MLPC 2025 Task 2: Data Exploration Report

Team Fumbling

Abdalaziz Ayoub **Abdulkarim Al Jamal** **Beibarys Abissatov** **Jeronim Bašić**

Contributions

Beibarys Abissatov answered the questions in task 1 (Case Study) & first part of task 2 (Annotation quality). Abdulaziz Ayoub did the rest of the coding for task 2, as well as task 4 (Text Features). Jeronim Bašić created the Github repository and did the coding for task 3 (Audio Features), Abdulkarim Al Jamal oversaw all tasks and prepared the report/presentation. In task 5, final conclusions were drawn from the entire team.

1 Case Study

To start things off, we try to get a feel for the dataset by exploring different audio recordings. We settled on two recordings, namely 185070.mp3 & 637068.mp3, both of which had multiple annotations made by several annotators.

1.1 Identify similarities or differences between temporal and textual annotations from different annotators.

Looking at the first file, we see that one annotator assumed the sound to be “A person blowing into a microphone”, whereas another annotator described it as a “whistling howl of the wind”, which indeed more accurate for that respective audio file. Temporal annotations for the 2 sound bites were quite similar for both annotators, only differing by < 100 ms.

	task_id	filename	annotator	text	onset	offset	filename_unsafe
2973	161982141	185070.mp3	368090603076295830415716083237311379068312527 1...	A person loudly blowing into a microphone	0.145015	6.145000	185070_howling_windwav.mp3
7241	161982141	185070.mp3	368090603076295830415716083237311379068312527 1...	A person loudly blowing into a microphone	9.679734	15.788481	185070_howling_windwav.mp3
7841	161982141	185070.mp3	630409381365151609149664936059703401607179984 8...	Eerie, whistling howl of the wind	9.706305	15.788503	185070_howling_windwav.mp3
29409	161982141	185070.mp3	630409381365151609149664936059703401607179984 8...	A eerie, whistling howl of the wind	0.126056	6.224011	185070_howling_windwav.mp3

The second file was annotated much more, producing a data frame too large to include in the report. Here, temporal differences were more significant for both annotators, as one individually annotated each beat produced by a drum (although <1 second apart) while the other recognized this as a repeating sound and used a single region to annotate it.

1.2 To what extent do the annotations rely on or deviate from keywords and textual descriptions in the audio’s metadata?

In file 185070.mp3’s metadata, we see the keywords “owling, whistling, wind” and a description of “howling wind”, which was similar to the one of the annotator’s work, suggesting a reliance on these keywords. while the other annotator deviated from these keywords or did not use them.

The second file 637068.mp3 contains keywords such as “atmosphere, Nanterre, field-recording, France, guitar, etc.” and

a description of “One-man-band playing ronroco and drums, indoor. I made this recording during Parades Festival, which is held in Nanterre (near Paris, France) every year at the beginning of June…”. These provide useful but also unnecessary information to the annotators, although they seem to have focused only on audible information.

1.3 Was the temporal and text annotations done according to the task description?

In the first file, both annotations were done according to the task description. However, in the second file, the annotator who non-descriptively annotated each individual instance of a repeating drum sound doesn't fulfill the criteria of a good textual annotation, unlike the other annotator who made fewer, but more descriptive annotations for the different instruments.

2 Annotation Quality

2.1 Audio recordings annotated by multiple annotators

Here, we only view audio recordings annotated by multiple annotators in order to compare them.

2.1.1 How precise are the temporal annotations?

We analysed annotation data to identify overlapping annotation durations among files. We first calculated the annotation duration by subtracting the onset from the offset, then we detected the instances of unique overlapping between annotations using a precision threshold of 100 ms, to achieve the following results: “Total number of files with 2 annotators: 725; Maximum number of overlapping regions: 20 (for audio file: 560530.mp3); Number of files with at least 1 overlapping region with 2 annotators: 446 Total number of files with 3 annotators: 6 Maximum number of overlapping regions: 6 (for audio file: 582364.mp3); Number of files with at least 1 overlapping region with 3 annotators: 4.”, meaning there is an agreement for 61.5% of the cases with two annotations within the given threshold. This is not very precise but perhaps explained by files, such as 560530.mp3, containing complex or multiple sound events, which can be interpreted differently by the several annotators.

