Error Analysis of NLP Models and Non-Native Speakers of English Identifying Sarcasm in Reddit Comments

Oliver Cakebread-Andrews, Le An Ha, Ingo Frommholz, Burcu Can
University of Wolverhampton, University of Stirling
O.P.Cakebread-Andrews@wlv.ac.uk, halean@yahoo.com, ifrommholz@acm.org, burcu.can@stir.ac.uk

Abstract

This paper summarises the differences and similarities found between humans and three natural language processing models when attempting to identify whether English online comments are sarcastic or not. Three models were used to analyse 300 comments from the FigLang 2020 Reddit Dataset, with and without context. The same 300 comments were also given to 39 non-native speakers of English and the results were compared. The aim was to find whether there were any results that could be applied to English as a Foreign Language (EFL) teaching. The results showed that there were similarities between the models and non-native speakers, in particular the logistic regression model. They also highlighted weaknesses with both non-native speakers and the models in detecting sarcasm when the comments included political topics or were phrased as questions. This has potential implications for how the EFL teaching industry could implement the results of error analysis of NLP models in teaching practices.

Keywords: Sentiment Analysis, Sarcasm Detection, TEFL

1. Introduction

Accurate detection of sarcasm can be essential for a variety of tasks, such as sentiment analysis, preventing cyberbullying and deciphering the legitimacy of intent with online reviews (Maynard and Greenwood, 2014; Rosenthal et al., 2014; Hee et al., 2018). However, despite its prolific use, sarcasm detection continues to be an issue for both humans and natural language processing (NLP) models alike (Wallace et al., 2014; Singh and Sharma, 2023). As sarcasm can be difficult even for native speakers of the language (Abercrombie and Hovy, 2016), unsurprisingly non-native speakers struggle even more to understand whether an utterance is sarcastic or not (Peters et al., 2016). This study aims to identify similarities and differences between three NLP models on the one hand and non-native speakers (NNS) of English on the other. As sarcasm is challenging to detect for both groups (Wallace et al., 2014), we theorise that by analysing the similarities, differences and errors that occurred between the two groups, ways to both improve model training and English as a Foreign Language (EFL) education of NNS could be highlighted. More specifically, if there was any overlap between the two groups in terms of what features of sarcastic utterance lead to higher or lower rates of successful sarcasm detection, then these features could be used for explicit model training or English language instruction. NNS were chosen over native speakers due to the fact that they process sarcasm mainly with semantic context, rather than prosodic features as native speakers do (Peters et al., 2016; Abercrombie and Hovy, 2016). In addition, there have already been some studies with native speakers, such as Abercrombie and Hovy 2016 and Farha et al. 2022b.

Several interesting patterns and trends were discovered when comparing the results of the models and the NNS. This study has found that punctuation can contribute to errors in both the NLP models' and NNS's predictions, but exclamation marks had a bigger effect on the models and question marks on the NNS. Spelling errors on the other hand have a much larger effect on models' ability to correctly predict sarcasm, suggesting that NNS are still better at ignoring those kinds of issues. This is consistent with the challenges highlighted by Singh and Sharma (2023) with NLP models. The errors in the models tended to be more evenly spread among the categories as opposed to the NNS who excelled in some areas, but were much weaker in others.

This paper uses natural language processing models with the conversational context provided in the FigLang 2020 Reddit Dataset (Ghosh et al., 2020) to attempt to compare the ability of humans and NLP models in detecting sarcasm. The advantages of using a dataset that had been compiled and used in other studies are that it has pre-labelled data that has been shown to be effective at producing useful data (Chowdhury and Chaturvedi, 2021). As far as could be ascertained by the author, this is the first study to investigate the differences in abilities of NNS of English and NLP models at detecting sarcasm and draw conclusions on the application of the results of these comparisons. In this paper there is no distinction made between verbal irony and sarcasm, as can be found in similar studies (Ghosh et al., 2020; Lee et al., 2020; Eke et al.,

LREC-COLING 2024, pages 6247–6256
20-25 May, 2024.

2021). Our research tries to answer the following research questions:

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

2. Related Work

2.1. Sarcasm Detection

Earlier attempts at sarcasm detection tended to use rule-based methods using linguistic or semantic features such as hyperbole (Bharti et al., 2015, 2018), or punctuation marks and interjections (Tsur et al., 2010). While they laid down important groundwork, they were often resource intensive and had low levels of accuracy. However, even recent works will often use models that focus on just sentence- or utterance-level features, despite the fact that humans also require context to understand whether something is sarcastic or not.

