annotator agreement (Krippendorff's α) for ratings used in subtask 1a, 1b and 2

The dataset was split 80:10:10 for training, development and test sets. The texts and annotations will continue to be available on the Codalab website, and the tweet ids, and usernames will be retained for non-commercial research use, in line with the Twitter Academic Developer Policy.

4 Task Description and Evaluation

We divided our tasks into four subtasks.

Task 1a: Humor Detection

This was a binary classification task to detect, given a text, if the majority label assigned to it was humorous or not. This was evaluated using F-score for the humorous class and overall accuracy

$$Accuracy = \frac{C}{N}$$

$$+‘213‘ - *‘F_1 = 2 * \frac{Precision \times Recall}{Precision + Recall}$$

Task 1b: Humor Rating Prediction

This was a humor rating regression task. Participants predicted the average rating given to texts from 0-5. Texts which had not been labelled as humorous by our annotators did not have a humor rating, and predictions for these texts were not counted towards the final score by our scoring system. The metric for this task was root mean squared error (RMSE).

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{N}\right)^2}$$

Task 1c: Humor Controversy Detection

This task was also a binary classification task to predict whether the humor ratings given to the text showed it to be controversial or not. This was based on the variance in the ratings being higher or lower than the median variance in the training set humor ratings. This was also evaluated using F-score and accuracy.

Task 2: Offense Detection

This was an offense rating regression task. Unlike the humorous task, this rating was not dependent on the text having been labelled as humorous. All annotator ratings were considered, and each text had a rating from 0-5. The metric was RMSE.

5 Benchmark Systems

We created simple, linear benchmarks using sklearn (Pedregosa et al., 2011) for the classification tasks which consists of a Naive Bayes classifier with bag of words features. For the regression tasks, we used a support vector regressor with term-frequency inverse document frequency features.

We also built a BERT-base classification/regression model which was run for one epoch, with a batch size of 16 and a learning rate of 5e-5, for all sub-tasks. As this system out-performed the linear benchmarks on all sub-tasks, we refer to this as the baseline in the rest of the paper.

6 Participant Systems

6.1 Overview

In total 63 teams submitted systems for the different tasks: 58 for task 1a, 50 for task 1b, 36 for task 1c and 48 for task 2. Tables 5, 6, 7 and 8 show the highest results for each task, with performance broken down by subsets of texts from the Kaggle jokes dataset and from Twitter. -*/

Team	Acc	F1	Kaggle F1	Twitter F1
PALI	0.9820	0.9854	0.9949	0.9811
stce	0.9750	0.9797	0.9871	0.9764
DeepBlueAI	0.9600	0.9676	0.9949	0.9551
SarcasmDet	0.9600	0.9675	0.9949	0.9548
mengyuan_jiayi	0.9590	0.9667	0.9871	0.9574
stevenhuahua	0.9580	0.9666	0.9949	0.9538
zain	0.9580	0.9663	0.9949	0.9534
EndTimes	0.9570	0.9655	0.9897	0.9545
MagicPai	0.9570	0.9653	0.9897	0.9542
Meizizi	0.9570	0.9653	0.9871	0.9554
mmmm	0.9560	0.9647	0.9923	0.9523
baseline (BERT)	*0.911*	*0.9283*	*0.9949*	*0.8978*
baseline (Linear)	*0.8570*	*0.8840*	*0.9792*	*0.8410*

Table 5: Results of the top performing systems for participants of task 1a (humor detection), showing F1 and accuracy for the whole test set, and F1 for Kaggle texts only and tweets only.

6.2 Highest Ranking Systems

The top-ranking teams were selected based on F-score, in the case of a tie in accuracy score. The top-10 made extensive use of pre-trained language models such as BERT, ERNIE 2.0 (Sun et al., 2020), ALBERT (Lan et al., 2019), DeBERTa (He et al., 2020) or RoBERTa (Liu et al., 2019). Ensembling these models by majority voting or averaging scores proved to be a popular and useful approach.

Team	All	Kaggle	Twitter
abcbpc	0.4959	0.4544	0.5141
mmmm	0.4977	0.4554	0.5162
Humor@IITK	0.5210	0.4702	0.5430
YoungSheldon	0.5257	0.4587	0.5541
IIITH	0.5263	0.4821	0.5456
fdabek	0.5271	0.4836	0.5462
Amherst685	0.5339	0.4584	0.5656
-*/ gerarld	0.5393	0.4857	0.5625
CS-UM6P	0.5401	0.4927	0.5608
SarcasmDet	0.5446	0.5001	0.5641
baseline (BERT)	*0.8000*	*0.4803*	*0.9117*
baseline (SVM)	*0.8609*	*0.7157*	*0.9205*

Table 6: Results of the top performing systems for participants of task 1b (humor rating), showing RMSE for whole test set, for Kaggle texts only and tweets only.

Team	Acc	F1	Kaggle F1	Twitter F1
PALI	0.4943	0.6302	0.6667	0.6118
mmmm	0.4699	0.6279	0.6621	0.6109
SarcasmDet	0.4699	0.6270	0.6552	0.6130
EndTimes	0.4602	0.6261	0.6598	0.6097
DeepBlueAI	0.4650	0.6257	0.6621	0.6078
CS-UM6P	0.4537	0.6242	0.6598	0.6070
CHaines	0.4537	0.6242	0.6598	0.6070
Ferryman	0.4537	0.6242	0.6598	0.6070
IIITH	0.4537	0.6242	0.6598	0.6070
abcbpc	0.4537	0.6242	0.6598	0.6070
fdabek	0.4537	0.6233	0.6598	0.6057
YoungSheldon	0.4780	0.6210	0.6545	0.6049
Humor@IITK	0.4520	0.6209	0.6574	0.6033
RoMa	0.4732	0.6197	0.6503	0.6042
baseline (BERT)	*0.4731*	*0.6232*	*0.6574*	*0.6060*
baseline (SVM)	*0.4374*	*0.4624*	*0.4804*	*0.4529*

Table 7: Results of the top performing systems for participants of task 1c (humor controversy), showing F1 and accuracy for the whole test set, and F1 for kaggle texts only and tweets only.

Similarly, many teams experimented with single and multi-task learning setups, and multi-task models tended to be more successful across sub-tasks. Further improvements were achieved with domain adaptation strategies and adversarial training.

6.2.1 DeepBlueAI (Song et al., 2021)

DeepBlueAI achieved high performance in subtasks 1a and 2. This team used stacked transformer models, which used the majority vote (in the case of classification) or the average prediction (for regression) from a RoBERTa and an ALBERT model. They optimized the performance of these PLMs with a number of techniques. First, they employed task-adaptive fine-tuning (Gururangan et al., 2020) by continuing pre-training on the text of the Hackathon data. They then augmented the dataset by using pseudo-labelling to generate labels for the test set, and added these to the training data. Then, after encoding the input, they used adversarial training (Miyato et al., 2016), e.g. the addition of perturbations to the embedding layer, to improve generalization. The predictions were produced after Multi Sample Dropout was applied. This approach achieved third place in task 1a and first place in task 2.

Team	All	Kaggle	Twitter
DeepBlueAI	0.4120	0.7607	0.2647
mmmm	0.4190	0.7757	0.2677
HumorHunter	0.4230	0.7742	0.2765
abcbpc	0.4275	0.7942	0.2712
fdabek	0.4406	0.7915	0.2979
stevenhuahua	0.4454	0.8019	0.2999
megatron	0.4456	0.8021	0.3001
MagicPai	0.4460	0.8113	0.2948
ES-JUST	0.4467	0.8065	0.2993
SarcasmDet	0.4469	0.8264	0.2861
baseline (BERT)	*0.5769*	*1.0141*	*0.4042*
baseline (SVM)	*0.6415*	*1.0908*	*0.4710*

Table 8: Results of the top performing systems for participants of task 2 (offense rating), showing RMSE for whole test set, for kaggle texts only and tweets only.

6.2.2 abcbpc (Pang et al., 2021)

This team deployed ERNIE 2.0 in a multi-task setup with task-specific gradients and loss for each sub-task. Using a cross-validation approach, they fine-tuned their model on each fold of data and took the average, or majority decision of their best-performing models as their predictions. Experiments demonstrated that their multi-task setup performed better than single-task learning with ERNIE 2.0, and they achieved the best score in task 1b.

6.2.3 Humor@IITK (Gupta et al., 2021)

This team also experimented with single-task and multi-task learning on pre-trained language models. They implemented two ensembling methods: in the single-task setup, they concatenated the embeddings produced by BERT, RoBERTa, ERNIE 2.0, DeBERTA and XLNET. In the multi-task setup, they used vote-based classification, or a weighted aggregate of outputs for the regression tasks. They also implemented an ensemble comprising a weighted average of best single-task and multi-task models, which achieved third place on task 1b. Interestingly, this team's experiments on data augmentation, e.g. generating slightly different variations of the input sentences, disimproved performance. The team hypothesize that the impact of both humor and offense often hinges on the choice of specific words, and replacing these words with synonyms may undermine the humorous or offensive effect.

6.2.4 SarcasmDet (Faraj and Abdullah, 2021)

For tasks 1a, 1b and 2, this team used either BERT or RoBERTa models with different hyperparameters, and used an ensemble of these models to make predictions with hard (e.g. majority or average) voting. Interestingly, for task 1c, in which they placed third, they used a rule, that if the humor rating predicted for a text was greater or equal to 3, they labelled the text as controversial.

6.2.5 HumorHunter (Xie et al., 2021)

This team used DeBERTa with an embedding table which took into account the relative position of each token in the sentence. In an error analysis, they noted that texts with a question and answer were more often misclassified as humorous, possibly because this mimics the structure of a setup and punchline.

6.2.6 Others

PALI and stce, the top-ranking teams in task 1a, both used an ensemble of RoBERTa large, and ERNIE 2.0, but declined to submit a paper outlining further details. Similarly, the team named mmmm, which placed 2nd in both task 1b and 1c, did not furnish details of their approach.

6.3 Trends

6.3.1 Domain Adaptation

Given that the majority of the data was sourced from Twitter, several teams implemented domain adaptation strategies at different stages of their pipeline. YoungSheldon (Sharma et al., 2021) used the Ekphrasis (Baziotis et al., 2017) toolkit, which is designed for Twitter-specific preprocessing. DLJUST (Al-Omari et al., 2021) also used it in their preprocessing pipeline, and found that this achieved better results, when used in combination with some further manual spelling correction.

Domain-specific models also showed some performance improvements. UPB (Smădu et al., 2021) used BERTweet (Nguyen et al., 2020), a transformer-based language model trained on tweets for their embedding layer, and DLJUST found that this model gave slightly better performance than RoBERTa on subtask 1a, but not on the regression tasks.