2.1.2 How similar are the text annotations that correspond to the same region?

Here we calculated the cosine similarity of text embeddings for sound event annotations overlapping within the mentioned threshold to get the following results: “mean similarity with 2 annotators: 0.4271726608276367; mean similarity with 3 annotators: 0.3451352119445801”, which makes for some level of agreement, although lesser among triple-annotator files.

2.2 Complete dataset of audio recordings

2.2.1 How many annotations did we collect per file? How many distinct sound events per file?

We visualized the number of annotations per file with a histogram (bins=30). We see that a majority of the files contained <20 annotations, with around 6412 containing <5 annotations.

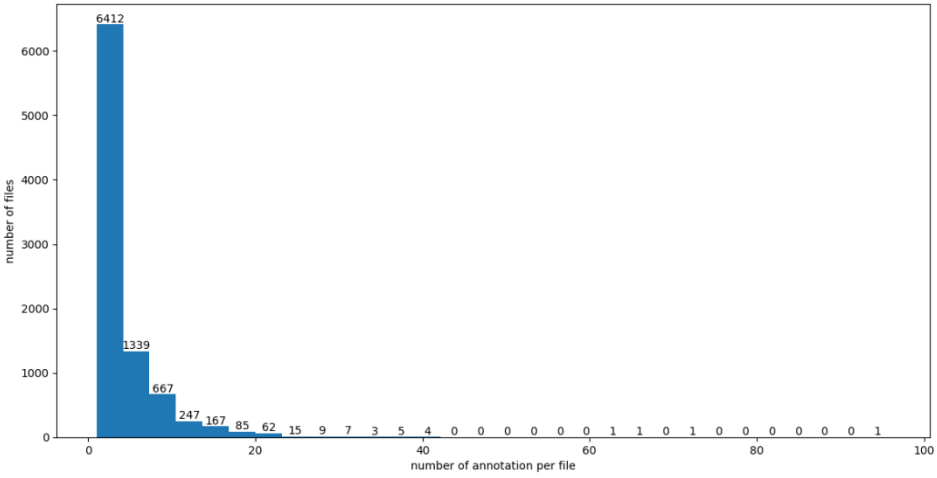
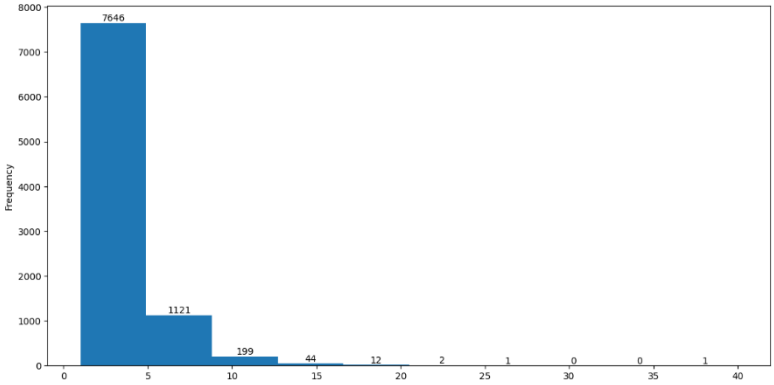


Figure 1: Histogram for no. of annotations per file

In order to calculate the number of distinct sound events, we make the following assumption: instances of repeating sound events (e.g. a dog barks loudly) will be annotated multiple times using the identical text. Therefore, we compute the number of unique textual annotations to get a ballpark figure of the number of distinct sound events and their number



of occurrences. Over 7500 files contained 1-5 sound events, with only 3 with 20> sound events.

Figure 2: Histogram for no. of annotations per file

2.2.2 How detailed are the text annotations? How much does the quality of annotations vary between different annotators?

To determine annotation quality, we calculated both the cosine similarity & dot product (identical results) to output a similarity score between the given embeddings of the audio annotations and the ones of the metadata for all annotators. We then calculated the length of characters in each annotation, creating box-plots for both these values, before plotting them against each other in an attempt to identify a correlation.

