Sarcasm Detection is Way Too Easy! 🙃
An Empirical Comparison of Human and Machine Sarcasm Detection

Ibrahim Abu Farha[1], Steven R. Wilson[2], Silviu Vlad Oprea[1], Walid Magdy[1,3]
[1] School of Informatics, The University of Edinburgh, Edinburgh, UK
[2] Oakland University, Rochester, MI, USA
[3] The Alan Turing Institute, London, UK
{i.abufarha,silviu.oprea,wmagdy}@ed.ac.uk, stevenwilson@oakland.edu

Abstract

Recently, author-annotated sarcasm datasets, which focus on *intended*, rather than *perceived* sarcasm, have been introduced. Although datasets collected using first-party annotation have important benefits, there is no comparison of human and machine performance on these new datasets. In this paper, we collect new annotations to provide human-level benchmarks for these first-party annotated sarcasm tasks in both English and Arabic, and compare the performance of human annotators to that of state-of-the-art sarcasm detection systems. Our analysis confirms that sarcasm detection is extremely challenging, with individual humans performing close to or slightly worse than the best trained models. With majority voting, however, humans are able to achieve the best results on all tasks. We also perform error analysis, finding that some of the most challenging examples are those that require additional context. We also highlight common features and patterns used to express sarcasm in English and Arabic such as idioms and proverbs. We suggest that to better capture sarcasm, future sarcasm detection datasets and models should focus on representing conversational and cultural context while leveraging world knowledge and common sense.

1 Introduction

Sarcasm is a form of verbal irony that is often used to express ridicule or contempt. Sarcasm is usually correlated with expressing an opinion in an indirect way where there would be a discrepancy between the literal and intended meaning of an utterance (Wilson, 2006). For example the sentence *"I love being ignored"* is an example of sarcasm, where there is a discrepancy between the positive surface meaning and the negative implied one.

Sarcasm is present on the social web and, due to its nature, it can be disruptive to computational systems that harness this data to perform tasks such as sentiment analysis, opinion mining, author profiling, and hate-speech detection (Liu, 2012; Rosenthal et al., 2014; Maynard and Greenwood, 2014; Van Hee et al., 2018). Rosenthal et al. (2014) show a significant drop in sentiment polarity classification performance when processing sarcastic tweets, compared to non-sarcastic ones. Such computational systems are widely deployed in industry, driving marketing, administration, and investment decisions (Medhat et al., 2014). While much of the computational sarcasm detection work focuses on the English language, in the context of Arabic, Abu Farha and Magdy (2021) show the effect of sarcasm on Arabic sentiment analysis systems, where the performance dropped significantly for the sarcastic tweets. As such, it is imperative to devise models for sarcasm detection in languages other than just English. Doing this, however, requires reliable evaluation datasets, and further, understanding where current models (and even humans) fail on these datasets. Most of the previous sarcasm datasets have been created using either distant supervision or manual labelling. Those approaches produce unreliable labels since rule-based systems can suffer from sampling bias, and third-party annotators do not know whether the intention of the author was to truly be sarcastic. Recently, new datasets that contain first-party labels have been released. One method to collect such data is reactive supervision, where conversational cues such as "I was being sarcastic" are used as labels of previous comments (Shmueli et al., 2020). This increases reliability by identifying texts in which the authors do claim to be sarcastic, but since texts are sampled according to predefined patterns, the data may be biased toward cases that required clarification.

Yet another method for collecting first-party labels even more directly is to ask authors to provide explicit annotations of their own texts (Oprea and Magdy, 2019; Abu Farha et al., 2022). This approach eliminates annotation proxies, further reduces sampling and annotator bias, and allows for

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5313 - 5324

the collection of additional data about each sarcastic text, such as explanations and rephrases.

Although more difficult than earlier sarcasm detection tasks, classification models have shown promising performance on first-party annotated datasets, as evidenced by a recent shared task (Abu Farha et al., 2022). However, there is currently no analysis of how *humans* would perform on these datasets. This kind of analysis can provide a range of benefits, including empirical insights into the difference in quality between third-party and first-party annotations. Additionally, most of previous works focused on building resources and detection models, without as much attention being paid to error analysis. These analyses are necessary to give insights about the limitations of the current best models, and pave the way to mitigate these limitations in the future. In this paper, we aim to fill this gap and answer the following research questions:

- **RQ1**: How do humans perform on author-annotated sarcasm detection tasks?
- **RQ2**: How does human performance on these tasks compare with state-of-the-art text classifiers?
- **RQ3**: What makes sarcasm challenging for both humans and classification models?

In this paper, we answer these questions by measuring both human and machine performance on iSarcasmEval's datasets (Abu Farha et al., 2022), which have first-party sarcasm labels. We make the following contributions: (1) we collect new human annotations for the iSarcasmEval datasets; (2) We analyse both humans' and state-of-the-art (SOTA) models' performance on both English and Arabic tasks, identifying cases where each succeeds and fails; (3) We analyse the error cases in order to determine the current limitations of sarcasm detection methods; and (4) we provide recommendations, based on our empirical evidence, for improving sarcasm detection models in the future.