Amherst685 (Gugnani et al., 2021) used intermediate fine-tuning to adapt a series of pre-trained models to the style of language used in humorous and offensive texts. They used two large humor datasets, and two offense datasets, to adapt a variety of transformer models to the task, however, they did not see performance gains from this. Similarly to DeepBlueAI, RoMa (Labadie et al., 2021) and IIITH (Raha et al., 2021) used task-adaptive pre-training, and the latter team saw performance improvements of 1-5%.

6.3.2 Data Augmentation/Perturbation

Similarly to DeepBlueAI, MagicPai (Ma et al., 2021) experimented with pseudo-labelling in order to increase the amount of data available. MagicPai also tried adversarial training by adding perturbations to the embedding layer, and along with Grenzlinie (Liu and Zhou, 2021) and UPB, found this to improve their transfer learning models' performance. Amherst685 tried backtranslation in order to generate more sample texts, however they found that this was not successful.

6.3.3 Contrasting Models and Task Setup

The majority of teams who contrasted RNNs with PLMs found that the latter was more suited to this task. ES-JUST (Bashabsheh and Alasal, 2021) found that RoBERTa performed better than RNNs and BERT. This finding replicates the ablation study by Morishita et al. (2020) in the 2020 SemEval task, which also demonstrated that RoBERTa performed better than other PLMs. However Tsia (Guan, 2021) found that RoBERTa was better suited to the regression task, and combining BERT+CNN gave better performance on the classification task. This contrasts with YoungSheldon, who achieved their best results with BERT-Base. Across all cases, we did not observe a single dominant architecture, indicating that the choice of hyperparamters and task setup played a large role in the results achieved by each team. However, teams like CS-UM6P (Essefar et al., 2021), who contrasted single and multi-task learning setups, found that the latter improved performance.

6.4 Other notable approaches

DUTH (Karasakalidis et al., 2021) produced a rigorous examination of different preprocessing approaches applied to data given to linear and neural models. They achieved an impressive 12th place on task 1b, with a combination of Light Gradient Boosting Machine (LGBM), XGBoost and Bayesian Ridge. They also achieved 12th place in task 1c using a combination of features such as POS-tagging, numerical features, a bigram term frequency inverse document frequency (TF-IDF) vectorizer as input to an LGBM model.

The utility of TF-IDF features was also seen in the transfer learning approaches as team hub also found that adding TF-IDF features improved the performance of their ALBERT/BERT+CNN models.

IIITH found that including lexical features such as letter and punctuation counts, named entities marking, identifying personal pronouns, wh-words and question marks, as well as a lexicon of hurtful

words (Hurtlex, Bassignana et al., 2018) improved the performance of their task-adaptively pre-trained RoBERTa model for detecting humor and predicting the rating, but that only the Hurtlex features improved offense detection, and neither of these improved controversy prediction.

7 Analysis and Discussion

7.1 Correlations between Tasks

As Table 9 indicates, humor rating is moderately correlated with humor controversy across the dataset. There are no discernible trends in offense rating and humor controversy. Interestingly, there is a moderate negative correlation between humor and offense rating overall, but this is not significant for the Twitter data, and becomes a much stronger negative correlation when we look at just the Kaggle data. This may have be a factor in the finding that multi-task setups tended to achieve better results that single-task systems. It may also suggest that in naturally occurring data, such as the Twitter texts, the relationship between humor and offense may be more subtle, and therefore more difficult to detect.

Task 1	Task 2	Overall	Twitter	Kaggle
Humor	Humor	0.15	0.14	0.18
Rating	Controversy	$p = 0.0001$	$p = 0.003$	$p = 0.009$
Offense	Humor	0.07	0.11	-0.02
Rating	Controversy	p =0.06	$p = 0.028$	$p = 0.82$
Humor	Offense	-0.156	-0.03	-0.42
Rating	Rating	$p = 0.0001$	$p = 0.51$	$p = 0.0011$

Table 9: Correlations between tasks, Pearson's r and p-value

7.2 Differences between Kaggle Texts and Tweets

As seen in tables 5, 6 and 7, the systems' performance for subtasks 1a, 1b and 1c seems to be consistently better for Kaggle texts than for tweets. One possible reason why systems are better at predicting humor from Kaggle texts, is that the Kaggle test set contains almost all humorous texts, while only about half of the tweets are considered humorous.

On the other hand, performance for task 2 is consistently better (lower RMSE) for tweets than for Kaggle texts, and the differences are sometimes very large. We noticed the distributions of offense ratings between Kaggle texts and tweets are very different, with tweets being more often classified as not offensive: more than 60% of the tweets have 0.1 offense rating or less (in a scale from 0 to 5), while less than 10% of the Kaggle texts do. This difference in distribution might in part come from differences in sampling methods, because some Kaggle texts were specifically selected to have certain offensive categories, while the tweets were selected at random. In order to check if the difference in scores could come from the difference in offense rating distributions, we resampled a subset of tweets from the Kaggle set and another one from the Twitter set, trying to keep a uniform offense rating distribution, and calculated task 2 scores for those subsets. The difference between scores for these new subsets was much lower for all teams, and even some of the teams got better scores for the Kaggle subset, which might be an indication that the sharp differences in score were caused by the difference in distributions.

7.3 Error Analysis: Humans and Machines vs Irony

Several interesting issues arise when analyzing the top-ten systems' errors. Irony continues to be a challenging problem, both at the annotation side, and the classification side. Several texts which were sourced from humorous accounts, and which had just less than a majority of annotator votes for humorous were classed as not-humorous in our dataset. In the following two examples, all of the top-10 systems classed this as humorous, and arguably, they are intended to be humorous, even though the majority of annotators technically did not class them as such.

1. What do you call a homosexual man on a wheel chair?
   A human being
2. It's almost like I gotta keep myself busy with random things like fluffing pillows just so I don't over eat.

The first example is an ironic subversion of a homophobic joke, using incongruity to undermine the anticipated punchline. While it is possible that the setup and punchline structure is what misled the system, similar question and answer structures were correctly classified.

The second example is arguably sarcasm, and all of the top systems classified it as humor, even though the annotators did not. However, there were several other texts which were classed as humorous

by the annotators, and which demonstrate traits of irony or sarcasm, were difficult to classify for the top teams, and produced mixed results:

1. If alcohol influences short-term memory, what does alcohol do?
2. How much should I rest between sets at the gym? I've been doing anywhere between 60 to 90 days to give my muscles a good chance to recover.

In terms of tasks 1b and 2, we analyzed the texts which proved most difficult to predict the humor and offense ratings for the top-10 systems. We calculated the mean average error (MAE) between the top 10 systems' predictions and the ground truth. We then examined the 75th percentile of MAE.

	Twitter	Kaggle
Humor	70%	30%
Offense	55.2%	44.8%

Table 10: Percentage of texts with highest MAE from the different sources

Interestingly, there was a disproportionately high number of Kaggle texts among the offensive texts whose rating was difficult to predict (44.8% while the Kaggle text make up only 20% of the data). A quick examination of these texts revealed there was a large number of ironic texts which were predicted to be highly offensive, although the ground truth did not reflect this, for example:

> Why do black people eat fried chicken? Because it tastes good.

7.4 Humor Controversy

As we were interested in the rule-based approach that team SarcasmDet took for this task, we investigated the upper-bound of success for any threshold-based heuristic which determines whether a text was controversial given the humor score alone. Figure 2 shows the hypothetical F1-score and accuracy that could be achieved by such a system. Assuming a perfect score on humor rating prediction, if teams assigned a controversial label for any text with a humor rating of over 2, they could achieve first place in this task in terms of accuracy with a score of 0.580. A threshold of 1.45 given perfect knowledge of the humor labels would result

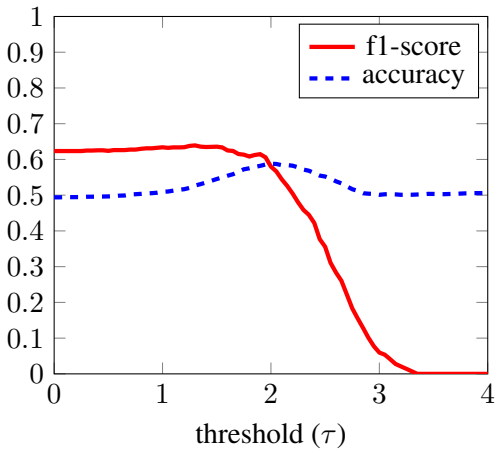


Figure 2: For varied values of a threshold, τ, accuracy and f1-score achieved by a hypothetical model predicting the label *controversial* for all texts in the test set with ground-truth humor score $> \tau$. Note that participants did not have access to these ground-truth scores for the test set, making these results an upper-bound for this type of threshold-based approach.

in a leaderboard-topping F1-score of 0.635. However, the teams that took part did not obtain the perfect humor rating scores required for this simple rule to work so effectively, yet were still able to achieve similar scores on the task. This suggests that their systems were learning something, but that ultimately the task is a difficult one.

Although we aimed to increase inter-annotator agreement in this task's annotation procedure, by matching the origin of the texts and annotators, the agreement on humor ratings was low, and indeed the task which aimed to capture this controversy proved difficult.

8 Conclusion

We provided 10,000 texts annotated for humor and offense by a broad range of annotators. Transformer models were a dominant approach to this task, with the exception of the humor controversy task, which proved to be difficult for most teams, and in which a simple, rule-based system achieved one of the top-3 scores. As multi-task learning setups proved more effective than single-task learning demonstrates, this that there is some correlation between humor and offense detection. It was also interesting to note which model adaptations were useful and which were not. Finally, an analysis of the errors in humor analysis reveals some types of humor which may be captured inaccurately, even by the most powerful models.

Acknowledgments

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. The authors also wish to thank William J. Toner who acted as a last-minute Idea Bouncer.

References

Hani Al-Omari, Isra'a AbedulNabi, and Rehab Duwairi. 2021. DLJUST at SemEval-2021 Task 7: Hahackathon: Linking Humor and Offense Across Different Age Groups. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Salvatore Attardo. 2008. A Primer for the Linguistics of Humor. *The Primer of Humor Research*, 8:101–156.

Emran Al Bashabsheh and Sanaa Abu Alasal. 2021. ES-JUST at SemEval-2021 Task 7: Detecting and Rating Humor and Offensive Text Using Deep Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual Lexicon of Words to Hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Santiago Castro, Luis Chiruzzo, and Aiala Rosá. 2018. Overview of the HAHA Task: Humor Analysis Based on Human Annotation at IberEval 2018. In *IberEval@ SEPLN*, pages 187–194.