Zeng et al. (2019) used a Local Context Focus (LCF) mechanism for a sentiment classification task. Rather than looking at the overall "global" context, which they suggest has a negative effect on the accuracy of prediction, they focus on context words closer to the aspect. Their study focused on positive or negative sentiment rather than sarcasm, but highlights the relevance of both local and global context, that is understanding the comment within the context of the conversation or comment thread, as well as any relevant cultural context.

Chowdhury and Chaturvedi (2021), using the same FigLang 2020 dataset, implemented "commonsense" knowledge in order to improve the accuracy of sarcasm detection. In contrast to Zeng et al. (2019), they used more general contextual knowledge. However, rather than limiting it to the global context of the document, they expanded it to "general beliefs and world knowledge". In addition to the interesting direction of adding common sense knowledge as parameters into the models, this study also demonstrates the legitimacy of the use of the FigLang 2020 dataset in other studies.

Ghosh et al. (2018) conducted a highly in-depth study using contextual "turns" both before as well as after the comment. They used Reddit and Twitter data that had been tagged by the author, but used crowdsourcing to identify sarcasm in the Internet Argument Corpus. They discovered post-comment context was not useful for improving results, however sentence-level attention using pre-comment context in multiple-level LSTM achieved significant improvement. They also found that 41% of the time, the attention of the LSTM model was focused on the same part of the contextual sentence as the participants in the crowdsourcing in order to detect the trigger for sarcasm. On the other hand, there was less consistency when deciding which part of the sarcastic comment expressed the author's sarcastic intent, indicating this is a difficult task for humans as well as computers.

More recently, the SemEval 2022 Task 6 (Farha et al., 2022a) asked authors to create models for predicting both English and Arabic sarcasm. In contrast to the FigLang dataset, where sarcasm was assumed based on tags by the comment authors, the SemEval 2022 dataset was collecting by directly asking the authors to provide sarcastic statements. One point of note is that the top ranking submissions (Yuan et al., 2022) used RoBERTa and DeBERTa-based models, similar to this study and the top ranking submission for FigLang 2018. Finally, the second subtask was to assign the most appropriate category to each text (such as irony, sarcasm, satire etc.). All teams performed poorly in this subtask, which seems to lend further weight to the idea that these categories are not so clearly defined and should be treated the same as in this and similar studies (Ghosh et al., 2020; Lee et al., 2020; Eke et al., 2021).

2.2. Sarcasm Datasets

There have been many attempts at creating corpora of sarcastic data (Tsur et al., 2010; Bamman and Smith, 2015; Khodak et al., 2019; Ghosh et al., 2020). There is a mixture between researcher or third-party tagged data, i.e. sarcastic perception by someone other than the original creator of the utterance (Abercrombie and Hovy, 2016; Ghosh et al., 2018), and corpora that have been tagged according to the original creator i.e. sarcastic intention (Khodak et al., 2019; Ghosh et al., 2020).

Abercrombie and Hovy (2016) chose to have researchers annotate the data using the backgrounds and histories of the Twitter users who wrote the comments. However, despite the intense effort involved in this for the researchers, and the subsequent 60 native speakers who did the rating, they found that rater agreement was not particularly high. They concluded that anyone involved with assigning the tag of sarcasm or not requires some level of understanding of the "common ground" shared with the participants. This echoes Bamman and Smith (2015) who demonstrated that lower levels of familiarity with audiences and viewers of Tweets had a correlation with higher usage of explicit sarcasm tags.

On the other hand, there are clear benefits when it comes to the speed of data collection when using author tags to collect sarcastic data like with Ghosh

et al. (2020) or Farha et al. (2022a). However, earlier studies such as Tsur et al. (2010) and Bamman and Smith (2015) demonstrated that there can be biases when using this explicit markers to collect the data. Bamman and Smith (2015) found that while author tagged data led to higher levels of accuracy, authors tended to explicitly tag Tweets as sarcastic when there was less familiarity and mutual understanding with their audience. Therefore the types of Tweets that were collected by utilizing the tags tended to be to ones aimed at a more general audience and therefore more obvious sarcasm.
The general trend in recent studies has been towards the latter, and indeed the data in this study also makes use of author intention (Ghosh et al., 2020). Farha et al. (2022b) also specifically recommend not using third party annotated data, especially when the intention is the most important metric.

