Education Applications of Sarcasm Detection NLP Models in EFL Context

Humans vs Models

Oliver Cakebread-Andrews, August 17th, 2023
cakebread@kwansei.ac.jp

1. Background
2. Aims/Research Questions
3. Methodology
4. Results
5. Implications
6. Future Research

1. Background

Sarcasm Detection

Why it poses a challenge

- NLP models good at detecting positive/negative mood/valency
- Figurative language, specifically sarcasm/irony presents a problem (Bharti et al., 2016)
- Sarcasm is often opposite of what is written (Bharti et al., 2016)
- Usually requires contextual clues - body language, intonation
- Even for native speakers, text-based can be hard (Wallace, 2014) - Reddit uses /s and Twitter #sarcasm
- In this research, no distinction between verbal irony and sarcasm (Ghosh et al., 2020)

Background

Data Collection

- Many studies using social media:
  - Reddit (Ghosh et al., 2020; Mishra, Kaushik, & Dey, 2020)
  - Twitter (Avvaru et al., 2020; Ghosh et al., 2020)
- Majority of cases either:
  - use an existing corpus of sarcastic utterances (Khodak et al., 2018)
  - or more commonly (Eke et al., 2019), they gather their own unique data:
    - 4000 Tweets (Ghosh et al., 2020); 5000 Tweets (Jaiswal, 2020); 39000 Tweets (Ghosh & Veale, 2016); 900 Tweets (González-Ibánez et al., 2011))

Background

Languages

- Majority of studies into sarcasm detection continue to be based on English
- Some examples of other languages are:
  - Dutch (Kunneman et al., 2015); Spanish (Frenda & Patti, 2019; Ortega-Bueno et al., 2019); Romanian (Buzea et al., 2020); Arabic (Ranasinghe et al., 2019); Hindi (Jain et al., 2020)
- Very few studies using transfer learning with sarcasm detection:
  - (Chronopoulou et al., 2019; Tavan et al., 2020, Tada et al., 2022)

2. Aims

Aims/Research Questions

What was the purpose?

Primary Questions/Aims

1. What are the similarities and differences in sarcasm detection between NLP models and NNS of English?

2. How can these similarities and differences be applied to EFL education?

Secondary Questions/Aims

1. What improvements to EFL education and NLP models can be found from false positive and negative error analysis?

3. Methodology

Dataset

FigLang2020

- FigLang is a bi-yearly conference focusing on using machine learning and NLP models with figurative language datasets
- 2020 is the most recent sarcasm dataset
- Pre-labelled and large (training) - only used 300 examples (testing)
- Reddit and Twitter datasets - only used Reddit
- Includes context, but I didn’t

Participants

Selection and Variation

- In total, 39 participants
- 75% Japanese
- 33.3% “Fluent”, “41%” reasonable
- 60% hadn’t heard of Reddit
- Age range 18-48, average 29

BERT

Briefly

- Trained on 110 million parameters
- Uses Transformer
  - learns contextual relations between words
  - Bidirectional
- Makes use of Masked Language Modelling and
- Next Sentence Prediction
- Then Fine-tuning is used for specific tasks

NLP Models

What was used

- RoBERTa (Liu et al., 2019)
  - Based on a BERT model (Devlin et al., 2019)
  - RoBERTa made improvements (dynamic masking, removing NSP loss, large mini batches etc)
- Logistic Regression
  - Specifically useful in binary classification (0/1, Yes/No)
  - Uses sigmoid function
- DeBERTa (He et al., 2020)
  - Improved version of RoBERTa
  - Half the training data, disentangled attention and enhanced masked decoder

Analysis

Methods and Tools

- Term Frequency-Inverse Document Frequency (TF-IDF)
  - Normalized count where word count is divided by the number of documents it appears in
- Antconc (Anthony, 2022)
  - Concordancing software
  - Measure keyness of words in target vs reference corpora

4. Results

Results

Overall Statistics

	RoBERTa	LR/TF-IDF	DeBERTa
Accuracy	0.69	0.56	0.72
Precision	0.66	0.57	0.70
Recall	0.79	0.56	0.69
F1	0.72	0.56	0.71

Computed Results from Models

(Mean)	RoBERTa	DeBERTa	LR/TF-IDF	NNS
Overall	0.5	0.5	0.56	0.53
Sarcastic (W/o context)	0.6	0.47	0.52	0.43
Sarcastic (With context)	0.6	0.48	0.51	0.45
Not Sarcastic (W/o context)	0.38	0.58	0.63	0.65
Not Sarcastic (With context)	0.43	0.49	0.59	0.56

