HaHackathon: SemEval-2021 Task 7

Eli Bales, Pangbo Ban, Avani Pai, Hilly Steinmetz

Task Description

- HaHackathon is a shared task from SemEval 2021, aptly named for its focus on humor detection, humor rating, and controversy detection tasks.
- Our group is focusing on the binary humor classification task as our main task, and controversy detection as our adaptation.
- We chose this task as humor presents a unique challenge to NLP. Jokes often require extrinsic knowledge to understand their meaning.
    - Additionally, humor can often be used to thinly veil hate speech or bigoted language, therefore improving humor detection could contribute towards content moderation.
        - HaHackathon was the first shared task to combine humor and offense detection, to tackle the above idea of users using humor as a “mask” for hate speech.

Task Data

- The data for this task was collected from Twitter and the Kaggle Short Jokes dataset, with about a 60/40 split between jokes and non-jokes.
- For humor rating, offense rating, and controversy, the data was annotated by 20 annotators ages 18-70.
  - is_humor
  - humor_rating
  - humor_controversy
  - Offense_rating

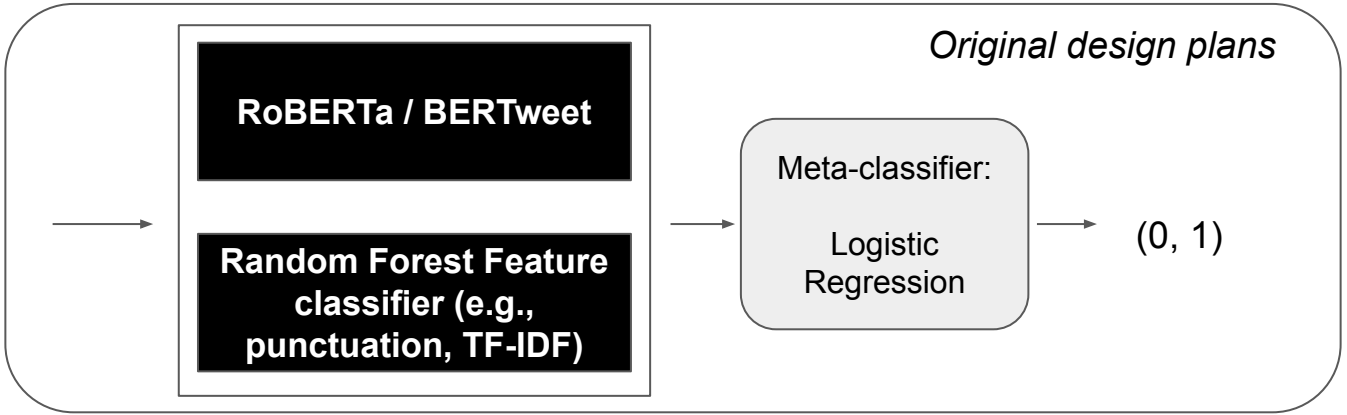
- **Is_humor** was decided by if the author intended the tweet/sentence to be funny. This removed discrepancies between what collectors thought was or was not humorous.
- **Ratings**, with a range of 0-4, were determined by the annotators, with the average being taken as the final result.
- **Controversy** was decided through variance in annotator ratings. If two different annotators gave a humor rating with a difference of over 1.79, then the joke was classified as “controversial.”

Task Data Examples

Text	Is humor	Humor rating	Humor controversy	Offense rating
The movie ‘Napoleon Dynamite’ only had a budget of $400,000. Jon Heder was initially paid $1,000 for his role as Napoleon.	0			0
“Whoever finds a friend, finds a treasure” - Cars	0			0
I won the "Most Secretive Guy" award in our office today. I can't tell you how much this award means to me.	1	2.2	0	0
What do you call bad breath that sneaks up on you? Ninjavitis	1	2.45	1	0
What did the Mexican say to the Italian? Que pasta?	1	2.32	0	0.85
In 2013, scientists implanted human brain cells in mice. The mice were 'statistically and substantially smarter than control mice.' They then created mouse-human hybrids by injecting baby mice...	0			0.4

Goals

1. Fine-tune RoBERTa for our classification task
2. Investigate whether fine-tuning BERTweet would improve performance
3. Create an stacked model that used a fine-tuned RoBERTa model and a classifier of features and information relevant to humor



System Architecture

- Fine-tuned a pre-trained RoBERTa and BERTweet models to classify sentences
- Sentences are inputted as raw text, tokenized
- A classifier head is trained on the sentences outputs

Our Approach

- With only 8,000 total labelled texts, we felt it necessary to use a pre-trained model and then fine-tune for our specific task. We split the data 80/10/10 for our training/dev/test set, so in total we had 6,400 documents for training.
  - We chose RoBERTa for its robustness, and BERTweet because much of the data comes from Twitter; other groups used RoBERTa, BERT, ERNIE 2.0, and ALBERT
- We minimally preprocessed the data and trained the base RoBERTa and BERTweet models for one epoch to not over-tune.
- Currently, we are working on implementing lexical features and ensembling with random forest to solve problems that arose in our error analysis.
  - Lexical features: NER, punctuation and letter counts, empathy ratings, hurtlex, TF-IDF
  - Feature classifiers

Our Approach

Khan et al., 2021

Table 3: Classification accuracy on lexical features for the Ubuntu dataset.