Our analysis shows that sarcasm detection is challenging for humans, who perform nearly as well as state-of-the-art models on their own, and even better when their annotations are combined through majority voting. However, human performance using third-party labels is still imperfect, and casts doubt on its utility as a source of ground truth for this task. We find that context and world knowledge are necessary to understand sarcasm in many cases. Thus, future works on sarcasm should focus on including this kind of information into datasets and leveraging it in detection models.

2 Related Work

Previous work on sarcasm detection falls into one of two branches: creating **datasets** (Ptáček et al., 2014; Khodak et al., 2018; Barbieri et al., 2014; Filatova, 2012; Riloff et al., 2013; Abercrombie and Hovy, 2016; Oprea and Magdy, 2020a) or creating **detection models** (Campbell and Katz, 2012; Riloff et al., 2013; Joshi et al., 2016; Wallace et al., 2015; Rajadesingan et al., 2015; Bamman and Smith, 2015; Amir et al., 2016; Hazarika et al., 2018; Oprea and Magdy, 2019). Most, but not all, of this work has focused on sarcasm detection in English. In this section, we cover the literature of sarcasm detection in both English and Arabic, since these are the languages represented in our dataset.

2.1 English Sarcasm Detection

Most of the previous work on sarcasm detection focused on either creating datasets or building detection models. The approaches used to create the datasets aim to mitigate the issues that might arise when collecting data in a specific way. Traditionally, distant supervision and manual labelling were used to collect sarcasm datasets. In distant supervision, text is considered sarcastic if it meets predefined criteria such as including a specific hashtag (e.g. #sarcasm, #irony) (Ptáček et al., 2014; Khodak et al., 2018), or being generated by specific accounts (Barbieri et al., 2014). However, this approach might lead to the inclusion of false positives. To mitigate that, manual labelling has been used, where sarcasm labels are provided by human annotators (Filatova, 2012; Riloff et al., 2013; Abercrombie and Hovy, 2016). As such, the labels represent *annotator perception*, which may actually differ from *author's intention*. Annotators might lack awareness of the contextual devices that, as linguistic studies suggest (Grice, 1975; Sperber and Wilson, 1981; Utsumi, 2000), could be essential for clarifying the sarcastic intention of the authors. (Shmueli et al., 2020) proposed a third method, reactive supervision, which aims to collect sarcastic examples based on the conversation dynamics, addressing some of these issues by using statements such as "I was being sarcastic" to automatically label texts. (Oprea and Magdy, 2020a) proposed to mitigate these issues by asking people to pro-

vide their own sarcastic sentences/tweets. Sarcasm was part of multiple shared tasks such as (Van Hee et al., 2018; Ghosh and Muresan, 2020). The most recent shared task is SemEval-2022 task 6 (iSarcasmEval) (Abu Farha et al., 2022), which proposed new English and Arabic datasets that were created using the approach proposed by Oprea and Magdy (2020a), and comprise the data that we use throughout this paper.

2.2 Arabic Sarcasm Detection

Arabic sarcasm has not received the same degree of attention as English. Most of the work on Arabic sarcasm data collection uses approaches similar to the ones used for English. Karoui et al. (2017) were the first to work on Arabic sarcasm/irony detection. They created a corpus of sarcastic Arabic tweets using distant supervision, where they used the Arabic equivalent of #sarcasm (#سخرية). Ghanem et al. (2019) organised a shared task on Arabic sarcasm/irony detection. They prepared their dataset using distant supervision and manual labelling. Abbes et al. (2020) also used distant supervision with manual labelling. Recently, there has been a growing interest in Arabic sarcasm with the shared task organised by Abu Farha et al. (2021). The dataset used for the task was built using third-party manual annotation. Arabic was also included in SemEval-2022 along with English in the iSarcasmEval shared task (Abu Farha et al., 2022). The dataset used for the Arabic task was collected using a similar approach to the English one, where the organizers asked the authors to provide sarcasm labels for their own texts. For detection systems, most of the previous work comes from submissions to the aforementioned shared tasks (Khalifa and Hussein, 2019; Abuzayed and Al-Khalifa, 2021; Alharbi and Lee, 2021; El Mahdaouy et al., 2021; Abu Farha and Magdy, 2021).

2.3 Analysis of Sarcasm Detection

It is clear that most of the work in this area has focused on how to improve data quality and how to mitigate the issues that would arise when using a specific approach. However, the literature lacks extensive analysis of which types of examples are easiest and most difficult to make accurate predictions about, yet there is some work in this direction. Some work has focused on analysing the effect of including context in sarcasm detection models (Oprea and Magdy, 2019; Abercrombie and Hovy, 2016; Wallace et al., 2014). Wallace et al. (2014) showed that annotators tend to need context to provide judgements about ironic content. They showed that there is a correlation between that and the misclassified cases. Oprea and Magdy (2019) explored the effect of contextual information to detect sarcasm, and Oprea and Magdy (2020b) analysed the effect of cultural background and age on sarcasm understanding. Their analysis indicates that age, English language nativeness, and country are significantly influential on sarcasm understanding and should be considered in the design of sarcasm detection systems. Similar results were confirmed in the case of spoken sarcasm, where Puhacheuskaya and Järvikivi (2022) found that having a foreign accent had a negative impact on irony understanding. In the context of Arabic, Abu Farha and Magdy (2022) show that dialect familiarity has an effect on sarcasm understanding. We add to this line of work by exploring sarcasm detection results through the lens of comparing human and machine labels in order to better understand factors related to each when making determinations about the sarcastic nature of text.

