Detection of Conversational Sarcasm

Ian Park: ian_park@sfu.ca
Hayyan Liaqat: hliaqat@sfu.ca
Austin Shaw: austins@sfu.ca

ABSTRACT

Detection of sarcasm is an ongoing problem in affective systems which rely on the understanding of an individual's true sentiment [10]. While there has been a significant amount of effort put towards detecting sarcasm in it's textual form [6], the research on multimodal sarcasm detection and specifically in-person sarcasm detection is scarce [6]. This paper proposes to look at sarcasm not in its textual form but as a social signal comprised of both vocal and facial indicators. To do this we have formed our own multimodal sarcastic dataset comprised of video clips from popular TV series and movies, and utilized a suite of classification algorithms of varying complexity to identify sarcastic expressions based on our proposed feature set. Performing classification in such a way has the benefit of circumventing any need for complex natural language processing. While only moderately effective, our research shows the potential of detecting conversational sarcasm on vocal and facial features alone and opens the door for further research to be explored.

KEYWORDS

sarcasm, datasets, neural networks, long short-term memory networks, affective systems, random forests, K-nearest neighbour, logistic regression, OpenFace, classification, feature extraction

1 INTRODUCTION

Sarcasm is a form of expression that at its core is derived from our use of language. A sarcastic phrase or comment is formed through the use of verbal or written irony to either mock another individual or display contempt [2]. It is believed that sarcasm was developed as a means to invoke criticism without coming off as aggressive [7]. Sarcasm often relies on the surrounding context to make sense of it. For example, the phrase *Wonderful weather we're having today!* would be a positive statement if the weather was indeed nice outside, but in the presence of rain the phrase could be interpreted as sarcastic.

It is these kinds of linguistic nuances and contextual properties that make sarcasm detection a challenge to any affective system. Even people do not develop the intuition to detect contextual sarcasm until the age of 7, and people with mental disorders may never be able to detect the social signal [7]. Due to the challenge of detection, sarcasm poses a huge threat to any system which relies on understanding the true sentiment of an individual [10]. Any automated customer service agent may frustrate the user if the bot cannot pick upon on the customers' sarcastic tone [10]. Thus it is critical that sarcasm detection is researched and implemented in any system designed to understand the true feelings of individuals it interacts with.

Because of the explosion in social media and other web focused services, much of the work done on sarcasm detection focuses on the textual modalities [10]. While this technique is useful in an online setting, it is not taking advantage of any potential physical cues that one might give in a sarcastic situation. Beyond the contextual irony, there exists a number of paralinguistic elements such as facial features and vocal qualities which help with sarcasm recognition [7]. Our team explores these paralinguistic elements as a means of detecting person-to-person conversational sarcasm without the ironic context. To do so we have created our own dataset of high quality video examples of sarcasm in its many forms. Utilizing early multimodal fusion on extracted facial and vocal features, we explore classification models of various complexities to determine the most effective method given the size of our dataset. Finally we reflect on our results, commenting on the nature of the dataset we have created and where any future work should focus on to improve on what was accomplished in this paper.

2 RELATED WORK

As mentioned previously, the majority of work relating to sarcasm detection is focused on its detection through textual and lexicographic means. A research team in 2016 surveyed the existing literature on automatic sarcasm detection and found that most of the sarcastic datasets publicly available consisted of short text examples, with long text examples being the second most common group of datasets [6]. A number of the short text datasets were formed using data from popular social media platforms such as twitter where the inclusion of hashtags provide an easy means of author driven annotation [6]. While rule based approaches were attempted, most of the research surveyed used supervised learning algorithms such as Support Vector Machines (SVMs) with a feature set derived from the sentiment, semantic similarity and emoticons from the text [6]. It was also noted that deep learning techniques have been gaining popularity with a few given examples of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) being used as means of sarcasm detection in recent years [6].

2.1 Multimodal Sarcasm Detection

Beyond the single modality of text, there have been a number of papers presented on the use of a multimodal feature set to improve sarcasm detection. One group in 2016 used the multimodal nature of image posts on various social media platforms [9]. Here the group trained both a SVM and neural network to identify sarcastic posts which involved both text and images [9]. The extracted features were combined in early fusion and it was found that the multimodal models outperformed their single modality counterparts [9].

In 2019 a group of researchers expanded the idea of multimodal sarcasm detection with the inclusion of video, speech and linguistic properties [2]. They developed what they dubbed the Multimodal

Sarcasm Detection Dataset (MUStARD) and used a SVM on multimodal features combined in early fusion to demonstrate the effectiveness of using multiple modalities [2]. This work provided the basis for our research as we expand upon the ideas put forward. Where our work differs is in the creation of the dataset, choice of modalities, and selection of features from the given modalities. The MUStARD dataset has a relatively low Kappa score of 0.5877 and upon inspecting the dataset a number of inconsistencies were identified [2]. Thus our team decided to create our own multimodal sarcastic dataset, picking out and editing some of the existing quality examples from the MUStARD set. Additionally, we go beyond using a single SVM and instead try a number of different models, both temporal and averaged, to try and fit the best classification algorithm to our dataset.

3 APPROACH

In this Section, we will go over our approach to the problem of multimodal sarcasm detection. This includes how we formed our multimodal dataset, what features we extracted from our examples and how we combined them in fusion, and a description of classification algorithms we used.

3.1 Data Collection

As demonstrated in the MUStARD dataset [2], the best choice of medium for extracting the desired modalities are video clips. Due to sarcasm and irony's prevalence in comedic television, our group collected short clips from many popular comedic dramas and sitcoms such as Friends, The Big Bang Theory, The Office, and Psych. Clips of Friends, which takes up a substantial 24% of the dataset, and other popular TV shows were gathered by downloading and editing popular YouTube compilations of moments from the hit sitcoms.

To download the YouTube videos we used the Python based command line tool *youtube-dl* and further edited the clips down using Microsoft's Photos tool (See Appendix B). In the process of editing, the context surrounding the sarcastic expression is removed. In addition to YouTube, clips were also extracted from the popular short clips site known as Vlipsy, which we had searched using the keyword *sarcasm*. The remaining clips were selected as the best quality examples of the MUStARD dataset and again further edited down. In total, we amassed 231 clips. All our data was collected from publicly available sources.

After data collection had been completed, each team member individually annotated portions of the dataset as either sarcastic or non-sarcastic. During the collection process, the clips were uniquely labeled from 0-230 and as each clip was annotated, it was either dropped into the sarcastic or non-sarcastic directory. After annotation, we had 127 clips labeled as sarcastic and 104 labeled as non-sarcastic behaviour. The dataset was scanned by multiple team members to ensure no duplicates had been collected. Our dataset is contained in both its raw and feature extracted form on our teams GitHub page (See Appendix A).

3.2 Feature Extraction

Previous research suggests that, in addition to the commonly known syntactic or lexical marker, both phonological and facial markers exist for detecting sarcasm [1]. Exploring this idea, both the facial and audio modalities are extracted to develop our feature set.

3.2.1 Facial Features. The important facial features for sarcasm detection include many of the Action Units (AUs) from the Facial Action Coding System (FACS), gaze direction, and head pose [1]. Our group used the open-source facial recognition software OpenFace to extract facial data for each frame in each clip (See Appendix B). OpenFace actually extracts both the intensity of the AUs (on a scale of 0-5) and their presence (0 for absent or 1 for present), but for our purposes we only use the intensity metric. Furthermore, if more than one face is detected in a clip, only the first face is considered and any frame with a sufficiently low confidence score is discarded. After preprocessing, if a clip has less than 10 frames, then it is discarded. In the case of our non-temporal models we take the average of each facial feature to obtain a single vector for each clip.

3.2.2 Audio Features. COVFEFE, a tool for feature extraction was used to extract the audio features [8]. The tool contains pipelines for extracting both temporal features, known as low-level descriptors, and features describing an entire audio file. The low-level descriptors split the audio into sequences, and generate features for each sequence. These sequences are similar to what frames are for a video. Approximately 80 features are extracted when the low-level descriptor pipeline is used. The other, non-temporal pipeline extracts about 1600 features. Some of the features extracted include Mel-frequency cepstral coefficients (MFCC), which is used for speech recognition, and pcm loundess, which is used for detecting high volume. Since the clips from our dataset were cut to contain only a single character speaking, we did not have to worry about isolating and extracting features from only cerain areas of the audio clips.

3.2.3 Early Fusion. Our modalities are combined in early fusion before being used to train our classification models. In the case of temporal models, our audio sequences turned out to be on average 3 times as large as our facial sequences. To remedy this we compressed our audio sequences by taking an average of neighbouring values in the sequence and constructing a new sequence which has been reduced in size by a given factor. For a factor of 3, audio sample x_i is averaged with the audio sample x_{i+1} and x_{i+2} and the resulting audio sequence is a third of the size. After performing this operation, the longer sequence out of the face and audio data is truncated so that their length is equal.

Early fusion for non-temporal models involves using principle component analysis (PCA) to reduce the feature set to 16 audio features and 12 facial features. These reduced vectors are then concatenated to form a feature vector of size 28.

3.3 Classifiers

3.3.1 Temporal Classifiers. Since both modalities are by nature temporal, we first looked at using a Long Short-Term Memory (LSTM) network. This variant on the RNN architecture is used as means of solving the vanishing gradient problem which usually plagues temporal based deep networks [4]. The LSTM was built using the open-source neural network library Keras on top of the machine learning platform TensorFlow (See Appendix B).

3.3.2 General Classifiers. Using the non-temporal audio and averaged facial features of each clip, several general purpose classifiers were trained and evaluated on our dataset. This included a Neural Network (NN), K-Nearest Neighbours (KNNs), Logistic Regression (LR), and a Random Forest classifier. The NN was again built using Keras while the Random Forest, LR, and KNN classifiers were created using the scikit-learn Python library (See Appendix B).

4 EXPERIMENTS AND RESULTS

In our experiments we aimed to demonstrate the potential of in-person sarcasm detection without the use of the textual modality or extended context beyond the uttered phrase of the sarcastic individual. Using a suite of classifiers we compare and contrast different methods to determine the most effective given our datasets size and composition.

4.1 Experimental Setup

For both the LSTM and NN there are a number of parameters to experiment with. We looked at the effects of changing parameters on the LSTM and noticed that in our case, there is essentially no change in the end accuracy of the model. Regardless, we went with an LSTM of 32 units, Adam optimizer with a learning rate of 0.0001, binary cross-entropy loss function, a batch size of 32, and epochs of 100. Additionally, early-stopping based on the validation accuracy of the LSTM was used to cut training short if the accuracy had not improved by at least 0.0001. This resulted in training for the LSTM being stopped at 9 epochs.

For the NN we went with a dense input layer of 32 units with a Rectified Linear Unit (ReLU) activation function, an additional dense layer of 8 units with again a ReLU activation function, and an output layer with a sigmoid activation function. Similar to the LSTM, our general classification NN used the Adam optimizer and a Binary Cross-entropy loss function. The NN was trained with 150 epochs. The other classifiers used a five-fold cross validation process. This process involved rotating the sets of training and test data such that by the end each classifier was comprehensively tested on on the entire dataset. For the KNN classifier we used a $K = 3$.

To measure the effectiveness of our classifiers we examine accuracy, precision, and recall as our metrics. Since our dataset is relatively symmetric (127 sarcastic clips vs 104 non-sarcastic clips), we will refer to accuracy as our main metric for classifier performance. But it is still important to know where the classification models struggle and as such we look at precision and recall as a means of diagnosing classifier performance.

4.2 Results

Performing training and validation on the models described above, we are given the results shown in Table 1. Based on our results, we can see that the Random Forest classifier performed better than the other classifiers. With an overall accuracy of 0.61, it scored 0.03 higher than any other classifier we tested. The LR performed only slightly worse to the Random Forest with an accuracy of 0.58. Both the LSTM and KNN classifiers tied in accuracy at 0.54. While the NN performed significantly worse than any other model with an accuracy of 0.45.

Classifier Performance			
Classifier	Accuracy	Precision	Recall
LSTM	0.54	0.54	**1.0**
NN	0.45	0.58	0.38
Random Forest	**0.61**	**0.61**	0.64
LR	0.58	0.60	0.59
KNN	0.54	0.56	0.63

Table 1: Performance measurements of our given classifiers, tested and trained on our multimodal sarcastic dataset

Another important result to highlight is the difference between accuracy, precision, and recall for each model. The Random Forest and LR classifiers both had scores which were fairly consistent across the three categories. Random Forest having the biggest disparity between the two with recall scoring 0.03 higher than accuracy or precision. Our KNN classifier favoured recall heavily with a score of 0.64, while our NN favoured precision with a score of 0.58 (0.2 higher than the model's recall score). The most interesting case is the LSTM classifier which had a perfect recall of 1.0. Given that the accuracy and precision are equal at 0.54, and that our data is comprised of 54% sarcastic clips, it would imply the model's strategy was to label each clip as sarcastic.

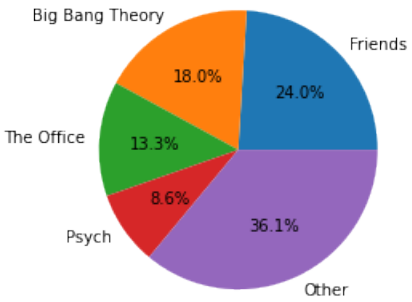

Figure 1: Dataset composition

4.3 Dataset Composition

In addition to evaluating the performance of our classification models, it is also important we discuss the composition of our dataset. As mentioned previously, our dataset contains 231 clips of which there is a roughly 54-46 split between sarcastic and non-sarcastic examples. Notice that in Figure 1, we can see the percentage distribution of clips from our various sources. Friends has the highest representation at 24%, but notice that not one show comprises a majority of our total dataset. Here the 'Other' category refers to clips which are either one-offs or only comprised of a few examples from any particular show or movie.

In Figure 2, we examine the distribution of Friends characters among our collected clips and notice that Chandler makes up a majority of total Friends clips. This is not surprising since Chandler is often seen as the comedic character in the group. Given that

Friends is our single largest repented show or movie, this means Chandler alone makes up 14% of clips in our dataset.

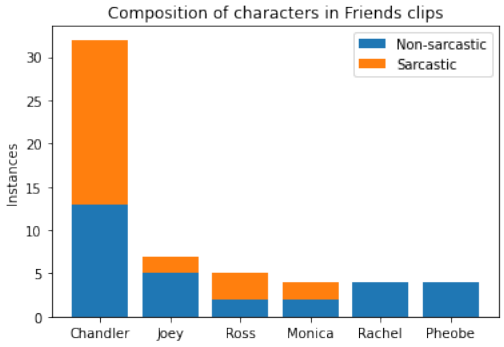

Figure 2: Representation of each of the main cast from Friends

5 DISCUSSION AND FUTURE WORK

5.1 Model Performance

Our classifiers saw mixed results when trained under our multimodal dataset. Our best performing classifier, random forests, saw a significant increase in accuracy over a majority pick baseline of 0.54 (+ 0.07). But the more complicated models such as our NN and LSTM performed significantly worse than this baseline or failed to develop any tangible ability to detect sarcasm. This is likely due to the relatively small size of our dataset, as deep neural networks require a large amount of data to train on. This is backed up by the fact that our LSTM was meant to train for 100 Epochs but due to our early-stopping, it only trained on 9.

Another concern that comes from our results is in which ways the models were accurate. Some of the classifiers trained on our dataset had a significantly higher recall versus precision. This implies a bias towards selecting any given expression as sarcastic. This is not a desirable trait of our classification models since, in reality, everyone uses sarcasm to different degrees [5]. Furthermore, an affective system which is more likely to assume sarcastic behaviour will perform poorly as a counsellor or customer service agent as it will not take the individuals' concerns seriously. This assumption is likely due to sarcastic clips taking up a majority of the total dataset.

5.2 Dataset

Looking further at our dataset we can see a number of biases that come out of its composition. Due to our focus on comedy TV as a sarcastic resource, most of our clips reflect the comedic style of these shows. The fact that Chandler from Friends is the subject of 14% of our total clips means that his sarcastic tendencies are highlighted in our dataset. For example, Chandler's classic emphasis of auxiliary verbs such as *be* or *get* as seen in many of his sarcastic phrases is useful for detecting his tendencies but would fail to recognize a deadpan sarcastic delivery common to many other less expressive characters [1]. While we gathered a large amount of clips from only a few source shows, we were sure to grab roughly equal examples of sarcastic and non-sarcastic expressions from each character. This attempts to mitigate the bias of associating one's voice or facial structure with sarcasm, and instead focus on picking up the social signals that come from a sarcastic expression.

Although our goal was to detect sarcastic expression through the use of non-textual modalities, some examples from our dataset highlight the difficulty of this task. The lack of distinction in facial features and tone of voice during a deadpan sarcastic delivery make it incredibly difficult to determine sarcasm without looking at what is being said. Additionally, sarcasm can be used in different contexts. It can be used as a playful mock, or in a state of complete disbelief or rage. This makes looking at facial and vocal features difficult since these can vary wildly between the two forms of sarcasm. See Appendix A for a list of examples mentioned in this Section.

5.3 Future Work

Our research gave us insight into the feasibility of multimodal sarcasm detection without the textual modality, but there is still much to improve on. To provide an accurate baseline for the performance of many deep neural networks, the size of the dataset needs to be increased. It would be hard to mitigate internal bias, but due to the long run time of many of the shows, examined in this paper, there is an opportunity to create a significantly larger dataset than the one presented.

Another area of study worth looking into would be multimodal sarcasm detection in different cultures and languages. There has been previous work done in sarcasm detection for a number of languages including Italian, Dutch, Czech, and Hindi but again, these works examine sarcasm through text [3]. It would be interesting to see multimodal datasets formed for different languages. From there, these different datasets could be combined to form a multi-linguistic multimodal sarcastic dataset which then could be examined using unsupervised learning to gain insight on the human and cultural tendencies which surround sarcasm.

6 CONCLUSION

Our research explored the potential of multimodal sarcasm detection without the textual modality. We formed our own multimodal dataset consisting of video clips of sarcastic behaviour from popular TV series and movies. Using this dataset, we explored sarcasm detection using audio and facial features combined in early fusion. Comparing a variety of classification models, we determined that a random forest classifier proved to be the most effective while more complex models, such as NNs and LSTMs require an increased sample size. We also note the many shortcomings of our dataset, the effects it had on our experiment, and observations made on the wide range of sarcasm behaviour. Finally, we note where our work can be expanded and further research is needed.

REFERENCES

[1] Salvatore Attardo, Jodi Eisterhold, Jen Hay, and Isabella Poggi. 2003. Multimodal Markers of Irony and Sarcasm. *Humor-international Journal of Humor Research - HUMOR* 16 (01 2003), 243–260. https://doi.org/10.1515/humr.2003.012

[2] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). *CoRR* abs/1906.01815 (2019). arXiv:1906.01815 http://arxiv.org/abs/1906.01815

[3] Abhijeet Dubey, Aditya Joshi, and Pushpak Bhattacharyya. 2019. Computational Sarcasm for Different Languages : A Survey.

[4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (12 1997), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735
[5] Stacey L. Ivanko, Penny M. Pexman, and Kara M. Olineck. 2004. How Sarcastic are You?: Individual Differences and Verbal Irony. *Journal of Language and Social Psychology* 23, 3 (2004), 244–271.
[6] Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2016. Automatic Sarcasm Detection: A Survey. *CoRR* abs/1602.03426 (2016). arXiv:1602.03426 http://arxiv.org/abs/1602.03426
[7] Maria Luisa Gorno-Tempini Maria Luisa Marc Sollberger Stephen M. Wilson Danijela Pavlic Christine M. Stanley Shenly Glenn Michael W. Weiner Bruce L. Miller Katherine P. Rankin, Andrea Salazar. 2009. Detecting Sarcasm from Paralinguistic Cues: Anatomic and Cognitive Correlates in Neurodegenerative Disease. *NeuroImage* 47 (06 2009), 2005–15. https://doi.org/10.1016/j.neuroimage.2009.05.077
[8] Majid Komeili, Chloé Pou-Prom, Daniyal Liaqat, Kathleen C. Fraser, Maria Yancheva, and Frank Rudzicz. 2019. Talk2Me: Automated linguistic data collection for personal assessment. *Plos One* 14, 3 (2019). https://doi.org/10.1371/journal.pone.0212342
[9] Rossano Schifanella, Paloma de Juan, Joel Tetreault, and LiangLiang Cao. 2016. Detecting Sarcasm in Multimodal Social Platforms. In *Proceedings of the 24th ACM International Conference on Multimedia* (Amsterdam, The Netherlands) *(MM '16)*. Association for Computing Machinery, New York, NY, USA, 1136–1145. https://doi.org/10.1145/2964284.2964321
[10] Xiaojun Wan Yitao Ca, Huiyu Cai. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2506–2515. https://doi.org/10.18653/v1/P19-1239

A DATASET AND EXAMPLES

Click on the following list items below for our dataset and examples.

- Multimodal sarcastic dataset (raw data)
- Multimodal sarcastic dataset (feature extracted)
- Example of Chandler's sarcastic tendencies.
- Example of comedic sarcasm.
- Example of deadpan sarcasm.
- Example of rage induced sarcasm.

B LINKS TO SOFTWARE PACKAGES

Click on the following list items below for links to the software packages used in this report.

- youtube-dl.
- Microsoft Photos.
- OpenFace.
- Keras.
- TensorFlow.
- scikit-learn.
- COVFEFE

C DATASHEETS FOR DATASETS

The following will answer the questions from Datasheets for Datasets which were not previously covered in our report.

Is any information missing from individual instances? No, there is no information missing.

Are relationships between individual instances made explicit. No, relationships between individual instances are not made explicit.

Are there recommended data splits? No, there are no recommended data splits.

Does the dataset contain data that might be considered confidential? No, the dataset uses publicly available footage.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

Does the dataset identify any subpopulations? No the dataset does not identify any subpopulations.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? Since everyone in the collected clips are payed actors for publicly released media it is fair to say that it is not possible to identify one or more natural persons in this dataset.

Does the dataset contain data that might be considered sensitive in any way? No, the dataset does not contain data that might be considered sensitive.

Over what timeframe was the data collected? The data was collected over the span of 2 weeks from July 1st to 15th.

Were any ethical review processes conducted (e.g., by an institutional review board)? No ethical review processes were conducted by any institutional review boards.

Does the dataset relate to people? While the dataset includes actors, it was retrieved from publicly available sources for which their consent is not required. Thus our dataset does not relate to people.

Has the dataset been used for any tasks already? No, only the work described in this paper utilizes this dataset.

What (other) tasks could the dataset be used for? Aside from sarcasm detection, this dataset could be used from analysis of popular TV shows or look at character specific tendencies.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? No, there is nothing about the way the dataset was preprocessed that would impact further use.

Are there tasks for which the dataset should not be used? No, there are no tasks which the dataset should not be used for.

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? No, the dataset will not be distributed to third parties outside the entity.

When will the dataset be distributed? With the submission of this report.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? . No, the dataset will not be distributed under a copyright or other IP license, and/or under ToU.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? Only those which apply to the Youtube terms of service found here.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? Only those which apply to the Youtube terms of service found here.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Please contact any of the emails given in this paper.

Is there an erratum? No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? The dataset will not be updated.

Will older versions of the dataset continue to be supported / hosted / maintained? Only one version of the dataset will be maintained.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? They can create a merge request on the GitHub page.