Classifier	Accuracy (%)
Random forest	95.4
SVC	73.0
Logistic regression	67.7
LinearSVC	67.1
MultinomialNB	66.4

Table 4: Classification accuracy on lexical features for the TripAdvisor (NYC) dataset.

Classifier	Accuracy (%)
Random forest	95.6
SVC	73.4
Logistic regression	62.7
LinearSVC	61.4
MultinomialNB	57

Results

	Baseline (linear)	RoBERTa	BERTweet	#10: Meizizi	#3: DeepBlueAI	#1: PALI
F1	0.8840	0.9372	0.9491	0.9653	0.9676	0.9854
Accuracy	0.8570	0.9237	0.9375	0.9570	0.9600	0.9820

- The top two submissions used both RoBERTa-large and ERNIE 2.0, but did not submit any further information on their system.
- DeepBlueAI used stack transformer models that utilized majority vote (classification) or average prediction (regression). They then used pseudo-labeling to generate labels for the test set and input that data in their training, as well as adding perturbations in the embedding layer to increase the generalization power of their model.
- Our models saw a significant jump over the baseline, but still need another 1-2% to reach the top 10 submissions.

Error Analysis

- BERTweet and RoBERTa did poorly on:
  - Sarcastic humor
  - Tongue-in-cheek non-humor
  - Less offensive humor

System	Humor	Offense
RoBERTa misclassified	2.00	0.66
BERTweet misclassified	1.98	0.61
Dataset	2.24	1.02

sentence	predicted	correct_label
Years from now, historians will look back on this period of American History and move to Canada.	not humor	humor
In the new James Bond movie, Bond apologizes to women for his behavior and is never seen again.	not humor	humor
I don't care how many times I see it, I will NEVER comprehend the fact that people have to use GoFundMe for medical bills in this country.	humor	not humor
Say it with me: The USPS is literally written into the Constitution.	humor	not humor

Error Analysis

- BERTweet and RoBERTa also have a difficult time with certain types of humor:
  - World-knowledge
  - “Linguistic” humor

sentence	predicted	correct_label
It costs $6 to visit the grave of Karl Marx.	not humor	humor
In 2018, a Missouri deer poacher was ordered to watch "Bambi" once a month for the entirety of his year-long prison sentence.	not humor	humor
There is a fine line between love and iove.	not humor	humor
Our attention spans these days are	not humor	humor

Successes

- Fine-tuning our model with just one epoch had highly accurate predictions
- Figured out how to use the Transformers and PyTorch packages
  - And use them together!
- Adapted existing RoBERTa model code to BERTweet
- Learned to design a late-fusion, stacked ensemble model on top of our fine-tuned BERT models
- Discovered that lexical features and other resources, on their own, do a decent job at the classification problem (~79% accuracy with SVM)

Issues

- Couldn’t implement our intended design in time
- PyTorch API was difficult to learn
  - Took a while to discover that we implemented the wrong loss function for the data
- Ensemble model size grew too large (nearly 80GB)
  - Decrease batch size and unnecessary padding, decrease random forest size, faulty environment?
- Working with different data structures on the transformers, scikit-learn, and PyTorch packages made it to difficult to implement this design in time
  - Difficult to find the right documentation for changing these data structures
- Ensemble designs required a lot of code refactoring
- Putting together our individual components led to new bugs

Related Readings

Ted Cohen. 1999. Jokes: philosophical thoughts on joking matters. University of Chicago Press, Chicago

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 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1125–1129, Online. Association for Computational Linguistics.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 105–119, Online. Association for Computational Linguistics.

Julia M. Taylor. 2014. Linguistic theories of humor. In Salvatore Attardo, editor, Encyclopedia of Humor Studies, volume 2, pages 455–457. SAGE Reference, Los Angeles, CA. Topic overview.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962v2.

Atif Khan, Muhammad Adnan Gul, Abdullah Alharbi, M. Irfan Uddin, Shaukat Ali, Bader Alouffi, "Impact of Lexical Features on Answer Detection Model in Discussion Forums", *Complexity*, vol. 2021, Article ID 2893257, 8 pages, 2021. https://doi.org/10.1155/2021/2893257

System Architecture (original design)