3 Methodology

3.1 Dataset

In this work, we use SemEval-2022 Task 6, iSarcasmEval, datasets (Abu Farha et al., 2022). The shared task includes three subtasks: (1) sarcasm detection (subtask A): given a text, determine whether it is sarcastic or nonsarcastic; (2) sarcasm category classification (subtask B): given a piece of text, determine which ironic speech categories it belongs to; and (3) pairwise sarcasm identification (subtask C): given a sarcastic text and its nonsarcastic rephrase, determine which is the sarcastic one. Subtasks A and C cover both English and Arabic, while subtask B is English only. In this work, as we aim to analyse the performance on the two languages, we use the test sets of subtasks A and C. The test sets for Task A consist of 1400 examples, while the sets for task C consist of 200 pairs, each containing a sarcastic text and its nonsarcastic rephrase, written by the same author.

3.2 Human Annotation

To analyse human performance, we decided to measure how humans would perform on the test sets and compare that to the performance of computational models that participated in the shared task.

To this end, we collected human annotations for the test sets using the Prolific[1] platform for the English dataset and Appen[2] for the Arabic one. Those are the same platforms as the original iSarcasmEval paper. The authors mention that they chose Appen due to the availability of native Arabic speakers, who are not available on Prolific, and we did try both and noticed the same. Thus, we followed the recommendation by the iSarcasmEval paper.

For each test set, we collected *5 annotations* for each item [3]. We allowed only native speakers of the annotated language to participate. Before starting the annotation process, each annotator is presented with test questions and only those who answer all the questions correctly would be allowed to participate in the annotation process. The test questions were sampled from a set of sentences that are clearly sarcastic/non-sarcastic. We used this approach to make sure that the annotators are not giving random answers and to avoid introducing any bias before the annotation. For the English datasets, the average percentage of votes that the majority label received for tasks A and C are 86% and 94%, respectively. For the Arabic dataset, these figures are 88% for task A and 94% for task C.

4 Results and Analysis

In this section, we compare the performance of humans against state-of-the-art models. The task organisers agreed to share these teams' detailed submissions, including individual predictions for each item of the test set, allowing us to perform this comparison. We consider comparing the human predictions with the top-performing system for each subtask, as well as with a combination of the top five performing systems using majority voting. However, in all cases, we find that using the output from the single top team for the subtask outperformed the combination of the top five. Therefore, we only compare the human predictions with the single[4] best performing model in each subtask.

For subtask A (English), Yuan et al. (2022) were ranked first with an $F_1^{\text{sarcastic}}$ of 0.605. They used an ensemble learning approach of three transformer-based models: RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and XLM-RoBERTa (Conneau et al., 2020). For Arabic, El Mahdaouy et al. (2022) were ranked first with an $F_1^{\text{sarcastic}}$ of 0.563. They used an ensemble of models based on MARBERT (Abdul-Mageed et al., 2021). For subtask C, Han et al. (2022) were ranked first on English with an accuracy of 0.870. They used an ensemble of ERNIE-M (Ouyang et al., 2021) and DeBERTa. For Arabic, the top team was (Zefeng et al., 2022) with an accuracy of 0.930. Their model is based on Arabic BERT (Safaya et al., 2020).

4.1 General performance

Table 1 shows the general performance on both tasks for English and Arabic. From the table, it is noticeable that when taking the majority vote from the human annotators, the performance for both languages and on both tasks is better than the models submitted by the top team who participated in the respective shared task. The scores achieved by humans when considering individual annotations, rather than majority vote, would have achieved second place in all tasks, indicating how challenging the tasks are, even for humans. When conducting McNemar's test (McNemar, 1947), the results show that the error distributions of humans and the machine model are different except for the Arabic subtask C (pairwise sarcasm identification). A deeper look into the nature of these errors is in the following section.

4.2 Performance analysis

Table 2 shows the annotation agreement (Cohen's Kappa) between humans' majority label and the top team in the respective task. These results demonstrate that although both human majority voting and the state-of-the-art machine-based methods achieved similar performance, they only have moderate agreement with one another for task A. On the other hand, there was substantial agreement between the human and machine annotations for task C. Based on these preliminary results, in this section, we examine cases where sarcasm was detected by humans and/or machines in order to further investigate both the differences and similarities between the sets of annotations. For the analysis, as mentioned previously, we consider humans' majority vote vs top team.

4.2.1 English

Figure 1 shows the quantitative difference between the human and machine generated labels for task

[1] https://prolific.co

[2] https://appen.com

[3] The data is available at: https://github.com/iabufarha/iSarcasmEval

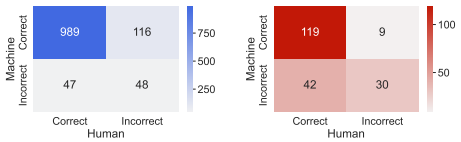
[4] Although these are already, in some cases, ensembles of several other models.