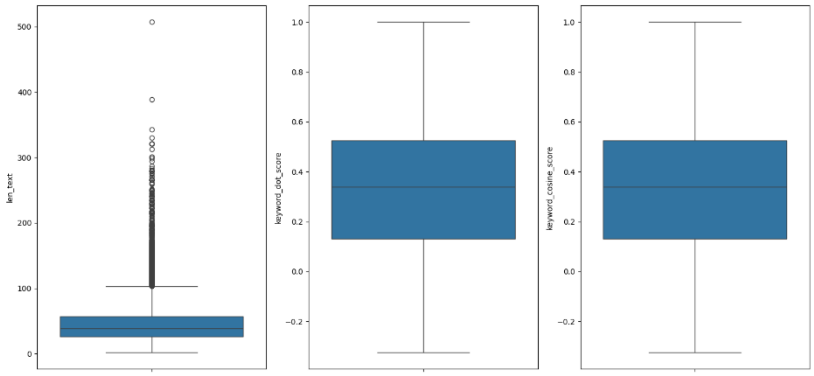


Figure 3: Box-Plots for summary statistics of annotations length & similarity score

3

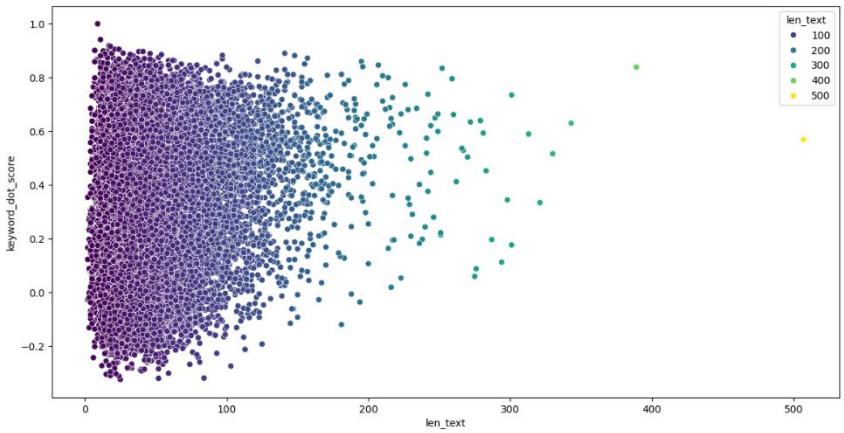


Figure 4: Scatter plot of of annotations length vs similarity score.

For annotation quality, we similarly used the cosine similarity scores for all annotators to calculate both their mean scores and standard deviation across each respective set of annotations and plotted this to attain the fifteen best/worst and most/least consistent annotators. We then identified them according to their annotator IDs

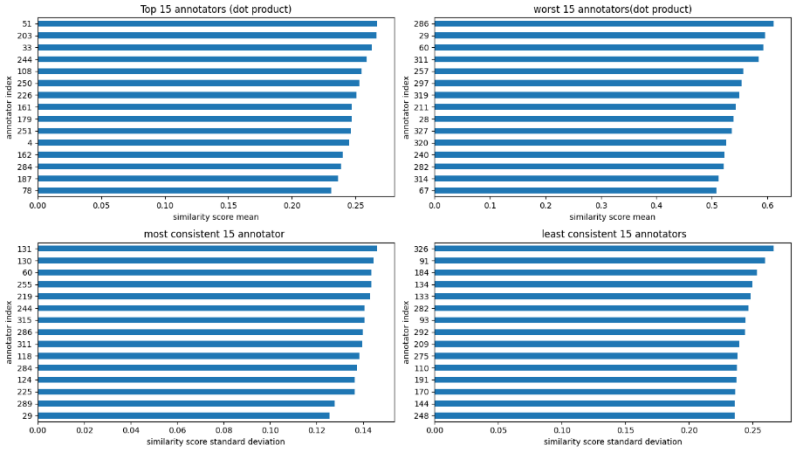


Figure 5: Annotation quality & consistency across annotators.

2.2.3 Are there any obvious inconsistencies, outliers, or poor-quality annotations in the data? Propose a simple method to filter or fix incorrect or poor-quality annotations (e.g., remove outliers, typos, or spelling errors).

We notice several outliers in the box-plots, with some annotations containing over 300 characters. However, according to the scatter, these did not necessarily register a higher similarity score. One annotation, consisting of a single word, achieved a perfect similarity score of 1. Upon further inspection, this annotation just so happened to exactly contain the only word listed in the metadata, hence being a perfect match. Most of the lower quality annotations also contained fewer characters. Therefore, a simple method at filtering such instances would be to exclude the quadrant of negative similarity scores containing <100 characters.