Luis Chiruzzo, Santiago 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 *IberLEF@ SEPLN*, pages 132–144.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Kabil Essefar, Abdellah El Mekki, Abdelkader El Mahdaouy, NABIL El Mamoun, and Ismail Berrada. 2021. CS-UM6P at SemEval-2021 Task 7: Deep Multi-Task Learning Model for Detecting and Rating Humor and Offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Dalya Faraj and Malak Abdullah. 2021. SarcasmDet at SemEval-2021 Task 7: Detect Humor and Offensive based on Demographic Factors using RoBERTa Pretrained Model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Zhengyi Guan. 2021. Tsia at SemEval-2021 Task 7:Detecting and Rating Humor and Offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Akshay Gugnani, Brian Zylich, Gabriel Brookman, and Nicholas Samoray. 2021. Amherst685 at SemEval-2021 Task 7: Joint Modeling of Classification and Regression for Humor and Offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Aishwarya Gupta, Avik Pal, Bholeshwar Khurana, Lakshay Tyagi, and Ashutosh Modi. 2021. Humor@IITK at SemEval-2021 Task 7: Large Language Models for Quantifying Humor and Offensiveness. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654*.

Tim Highfield. 2015. Tweeted Joke Lifespans and Appropriated Punchlines: Practices around Topical Humor on Social Media. *International Journal of Communication*, 9:22.

Jennifer Hofmann, Tracey Platt, Chloe Lau, and Jorge Torres-Marín. 2020. Gender Differences in Humor-Related Traits, Humor Appreciation, Production, Comprehension,(Neural) Responses, Use, and Correlates: A Systematic Review. *Current Psychology*, pages 1–14.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 Task 7: Assessing humor in Edited News Headlines. *arXiv preprint arXiv:2008.00304*.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. *arXiv preprint arXiv:1801.06146*.

Alexandros Karasakalidis, Dimitrios Effrosynidis, and Avi Arampatzis. 2021. DUTH at SemEval-2021 Task 7: Is Conventional Machine Learning for Humorous and Offensive Tasks enough in 2021? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Giselinde Kuipers. 2015. The Humor Divide: Class, Age and Humor Styles. In *Good Humor, Bad Taste*, pages 71–101. De Gruyter Mouton.

Roberto Labadie, Mariano Rodriguez, Reynier Ortega, and Paolo Rosso. 2021. Dual Transformer for Sentiment Analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*.

Renyuan Liu and Xiaobing Zhou. 2021. Grenzlinie at SemEval-2021 Task 7: HaHackathon Detecting and Rating Humor and Offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

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*.

Sharon Lockyer and Michael Pickering. 2005. *Beyond a Joke: The Limits of Humour*. Springer.

Jian Ma, ShuYi Xie, Jiang Lianxin, Ryan Stark, Mo Yang, and Jianping Shen. 2021. MagicPai at SemEval-2021 Task 7: Method for Detecting and Rating Humor Based on Multi Task Adversarial Training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

J. A. Meaney. 2020. Crossing the line: Where do demographic variables fit into humor detection? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 176–181, Online. Association for Computational Linguistics.

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.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to Laugh (Automatically): Computational Models for Humor Recognition. *Computational Intelligence*, 22(2):126–142.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial Training Methods for Semi-supervised Text Classification. *arXiv preprint arXiv:1605.07725*.

Terufumi Morishita, Gaku Morio, Shota Horiguchi, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 task 8: Simple but effective modality ensemble for meme emotion recognition. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1126–1134, Barcelona (online). International Committee 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.

Chao Pang, Xiaoran Fan, Weiyue Su, Xuyi Chen, Shuohuan Wang, Jiaxiang Liu, Xuan Ouyang, Shikun Feng, and Yu Sun. 2021. abcbpc at SemEval-2021 Task 7: ERNIE-based Multi-task Model for Detecting and Rating Humor and Offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Tathagata Raha, Ishan Sanjeev Upadhyay, Radhika Mamidi, and Vasudeva Varma. 2021. IIITH at

SemEval-2021 Task 7: Leveraging Transformer-based Humourous and Offensive Text Detection Architectures using Lexical and Hurtlex Features along with Task Adaptive Pretraining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

J. T. Rayz. 2017. In Pursuit of Human-Friendly Interaction with a Computational System: Computational Humor. In *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000015–000020.

Willibald Ruch. 2010. *The Sense of Humor: Explorations of a Personality Characteristic*, volume 3. Walter de Gruyter.

Mayukh Sharma, Ilanthenral Kandasamy, and Vasantha W B. 2021. YoungSheldon at SemEval-2021 Task 7: Fine-tuning Is All You Need. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10.

Răzvan-Alexandru Smădu, Dumitru-Clementin Cercel, and Mihai Dascalu. 2021. UPB at SemEval-2021 Task 7: Adversarial Multi-Task Learning for Detecting and Rating Humour and Offence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Bingyan Song, Chunguang Pan, Shengguang Wang, and Zhipeng Luo. 2021. DeepBlueAI at SemEval-2021 Task 7: Detecting and Rating Humor and Offense with Stacking Diverse Language Model-Based Methods. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Christina A Sue and Tanya Golash-Boza. 2013. 'It Was Only a Joke': How Racial Humour Fuels Colour-Blind Ideologies in Mexico and Peru. *Ethnic and Racial Studies*, 36(10):1582–1598.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A Continual Pre-training Framework for Language Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Julia Taylor and S Attardo. 2017. Computational Treatments of Humor. *The Routledge Handbook of the Linguistics of Humor. New York: Routledge*, pages 456–471.

Julia M Taylor and Lawrence J Mazlack. 2004. Computationally Recognizing Wordplay in Jokes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

Yubo Xie, Junze Li, and Pearl Pu. 2021. HumorHunter at SemEval-2021 Task 7: Humor and Offense Recognition with Disentangled Attention. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

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.

A Appendices

Table 11 displays the sources for the Twitter data, e.g. 80% of the texts

Username	Count	Username	Count
humurous1liners	924	BlkMentalHealth	37
joeljeffrey	692	mikewickett	35
UberFacts	632	BlackLoveAdvice	35
Dadsaysjokes	541	JNFUSA	35
GreysAnatomyMsg	402	JokesMemesFacts	34
ConanOBrien	340	MissyDuckWife	32
boonaamohammed	337	blackbodyhealth	32
Demented_Jokes	325	RobBenedict	31
thenatewolf	284	Boyfriend_Tips	30
DailyHealthFact	284	TheJimMichaels	29
Kasandd	219	realGpad	29
songs_Iyrics	203	EverBestFilms	27
Shen_the_Bird	187	NicoleB_MD	23
BadJokeCat	130	iGirlfriendTip	23
OURSELVES_BLACK	129	Grindr	23
SupereeeGO	124	MNateShyamalan	23
Mr_Truth_Hurts	112	kecia_ali	20
GayAdvicer	112	RobbyActually	19
Wizdomstweets	103	hardwick	19
TrippAdvice	102	RabbiHarvey	19
JensenAckles	97	taylorswift13	18
BunAndLeggings	93	PGATOURWives	17
MovieQuotesPage	90	tomhanks	15
annehelen	87	BlackGirlsSmile	15
YaGayAunties	83	curtisisbooger	11
mindykaling	74	evanmarckatz	11
RyanSeacrest	70	bosshogswife	11
murrman5	59	PenguinBooks	10
TheOkraProject	59	GuyStuffAdvice	10
benyahr	57	gaystarnews	10
thatonequeen	55	DrakeGatsby	9
ZaraRahim	52	offensivefcker	9
Oprah	52	outmagazine	9
michaelstrahan	43	therapy4bgirls	8
youknowwhenshe	42	ProBonoASL	4
Blackkidsswim	40	TheAdvocateMag	3
andreavsmoak	40		

Table 11: Twitter sources of data and number of texts sourced from each account

Table 12 shows the results of the top system for each team and for each task.

Team	Task1a F1	Task1a Acc	Task1b RMSE	Task1c F1	Task1c Acc	Task2 RMSE
PALI	**0.9854**	**0.9820**	-	**0.6302**	0.4943	0.9710
stce	0.9797	0.9750	-	-	-	-
DeepBlueAI	0.9676	0.9600	0.5607	0.6257	0.4650	**0.4120**
SarcasmDet	0.9675	0.9600	0.5446	0.6270	0.4699	0.4560
mengyuan_jiayi	0.9667	0.9590	0.5621	0.5814	0.5106	-
stevenhuahua	0.9666	0.9580	0.5831	0.4991	0.5626	0.4454
zain	0.9663	0.9580	0.5748	-	-	-
EndTimes	0.9655	0.9570	0.6539	0.6261	0.4602	0.4691
MagicPai	0.9653	0.9570	0.5572	-	-	0.4460
Meizizi	0.9653	0.9570	0.6136	-	-	-
mmmm	0.9647	0.9560	0.4977	0.6279	0.4699	0.4190
fdabek	0.9647	0.9560	0.5271	0.6233	0.4537	0.4406
Isra	0.9640	0.9550	-	-	-	-
DLJUST	0.9633	0.9540	0.5555	0.4813	0.5480	0.4822
IIITH	0.9616	0.9530	0.5263	0.6242	0.4537	0.4772
megatron	0.9612	0.9520	0.6307	-	-	0.4456
CS-UM6P	0.9606	0.9510	0.6360	0.6242	0.4537	0.4759
Amherst685	0.9604	0.9510	0.5339	0.4842	0.5220	0.4530
MLXG	0.9590	0.9490	2.1883	0.0000	0.5463	0.9587
abcbpc	0.9587	0.9480	**0.4959**	0.6242	0.4537	0.4275
StoneOpen	0.9583	0.9480	0.5470	0.5427	0.5561	0.4489
Humor@IITK	0.9581	0.9480	0.5210	0.6209	0.4520	0.4607
Ferryman	0.9581	0.9480	0.5651	0.6242	0.4537	0.4813
RoMa	0.9576	0.9480	0.5905	0.6197	0.4732	0.4532
HumorHunter	0.9572	0.9480	0.5510	0.6111	0.4764	0.4230
RedwoodNLP	0.9571	0.9460	0.5580	0.4883	0.5024	0.7229
UPB	0.9566	0.9470	0.6200	0.0000	0.5463	0.5318
ES-JUST	0.9564	0.9460	0.5709	0.4888	0.5545	0.4467
DeathwingS	0.9563	0.9460	0.5561	-	-	-
zeus_yao	0.9557	0.9450	-	-	-	0.4621
apostaremczak	0.9544	0.9440	0.8497	0.0000	0.4341	0.5625
LeoJ	0.9543	0.9430	2.1883	0.0000	0.5463	0.9587
CHAOYUDENG	0.9538	0.9410	-	-	-	-
gerarld	0.9532	0.9420	0.5393	0.4972	**0.5659**	0.4489
CS-UM6P	0.9506	0.9380	0.6360	0.6242	0.4537	0.4759
CSECU-DSG	0.9496	0.9380	0.6803	0.4423	0.5366	0.5395
YoungSheldon	0.9468	0.9330	0.5257	0.6210	0.4780	0.4500
DuluthNLP	0.9399	0.9260	0.6461	-	-	0.5059
pakawat.nk	0.9386	0.9240	0.5700	0.4683	0.5496	0.5368
Grenzlinie	0.9386	0.9250	0.6312	0.5455	0.5203	0.4761
bousselham	0.9368	0.9200	-	-	-	-
hub	0.9364	0.9210	0.6288	0.5591	0.5333	0.5027
ZYJ	0.9348	0.9210	0.7214	0.4603	0.4407	0.5204
xjh	0.9345	0.9180	0.6385	0.5205	0.5447	0.5151
Gulu	0.9341	0.9190	0.7405	0.5488	0.5561	0.5807
chenshi	0.9328	0.9160	0.6303	0.5547	0.5301	0.5422
UMUTeam	0.9325	0.9160	0.8847	0.5722	0.4650	0.8740
Han_Jiawei	0.9286	0.9120	0.5577	0.4904	0.5268	0.5187