Annotation	English		Arabic	
	Task A ($F1_{sarcastic}$)	Task C (Acc)	Task A ($F1_{sarcastic}$)	Task C (Acc)
Human (majority vote)	0.613	0.970	0.665	0.935
Human (individual-level)	0.523	0.819	0.525	0.909
Machine (SOTA)	0.605	0.870	0.563	0.930

Table 1: Results for humans with majority voting, humans individually, and the top performing system for both the sarcasm detection (task A) and pairwise identification (task C) for English and Arabic. $F1_{sarcastic}$ is the $F1$ score for the sarcastic class, the official metric used in the shared task.

Task	English	Arabic
Task A	0.52	0.49
Task C	0.72	0.77

Table 2: Cohen's kappa agreement between the human (majority vote) annotations and the predictions from the top performing system in the respective task.



(a) Non-sarcastic (b) Sarcastic

Figure 1: Prediction distribution for task A (English).

A in more detail. We can observe that for both sarcastic and non-sarcastic texts, the human and machine labels were correct most of the time. Humans outperform machines in identifying both types of texts, and the machines very rarely identified sarcastic texts that the humans missed, doing so only 9 times. The results for task C are shown in Figure 3a. Here, there are only 2 pairs where both humans and machines could not accurately select the sarcastic phrasing, and roughly a dozen mistakes were made by only either humans or the classification models. Next, we take a closer look at some of the specific examples detected correctly and incorrectly by humans and classification models.

Non-sarcastic

Examples of non-sarcastic English texts are shown in Table 3. Examples where both human and machine predictions are correct show sincere positive

Machine \ Human	Correct	Incorrect
Correct	1. An absolutely gorgeous winter's day. Happy Friday all. ☺ 2. I hate paying so much for gas.	3. good for her 4. The 5,000 cases figure in Scotland is partly down to backlog. But of course Sturgeon doesn't mention this...
Incorrect	5. Well it's a good question 🤔 6. Wow, Chelsea bean beaten by West Ham, that was a day I had been looking forward to for a while.	7. Well this is awesome news to wake up to! 8. Today, I lost a follower because I said I was unwell. Dontcha love Twitter.

Table 3: English non-sarcastic examples.

Machine \ Human	Correct	Incorrect
Correct	1. It makes me feel a lot safer knowing the MET Police don't investigate crimes after they happen. 2. Love it when someone with no mask chooses to sit next to me on the bus...😑	3. Really happy that the weather has stayed like this for the whole weekend 4. Wow, work is just so rewarding and fulfilling right now
Incorrect	5. if you listen carefully, you can hear me not carig 6. Politicians are so honest it melts my heart	7. Thoughts and prayers will actually work this time. Trust me 8. Biden is a great President like none other we have had

Table 4: English sarcastic examples.

and negative statements with no indication of any reversal of meaning (Table 3, items 1 and 2). Cases where only the human annotators incorrectly predicted sarcasm include those that cannot fully be resolved without additional information: "good for her" requires more conversational context to know whether "good" is sincere or sarcastic. Item 4 (Table 3) *could* be sarcastic if the person mentioned (Sturgeon) *did* in fact mention the backlog, so this example requires knowledge of actual events in a specific domain (The leader of the Scottish Government's reporting of COVID-19 cases). Humans outperformed the classifier when certain emojis that are sometimes associated with sarcasm were used (item 5) or world knowledge is required (e.g., knowledge about the general sentiment of football fans toward certain teams in item 6). Both humans and the machine model were likely to incorrectly predict sarcasm in cases where the text is overly positive and lacking context (item 7) or examples that *do* appear to contain some sarcasm ("Dontcha love Twitter" in item 8) that might have actually been incorrectly labelled as non-sarcastic by the author.

Sarcastic

In the set of sarcastic English texts (Table 4), we observed that many of the easiest to correctly detect were those that, on the surface form, made positive statements about government entities (item 1 in Table 3) and those with incongruity between the literal meaning of the text and emotional markers, such as the emoji in item 2. Also included in these cases were those that use typical sentence forms

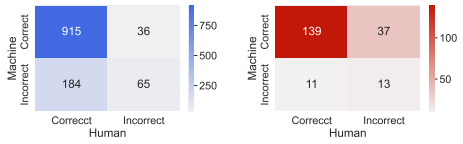
Machine \ Human	Correct	Incorrect
Correct	1. **S:** Wow the weather is practically tropical. **NS:** The weather outside is freezing today. 2. **S:** Oh yes, because allowing your dog to bite other animals & people is perfectly fine! **NS:** It absolutely not ok to allow your dog to bit other animals & people	3. **S:** Mohammed Salah isnt a bad player is he? **NS:** Mohammed Salah is one of the best in the world. 4. **S:** I love living at home now that I'm back from uni. Love the peace and quiet I get from my parents. **NS:** Wish I could return to living at uni so bad! I miss the peace and quiet, away from my parents.
Incorrect	5. **S:** Biden is a great President like none other we have had **NS:** Biden's ratings take another hit. 6. **S:** Another gorgeous day in sunny Accrington! NS: It's another wet and windy day in typically wet and windy Accrington.	7. **S:** My lovely husband just brought me this fantastic steam cleaner for my birthday, he is so thoughtful **NS:** Can you believe my husband thought it would be a good idea to buy me a steam cleaner for my birthday

Table 5: English pairs of sarcastic texts and their non-sarcastic rephrases.