2.3. Sarcasm Detection with Humans

First of all, studies such as Kreuz and Caucci (2007), Ghosh et al. (2018) and Abercrombie and Hovy (2016) looked at the ability of native speakers in understanding sarcasm. They provided a useful model for similar studies. All studies found that humans generally relied more on context than machines, but that labelling was significantly more important for the models. In addition, the studies demonstrated the benefit of not providing the participants with a definition of "sarcasm". Ghosh et al. (2018) also noted that the human participants tended to focus on the same position in the sentence as the models when deciding if an utterance was sarcastic or not, lending weight to the idea that it could potentially aid NNS.
Farha et al. (2022b) also recently investigated this area with the SemEval 2022 Task 6 and had similar conclusions. The studies highlighted the necessity of both including context in datasets, as well as creating models that took into account context for sarcasm detection. The FigLang 2020 dataset (Ghosh et al., 2020) does include context, and many of the participating teams used models that took context into consideration. Context in this case means the previous comments to which the target comment was replying, which was usually between three to five.

2.4. Sarcasm in EFL

In addition to the research into the ability of NLP models to detect sarcasm, as well as native speakers, there has been some, albeit limited, research into the ability of NNS to understand and learn how to identify sarcasm in EFL education.
Kim and Lantolf (2018) demonstrated reasonable success in teaching understanding of sarcasm in nine Korean university students. They used American TV shows with some element of sarcasm as in their classes along with additional contextual information that aided in the students' detection of sarcasm. This demonstrated that with explicit instruction and clear context, NNS could improve their sarcasm detection abilities. However, the kind of sarcasm that appears in TV shows is not completely natural and also includes a lot more visual contextual clues.
Shively et al. (2008) taught sarcasm to Spanish learners also by making use of movie scenes. One result of note was that language proficiency and accurate sarcasm detection in Spanish were correlated with a marginal statistical significance. It was therefore deemed prudent to include a question of English language ability in the data collection section of this study, as it was likely that an increased ability in English would result in a higher ability to detect sarcasm.
Finally, Prichard and Rucynski (2019) showed that the direct instruction of humour using satirical newspaper headlines resulted in statistically significant improvements in detection of sarcasm in English language. Additionally there was one unexpected result with the control group of native speakers who displayed a reduced ability to detect sarcasm with explicit instruction. This study does not deal with native speakers, but it highlights a potential future direction of study.
Most of the research does not consider the role or positions of NNS in sarcasm detection. Our contribution to the research is to demonstrate which areas of sarcasm detection cause issues for NNS and models. The results of this can be used to inform future educational materials for teaching English to NNS, and potential areas of improvement for models.

3. Methodology

3.1. Models

3.1.1. Preliminary Study

In order to establish a baseline level of accuracy, three different types of pretrained models and transformers were used. They were chosen as they had been used by teams in the FigLang 2020 (such as Dong et al. (2020)) conference and the SemEval 2022 Task 6 (such as Yuan et al. (2022)), or were closely related to them.

For the preliminary study, to test the viability of the research, a standard RoBERTa model (Liu et al., 2019), which has 12 layers, 768 hidden state size and 12 attention heads, with fine tuning, as well as a logistic regression model (LR) with term frequency-inverse document frequency (TF-IDF) vectorization were used on the utterances without context. RoBERTa implements optimised pre-training of BERT architecture to achieve higher rates of accomplishment in a variety of tasks.

Demographic	75% Japanese	25% Other NNS	
Age	18-48	Average: 29	
English Ability	Fluent: 33.3%	Reasonable: 41%	Weak or lower: 25.7%
Heard of Reddit?	No: 60%	Yes: 40%	

Table 1: Summary of participants (n=39)

It was preprocessed using Byte-Pair Encoding (BPE) as a tokenizer, and used the transformer library from HuggingFace with a cross entropy loss function. As for TF-IDF, TF is the proportion of a term occurrence in a document to the total occurrences of terms in a document, and IDF reflects the proportion of the number of documents a term appears in to the total number of documents. The resulting vector between 0 and 1 can provide insight into commonly found words within a certain dataset and was used to analyse the False Negative (FN)/False Positive (FP) results.

3.1.2. Main Study

After adjustments to the data and the data collection methods, stated below, a standard DeBERTa model (He et al., 2020) was also tested in addition to the two other previously mentioned models, to compare results and see if there were any improvements from the previous two models. DeBERTa is an improved version of RoBERTa that despite using half the training data, by using disentangled attention and an enhanced mask decoder results in a significant accuracy increase. Similar to RoBERTa, DeBERTa utilises 12 layers, 768 hidden state size and 12 attention heads.

3.2. Non-Native Speakers

A reasonably varied group of NNS participants took part in the study as summarised in Table 1. Participants were all volunteers recruited through social media. There was no compensation offered, which all participants knew and agreed to from the outset. Their English level was self-reported, which was appropriate for this study. Conducting a standardised test on all of the participants was outside the scope of this research. In addition, self-assessed language levels are commonly used in other studies (Reuland et al., 2009; Edele et al., 2015) and have been shown to be generally reasonably accurate (Ross, 1998; Diamond et al., 2014).