3 Audio Features

3.1 Which audio features appear useful? Select only the most relevant ones or perform a down projection for the next steps.

4

We determined the most useful audio features based on their variance and correlation to each other, as ones with higher variance are generally more informative, and using feature combinations with low correlation eliminates redundancy.

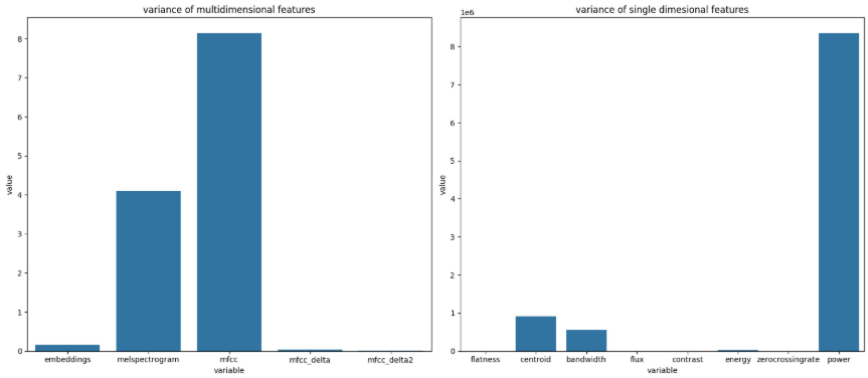


Figure 6: Variance of single & multi-dimensional features

According to variance, the most useful multidimensional features are MFCC & Melspectogram, but since they are highly correlated, only one of them should be chosen to train the model. For single-dimensional features: power & centroid are the most informative

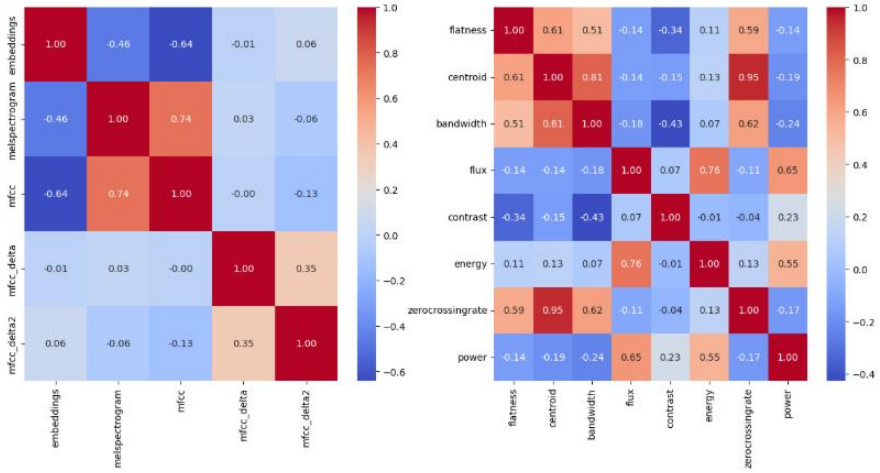
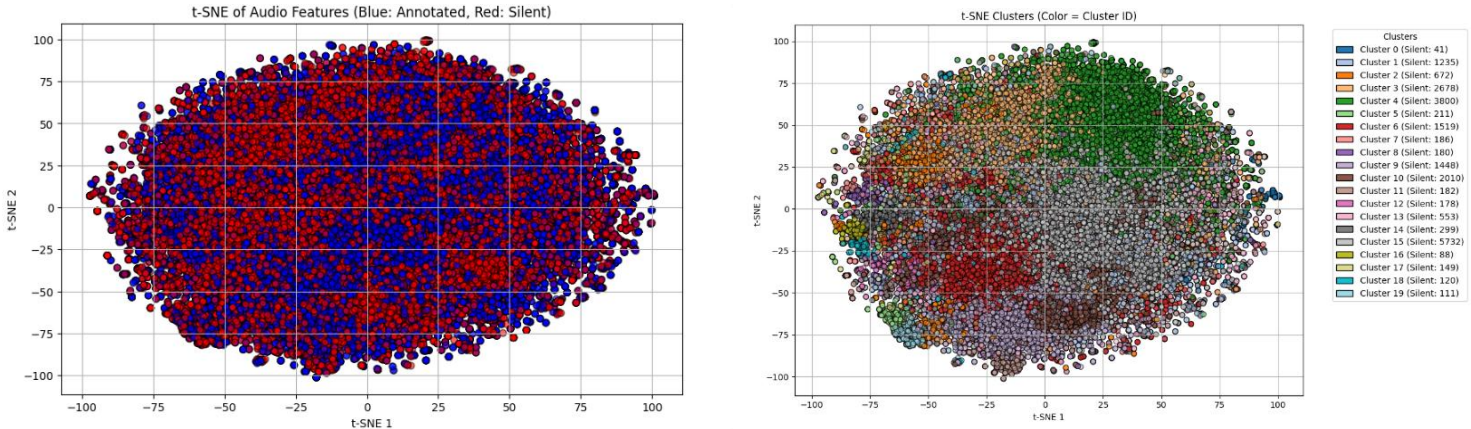