for English-language sarcasm (e.g., "I just love it when..."). Where the text classifier succeeded and humans failed to detect sarcasm, the texts typically require more context to make a prediction. Items 3 and 4 both express a positive sentiment on the surface which *could* be true. However, as suggested by theories of sarcasm (Kreuz and Glucksberg, 1989) and psycholinguistic research (Pexman and Olineck, 2002), sarcasm may more commonly be used to express negative meanings using positive surface forms due to societal preferences toward positive statements, and therefore detection models may be picking up on this connection between positive surface forms and sarcasm, causing them to predict sarcasm in these ambiguous yet positive instances. Humans excelled in cases where typos ("carig" in item 5) or idioms were used ("it melts my heart" item 6) that might have been more difficult to the text classification model to generate accurate representations for. Both humans and the machine model also struggled in some cases where phrases with a noncompositional meaning were used ("thoughts and prayers" in item 7) or where information about the user's stance toward a political figure (Biden) needs to be known in order to correctly detect non-sarcasm (item 8).

Pairwise identification.

Both humans and the model were better able to identify sarcasm when the sarcastic text was paired with a non-sarcastic rephrase, with the sarcastic text being identified the vast majority of the time. The main cause for errors, when they did occur, were cases that required more information about either the authors of the text (items 4, 5, and 7 in Table 5) or world knowledge. For example, to classify these correctly, it is important to have knowledge of the typical weather patterns in Accrington, England (item 6), or the rankings of international football



(a) Non-sarcastic (b) Sarcastic

Figure 2: Prediction distribution for task A (Arabic).

players, without which, saying that Salah is "one of the best in the world" might appear as a sarcastic exaggeration (item 3).

4.2.2 Arabic

For Arabic, similar to English, we analyse the sarcastic/non-sarcastic cases that were detected by human annotators and/or the respective model. Figure 2 shows the distribution of the examples for each case.

Non-sarcastic

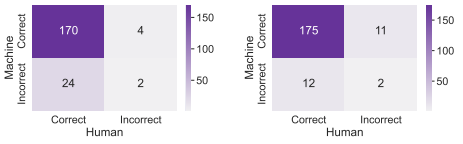
Table 6 shows some non-sarcastic examples for all the possible scenarios. When looking at the examples that were detected by both humans and machine, i.e. easy cases, we noticed direct sentences with common words that are used frequently (Table 6, items 1 and 2). Also, given that Arabic has free word-order syntax, the structure of these sentences is also clear and direct, without any changes to the traditional word-order of an Arabic sentence. These examples are also quite close to modern standard Arabic (MSA), without many dialectal words or spelling variations. The examples that neither humans nor the model detected were sentences that included wordplay or some changes to the spelling of the word, an example of this is item 7 (Table 6). The author changed the structure of the words غرفة (room) and بيت (house) to غرفوفي and بيتوتي (dialectal derivation to mean someone who love being in the room/house). Also, sentences that included sarcasm along with the non-sarcastic rephrase, were misclassified (Table 6, item 8). The annotators considered such cases to be sarcastic despite the annotation instructions asking them to consider sentences with indirect expressions to be sarcastic. Humans were better at detecting non-sarcastic sentences that contain descriptions that appear metaphorical, and in some cases they are, but are so commonly used that humans consider them direct descriptions (Table 6, items 5 and 6). The model performed better than humans for sentences that contain exaggerated complaints or descriptions.

Machine	Human: Correct	Human: Incorrect
Correct	1. ليت اكون شخص ما يفكر كثير I wish I was a person who doesn't think a lot 2. أحب لما انسأل عن كل التفاصيل I like when I am asked about all the details	3. قاعدة اكل شيبس واتحسر على ايامي الضايعة اف I am eating snacks and bemoaning lost days 4. جماعه ممكن لما حد يضايقني انا اللي از عل مش هو ؟ Would it be possible when someone annoys me that I get annoyed and not them?
Incorrect	5. راسي رح بنفجرر My head will explode 6. لا انا البيوم الضغط عندي بيرتفع No! it seems that I will have a high blood pressure today	7. تعديت مرحله البيتوتي صرت غرفوقي I moved from someone who stay at home to someone who stays in the room 8. شعره سئ و يشبهه المكنسة His is hair is bad and looks like a broom

Table 6: Arabic non-sarcastic examples.

Machine	Human: Correct	Human: Incorrect
Correct	1. شعب بيتعاطى كلور A population addicted to chlorine 2. الي ميعرفش يقول عدس This who doesn't know says lentil (proverb) 3. قده قد الفأرة He is the size of a mouse	4. طول البال بيد جبال patience destroys mountains (Egyptian proverb) 5. بدري!! لليش مروح هلكيت ... خليك اخرى شوي too early! why are you here, stay a little bit more (uses specific terms from some regions in Palestine)
Incorrect	6. عايزينا نتعلم عن بعد واحنا أصلا مش بنفهم عن قرب They want us to learn remotely while we don't understand in person 7. يستهلك المواطن الأردني ٧٩٪ من حسناته أثناء قيادة المركبة في كافة شوارع المملكة الحبيبة The Jordanian citizen loses 79% of his good deeds while driving in the beloved Kingdom streets	8. ابغي تعارف جاد I am looking for a serious relationship 9. عد زوال هذه الأزمة إن شاء الله، ستشهد مصر اعظم موسم افراح عرفه التاريخ After this struggle, Egypt will see the largest weddings season

Table 7: Arabic sarcastic examples.