3.3. Data Collection

3.3.1. Pilot Study

The test data on NNS was collected through Google Forms. In the trial stage, after an initial briefing about the contents of the questionnaire, participants were shown 25 Reddit comments from a variety of subreddits such as politics, religion, sports, technology and had to decide whether they were “sarcastic” or “not sarcastic”. If they were unsure, they were prompted to choose “not sarcastic”, due to participants’ tendency of regularly choosing “I don’t know” when the option was available. Although this means there could be a bias with the results trending towards "not sarcastic", this was in line with similar studies such as Abercrombie and Hovy (2016) and Kreuz and Caucci (2007). The justification being that if participants are not perceiving it to be sarcastic, then they should choose the option that best reflected that i.e. "not sarcastic". They were then shown a further 25 comments, this time with the context shown to them. The 50 comments were the same as the models were tested on. The comments were the first 25 examples of both sarcastic and non-sarcastic comments from the FigLang 2020 test dataset. Examples of some of the comments can be seen in Figure 1 and Figure 2.

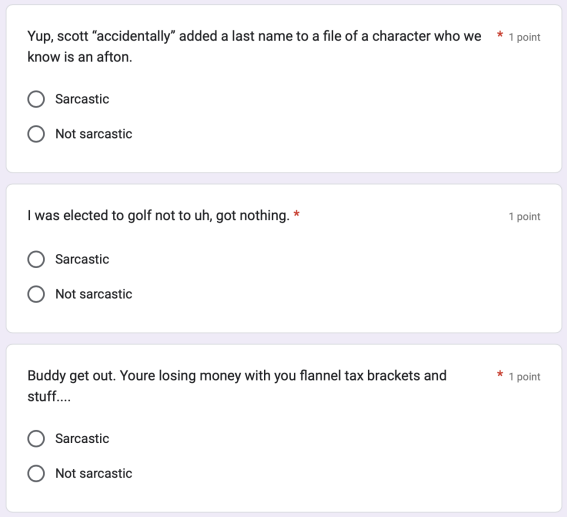

Figure 1: Example comment without context

3.3.2. Main Study Data Collection

Unfortunately, both native and NNS of English took around 30-40 minutes to complete the questionnaire, which was longer than expected. Additionally, by only conducting the study on 50 comments, it limited the scope of the experiment. Therefore the total number of comments was increased to 300, however each participant was only given 15 comments with, and without, context for a total of 30. There were 10 sets of 30 comments, which meant there were fewer participants per set of 30, but more data to work with. Once again, the 300 comments were the first 150 sarcastic and non-sarcastic comments from the test dataset.
All three of the models, RoBERTa, DeBERTa and

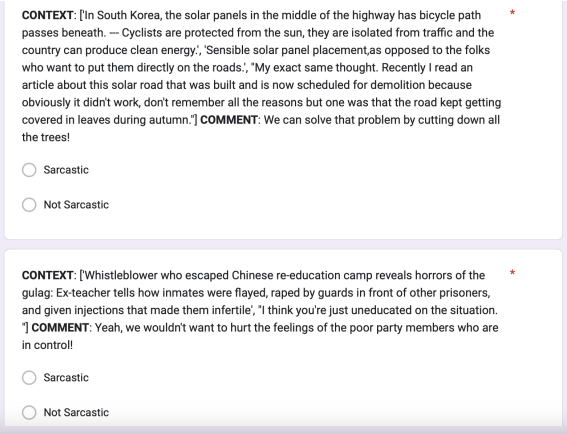

Figure 2: Example comment with context

	Acc.	Recall	Prec.	F1
RoBERTa	0.67	**0.83**	0.62	0.71
LR/TF-IDF	0.56	0.55	0.57	0.57
DeBERTa	**0.73**	0.74	**0.72**	**0.73**

Table 2: Results of second model runs on 300 comments

the logistic regression model, were trained on the full training set from FigLang 2020 of around 3500 comments, which had been labelled as sarcastic or not sarcastic - sarcastic labels were attached by the original authors of the comments. Initially, just the RoBERTa and logistic regression models were tested on the same 50 comments (25 sarcastic, 25 not sarcastic) presented to the preliminary NNS cohort. After the decision to change to 300 comments, all three of the models were run on the same 300 comments as the NNS.

