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 focused on the **binary humor classification** task as our main task, and **controversy detection** as our adaptation.
  - Humor classification was decided by the author of the tweet/joke
  - Controversy was automatically decided by the average difference of the ratings the annotators gave each joke. If the difference between two ratings was greater than a certain threshold, then the joke was deemed controversial.
- Motivated to tackle controversy detection as it has many useful downstream applications in moderation, and the task seemed inherently challenging as the top scores were about 0.49 accuracy and 0.63 F1.

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

From D3 -> D4 we wanted to:

1. Incorporate linguistically motivated features.
    a. By doing so, we hoped to tackle problems identified in our error analysis for D3, such as lack of world knowledge.
    b. Such features include Hurtlex, Punctuation Count, NER, etc.
2. Experiment with ensemble architectures to improve performance.
    a. Many of the highest ranking systems for our shared task used ensembling to some capacity, often using majority voting from numerous models.

Our Revised Approach

- Lexical features!
  - NER, punctuation and letter counts, Empathy ratings (Fast et al 2016), Hurtlex (Bassignana et al 2018), TF-IDF
  - Feature classifiers: Random Forest, NN classifier
- Feature dimension reduction!
  - Mutual information
  - Principal component analysis
- Training methods!
  - K-fold training of two classifiers and a meta-classifier
  - Late fusion of two classifiers

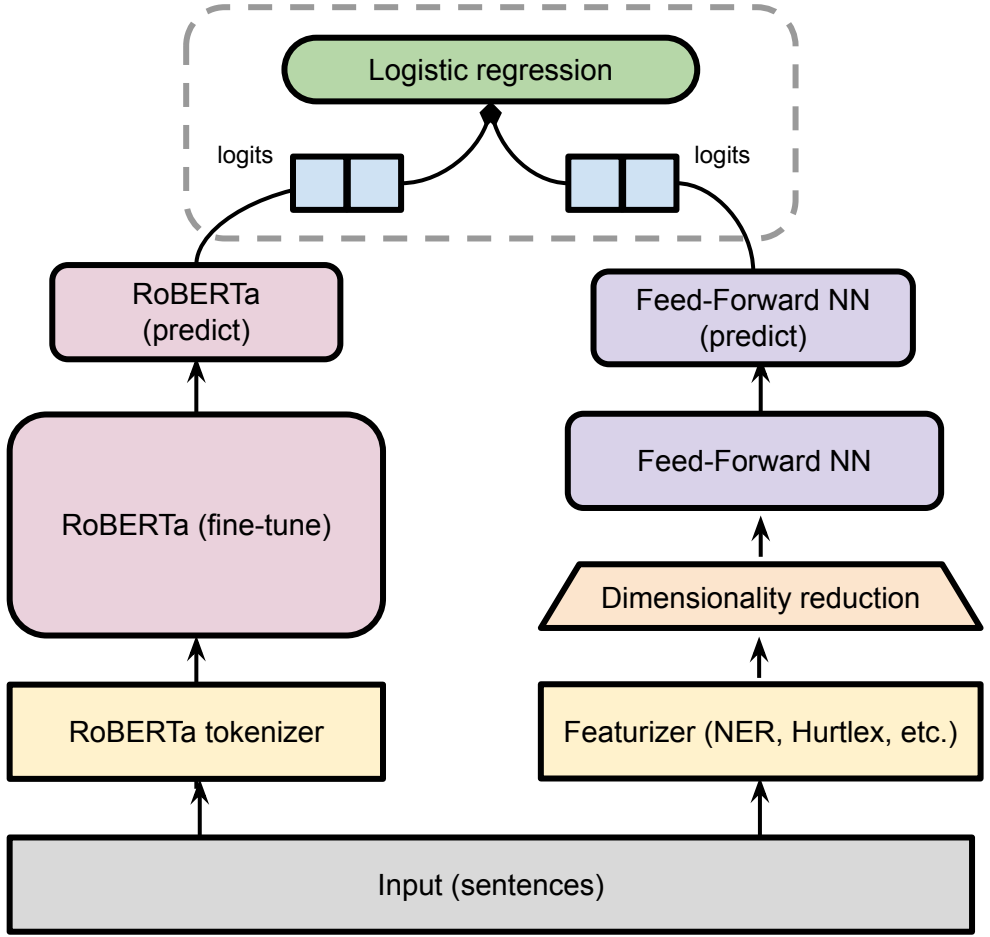
D3 System Architecture

- Fine-tuned 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

Final D4 System Architecture

After several experiments, decided on the following neural network ensemble architecture

- RoBERTa fine-tuned on adaptation and primary tasks
- Trained in k-folds splits to avoid overfitting
- A feed-forward neural network with 3 hidden layers
- Logistic regression with trained on logit outputs



Results - Primary Task

	Dev Set		Eval Set	
	D3 (RoBERTa)	D4	D3 (RoBERTa)	D4
F1	0.9372	0.951417	n/a	0.962739
Accuracy	0.924	0.94	n/a	0.953807

- On the dev set, we noticed a slight increase in accuracy and F1 between D3 and D4. The additional features we added for our adaptation also helped with our primary task.
- When we finally tested on the evalset, we were surprised to find our model did even better than our test set. Our numbers for the eval set put us about 0.002 off from the top 10 leaderboard for this shared task.

Results - Adaptation Task

	Dev Set		Eval Set	
	BERT Baseline (Meaney et. al)	D4	BERT Baseline (Meaney et. al)	D4
F1	n/a	0.597359	0.6232	0.605442
Accuracy	n/a	0.505071	0.4731	0.530364

- Our model's accuracy is in the top 10% compared to the shared task leaderboard, but our F1 doesn't break the top 10. Additionally, the BERT baseline data provided by the creators of the shared task places themselves at rank 11 for F1.
- It should be noted that our eval set and the paper's test set were different: we used 10% of our training data

Error Analysis - Controversy

- Seems to perform better at detecting puns
- Not great at detecting whether offensive will result in a controversial rating

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

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

sentence	predicted	correct_label
Dad: What's a lion and a witch doing in your wardrobe Me: it's Narnia Business	controversial	controversial
Shark Tank idea: a microwave that will self-destruct if someone tries to use it to cook fish.	not controversial	controversial
If pronouncing my B's as V's makes me sound Russian Then soviet.	controversial	not controversial
I can't believe I forgot to go to the gym today. That's 7 years in a row now.	not controversial	not controversial

Successes

- Fully incorporated various lexical features and feature dimension reduction techniques
- Implemented several ensembles and got to see which fared the best
- Achieved gains in primary task via neural ensemble architecture revisions (and other architectures)
- Relatively good performance on adaptation task

Issues

- A Git merge of two feature branches caused a lot of last minute issues
  - Not fun!
- “Controversial” humor encompassed a wide range of categories: including corny, offensive, or not clever
- Work got more and more specialized, making it difficult to work off of each others’ code
- Difficult to keep the repos/environments organized

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 Eval- uation (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

Fast, E., Chen, B., & Bernstein, M. S. (2016, May). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4647-4657).

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018* (Vol. 2253, pp. 1-6). CEUR-WS.

Additional slides (if extra time)

Error Analysis

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

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