Zehao_Liu	0.9241	0.9060	-	-	-	-
Team KGP	0.9233	0.9030	0.5694	0.5628	0.5301	0.5800
Tsia	0.9205	0.8960	0.7010	0.4271	0.5593	0.5419
chilai1996	0.9177	0.8970	2.1883	0.0000	0.5463	0.9587
ayushnanda14	0.9081	0.8840	2.1883	0.0000	0.5463	0.9587
DUTH	0.8942	0.8720	0.5507	0.5990	0.4732	0.5819
baseline	*0.8840*	*0.8570*	*0.8609*	*0.4624*	*0.4374*	*0.6415*
LOLASING	0.8704	0.8490	-	-	-	0.7106
CHaines	0.8504	0.8170	0.5762	0.6242	0.4537	0.6473
AlviIshmam	0.8489	0.8160	-	-	-	-
milad.sayadamooz	0.6290	0.5270	2.5497	0.0000	0.5463	0.9587
FII Funny	0.0630	0.0780	0.5598	0.4752	0.5008	0.4788
Paima	-	-	0.5701	-	-	0.4655
abhideepmitra	-	-	1.0343	0.5366	0.4612	-
justglowing	-	-	-	-	-	0.6347

Table 12: Top system for each participant for all subtasks.

Don't Take It Personally: Analyzing Gender and Age Differences in Ratings of Online Humor

J. A. Meaney[1( )], Steven R. Wilson[1,2], Luis Chiruzzo[3], and Walid Magdy[1]

[1] School of Informatics, University of Edinburgh, Edinburgh, UK

[2] Oakland University, Rochester, MI, USA

[3] Universidad de la República, Montevideo, Uruguay

Abstract. Computational humor detection systems rarely model the subjectivity of humor responses, or consider alternative reactions to humor - namely offense. We analyzed a large dataset of humor and offense ratings by male and female annotators of different age groups. We find that women link these two concepts more strongly than men, and they tend to give lower humor ratings and higher offense scores. We also find that the correlation between humor and offense increases with age. Although there were no gender or age differences in humor detection, women and older annotators signalled that they did not understand joke texts more often than men. We discuss implications for computational humor detection and downstream tasks.

Keywords: Computational humor · Offense detection · Online texts · Demographics

1 Introduction

Computational Humor Detection is a fast-growing area of research and has produced at least one humor detection challenge per year since 2017 with Hashtag Wars in SemEval 2017, [18], the Spanish-language HAHA task in Iberlef 2018 [4] and 2019 [5], Assessing Humor in Edited Headlines in 2020 [11] and HaHackathon in 2021 [17]. With participation in these challenges increasing year on year, organisers are beginning to refine their conception of humor, and to incorporate some of the vast, inter-disciplinary findings of the broader humor research community.

One vital branch of this research is that humor is known to vary along the lines of demographic characteristics. Factors such as age [14], gender [10], personality [21] and other demographic variables all modulate responses to humor. Humor tasks have struggled to incorporate such demographic awareness into

F. Hopfgartner et al. (Eds.): SocInfo 2022, LNCS 13618, pp. 20–33, 2022.
https://doi.org/10.1007/978-3-031-19097-1_2

their tasks, and instead tend to average over all humor ratings - which removes nuance and subjectivity from the data [16], as well as possibly decreasing the generalizability of humor detection systems.

A second salient finding from the broader humor literature is that humor is closely linked to offense [15] and indeed, can be used as a mechanism to mask hateful or offensive content. Several competitions have modelled hate speech [1] [27], which is related to offense, but HaHackathon was the first humor detection competition to co-model humor and offense. As the concept of offense is less tangible than humor, it was split in two:

1. *General offense* meaning that a text targets a group of people simply because they belonged to that group and/or is likely upsetting to a lot of people.
2. *Personal offense*, targeting a group that the reader belongs to or cares about.

Although the annotators of this dataset provided demographic data about their age and gender, this was not released as part of the humor detection task, and this is the first analysis of the impact of these age and gender on the humor and offense ratings in this large dataset. The analysis aims to uncover if humor and offense are as meaningfully linked in big datasets as they are in small-N studies, while validating evidence that there are gendered differences in the distribution of humor ratings [24], as well as tolerance of aggressive humor.

As in [10], we are mindful of the use of *gender* to specify a cultural phenomenon, indicating men and women as socially-defined groups, rather than a biological distinction.

1.1 Related Work

Gender and age differences have been the subject of many studies in the fields of psychology, sociology, education, and management studies. Svebak et al. [24] found that "overall humor scores" were higher for men than they were for women. However, it should be noted that "overall humor" was narrowly assessed, using only three items, with each representing one of the dimensions of the Situational Humor Questionnaire. The same work reported that humor appreciation declines with age: the mean scores for total sense of humor on average declined across the age cohorts from highest score in the 20s to lowest score among those aged 70. More recently, an Italian study of covid-related humor [3] reported that increasing age, as well as being female was related to finding pandemic humor more aversive and less funny.

In terms of gender differences, perhaps the most replicated result is that men tolerate aggressive humor more than female respondents do [10]. Proyer and Ruch [20] report that men tended to score higher on kagelaticism - the joy of laughing at others, which suggests that as long as a joke does not target men explicitly, it may be offensive towards other groups, without impacting men's humor ratings. Interestingly, Knegtmans et al. [12] found that participants whose social power had been manipulated to place them in a high-power state rated jokes which targeted others as less offensive, and gave higher humor ratings.

No differences in the appreciation of nonsense humor [13], or neutral jokes [7] were found.

1.2 Research Questions (RQ)

1. Is there a correlation between annotators' perceptions of humor and offense? Does this vary by age and gender?
2. Are there differences in humor *detection* and *comprehension* between groups?
3. Are there differences in the distributions of humor and offense ratings between groups?

Using a dataset of >120k ratings of humor and offense [17], we find a slight negative correlation between humor and offense, which varies as a function of gender and age. The negative link between humor and offense increases as annotators age. We also find a stronger correlation between general offense and humor for women, but male annotators only linked these concepts when they signalled that they were personally offended. There were no significant differences between groups when it came to correctly identifying texts as jokes (i.e. humor detection), but there were differences when it came to humor comprehension. More women than men indicated that they did not get a joke, and women of all age groups had higher rates of using the label "I don't get it" than men of all age groups. In terms of the distribution of ratings, women were more likely to use lower humor ratings and higher offense ratings, while men showed the opposite trend. In terms of age groups, the oldest group tended to report that they didn't get a joke more than any other group, while annotators ages 26–40 were least likely to use this label, and also gave the highest humor ratings overall. Older groups were more likely to use higher ratings of general and personal offense, while younger annotators were less likely to use these.

2 Dataset Description

The dataset features the texts and ratings used in the humor and offense shared task HaHackathon at SemEval 2021 [17]. Including non-humorous texts, this comprises 202,369 ratings of 10,000 texts. Each text has an average of 20.2 ratings, with no text having fewer than 17 votes. There were 1,821 unique annotators (mean age 40.45 years, SD = 15.64 years), and each annotator rated an average of 111.13 texts. The highest number of texts rated by one person was 307.

Of the 10,000 texts in the dataset, 2,000 were sourced from the Kaggle Short Jokes Dataset[1]. Half of the Kaggle texts were selected because they referred to one of the common targets of online hate speech outlined by Silva et al. [22], e.g. women, members of the LGBT community, religious/racial minorities, and this target was the butt of the joke. These texts were deemed likely to elicit ratings

[1] https://github.com/amoudgl/short-jokes-dataset.

of offense from some annotators. The other half of the Kaggle texts referred to a common hate speech target, but did not make it the butt of the joke.

The other 8,000 texts were sourced from Twitter, from a mix of humorous and non-humorous accounts. Amongst the non-humorous accounts, there were several which advocate for, or provide information to common targets of hate speech. This ensured that mentions of these targets were not limited to humorous texts only.

Annotators were asked up to three sets of questions about each text: one related to humor and two related to offense.

1. **Humor detection/rating:** annotators were asked if the intention of the text was to be humorous. This binary response question was aimed at gauging the genre of the text, and annotators were asked not to judge based on whether they found it funny, but whether it contained indicators of the humor genre, e.g. a setup and punchline, puns, absurd content, etc. If the annotator selected 'yes', they were asked to rate how funny they found it from 1–5. There was also the option to select 'I don't get it' if the text was identified as humorous, but the humor was not understood. If the annotator selected 'no', they were not asked any further questions about this text.
2. **General offense detection/rating:** If a text had been labelled as humorous, annotators were asked if they thought the text targeted a group simply because they belonged to a group, or if they thought the text would be offensive to a large number of people. In the case of a 'yes' response, they were asked how generally offensive they thought the text was from 1–5.
3. **Personal offense:** If a text had been labelled as humorous, we asked annotators if they were personally hurt by the text, or were hurt on someone else's behalf, and if so, to rate how much from 1–5.

The pool of annotators comprised 4 age groups: 18–25, 26–40, 41–55, 56–70. In order to avoid a lack of shared cultural knowledge, all annotators were native English speakers and citizens of the United States. Although we aimed to be inclusive of diverse genders, the dataset included only four annotators who preferred not to disclose their gender. As they rated a total of 384 texts, they were excluded from the gender analysis, for reasons of statistical power.

Annotators provided informed consent before beginning the annotation, and the procedure was approved by the Ethics Committee of the corresponding author's institution. Other demographic data about the annotators, such as gender and personality traits, was also provided as part of the dataset.