4. Experiments

The initial trial run of 50 comments was run on just the RoBERTa and the logistic regression model. However, after the feedback from the initial test participants leading to an increase in the number of comments to 300, the models were re-tested with the new data and the results are shown in Table 2. The best scores have been bolded. For this stage, once again initially only RoBERTa and the logistic regression model were used, but later the DeBERTa model was also run. While not the purpose of this study, the results do show that DeBERTa outperformed the other models in sarcasm detection, despite being pre-trained on half the data as RoBERTa, as can be seen in Table 2, which is in line with similar studies at SemEval 2022 (Yuan et al., 2022).

In comparison, the top results from the original FigLang 2020 study were higher than the best ones in this study. The team called *"miroblog"* (Lee et al., 2020) used BERT + BiLSTM + NeXtVLAD + Context Ensemble + Data Augmentation in their approach and scored around 0.83 in all metrics. *"andy3223"* (Dong et al., 2020) used a RoBERTa-based model, which is closest to the one used in this study and scored around 0.75 in all metrics. These results are to be expected - the purpose of this study was not to outperform state-of-the-art models, and indeed the models used in FigLang 2020 were purpose-built for this data set, meaning they may have fewer general applications. The models used in this study were the basic versions using the original parameters found on repositories such as HuggingFace, and therefore can be replicated more easily. Further, the objective was to discover similarities with NNS, and therefore the accuracy was of secondary importance to analysis finding out crossover points with NNS. This also applies to the different levels of accuracy within the three models used in this study - while RoBERTa and DeBERTa outperformed the logistic regression model, it is still useful for the purpose of predicting NNS abilities.

5. Results

The results of the study showed that there were several areas where the models and the human participants differed from each other as well as areas in which they fared similarly well. Despite that, in the initial run, prior to the inclusion of the DeBERTa model, there were only two comments out of the 300 that had 100% agreement between the participants and both of the model. It is also of note that both comments were predicted incorrectly (correct answer in parentheses):

"But what if it's my birthday today" (Sarc - no context)
"He's too busy selling off all his slaves to pay the debts on his lavish lifestyle to notice." (Not Sarc - with context)

With the results of the DeBERTa model included however, there were no comments that had 100% agreement as the DeBERTa correctly identified both sentences.

After the models were run and the questionnaires were completed, the errors - both FP and FN were analysed using two methods. The first was by measuring the keyness of the FP/FN corpora against the corpus of the full comments, conducted through software called "AntConc" (Anthony, 2022). AntConc is concordancing software that generates a list of keywords ranked by *keyness*. The keyness k of a word is generated using a log-likelihood statistical measure - the higher the number, the higher the relative frequency of the word:

$$k = 2\left(O_t \ln\left(\frac{O_t}{E_t}\right) + O_r \ln\left(\frac{O_r}{E_r}\right)\right)$$

with O_t and O_r being the observed target word frequency in the target corpus and reference corpus, respectively, and E_t and E_r being the expected target word frequency in the target and reference corpus, respectively. The words that ranked highly in keyness were then analysed in their context to identify any common patterns.
The second method used the results from the TF-IDF vectorization of the logistic regression model, which also highlighted some of the most commonly appearing words in the FP and FN. These were also manually analysed for common themes or patterns, by comparing the results of the models and the NNS.

5.1. Non-Native Speakers

Analysing the results using AntConc provided several insights. First of all, the top 5 content words in both FP and FN ranked by keyness from the NNS is presented in Table 3. Content words in this context means that actually provide meaning to the sentence, and do not include preposition, articles etc such as "the" or "and".

FP	Keyness	FN	Keyness
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

Table 3: Keyness of FP/FN from NNS

Highlighted in bold are keywords that fit into certain patterns that could be found in the two categories. In the False Positive category, there were several keywords related to politics. Of the top five, three of them were politically-related when looking at them in context. For example, while "Brazil" on its own doesn't have any political implications, in the full context it was talking about a politicians actions in Brazil. On the other hand, within the False Negatives, there was a tendency for keywords and topics to be "ordinary". Again, as an example "IKEA" was inside of a comment chain that was simply talking about furniture. This suggests that, at least with the NNS in this study, politically-related comments are more likely to appear sarcastic, and everyday topics are more likely to appear not sarcastic.

There were 26 comments that were identified incorrectly by 100% of the NNS, 20 of which were sarcastic and 6 were non-sarcastic, suggesting that the NNS struggled more with identifying sarcasm correctly. While some could be attributed to mislabelled data, there were clearly some patterns among the errors. The comments that had little other context and looked like normal questions or statements caused problems for the NNS, such as:

"Wasn't his post deleted and his account banned?!"
"Nah stay away from Oregon, that place is terrible"

In fact, in general questions appeared to cause issues for the NNS - 10/26 of 100% incorrect comments had questions, against 4/41 of the 100% correct. Once again, it appears that at least for the NNS, the more "normal" a statement appears, the more likely it is that they will identify it incorrectly.