(a) English (b) Arabic

Figure 3: Performance of human vs machine on task C.

Humans assumed these cases to be sarcastic, but in reality, the meaning is expressed directly (Table 6, items 3 and 4).

Sarcastic

Table 7 shows some sarcastic examples for the possible scenarios. The mostly commonly used and easily detected sarcastic sentences were those that contained the usage of words in uncommon contexts like item 1 (Table 7). In this example, the author used the word كلور (chlorine) in the context of addiction, which does not match the reality as humans cannot consume chlorine or be addicted to it. Proverbs and idioms (Table 7, items 2 and 4) were common among this set and it seems that Arabic speakers tend to rely on these meaning-dense phrases to express their feelings towards something or reply to someone. Another common pattern to express sarcasm is to use animals or objects or their attributes as adjectives to describe someone or something, e.g., item 3 (Table 7). On the other hand, the most challenging sarcastic examples were those that require specific context, either culturally or based on the sociocultural background and personality of the speaker (Table 7, items 8 and 9). Also, understanding of the dialect plays an important role here. Humans were better at detecting sarcasm that is expressed using complex metaphors that require specific world knowledge (Table 7, items 6 and 7). The model was better at detecting sarcasm that is expressed using dialect-specific words (Table 7 item 5).

Pairwise identification

Like the English-language versions of the tasks, pairwise identification was easier than sarcasm detection. There were only two sentences that neither humans nor the model detected, the first one a Maghrebi dialect sentence (Table 8, item 7) and it seems that the annotators are not familiar with this dialect. For the model, the reason for this is probably the fact that Maghrebi dialect is the one with fewest examples in the iSarcasmEval's training data. The other sentence (Table 8, item 8) included extremely implicit sarcasm, which can be considered present in both the sarcastic and non-sarcastic rephrases. It seems that this case caused confusion for both humans and the model. For the other cases, no clear pattern was observed.

Machine	Human: Correct	Human: Incorrect
Correct	1. **S:** معنديش فكه تاخد برميل بنزول Don't you have change so you take on barrel of oil **NS:** سعر البترول يصل لادني مستوي Oil prices reach all time low 2. **S:** اجى حبيب قلبك Your beloved one has arrived **NS:** اجى الحدا الي بتكرهه The one you hate has arrived	3. **S:** يطعمك الحج والناس راجعة May you go for pilgrimage while people are coming back **NS:** هلأ لحسيت ع حالك You just knew! (it is too late) 4. **S:** ابو تريكة امير القلوب Aboutrika is the prince of hearts **NS:** أبو تريكة كل الناس بتحبه Everybody loves Aboutrika
Incorrect	5. **S:** ارحمنا يا مركز الكون Oh, center of the universe, excuse us! **NS:** ما جبنا سيرتك أصلا We didn't even mention you! 6. **S:** بات مع الدجاج صبح اقاقي He slept in a hen house and woke up clucking **NS:** تخليت على المبادئ نتعك فيساع You left your principles quickly	7. **S:** لوكان حراث بحرث وطنه If he was a good farmer, he would farmed his land **NS:** حاط روحك قادر و نت صفر You think you are string (or good), but you are a zero 8. **S:** اول عقوبه يحصل عليها السعودي في مرحلة البلوغ صورته في بطاقة الأحوال The first punishment a Saudi gets when becoming an adult is his ID photo **NS:** اول عقوبه يحصل عليها السعودي عند البلوغ صورته القبيحة في البطاقة The first punishment a Saudi gets when becoming an adult is his ugly ID photo

Table 8: Arabic pairs of sarcastic texts and their non-sarcastic rephrases.

4.3 Thematic Error Analysis

The previous section provided a general overview and discussion of the nature of errors made by both humans and the machine model. To better categorize and quantify these errors we annotated the sarcastic sentences in the test set of subtask A (sarcasm detection) for both languages according to the set of themes that we discovered in our initial

error analysis. Tables 9 and 10 provide detailed statistics of the available themes in each language.

Theme	N	H_{Er}	M_{Er}
Requires context about speaker	50	14	29
Incongruity between text and emotional markers	45	5	8
Incongruity between text and implied sentiment	42	1	6
Requires world knowledge	31	13	15
Contains non-compositional meaning	14	3	6
Idioms	9	2	4
Positive comment about government	6	0	1
Contains misspelling/typo	4	1	2

Table 9: Sarcasm themes among sarcastic English sentences. N: number of examples, H_{Er}: human error, and M_{Er}:machine error

Theme	N	H_{Er}	M_{Er}
Idioms	58	13	7
Proverbs	45	10	0
Referencing specific context, world knowledge	45	15	12
Complex metaphors, world knowledge	45	11	8
Dialect specific words	21	8	1
Referencing animals or objects	11	0	0
Words in uncommon context	8	0	0

Table 10: Sarcasm themes among sarcastic Arabic sentences. N: number of examples, H_{Er}: human error, and M_{Er}:machine error

For English, humans misclassified 39 sarcastic texts in total, while the machine model misclassified 72 of the sarcastic texts. 69% of the human errors were due to lack of contextual information such as context about the speaker (36%) and world knowledge (33%). The machine model also struggled with the same kind of examples which caused 62% of the errors as follows: context about the speaker (40%) and world knowledge (21%)

For Arabic, humans misclassified 50 sarcastic texts while the machine model misclassified 24 sarcastic texts. Most of the human errors are due to a lack of world knowledge (52%), idioms (26%) and proverbs (20%). The machine model was mostly affected by a lack of world knowledge which caused 83% of the errors.

5 Discussion

Here we revisit and answer our research questions and provide some additional discussion.

RQ1: How do humans perform on author-annotated sarcasm detection tasks? Human annotations from non-authors of the text are vastly different from the labels provided by the authors themselves. This suggests that there is an important difference between *intended* and *perceived* sarcasm, as suggested by Oprea and Magdy (2020a), who also argue that first-party annotations are more reliable as being sarcastic is an intentional act. Not only for sarcasm, but also for which intention and perception may not be consistent, the use of third-party annotations has serious implications for the reliability of our datasets' ground truth: as shown in our results, there are cases where these annotations do not align with the labels provided by the texts' authors themselves.

RQ2: How does human performance on these tasks compare with state-of-the-art text classifiers? We found that on their own, humans performed almost as well as the state-of-the-art sarcasm detection systems, but when working together using majority voting, humans achieve the best results. On the other hand, using majority voting for the systems led to worse performance. This suggests that humans provide complementary knowledge when it comes to the task of sarcasm detection, while the text classification models' predictions typically have high overlap with one another.

RQ3: What makes sarcasm challenging for both humans and classification models? For English, we observed that many of the less challenging sentences were those with clear sentiment incongruity, e.g., "I love failing exams", or sentences criticising governmental figures by sarcastically presenting them in overly positive contexts. For Arabic, idioms and proverbs were quite common (around 23% of sarcastic sentences in the test set) and detected easily by humans and the models. It seems that Arabic speakers rely heavily on proverbs and the dense meaning they contain, while having a shared cultural context. The other common Arabic sarcasm pattern is to say " You look like/act animal/object", which is often used for derogatory remarks.

However, the most challenging cases on both languages were the sentences that require additional context. This can be in the form of conversational context, information about the author of the text, world knowledge, and dialectal awareness in the case of Arabic. For example, the model was better than humans in detecting sarcasm in sentences that used local words which some of the humans may be unfamiliar with. Given that MARBERT(Abdul-Mageed et al., 2021) was trained on 1B tweets, it seems that it had better coverage of some specific dialects than the annotators. Interestingly, this model was able to classify Levantine examples better than the Maghrebi ones, which could be attributed to a possible bias in the training data.

The fact that missing context led to a large num-

ber of error cases implies that it is necessary for detection systems to have representations for world knowledge and cultural background, and be aware of the language/dialect of the sarcastic utterance. Progress in this direction is possible: for example, dialectal awareness can be addressed by using language models that are trained on a large dialectal variety.

6 Recommendations

Based on analyses and discussions in the previous sections, we make the following recommendations:

Avoid 3rd-party annotations. We should re-evaluate third-party annotation as a method to create datasets of subjective content, particularly when author intention is important. The analysis of the performance of human annotators on two languages, English and Arabic, shows that their performance is comparable to state-of-the-art models. However, the performance of both the models and humans still has much room for improvement.

Develop models that incorporate context. In order to improve the performance of detection models, we need to better incorporate contextual information such as cultural references, author tendencies, world knowledge, and dialect awareness. The need for this has been demonstrated through the fact that both humans and models failed to detect sarcasm that relied on such information.

Include contextual features in shared datasets. Further, in order to train these models, sarcasm detection datasets that contain a wealth of contextual information should be created and released, especially conversational and author-level information which cannot be obtained from external knowledge bases. However, releasing this type of data brings new challenges in the space of privacy, as conversations contain texts written by other authors who may not have consented to sharing their content, and including more author-level information may lead to deanonymization and loss of privacy.

Build accurate representations of idioms and proverbs. More focus should be put into building accurate representations of idioms and proverbs, which are extensively used in sarcastic communication, especially by Arabic speakers.

7 Conclusions

In this paper, we analyse human performance on sarcasm detection and compare it to state-of-the-art detection models. We use SemEval's 2022 task 6 datasets, which has first-party sarcasm labels for both English and Arabic texts. Our analyses show that sarcasm detection is challenging for humans on both languages with performance only slightly better than trained models, and only when using majority voting between the human predictions. The low human performance emphasizes the subjective nature of sarcasm and indicates that third-part labels for subjective tasks are noisy. Consequently, we urge the community to re-evaluate third-party annotations for extremely subjective tasks, such as sarcasm, and use first-party labels. We conduct a thorough error analysis, revealing that the most challenging sarcastic sentences are those that require additional contextual information to accurately resolve, suggesting that future work focus on creating context-rich datasets and models with the ability to adequately leverage contextual information. Our analyses show that idioms and proverbs common linguistic tools used to express sarcasm, especially in Arabic, yet trained models often struggle with examples that contain them.