Figure 7: Correlation Heat Map

3.2 Extract a fixed-length feature vector for each annotated region as well as for all the silent parts in between. The most straightforward way to do this is to average the audio features of the corresponding region over time, as shown in the tutorial session.

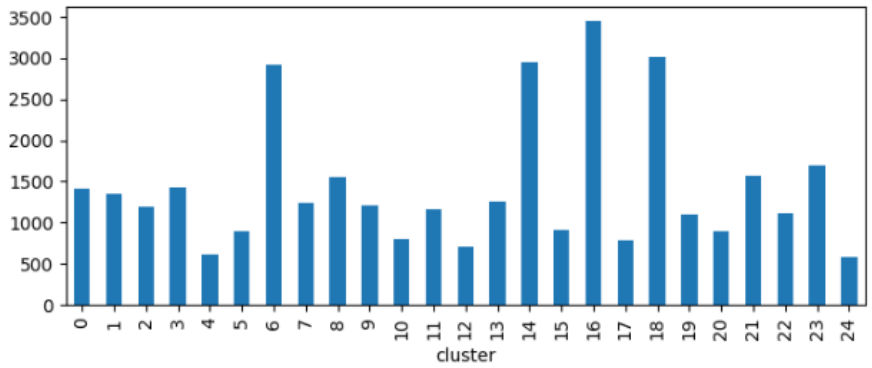
For each annotated and silent region, MFCC features were averaged over time using onset/offset boundaries. Silent segments were identified as gaps between annotations. All features were scaled using a standard scaler fit on the entire dataset. This resulted in fixed-length vectors representing both annotated and silent regions, enabling consistent downstream analysis like clustering.

3.3 Cluster the audio features for the extracted regions. Can you identify meaningful clusters of audio features? Do the feature vectors of the silent regions predominantly fall into one large cluster



computed. Most silent regions grouped into one dominant cluster, indicating similar acoustic characteristics, while annotated regions were more spread out across multiple clusters, reflecting their variability.

Figure 8: TSNE of normalized audio features & clusters



4 Text Features

4.1 Cluster the text features. Can you find meaningful clusters?

Here we clustered the text features with 25 cluster centers. We then apply a labeling function to find meaningful clusters

Figure 9: Text Feature Clusters

4.2 Design a labeling function for classes dog and cat. Do the annotations labeled as dog or cat sounds form tight clusters in the text and audio feature space?

Using the simple labeling function that checks for the following keywords ["dog","cat","meow","bark"], we filter the text features and notice 2 clusters, namely clusters 7 and 12.

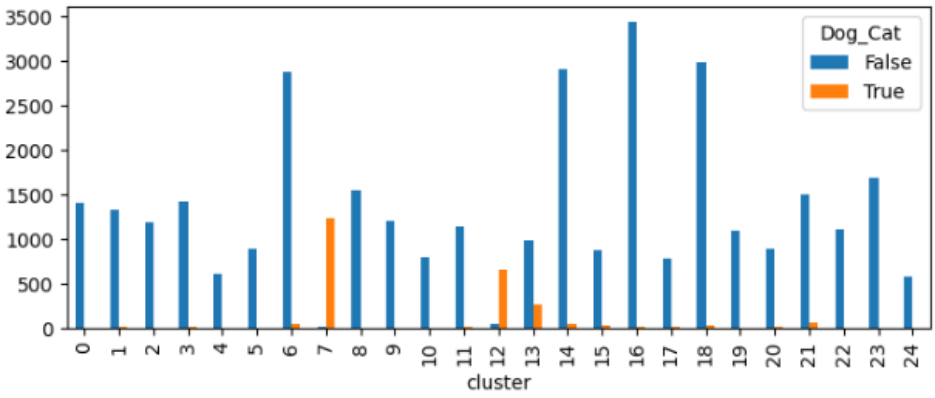


Figure 10: Text Feature Clusters after labeling

6

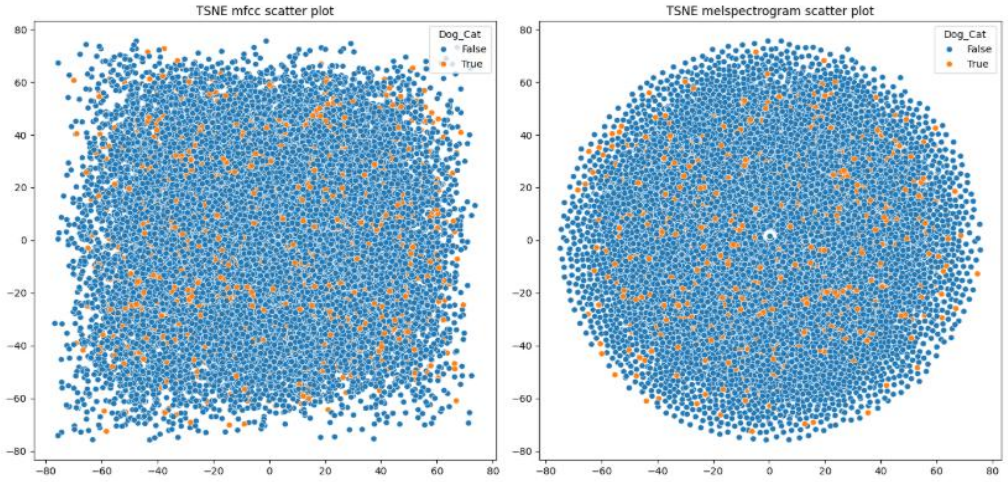
We then apply TSNE to visualize the clusters. We can clearly see that dog and cat labeled samples based on our labelling function form tight clusters.

Figure 11: TSNE of text clusters

4.3 How well do the audio feature clusters align with text clusters?

After visualizing each annotated region, the audio features do not form tight cluster and therefore do not align with text feature clusters. We believe this may be due to the presence of background noise in the annotated regions containing dogs and cats which may affect the audio features in general.

Figure 11: TSNE of audio clusters

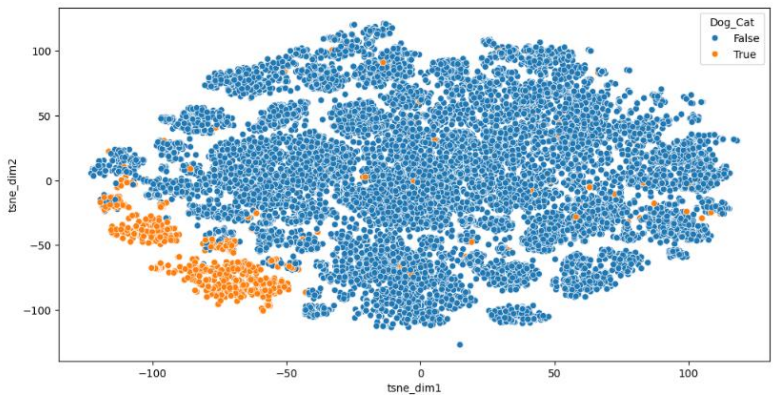


5 Conclusions

5.1 Is the dataset useful to train general-purpose sound event detectors?

According to the results of the previous task, there are no meaningful clusters in audio features for a given label, so it may be difficult to learn patterns from merely the audio feature space, but perhaps combining it with the text features may provide viable training data.

5.2 Which biases did we introduce in the data collection and annotation phase?



There may be a bias introduced when relying on the metadata (i.e. including information from metadata that cannot detected in audio, e.g. location) which may hinder model learning. Furthermore, the variation in annotation quality due to non-descriptive, inconsistent, annotations or the inclusion of several distinctive sound events in single annotation, will also have a negative effect on the training data. Lastly, language bias due to non-native English speakers, as well as the vocabulary of a particular age group amongst annotators may play a role.

8