5.2. Models

As for the models, the results of the errors were also analysed using Antconc as seen in Table 4. The words that appeared high in the NNS's errors have been bolded - some were outside the top five, but still ranked within the top ten.

FP	Keyness	FN	Keyness
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

Table 4: Keyness of FP/FN from models

Table 3 and Table 4 demonstrate of the key findings, that there are similar patterns between the models and the humans in terms of the categories of words that were likely to appear. More specifically, **content words related to political topics were more likely to appear in the FP**, such as "feminist", and further down the top ten list, "Brazil". Although these words are not political by themselves, in the context of the comment where they were written, they became political. On the other hand, **non-political, normal topics such as "IKEA" again appeared in the FN**. While it is not clear *why* these categories were the ones that caused difficulties for the NNS and the models, there is a pattern that could begin to inform other research.

As for the distribution of errors between sarcastic and non-sarcastic, unlike the NNS, the models had an unexpectedly regular patterned spread as seen in Table 5. There was a very even split between the FP and FN. It clearly shows that on average these models were considerably better at determining what was sarcastic than

what was not sarcastic. This is possibly due to predictions being more heavily weighted because of certain variables within the model itself. Potentially using the results of the error analysis in this study, further models could have their parameters adjusted for increases in accuracy.

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

Table 5: Distribution of errors

Further analysis of the comments that came up in the errors found that exclamation marks appeared in the FN more regularly than the FP, but question marks did not appear to have any major trend either way. It also appeared that there were more likely to be spelling errors in the FN/FP of the models than the NNS. This is in keeping with similar studies (Tsur et al., 2010) that punctuation can affect the accuracy of models' predictions.

Finally, one other pattern of particular note was the tendency of the logistic regression model to be closer to the prediction capability of the NNS than the other models. One area that this is apparent is in the accuracy as seen in Table 6. The highest scores are in bold, and the closest scores to the NNS are in italics.
In three out of the five categories, the logistic regression model performed closer to the NNS, and DeBERTa in the remaining two, when compared with RoBERTa. In particular, the logistic regression model performed closest to the NNS when predicting something was not sarcastic, and indeed they had the high scores for this category too. DeBERTa performed closest to the NNS in the sarcasm categories, in this case DeBERTa and the NNS had the lowest scores. This suggests there could be certain implications for using NLP models to predict the results of NNS's language abilities, though it certainly needs further research to be able to confidently state how.

Comment group	RoBERTa	DeBERTa	LR	NNS
Overall	**0.69**	**0.69**	*0.56*	*0.53*
Sarc (w/o context)	0.74	**0.81**	*0.51*	*0.43*
Sarc (with context)	0.7	**0.75**	*0.51*	*0.45*
Not Sarc (w/o context)	**0.7**	0.58	*0.63*	*0.65*
Not Sarc (with context)	**0.63**	0.6	*0.59*	*0.56*

Table 6: Accuracy of models and NNS

6. Discussion

The first research question asked "What are the similarities and differences in sarcasm detection between NLP models and NNS of English?". The study has shown that with regards to general patterns of detection, models appeared to be better than the NNS at determining whether a comment was sarcastic or not sarcastic, with RoBERTa and DeBERTa being better than the logistic regression at that task. While all the models, and DeBERTa in particular, were better at concluding if a statement was definitely sarcastic, the NNS performed at a similar level to the models when deciding if something was *not* sarcastic. In addition, the logistic regression models tended to perform closer to NNS than DeBERTa and RoBERTA, which both generally performed better. Therefore it seems to be the case that the largest differences between models and NNS are in tasks involving if a comment is sarcastic, and the similarities are when the task is deciding if a comment is not sarcastic. In addition, the model most similar to the NNS is the logisitic regression model.
As for patterns within the error analysis, political topics were often in the FP errors and "normal" topics were more likely to be in the FN errors. This was true for both the NNS and the models. However it should be noted that one issue with the data was that it was self-labelled by the author but not mandatory for them to do so, i.e. they might think their comment is so obviously sarcastic, it did not need tagging. So it is entirely possible that some of these were correctly predicted, but did not match the label due to mis-labelling.
The second and third research questions asked "How can these similarities and differences be applied to language education?" and "What improvements to English education and NLP models can be found from false positive and negative error analysis?". From the results of this study, several suggestions can be made to improve the accuracy of sarcasm detection for models through parameters, and improve the English education of NNS:

Models

1. Adjustments for political topics - certain key words could be slightly negatively adjusted. Models have a tendency to over-tag political topics as sarcastic.