Limitations

The main limitation of our work is that with our analysis we only considered the performance of top teams in iSarcasmEval. The approaches use the most recent pretrained language models, but the results might differ slightly if compared with other models or other datasets. Another limitation is that for Arabic, we did not filter based on dialect and we matched the annotators with entries randomly. Dialect awareness might have an effect and should be investigated. Finally, we did not try to match examples to different age groups and previous studies suggest that this can have an effect on how people understand sarcasm (Oprea and Magdy, 2020b).

Acknowledgements

This work was partially supported by the Defence and Security Programme at the Alan Turing Institute, funded by the UK Government; the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1); the University of Edinburgh; and The Financial Times. We also thank Balquis Shalabi from Birzeit University, who helped with the thematic annotation.

References

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. DAICT: A dialectal Arabic irony corpus extracted from Twitter. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6265–6271, Marseille, France. European Language Resources Association.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

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 ACL 2016 Student Research Workshop*, pages 107–113. ACL.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 21–31, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2021. A comparative study of effective approaches for arabic sentiment analysis. *Information Processing & Management*, 58(2):102438.

Ibrahim Abu Farha and Walid Magdy. 2022. The effect of arabic dialect familiarity on data annotation. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ibrahim Abu Farha, Silviu Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Abeer Abuzayed and Hend Al-Khalifa. 2021. Sarcasm and sentiment detection in Arabic tweets using BERT-based models and data augmentation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 312–317, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Abdullah I. Alharbi and Mark Lee. 2021. Multi-task learning using a combination of contextualised and static word embeddings for Arabic sarcasm detection and sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 318–322, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mario J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *CoNLL*, pages 167–177. ACL.

David Bamman and Noah A. Smith. 2015. Contextualized sarcasm detection on twitter. In *ICWSM*, pages 574–577. AAAI Press.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2014. Italian irony detection in twitter: a first approach. In *CLiC-it*, page 28. AILC.

John D Campbell and Albert N Katz. 2012. Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes*, 49(6):459–480.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Abdelkader El Mahdaouy, Abdellah El Mekki, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. Deep multi-task model for sarcasm detection and sentiment analysis in Arabic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 334–339, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Abdelkader El Mahdaouy, Abdellah El Mekki, Kabil Essefar, Abderrahman Skiredj, and Ismail Berrada. 2022. Cs-um6p at semeval-2022 task 6: Transformer-based models for intended sarcasm detection in english and arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*. ELRA.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. 2019. Idat at fire2019: Overview of the track on irony detection in arabic tweets. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 10–13.

Debanjan Ghosh and Smaranda Muresan. 2020. Figlang2020 - sarcasm detection shared task.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press.

Yaqian Han, Yekun Chai, Shuohuan Wang, Yu Sun, Hongyi Huang, Guanghao Chen, Yitong Xu, and Yang Yang. 2022. X-pudu at semeval-2022 task 6: Multilingual learning for english and arabic sarcasm detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. In *COLING*, pages 1837–1848. ACL.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? In *EMNLP*, pages 1006–1011. ACL.

Jihen Karoui, Farah Banamara Zitoune, and Veronique Moriceau. 2017. Soukhria: Towards an irony detection system for arabic in social media. *Procedia Computer Science*, 117:161–168.

M. Khalifa and Noura Hussein. 2019. Ensemble learning for irony detection in arabic tweets. In *FIRE*.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Roger J. Kreuz and Sam Glucksberg. 1989. How to Be Sarcastic: The Echoic Reminder Theory of Verbal Irony. *Journal of Experimental Psychology: General*, 118(4):374–386.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

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. *CoRR*, abs/1907.11692.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4238–4243, Reykjavik, Iceland. European Language Resources Association (ELRA).

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, Florence, Italy. Association for Computational Linguistics.

Silviu Oprea and Walid Magdy. 2020a. isarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Silviu Vlad Oprea and Walid Magdy. 2020b. The effect of sociocultural variables on sarcasm communication online. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Penny M Pexman and Kara M Olineck. 2002. Understanding Irony: How Do Stereotypes Cue Speaker Intent? *Journal of Language and Social Psychology*, 21(3):245–274.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *COLING*, pages 213–223. ACL.

Veranika Puhacheuskaya and Juhani Järvikivi. 2022. I was being sarcastic!: The effect of foreign accent and political ideology on irony (mis) understanding. *Acta Psychologica*, 222:103479.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *WSDM*, pages 97–106. ACM.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, pages 704–714. ACL.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Boaz Shmueli, Lun-Wei Ku, and Soumya Ray. 2020. Reactive Supervision: A New Method for Collecting Sarcasm Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2553–2559, Online. Association for Computational Linguistics.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction. *Philosophy*, 3:143–184.

Akira Utsumi. 2000. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, 32(12):1777–1806.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Byron C. Wallace, Do Kook Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1044, Beijing, China. Association for Computational Linguistics.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland. Association for Computational Linguistics.

Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743.

Mengfei Yuan, Zhou Mengyuan, Lianxin Jiang, Yang Mo, and Xiaofeng Shi. 2022. stce at semeval-2022 task 6: Sarcasm detection in english tweets. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Li Zefeng, Yu Bingjie, Tuerxun Tunike, Li Zhaoqing, and Wang Yuhan. 2022. Naive at semeval-2022 task 6. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.