Average Scores - 0 is mistake, 1 is correct

Common Errors - Humans

Checking by Keyness

False Positive	Keyness (Likelihood)	False Negative	Keyness (Likelihood)
McCabe	28.6	Offender	24.4
Twitter	22.3	**IKEA**	23.2
Feminist	20.7	**Fund**	18.1
See	19.1	**Avocados**	14.6
Brazil	14.8	**Business**	14.2

Common Errors - Humans

Summary and Example Sentences

- 100% incorrect - 20/26 Sarc, 6/26 Not Sarc
- Most of Not Sarc mistakes included swear words
- Some possibly mislabeled - *“He’s too busy selling off all his slaves to pay the debts on his lavish lifestyle to notice.”*
- No context often just looked like questions/statements - *“Wasn’t his post deleted and his account banned?!”*/*“Nah stay away from Oregon, that place is terrible”*
- Also, more questions on average than 100% correct list (10/26 vs 4/41)

Common Errors - Models

Checking by Keyness

False Positive	Keyness (Likelihood)	False Negative	Keyness (Likelihood)
See	21.8	She	7.9
Run	16.5	**Business**	7.4
Twitter	14.7	Work	5.4
Weather	13.6	**Ikea**	5.2
Feminist	12.8	Trees	5.2

Common Errors - Models

Example Sentences

	Sarcastic	Non-sarcastic	Total
All models incorrect	0	31	31
One or two models	118	119	237
All models correct	32	-	32

- Exclamation marks often appeared (Sarc) - *“god damn billionaires!! it's not always about you!” - “That was intended so they can show their stealth drone and claim "no one can see it!!"”*
- Again, mislabelling a potential problem - *“Or you can upgrade to the deluxe package for $70 to also receive polio and a life long warranty of a **cool metal box**!”*

TD-IDF

Logistic Regression Top Words

Token	TF-IDF Score	Token	TF-IDF Score
Number	1	Mean	1
Man	1	Marijuana	0.89
Far	1	Island	0.86
Approval	1	Lol	0.85

Similarities and Differences

NLP vs NNS

- *“But what if it’s my birthday today” (Sarc)*
- *“He’s too busy selling off all his slaves to pay the debts on his lavish lifestyle to notice.” (Not Sarc)*
- Exclamation marks, and to some extent questions marks, can confuse both, but more so models
- Models are much more evenly spread than humans
- Models are more influenced by spelling mistakes
- Mislabeling also a problem for both

Summary of Results

Key takeaways

- Models and NNS have some similar areas they struggle and succeed with
- NNS tended to be better at determining when something definitely wasn’t sarcastic, followed by logistic regression model
- In fact, logistic regression tended to be closer to NNS in predictions, DeBERTa second
- Generally political topics appear in false positives
- Generally “normal” topics appear in false negatives
- RoBERTa generally outclassed NNS

5. Implications

How is this useful?

Potential application of the results in the classroom

- Sarcasm itself - limited use within classroom (sarcasm lesson??)
  - However, pattern grammar and patterns of figurative language
  - Particularly useful for advanced writing/reading classes
  - FigLang looks at many areas - 2022 was euphemisms
- Targeted error analysis (making use of big data and machine learning)
  - Writing corpora? (Collected in Kwansei Gakuin University)
- Fine-tuning chat bots
- Running texts through these models to highlight likely areas of difficulty

6. Future Research

Next Focus

Memes? Why not Japanese?

- Memes - next step in multi-modal analysis
  - VisualBERT
  - UNITER
  - TxtBERT with ImgBERT
- Japanese sarcasm - lacking resources and cultural differences
  - No corpus of Japanese data with sarcasm tags
  - Japanese Twitter/Reddit doesn’t have anything like #sarcasm or /s
- Making clear lesson plans to use the results of such research
- Making an easy to access and use version so teachers can do it themselves

Thank you for listening

Do you have any questions?

References

- Anthony, L. (2022). AntConc (Version 4.1.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software
- Bharti, S. K., Vachha, B., Pradhan, R. K., Babu, K. S., & Jena, S. K. (2016). Sarcastic sentiment detection in tweets streamed in real time: a big data approach. *Digital Communications and Networks*, *2*(3), 108-121.
- 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*.
- Ghosh, D., Vajpayee, A., & Muresan, S. (2020). A report on the 2020 sarcasm detection shared task. *arXiv preprint arXiv:2005.05814*.
- IBM. (2020, July 2). *Natural Language Processing (NLP).* IBM. https://www.ibm.com/cloud/learn/natural-language-processing
- Lee, H., Yu, Y., & Kim, G. (2020). Augmenting data for sarcasm detection with unlabeled conversation context. *arXiv preprint arXiv:2006.06259*.
- 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*.

If you would like to help out (and test your own sarcasm detection abilities!) please use this QR code:

If you would like to share your opinions on using corpora in the language classroom, then please use this QR code: