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
{jameaney, steven.wilson}@ed.ac.uk
{alopez, wmagdy}@inf.ed.ac.uk
luischir@fing.edu.uy

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.

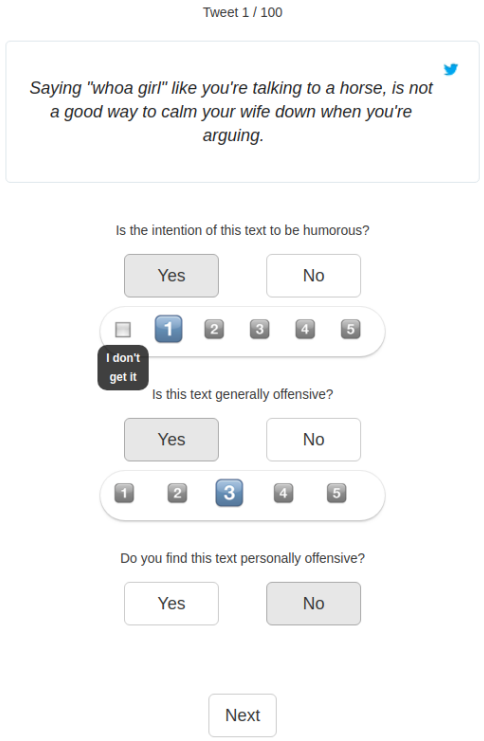

Figure 1: Screenshot from the tool used to annotate the texts.

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

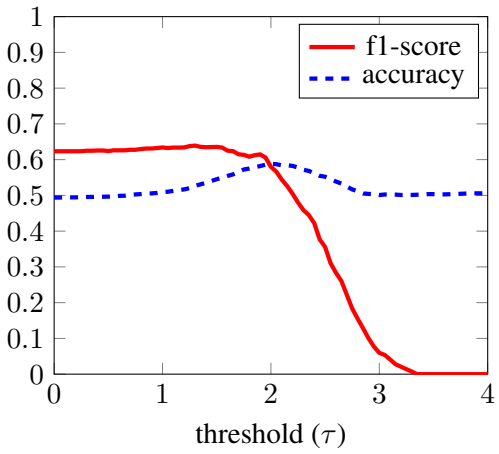

Figure 2: For varied values of a threshold, τ, accuracy and f1-score achieved by a hypothetical model predicting the label *controversial* for all texts in the test set with ground-truth humor score $> \tau$. Note that participants did not have access to these ground-truth scores for the test set, making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