**Evaluation of Annotation Strategies:**

The literature review conducted highlighted validity concerns with regards to annotation and sampling. Sarcastic tweets were primarily sourced across the surveyed literature through identification of key hashtags such as ‘#sarcasm’. This strategy was shown to result in an unrepresentative sample of sarcastic tweets. This is likely to reduce performance on sarcasm detection models and thus true performance of such models is difficult to assess. One study found that only 15% of tweets labelled as sarcastic using this methodology were true labels- highlighting significant shortcomings to this strategy.

Datasets with a more robust sampling methodology exist, however there is little data available in these datasets which contain emojis. This works proposes a survey which aims create a more representative sample of sarcastic online content containing emojis. Previous work has been conducted to improve upon this poor annotation strategy however the majority of the tweets collected do not contain emojis; the iSarcasm dataset collected self-reported sarcastic and non-sarcastic tweets from participants, alongside a rephrase of the tweet in a more literal style. The work was subject to quality control by a linguistics professional therefore it is likely that the results are more representative of organic sarcastic content than the previously discussed sampling strategy.

**Primary Research Methodology:**

This primary research aims to adapt this dataset to evaluate emoji use patterns in sarcastic and non-sarcastic tweets. The adaptation will consist of the addition of emojis to known sarcastic and non-sarcastic text by survey participants. The methodology will collect quantitative data regarding demographics of the sampled population in addition to quantitative data regarding emoji use in sarcastic and control content. The components of the survey including sampling strategies, question selection and format and design optimisation are discussed in section X.

The goal of this work is to generate a dataset of verified sarcastic and control data which is richer in emojis. There are some limitations to this strategy; these are tweets where the author did not originally use emojis. While the use of emojis is known to be systematic in nature, it is a reasonable assumption that two people create a tweet with matching sentiment and pragmatic intent, where one uses emojis and the other does not. A more optimal approach would ask participants to self-report a sample of their tweets containing emojis as sarcastic and non-sarcastic to generate a dataset rather than providing prompts due to the possibility of irrelevant text to the participant reducing the quantity of usable results for each participant. The issue of relevance is addressed by enabling participants to not assign classifications to text which is not relevant or understood by them. The alternative of submission of sample text containing emojis with classifications of sarcastic and non-sarcastic would likely yield more relevant results to the participant however this approach significantly increases the effort required from participants which would decrease response rate and possibly reduce the likelihood of the task being completed as instructed.

**Survey outcomes**

*Demographic Analysis*

Raw results for the survey can be found in section X. The survey yielded 87 responses which can be deconstructed as follows:

A screenshot of a video game

Description automatically generated

*Figure X* Survey responses deconstructed by demographics.

Responses were not distributed evenly across subsets of the population, with gender skewing towards female and younger individuals. The shape of the age-related data is logical given that emojis are used disproportionately by younger people. No individuals were identified for participation over 65 years of age, which is likely attributable to the same observation. The age distribution of the survey does however seem to align reasonably well with the distribution of Twitter users. Given the age bins are not aligned it is plausible to conclude that the survey respondents follow a very similar age distribution to organic Twitter users, a result which may be desirable in this context. While a positive result has been observed for age distribution, gender alignment to Twitters user-base is not closely aligned with a global gender distribution skewing 70.4% towards male users. English speaking countries with reportable data cite a greater proportion of female users (averaging 41.5%). Given the contradictory skew in gender representation within the survey data, evaluation must be carried out to identify any differences which correlate to gender.

A blue graph with black border

Description automatically generated

*Figure X* Age distribution of Twitter users.

*Evaluation of Classifications*

Given the 1:1 split of speculative labels for each survey, the distribution of sarcastic and non-sarcastic assignments is encouraging for the overall quality of the data returned by the participant population due to one noteworthy implication; although participants were explicitly informed of the surveys focus on their assessment of sarcasm within the text prompts, this does not appear to have impacted the classifications. This point addresses a consideration with regards to survey validity; the comprehension of the underlying nature of a study could conceivably introduce variance in the subjects’ responses. Hence, the potential for such an influence warrants consideration in the evaluation of the implications of any conclusions drawn.

The distribution of data within figure X additionally lends credibility to the data with respect to its incongruent skew in gender distribution compared to the Twitter user-base; this parameter is uninfluenced by gender where usable classifications are applied. Both gender and age were found to have minimal impact on the rates of sarcasm reported in responses.

A blue and orange bars

Description automatically generated A colorful graph with black border

Description automatically generated

*Figure X* Classification results breakdown by gender and age.

A screenshot of a graph

Description automatically generated

*Figure X* Distribution of labels assignments for each survey response. Where ‘I don’t know’ was selected 15.84% of the time (approximately 1-2 times per survey). Considering only sarcastic and non-sarcastic labels and controlling for unassigned text 68% of respondents answered within 2 responses of a 1:1 ratio, aligning closely with speculative labels.

Inter-annotator patterns may be evaluated through the consideration of the quality control questions. Where text prompt is controlled, comparative work between participants is more intuitive. Agreement was not achieved in many cases, which is expected where context or the individuals’ personal beliefs are of greater relevance to the perception of the text prompt. Consider the following prompt where the beliefs of the participant were likely to have an impact on their response:

*Vaccine dose 1. Thank you, science.*

Responses labelled sarcastic:

*Vaccine dose 1. Thank you, science.* 💀

*Vaccine dose 1. Thank you, science.* 😜

Responses labelled non-sarcastic:

*Vaccine dose 1. Thank you, science.*👍

*Vaccine dose 1. Thank you, science.*👍

*Vaccine dose 1. Thank you, science.* 💛

However, in cases where the text prompt does not represent content that has ties to personal beliefs agreement between annotators was high:

Text prompt:

*was not aware that Crocs were appropriate business casual attire.*

Responses were universally labelled sarcastic:

*was not aware that Crocs were appropriate business casual attire.*

*was not aware that Crocs were appropriate business casual attire.* 😒

*was not aware that Crocs were appropriate business casual attire.* 😅

*was not aware that Crocs were appropriate business casual attire.*😂

*was not aware that Crocs were appropriate business casual attire.* 🙄

*was not aware that Crocs were appropriate business casual attire.* 🙄

*was not aware that Crocs were appropriate business casual attire.*🤔

Such an observation indicates that successful sarcasm detection models must implement highly sophisticated models which can weight the degree to which topics are tied to beliefs which are polarised across the population. This adds additional complexity to the necessary context awareness, making this a challenging problem set to overcome.

**Statistical Analysis of Emoji Use Frequency in Sarcastic Content**

The following work aims to establish which structural, or sentiment parameters have statistical significance when identifying sarcasm in short form text prompts. Section X discusses the selection methodology for statistical tests selected in each case.

*Table X* Statistical Analysis of Potential Structural Markers of Sarcasm.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sarcastic Content** | **Non-Sarcastic Content** | **Unknown Content** | **Statistical Significance between Sarcastic and Non-Sarcastic Labels** |
| **Frequency of Emoji Use (Average Emojis per String)** | | | |
| 0.980 | 0.607 | 0.297 | Yes |
|  |  |  |  |
|  |  |  |  |