2. Data should also be checked for spelling errors (a hard task with larger datasets, but could have a big impact).

3. Punctuation - often this is removed from datasets anyway. However it seemed to help as often as it harmed, so it is hard to say whether this should be adjusted. Both models and NNS struggled with question marks, so that could be worth consideration.

NNS

1. Spelling errors are clearly not a major issue even for NNS. While these created issues for the models, the NNS didn't appear to be affected by them.

2. Models are clearly good predictors of what areas are difficult for NNS - this study highlighted political topics tending towards FP and normal topics tending towards FN for models and NNS. Model predictions could be useful aides in the language classroom.

3. Unlike spelling, punctuation, in particular question marks and excessive uses, seemed to sway NNS towards certain directions. This should be considered when teaching too.

In particular, the overlap between the logistic regression model and the NNS' predictions was an unexpected, and potentially very useful finding. Ghosh et al. (2018) noted that native speaker humans tended to look at the same areas of context as the models. However, it is difficult to suggest with any level of certainty why this model was the most similar. While all of the models and the NNS struggled and succeeded with similar types of comments (e.g. the political topics appearing more commonly in false positives), the logistic regression models had more similarities the mean scores of the different sections (e.g. non-sarcastic comments). It is likely they that the logistic regression model did look at the same context areas as the NNS as suggested by Ghosh et al. (2018).
This study does not have the scope to be able to make further suggestions on how to utilise this finding, but it is something to be investigated further for practical applications. The implications of this can go both ways - NLP models could be useful in predicting what areas NNS would struggle to understand when learning a second language. The results of running similar tests could guide future teaching practices. On the other hand, studying further where NNS struggle to correctly understand certain areas of language could highlight potential weaknesses or variables that could be added to models. Similarly, the fact that political topics and political-related vocabulary had an effect on predictions also highlights a weakness in all of the models' abilities.

Finally, incorrect labelling was an issue for both groups - which was in line with Abercrombie and Hovy (2016) - checking all of the labels in a large dataset is impossible. This possibly could be done for a second attempt after error analysis of FP/FN, i.e. do a preliminary run through of the data with a smaller subset of participants and models. From the results of an error analysis, a further round of checking could be implemented.

7. Limitations

This study had limitations in a number of areas. First of all was the small participant number - with just 39 participants, it was difficult to get enough participants for each set of 30 comments. Essentially, more participants would have created a lot more useful data from which to extrapolate results. In addition, having all participants take a standardised test to assess their English skills could have resulted in more accurate analysis and comparisons. These participants were also heavily skewed towards Japanese (75%), so that could lead to some cultural biases in the NNS results. Finally, the participants were not compensated, so it is possible that they were not as motivated to fully engage with all of the tasks.

This study used high performing models, with RoBERTa and DeBERTa, however they were just the basic versions of the models with limited parameters and not as accurate as the state-of-the-art models used in FigLang 2020. The purpose of this study was not to try and achieve the highest level of accuracy, instead it was to analyse the errors of the models and compare them with NNS. Despite this, potentially different or more useful results could have been obtained by using state-of-the-art models.

Further, the analysis methods of just keyness and TF-IDF were quite limited - additional analysis methods would have provided additional insights. Finally, the issues with the data, such as mislabelling, was an unavoidable, but nonetheless counter-productive issue that would be good to deal with for future research. This has been mentioned in other recent papers such as Singh and Sharma (2023).

8. Conclusions

This study shows there are areas of clear overlap between NLP models and NNS ability to detect sarcasm, both in the content of the comments - such as political topics being more likely to be tagged as sarcasm - and the linguistic features -

such as exclamation marks and questions. These weak areas for both NLP models and NNS suggest areas of improvement for the models, and areas to focus teaching English to NNS.

For future studies, it could be of value to widen the scope to more areas of figurative language and sentiment analysis to observe what else NLP models struggle with, and then implement those areas into a curriculum for English teaching. If it is found to successful improve English ability versus a control group, then using a combination of error analysis of the results from NLP models with English teaching could be implemented in more classroom situations, as well as for self-study.

In addition, another area for future studies would be to use the results of the error analysis to improve the models' abilities. If the areas of weakness, such as marginally reducing the likelihood of predicting a political topic as sarcastic, are added as additional parameters in the fine-tuning stage, it could lead to improved results of sarcasm detection, as well as other areas of sentiment analysis.

9. References

Gavin Abercrombie and Dirk Hovy. 2016. Putting Sarcasm Detection into Context: The Effects of Class Imbalance and Manual Labelling on Supervised Machine Classification of Twitter Conversations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop*, pages 107–113, Berlin.