3 Methodology

Given that the humor and offense annotations were reported using an ordinal scale, for RQ1, we used the Spearman rank correlation [23] to report the correlations between these variables. The Spearman rank correlation is a generalisation of the Pearson correlation which is used for discrete and ordinal data which captures the strength and direction of the relationship between two variables by

ranking the values of each variable, summing the square differences and calculating the covariance of the ranks. This returns a correlation coefficient, ρ, ranging from -1 to $+1$, the magnitude of which indicates the strength of the relationship and the sign signifies the direction. It also returns a p-value - the probability that the value of the coefficient could occur under the null hypothesis.

To answer RQ2, we calculated the proportion of annotators from each group (i.e., gender or age group) that mislabeled (failed to *detect*) or misunderstood (failed to *comprehend*) each text. The resulting distributions were non-normal, so we chose non-parametric tests, which do not assume an underlying distribution. As we have only two values for gender in the dataset, we used a Wilcoxon Signed Rank test [26] to examine the null hypothesis that the samples from male and female annotators came from the same distribution. This is similar to a paired t-test, and it ranks the absolute value of the pairs of differences to calculate the test statistic, w. With this test, we report the Common Language Effect Size (CLES): the proportion of pairs where the values for one group are higher than the other.

For more than two groups, i.e., our age variable, which had four bins, we use the Friedman test [8], which is similar to a repeated measures ANOVA. Again values are ranked and the test compares the mean rank of each group for statistical significance. In the case of a significant result, we ran post hoc pairwise Wilcoxon tests. We used the Bonferroni correction to adjust the p-values for multiple comparisons, reducing the risk of false positive results.

For RQ3, we first used the Wilcoxon and Friedman tests to determine if one group tended to give higher or lower ratings than another. We then used a chi-square test of homogeneity to examine how the distributions differed from each other. This test determines if the frequencies of each possible value of the dependent variable are distributed in the same way across the different groups. The test calculates the expected frequencies of each rating by each group by multiplying the number of annotators in each group by the true probability that any annotator would pick each answer. This expected frequency is then compared to the observed frequency.

4 Results

4.1 RQ1: Is There a Correlation Between Humor and Offense?

For the following analysis, we excluded texts which had been labelled as 'not humorous' by our annotators, and removed outliers (e.g. texts that had fewer than 3 humor ratings). This left 121,622 ratings of 6,918 texts.

Overall, there was a small negative correlation between humor and general offense ($\rho = -0.13$, $p < 0.05$), and this grew stronger for humor and personal offense ($\rho = -0.19$, $p < 0.05$), which suggests that offensive content is negatively related to humor appreciation. There was a strong correlation between general and personal offense ($\rho = 0.60$, $p < 0.05$), indicating that these concepts are linked, but are not identical.

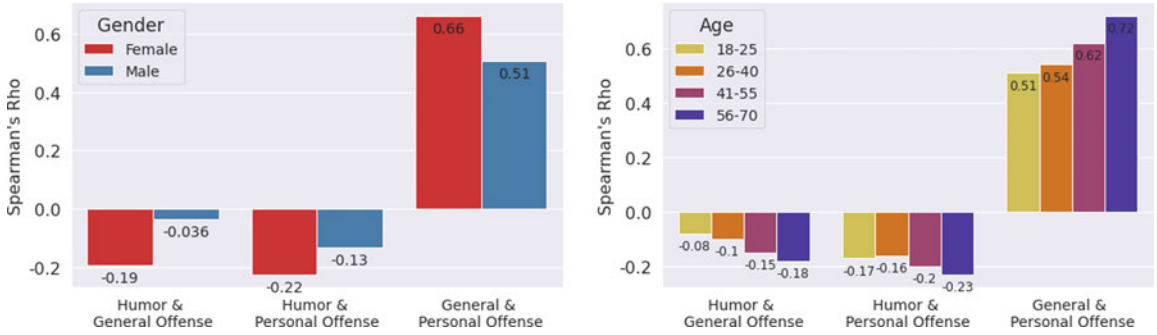


Fig. 1. Correlations between humor and offense by gender and age

Correlations Between Ratings by Gender. When examining the correlations between ratings split by gender, an interesting trend emerged (Fig. 1). There was almost no relationship between humor and *general* offense for men, however *personal* offense ratings were negatively correlated with humor ratings. Conversely, for female annotators, both types of offense were more strongly correlated with a reduced humor rating for female annotators.

Correlations by Age. A second interesting trend emerged in terms of age: the older the annotators were, the stronger the negative link between general *and* personal offense on humor ratings was (Fig. 1). The oldest group had the most prominent negative correlation between humor and both types of offense, as well as the strongest correlation between the two offense metrics.

Correlations by Age and Gender. Although splitting 20 ratings per text into 8 groups (for four age groups by two gender groups) would cause issues of data sparsity and statistical power, we noted that the trend of an increasingly negative correlation between humor and offense continues when this is broken down by age and gender (Fig. 2). Female annotators relate lower humor scores to higher offense scores increasingly with age, and this trend is much less pronounced in male annotators.

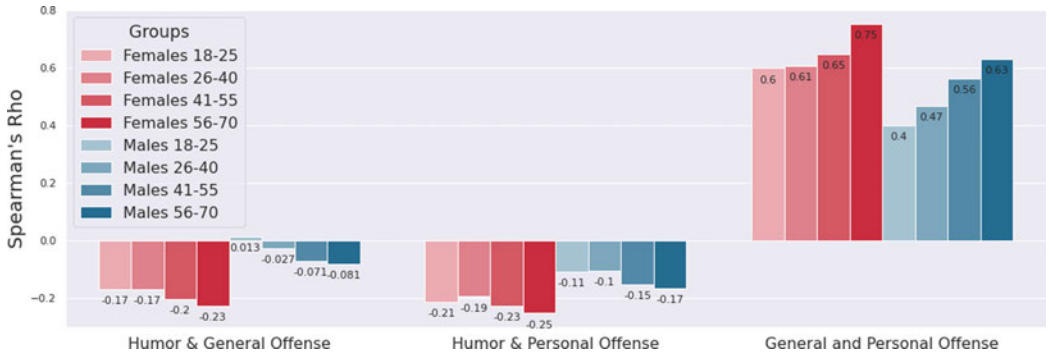


Fig. 2. Correlations between humor and offense by age and gender

4.2 RQ2: Are There Differences in Humor Detection and Comprehension Between Groups?

Humor Detection. To investigate differences in annotators' humor *detection*, we looked at the proportion of male and female annotators who labelled each text from the Kaggle data as 'not humorous'. We confined this analysis to the Kaggle data because all texts in this dataset was intended to be humorous, and should have been labeled as such. A paired Wilcoxon signed rank test showed that there was no significant difference between groups ($z = 134201.0$, $p = 0.29$) (Table 1).

Table 1. Mislabeling and misunderstanding in the Kaggle Jokes

	Male	Female
Proportion of annotations from each group	42.92%	57.08%
'Not Humor' ratings from each group	3.79%	3.70%
Unique texts with 1+ label of not-humorous	22.91%	25.42%
'I don't get it' ratings from each group	5.77%	7.35%
Unique texts with 1+ rating of 'I don't get it"	33.04%	45.81%

We used a similar procedure to test if there were significant differences between age groups in terms of humor detection. A Friedman test showed that there were no significant differences between groups ($\chi^2 = 6.976$, $p = 0.07$).

Humor Comprehension. After labeling a text as humorous, one of the options for humor rating was 'I don't get it'. This indicated that the annotator had recognized that the text was intended to be humorous, but that they lacked the knowledge to fully understand the joke. We first looked at the Kaggle dataset, and calculated the number of 'I don't get it' votes from men and women, as a proportion of the total votes per text from each group. A paired Wilcoxon signed rank test showed that there was a significant difference between groups ($z = 214403.0$, $p < 0.05$). We used Pingouin [25] to calculate the Common Language Effect Size (CLES), i.e. the proportion of pairs where the proportion of 'I don't get it' ratings provided by female annotators is greater than the proportion of male annotators who gave that rating. The resulting CLES of 0.5540 indicates that a larger proportion of female annotators indicated that they did not get the joke in 55.45% of pairs. When looking at the data from Twitter, women still admit to not getting the joke more than men ($z = 2298680.0$, $p < 0.05$), but the effect is less pronounced, CLES = 0.5223.

We examined differences between age groups in terms of humor detection. A Friedman test showed that there were no significant differences between groups ($\chi^2 = 0.0012$, $p = 0.06$).

4.3 RQ3: Are There Differences Between Groups in Distributions Humor and Offense Ratings?

When looking at the distribution of ratings across the 6 possible values (1–5 and *'I don't get it'*) for the entire dataset (both Kaggle and Twitter), a χ^2 test of homogeneity demonstrated that there were significant differences between the distributions of humor ratings between men and women ($\chi^2 = 202.25$, $p < 0.05$) and showed that women were more likely to select 'I don't get it', while men were more likely to use higher ratings. We also explored if this difference translated into different average humor ratings per text and a Wilcoxon signed rank showed that men gave significantly higher ratings than women on humor ($z = 9684516.5$, $p < 0.05$) and the CLES score of 0.5333 indicated that men gave higher humor ratings in 53.33% of pairs.

For general offense, a χ^2 test of homogeneity showed significant differences between groups ($\chi^2 = 430.85$, $p < 0.05$), and examining the expected versus observed counts showed that the trend seen in the humor ratings was reversed: men were more likely to choose low offense ratings and women were more likely to select higher values. In terms of averaged general offense ratings, group differences were significant ($z = 4260050.5$, $p < 0.05$, CLES = 0.4704), with men giving higher offense ratings in 47.04% of pairs.

Similarly, for personal offense, a χ^2 test of was significant ($\chi^2 = 1195.94$, $p < 0.05$) with a more pronounced trend showing that women were more likely to select a high personal offense rating, and men systematically under-selected high ratings. This led to significant differences in the average personal offense ratings per text, where men gave higher personal offense scores in only 41.5% of pairs ($z = 1234096.5$, $p < 0.05$, CLES = 0.4146).

When looking at age groups, a χ^2 test showed significant differences in humor ratings between age groups ($\chi^2 = 239.98$, $p < 0.05$). The oldest group, 56–70, were most likely to report 'I don't get it', while annotators aged 26–40 were least likely to use this, and most likely to give high ratings. In terms of general offense, there were significant group differences ($\chi^2 = 540.936$ $p < 0.05$), and annotators 18–40 were more likely to give lower general offense ratings, while those aged 41–70 used fewer low ratings than expected, and the group ages 56–70 was most likely to give the highest possible offense rating of 5. Group differences were more pronounced in personal offense ratings ($\chi^2 = 1387.43$, $p < 0.05$) where the two youngest groups gave consistently lower than expected ratings of personal offense, while the older group gave consistently higher ratings. This resulted in significant differences in the average personal offense scores between groups ($\chi^2 = 38.223$, $p < 0.05$).

5 Qualitative Analysis

The negative correlation for female annotators between humor and general offense, which was uncovered in the above analysis, is succinctly illustrated in Fig. 3. Texts which are offensive for women tend to earn a lower humor rating, while general offense is more tolerated by men.

Fig. 3. Relationship between humor and offense by gender

To examine what type of texts male and female annotators differed on with regard to general offense ratings, we selected the top 40 texts where there was at least a 1.5 point difference between the mean general offense score given by male and female annotators. We labeled the topic or target of the texts and five annotators rated whether the content was aggressive or not. Annotators were instructed that a text should be deemed aggressive if it contained violent content or used racial slurs, and inter-annotator agreement was relatively high (Fleiss's $\kappa = 0.3815$) (Fig. 2).

Table 2. Sample texts where annotators differed on general offense

Text	Humor		G. Offense	
	Female	Male	Female	Male
Why are the labia on Japanese women oriented sideways instead of vertically? Goes better with their eyes	1.0	2.2	4.2	1.3
In my spare time, I help blind kids I mean the verb, not the adjective	1.3	2.0	2.2	0.17
Two condoms walk by a gay bar... One says to the other, "Wanna go inside and get shitfaced?"	2.6	1.6	0.85	2.4
What did the Jewish pirate say when he heard his wife died? Argh, shiva me timbers	1.6	1.6	1.0	2.1

There was a sizeable overlap of topics, with women finding texts about the LGBT community more offensive than men, while male annotators found texts about religion more generally offensive. The texts that were offensive to women tended to be aggressive, while men were more tolerant of this. Interestingly, men selected several texts which were not intended to be jokes (e.g. were drawn

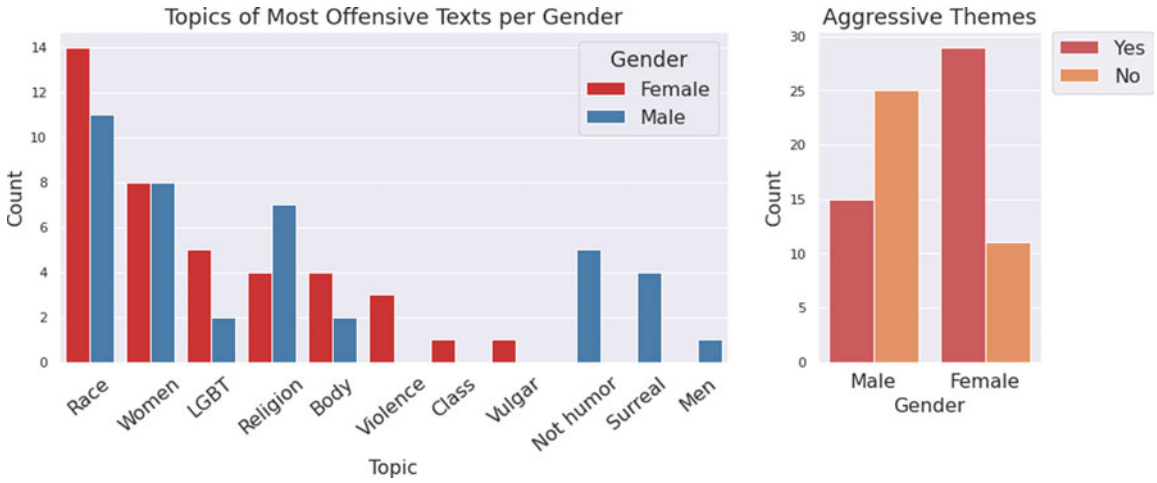


Fig. 4. Topics and aggression where gender groups disagreed on general offense ratings

from accounts supporting targets of hate speech) as both humorous and offensive (Fig. 4).

We followed a similar procedure to examine the texts where offense ratings from different age groups differed from each other. We compared the mean general offense rating from each group to the average general offense rating from the other 3 groups combined, and looked at the top 40 texts where there was at least a 1.5 point difference. Several topics predominate, namely race, women, body (e.g. disability, body weight). The texts rated as more generally offensive by younger groups focused on these topics, but as age increased, so did the variety of topics featured. The texts selected by group 1 (the youngest group) featured more which were aggressive in nature, but as age increased, aggression was less linked to offense (Fig. 5).

6 Discussion

We used a large dataset of texts rated for humor and offense, along with some demographic information about the annotators to explore differences between age and gender groups. We looked at how the groups link humor and offense, differences in humor detection and comprehension, as well as differences in the distributions of ratings.

RQ1: We found that female annotators negatively link humor and offense more strongly than men. Male annotators do not link general offense with diminished humor ratings. In fact, they link humor and offense to a lesser extent, and only when personally offended.

As regards age groups, the correlation between humor and offense was weakest in the youngest group, and grew steadily with age - as did the link between general and personal offense.

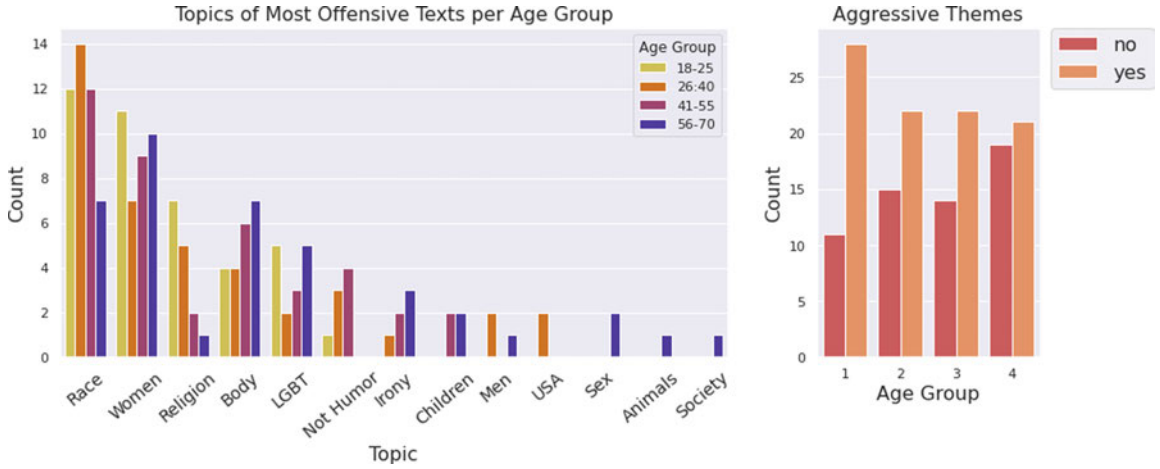


Fig. 5. Analysis of topics and aggression where age groups disagreed on general offense ratings

RQ2: There were no differences in gender or age groups in terms of humor detection. However, when it came to humor comprehension, women selected 'I don't get it' more often than men.

RQ3: In terms of the distributions of ratings, women gave lower humor ratings and higher offense ratings, while men showed the opposite trend. Amongst the age groups, annotators 26–40 gave the highest ratings and the fewest reports of 'I don't get it'. In line with findings from RQ1, younger groups gave lower offense ratings and older groups reported higher offense.

Some of the findings above are well attested in the humor literature, albeit in smaller-N studies. Hofmann et al. [10] report that men's tolerance of aggressive humor is one of the most consistent findings in the humor field, with seven out of eight studies mentioned replicating this result. Our qualitative work shows that in the texts on which men and women differed most on general offense, aggression featured more prominently for women. Perhaps relatedly, Proyer and Ruch [20] report that men score higher on katagelasticism - the joy of laughing at others. This may be reflected in the fact that general offense does not diminish male annotators' humor ratings, only personal offense does.

A more surprising result is the increasingly strong negative correlation between humor and offense as age progressed. This contradicts the oft-touted idea of *Generation Snowflake*, which contends that those born after 1995 tend to be the most overly reactive to offensive material [9]. The older age groups - 40–55 and 56–70 - gave higher ratings of offense than their younger counterparts, and our qualitative analysis indicated that the older groups gave higher offense ratings to a wider variety of topics.

The finding that women used the 'I don't get it' label more than men is a result that may benefit from some contextualisation from the humor literature. Bell [2] found that when shown incomprehensible jokes, women tended to explicitly state that they did not get it, while men implicitly signaled it by asking concept-checking questions. It is not possible to know whether this was the

case here, but it is true that the qualitative results uncovered that men were mentioning not humorous texts as both humorous and offensive.

6.1 Implications

Given the gender and age group differences in ratings of humor and offense, it is evident that humor detection systems which average over all annotators' ratings fail to model the subjectivity that is inherent to this task. These systems may not generalise well on downstream tasks, such as content moderation, and may not be effective at moderating aggressive content if they are tuned to men's preferences, or alternatively may be more restrictive if tuned to women's preferences. Furthermore, as sociologists have pointed out [15], the line between humor and offense is continually under revision in most societies, therefore not only are these responses subjective, but they are a moving target. We should focus on incorporating frameworks to include demographic knowledge in our systems, which can constantly be updated to reflect society's changing definitions of humor and offense.

6.2 Limitations

It is a limitation that the dataset did not afford the opportunity to explore the interaction between age and gender. As each text has approximately 20 annotations per text, splitting these into 8 groups to model age and gender would not have provided sufficient statistical power. Similarly, it is a limitation that there were insufficient annotations from gender non-conforming annotators, as there is a dearth of literature on their reactions to humor and offense. The lack of annotators that self-identify with genders other than female and male has been noticed in the past in different tasks as well [6,19].

A final constraint is that we are modelling only one half of the humorous interaction - the recipient of the joke. Excluding the teller of the joke can deny the recipient some important context needed to enjoy the joke, and different tellers can mitigate the responses. Future work should include this dimension.

7 Conclusion

We present the first analysis of the demographic data provided with the HaHackathon data - a large dataset used to train systems for computational humor detection. Our findings indicate that women negatively link humor to offense, while men only do so if they are personally offended. Links between humor and offense grew with age. There were no differences in humor detection by gender or age groups, but women and older annotators indicated that they did not understand jokes more than men. Distributions of humor and offense ratings replicated findings from humor research, namely that men gave higher humor ratings and lower offense ratings. We hope that these findings will inform future frameworks for computational humor detection and dataset creation.