Laurence Anthony. 2022. AntConc.

David Bamman and Noah A Smith. 2015. Contextualized Sarcasm Detection on Twitter. Technical report.

Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. 2015. Parsing-based sarcasm sentiment recognition in Twitter data. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*.

Santosh Kumar Bharti, Reddy Naidu, and Korra Sathya Babu. 2018. Hyperbolic Feature-based Sarcasm Detection in Tweets: A Machine Learning Approach. In *2017 14th IEEE India Council International Conference, INDICON 2017*.

Somnath Basu Roy Chowdhury and Snigdha Chaturvedi. 2021. Does Commonsense help in detecting Sarcasm? *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 9–15.

Lisa Diamond, Sukyung Chung, Warren Ferguson, Javier Gonzalez, Elizabeth A. Jacobs, and Francesca Gany. 2014. Relationship between self-assessed and tested non-english-language proficiency among primary care providers. *Medical Care*, 52(5).

Xiangjue Dong, Changmao Li, and Jinho D. Choi. 2020. Transformer-based Context-aware Sarcasm Detection in Conversation Threads from Social Media. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 276–280.

Aileen Edele, Julian Seuring, Cornelia Kristen, and Petra Stanat. 2015. Why bother with testing? The validity of immigrants' self-assessed language proficiency. *Social Science Research*, 52.

Christopher Ifeanyi Eke, Azah Anir Norman, and Liyana Shuib. 2021. Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model. *IEEE Access*, 9:48501–48518.

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven R Wilson, and Walid Magdy. 2022a. SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. Technical report.

Ibrahim Abu Farha, Steven R. Wilson, Silviu Vlad Oprea, and Walid Magdy. 2022b. Sarcasm Detection is Way Too Easy! An Empirical Comparison of Human and Machine Sarcasm Detection. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5313–5324.

Debanjan Ghosh, Alexander R Fabbri, and Smaranda Muresan. 2018. Sarcasm Analysis Using Conversation Context. *Computational Linguistics*.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A Report on the 2020 Sarcasm Detection Shared Task. *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *International Conference on Learning Representations*.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. Technical report.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2019. A large self-annotated corpus for sarcasm. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*.

Jiyun Kim and James P. Lantolf. 2018. Developing conceptual understanding of sarcasm in L2 English through explicit instruction. *Language Teaching Research*, 22(2):208–229.

Roger J Kreuz and Gina M Caucci. 2007. Lexical Influences on the Perception of Sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4.

Hankyol Lee, Youngjae Yu, and Gunhee Kim. 2020. Augmenting Data for Sarcasm Detection with Unlabeled Conversation Context. *Proceedings of the Second Workshop on Figurative Language Processing*, pages 12–17.

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*.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. Reykjavik. European Language Resources Association (ELRA).

Sara Peters, Kathryn Wilson, Timothy W. Boiteau, Carlos Gelormini-Lezama, and Amit Almor. 2016. Do you hear it now? A native advantage for sarcasm processing. *Bilingualism*, 19(2):400–414.

Caleb Prichard and John Rucynski. 2019. Second language learners' ability to detect satirical news and the effect of humor competency training. *TESOL Journal*, 10(1).

Daniel S. Reuland, Pamela Y. Frasier, Matthew D. Olson, Lisa M. Slatt, Marco A. Aleman, and Alicia Fernandez. 2009. Accuracy of self-assessed Spanish fluency in medical students. *Teaching and Learning in Medicine*, 21(4).

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. Technical report.

Steven Ross. 1998. Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, 15(1).

Rachel L Shively, Mandy R Menke, and Sandra M Manzón-Omundson. 2008. Perception of Irony by L2 Learners of Spanish. *Issues in Applied Linguistics*, 16(2).

Bhuvanesh Singh and Dilip Kumar Sharma. 2023. A Survey of Sarcasm Detection Techniques in Natural Language Processing. In *2023 6th International Conference on Information Systems and Computer Networks, ISCON 2023*. Institute of Electrical and Electronics Engineers Inc.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM-A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Byron C Wallace, Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans Require Context to Infer Ironic Intent (so Computers Probably do, too). In *52nd Annual Meeting of the Association for Computational Linguistics*, pages 512–516. Association for Computational Linguistics.

Mengfei Yuan, Mengyuan Zhou, Lianxin Jiang, Yang Mo, and Xiaofeng Shi. 2022. stce at SemEval-2022 Task 6: Sarcasm Detection in English Tweets. Technical report.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. LCF: A Local context focus mechanism for aspect-based sentiment classification. *Applied Sciences (Switzerland)*, 9(16).