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

1. Basile, V., et al.: SemEval-2019 task 5: multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 54–63 (2019)
2. Bell, N.D.: Responses to incomprehensible humor. J. Pragmat. **57**, 176–189 (2013)
3. Bischetti, L., Canal, P., Bambini, V.: Funny but aversive: a large-scale survey of the emotional response to Covid-19 humor in the Italian population during the lockdown. Lingua **249**, 102963 (2021)
4. Castro, S., Chiruzzo, L., Rosá, A.: Overview of the HAHA task: humor analysis based on human annotation at IberEval 2018. In: IberEval@ SEPLN, pp. 187–194 (2018)
5. Chiruzzo, L., Castro, S., Etcheverry, M., Garat, D., Prada, J.J., Rosá, A.: Overview of haha at IberEval 2019: humor analysis based on human annotation. In: IberLEF@ SEPLN (2019)
6. Excell, E., Moubayed, N.A.: Towards equal gender representation in the annotations of toxic language detection. In: Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, pp. 55–65 (2021)
7. Ferstl, E.C., Israel, L., Putzar, L.: Humor facilitates text comprehension: evidence from eye movements. Discourse Process. **54**(4), 259–284 (2017)
8. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. **32**(200), 675–701 (1937)
9. Haidt, J., Lukianoff, G.: The coddling of the American mind: how good intentions and bad ideas are setting up a generation for failure. Penguin UK (2018)
10. Hofmann, J., Platt, T., Lau, C., Torres-Marín, J.: Gender differences in humor-related traits, humor appreciation, production, comprehension, (neural) responses, use, and correlates: a systematic review. Curr. Psychol. 1–14 (2020)
11. Hossain, N., Krumm, J., Gamon, M., Kautz, H.: SemEval-2020 task 7: assessing humor in edited news headlines. arXiv preprint arXiv:2008.00304 (2020)
12. Knegtmans, H., Van Dijk, W.W., Mooijman, M., Van Lier, N., Rintjema, S., Wassink, A.: The impact of social power on the evaluation of offensive jokes. Humor **31**(1), 85–104 (2018)
13. Köhler, G., Ruch, W.: Sources of variance in current sense of humor inventories: how much substance, how much method variance? (1996)
14. Kuipers, G.: The humor divide: class, age and humor styles. In: Good Humor, Bad Taste, pp. 71–101. De Gruyter Mouton (2015)
15. Lockyer, S., Pickering, M.: Beyond a Joke: The Limits of Humour. Springer, London (2005). https://doi.org/10.1057/9780230236776
16. Meaney, J.: Crossing the line: where do demographic variables fit into humor detection? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 176–181 (2020)
17. Meaney, J., Wilson, S., Chiruzzo, L., Lopez, A., Magdy, W.: SemEval 2021 task 7: hahackathon, detecting and rating humor and offense. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 105–119 (2021)

18. Potash, P., Romanov, A., Rumshisky, A.: SemEval-2017 task 6:# hashtagwars: learning a sense of humor. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 49–57 (2017)
19. Prabhakaran, V., Davani, A.M., Diaz, M.: On releasing annotator-level labels and information in datasets. In: Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop, pp. 133–138 (2021)
20. Proyer, R.T., Ruch, W.: Enjoying and fearing laughter: personality characteristics of gelotophobes, gelotophiles, and katagelasticists. Psychol. Test Assess. Model. **52**(2), 148–160 (2010)
21. Ruch, W.: The Sense of Humor: Explorations of a Personality Characteristic, vol. 3. Walter de Gruyter, Berlin (2010)
22. Silva, L., Mondal, M., Correa, D., Benevenuto, F., Weber, I.: Analyzing the targets of hate in online social media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 10 (2016)
23. Spearman, C.: The proof and measurement of association between two things. Am. J. Psychol. **15**(1), 72–101 (1904)
24. Svebak, S., Martin, R.A., Holmen, J.: The prevalence of sense of humor in a large, unselected county population in Norway: relations with age, sex, and some health indicators (2004)
25. Vallat, R.: Pingouin: statistics in python. J. Open Source Softw. **3**(31), 1026 (2018)
26. Wilcoxon, F.: Individual comparisons by ranking methods. In: Biometrics Bulletin, no. 6, vol. 1, pp. 80–83 (1945). http://www.jstor.org/stable/3001968
27. Zampieri, M., et al.: SemEval-2020 task 12: multilingual offensive language identification in social media (OffensEval 2020). arXiv preprint arXiv:2006.07235 (2020)

Bibliography

Aillaud, M., & Piolat, A. (2012). Influence of gender on judgment of dark and nondark humor. *Individ. Differ. Res.*, *10*(4), 211–222.

Al-Omari, H., AbedulNabi, I., & Duwairi, R. (2021). DLJUST at SemEval-2021 Task 7: Hahackathon: Linking Humor and Offense Across Different Age Groups. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing.*

Annamoradnejad, I., & Zoghi, G. (2020). Colbert: Using bert sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*.

Attardo, S., & Raskin, V. (1991). Script theory revis (it) ed: Joke similarity and joke representation model.

Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Evaluation measures for ordinal regression. In *2009 ninth international conference on intelligent systems design and applications* (pp. 283–287).

Barbieri, F., & Saggion, H. (2014). Automatic detection of irony and humour in twitter. In *Iccc* (pp. 155–162).

Bashabsheh, E. A., & Alasal, S. A. (2021). ES-JUST at SemEval-2021 Task 7: Detecting and Rating Humor and Offensive Text Using Deep Learning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing.*

Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017, August). DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 747–754). Vancouver, Canada: Association for Computational Linguistics.

Bell, N. D. (2013). Responses to incomprehensible humor. *Journal of Pragmatics*, *57*, 176–189.

Bischetti, L., Canal, P., & Bambini, V. (2021). Funny but aversive: A large-scale survey of the emotional response to covid-19 humor in the italian population during the lockdown. *Lingua*, *249*, 102963.

Bishop, P. A., & Herron, R. L. (2015). Use and misuse of the likert item responses and other ordinal measures. *International journal of exercise science*, *8*(3), 297.

Buscaldi, D., & Rosso, P. (2007). Some experiments in humour recognition using the italian wikiquote collection. In *International workshop on fuzzy logic and applications* (pp. 464–468).

Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, *140*, 325–331.

Castro, S., Chiruzzo, L., & Rosá, A. (2018). Overview of the HAHA Task: Humor Analysis Based on Human Annotation at IberEval 2018. In *Ibereval@ sepln* (pp. 187–194).

Chi, N., & Chi, R. (2021). Redwoodnlp at semeval-2021 task 7: Ensembled pretrained and lightweight models for humor detection. In *Proceedings of the 15th international workshop on semantic evaluation (semeval-2021)* (pp. 1209–1214).

Chiruzzo, L., Castro, S., Etcheverry, M., Garat, D., Prada, J. J., & Rosá, A. (2019). Overview of HAHA at IberLEF 2019: Humor Analysis based on Human Annotation. In *Iberlef@ sepln* (pp. 132–144).

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

Consortium, B., et al. (2007). British national corpus. *Oxford Text Archive Core Collection*.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international aaai conference on web and social media* (Vol. 11, pp. 512–515).

Dawes, J. (2008). *Do data characteristics change according to the number of scale points used? an experiment using 5-point, 7-point and 10-point scales: Https://doi. org/10.1177/147078530805000106, 50 (1), 61–77.*

Detlefsen, N. S., Borovec, J., Schock, J., Jha, A. H., Koker, T., Di Liello, L., ... Falcon, W. (2022). Torchmetrics-measuring reproducibility in pytorch. *Journal of Open Source Software*, *7*(70), 4101.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Diao, Y., Lin, H., Yang, L., Fan, X., Wu, D., & Xu, K. (2020). Crga: Homographic pun detection with a contextualized-representation: Gated attention network. *Knowledge-Based Systems*, *195*, 105056.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology*, *41*(1), 417–440.

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 english words. *Behavior research methods*, *50*, 1116–1124.

Ermilov, A., Murashkina, N., Goryacheva, V., & Braslavski, P. (2018). Stierlitz meets svm: humor detection in russian. In *Artificial intelligence and natural language: 7th international conference, ainl 2018, st. petersburg, russia, october 17–19, 2018, proceedings 7* (pp. 178–184).

Essefar, K., Mekki, A. E., Mahdaouy, A. E., Mamoun, N. E., & Berrada, I. (2021). CS-UM6P at SemEval-2021 Task 7: Deep Multi-Task Learning Model for Detecting and Rating Humor and Offense. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing.*

Excell, E., & Moubayed, N. A. (2021). Towards equal gender representation in the annotations of toxic language detection. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 55–65.

Falcon, W., & The PyTorch Lightning team. (2019, March). *PyTorch Lightning.* Retrieved from `https://github.com/Lightning-AI/lightning` doi: 10 .5281/zenodo.3828935

Faraj, D., & Abdullah, M. (2021). SarcasmDet at SemEval-2021 Task 7: Detect Humor and Offensive based on Demographic Factors using RoBERTa Pre-trained Model. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing.*

Faruqi, F., & Shrivastava, M. (2018). Is this a joke?": A large humor classification dataset. In *15th international conference on natural language processing* (p. 109).

Fedor, N., Gaynor, F., & Reik, T. (1950). Freud: dictionary of psychoanalysis. *Academic Medicine*, *25*(6), 458.

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications* (pp. 231–243). Springer.

Fortuna, P. C. T. (2017). Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes.

Frangidis, P., Georgiou, K., & Papadopoulos, S. (2020). Sentiment analysis on movie scripts and reviews: Utilizing sentiment scores in rating prediction. In *Ifip international conference on artificial intelligence applications and innovations* (pp. 430–438).

Freedman, D., Pisani, R., & Purves, R. (2007). Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, *32*(200), 675–701.

Garimella, A., Banea, C., & Mihalcea, R. (2017, September). Demographic-aware word associations. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2285–2295). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D17-1242` doi: 10.18653/v1/D17-1242

Greengross, G. (2013). Humor and aging-a mini-review. *Gerontology*, *59*(5), 448–453.

Gu, K., & Budhkar, A. (2021, June). A package for learning on tabular and text data with transformers. In *Proceedings of the third workshop on multimodal artificial intelligence* (pp. 69–73). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/2021.maiworkshop-1.10` doi: 10.18653/v1/2021.maiworkshop-1.10

Gugnani, A., Zylich, B., Brookman, G., & Samoray, N. (2021). Amherst685 at SemEval-2021 Task 7: Joint Modeling of Classification and Regression for Humor and Offense. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing.*

Gupta, A., Pal, A., Khurana, B., Tyagi, L., & Modi, A. (2021). Humor@IITK at SemEval-2021 Task 7: Large Language Models for Quantifying Humor and Offensiveness. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint arXiv:2004.10964*.

Haidt, J., & Lukianoff, G. (2018). *The coddling of the american mind: How good intentions and bad ideas are setting up a generation for failure*. Penguin UK.

Hanson, M. (2021). Average cost of college & tuition. *Education Data Initiative. https://educationdata. org/average-cost-of-college*.

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654*.

Hempelmann, C. F. (2004). Script opposition and logical mechanism in punning.

Highfield, T. (2015). Tweeted Joke Lifespans and Appropriated Punchlines: Practices around Topical Humor on Social Media. *International Journal of Communication*, *9*, 22.

Hofmann, J., Platt, T., Lau, C., & Torres-Marín, J. (2020). Gender Differences in Humor-Related Traits, Humor Appreciation, Production, Comprehension,(Neural) Responses, Use, and Correlates: A Systematic Review. *Current Psychology*, 1–14.

Hossain, N., Krumm, J., Gamon, M., & Kautz, H. (2020). SemEval-2020 Task 7: Assessing humor in Edited News Headlines. *arXiv preprint arXiv:2008.00304*.

Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 752–762).

Jaiswal, A., Mathur, A., Mattu, S., et al. (2019). Automatic humour detection in tweets using soft computing paradigms. In *2019 international conference on machine learning, big data, cloud and parallel computing (comitcon)* (pp. 172–176).

Jiang, T., Li, H., & Hou, Y. (2019). Cultural differences in humor perception, usage, and implications. *Frontiers in psychology*, *10*, 123.

Kalliny, M., Cruthirds, K. W., & Minor, M. S. (2006). Differences between american, egyptian and lebanese humor styles: Implications for international management. *International Journal of Cross Cultural Management*, *6*(1), 121–134.

Kazarian, S. S., & Martin, R. A. (2004). Humour styles, personality, and well-being among lebanese university students. *European journal of Personality*, *18*(3), 209–219.

Khandelwal, A., Swami, S., Akhtar, S. S., & Shrivastava, M. (2018). *Humor detection in english-hindi code-mixed social media content : Corpus and baseline system.*

Knegtmans, H., Van Dijk, W. W., Mooijman, M., Van Lier, N., Rintjema, S., & Wassink, A. (2018). The impact of social power on the evaluation of offensive jokes. *Humor*, *31*(1), 85–104.

Kotthoff, H. (2006, January). Gender and humor: The state of the art. *J. Pragmat.*, *38*(1), 4–25.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Kuipers, G. (2015). The Humor Divide: Class, Age and Humor Styles. In *Good humor, bad taste* (pp. 71–101). De Gruyter Mouton.

Kuipers, G. (2017). Humour styles and class cultures: Highbrow humour and lowbrow humour in the netherlands. In *The anatomy of laughter* (pp. 58–69). Routledge.

Ladhak, F., Durmus, E., Suzgun, M., Zhang, T., Jurafsky, D., Mckeown, K., & Hashimoto, T. B. (2023). When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th conference of the european chapter of the association for computational linguistics* (pp. 3198–3211).

Lampert, M. D., & Ervin-Tripp, S. M. (1998). Exploring paradigms: The study of gender and sense of humor near the end of the 20th century. *The sense of humor: Explorations of a personality characteristic*, *3*, 231–270.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, *18*(1), 559–563.

Leung, S.-O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point likert scales. *Journal of social service research*, *37*(4), 412–421.

Liu, L., Zhang, D., & Song, W. (2018). Exploiting syntactic structures for humor recognition. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1875–1883).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lockyer, S., & Pickering, M. (2005). *Beyond a joke: The limits of humour*. Springer.

Ma, J., Xie, S., Lianxin, J., Stark, R., Yang, M., & Shen, J. (2021). MagicPai at SemEval-2021 Task 7: Method for Detecting and Rating Humor Based on Multi Task Adversarial Training. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing.*

Mahajan, R., & Zaveri, M. (2020). Humor identification using affect based content in target text. *Journal of Intelligent & Fuzzy Systems*, *39*(1), 697–708.

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of research in personality*, *37*(1), 48–75.

McCulloch, G. (2020). *Because internet: Understanding the new rules of language*. Penguin.

Meaney, J. (2020). Crossing the Line: Where do Demographic Variables Fit into Humor Detection? In *Proceedings of the 58th annual meeting of the association for computational linguistics: Student research workshop* (pp. 176–181).

Mihalcea, R., & Strapparava, C. (2005). Computational laughing: Automatic recognition of humorous one-liners. In *Proceedings of cognitive science conference* (pp. 1513–1518).

Mihalcea, R., & Strapparava, C. (2006). Learning to Laugh (Automatically): Computational Models for Humor Recognition. *Computational Intelligence*, *22*(2), 126–142.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miyato, T., Dai, A. M., & Goodfellow, I. (2016). Adversarial Training Methods for Semi-supervised Text Classification. *arXiv preprint arXiv:1605.07725*.

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory and discussion. *ACM Journal of Data and Information Quality*.

Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2016). Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4920–4928).

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145–153).

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

Pandey, C. K., Singh, C., & Mangla, K. (2021, August). EndTimes at SemEval-2021 task 7: Detecting and rating humor and offense with BERT and ensembles. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)* (pp. 1215–1220). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pang, C., Fan, X., Su, W., Chen, X., Wang, S., Liu, J., ... Sun, Y. (2021). abcbpc at SemEval-2021 Task 7: ERNIE-based Multi-task Model for Detecting and Rating Humor and Offense. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Potash, P., Romanov, A., & Rumshisky, A. (2017). SemEval-2017 Task 6: #Hashtagwars: Learning a Sense of Humor. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 49–57).

Prabhakaran, V., Davani, A. M., & Diaz, M. (2021). On releasing annotator-level labels and information in datasets. *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, 133–138.

Proyer, R. T., & Ruch, W. (2010). Enjoying and fearing laughter: Personality characteristics of gelotophobes, gelotophiles, and katagelasticists. *Psychological Test and Assessment Modeling*, *52*(2), 148–160.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872–1897.

Raskin, V. (1985). *Semantic mechanisms of humor. dordrecht–boston–lancaster: D.* Reidel Publishing Company.

Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, *74*, 1–12.

Ruch, W. (2010). *The Sense of Humor: Explorations of a Personality Characteristic* (Vol. 3). Walter de Gruyter.

Ruch, W., McGhee, P. E., & Hehl, F.-J. (1990). Age differences in the enjoyment of incongruity-resolution and nonsense humor during adulthood. *Psychology and aging*, *5*(3), 348.

Sakai, T. (2021). Evaluating evaluation measures for ordinal classification and ordinal quantification. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 2759–2769).

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schaier, A. H., & Cicirelli, V. G. (1976). Age differences in humor comprehension and appreciation in old age. *Journal of Gerontology*, *31*(5), 577–582.

Sharma, M., Kandasamy, I., & B, V. W. (2021). YoungSheldon at SemEval-2021 Task 7: Fine-tuning Is All You Need. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing.*

Shi, X., Cao, W., & Raschka, S. (2023). Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 1–15.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the Targets of Hate in Online Social Media. In *Proceedings of the international aaai conference on web and social media* (Vol. 10).

Sjöbergh, J., & Araki, K. (2007). Recognizing humor without recognizing meaning. In *Applications of fuzzy sets theory: 7th international workshop on fuzzy logic and applications, wilf 2007, camogli, italy, july 7-10, 2007. proceedings 7* (pp. 469–476).

Smădu, R.-A., Cercel, D.-C., & Dascalu, M. (2021). UPB at SemEval-2021 Task 7: Adversarial Multi-Task Learning for Detecting and Rating Humour and Offence. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing.*

Song, B., Pan, C., Wang, S., & Luo, Z. (2021). DeepBlueAI at SemEval-2021 Task 7: Detecting and Rating Humor and Offense with Stacking Diverse Language Model-Based Methods. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing.*

Spearman, C. (1904). American journal of psychology 15. *The Proof and Measurement of Association Between two Things*(1), 72–101.

Stanley, J. T., Lohani, M., & Isaacowitz, D. M. (2014). Age-related differences in judgments of inappropriate behavior are related to humor style preferences. *Psychology and aging*, *29*(3), 528.

Suárez, J. L., García, S., & Herrera, F. (2021). Ordinal regression with explainable distance metric learning based on ordered sequences. *Machine Learning*, *110*(10), 2729–2762.

Sue, C. A., & Golash-Boza, T. (2013). 'It Was Only a Joke': How Racial Humour Fuels Colour-Blind Ideologies in Mexico and Peru. *Ethnic and Racial Studies*, *36*(10), 1582–1598.

Sultana, A., Ayman, N., & Chy, A. N. (2021, August). CSECU-DSG at SemEval-2021 task 7: Detecting and rating humor and offense employing transformers. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)* (pp. 1204–1208). Stroudsburg, PA, USA: Association for Computational Linguistics.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2020). Ernie 2.0: A Continual Pre-training Framework for Language Understanding. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 8968–8975).

Svebak, S., Martin, R. A., & Holmen, J. (2004). The prevalence of sense of humor in a large, unselected county population in norway: Relations with age, sex, and some health indicators.

Taylor, J. M., & Mazlack, L. J. (2004). Computationally Recognizing Wordplay in Jokes. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 26).

Tsai, P.-H., Chen, H.-C., Hung, Y.-C., Chang, J.-H., & Huang, S.-Y. (2021). What type of humor style do older adults tend to prefer? a comparative study of humor style tendencies among individuals of different ages and genders. *Current Psychology*, 1–12.

Vallat, R. (2018). Pingouin: statistics in python. *Journal of Open Source Software*, *3*(31), 1026.

van den Beukel, S., & Aroyo, L. (2018). Homonym detection for humor recognition in short text. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 286–291).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Veatch, T. C. (1998). A theory of humor.

Venkata Raju, K., & Sridhar, M. (2020). Based sentiment prediction of rating using natural language processing sentence-level sentiment analysis with bag-of-words approach. In *First international conference on sustainable technologies for computational intelligence: Proceedings of ictsci 2019* (pp. 807–821).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. biometrics bulletin 1, 6 (1945), 80–83. *URL http://www. jstor. org/stable/3001968*.

Winters, T., & Delobelle, P. (2020). *Dutch humor detection by generating negative examples.*

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... others (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.

Zhang, R., & Liu, N. (2014). Recognizing Humor on Twitter. In *Proceedings of the 23rd acm international conference on conference on information and knowledge management* (pp. 889–898).

Zhou, M., Ma, J., Yang, H., Jiang, L., & Mo, Y. (2021). Sequential attention module for natural language processing. *arXiv preprint arXiv:2109.03009*.

Zhu, H., Shan, H., Zhang, Y., Che, L., Xu, X., Zhang, J., ... Wang, F.-Y. (2021). Convolutional ordinal regression forest for image ordinal estimation. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(8), 4084–4095.