A New Task: Deriving Semantic Class Targets for the Physical Sciences

Micah Bowles[1][*] **Hongming Tang**[2] **Eleni Vardoulaki**[3]
Emma L. Alexander[1] **Yan Luo**[4] **Lawrence Rudnick**[5] **Mike Walmsley**[1]
Fiona Porter[1] **Anna M. M. Scaife**[1,6] **Inigo Val Slijepcevic**[1] **Gary Segal**[7,8]

[1]Jodrell Bank Centre for Astrophysics, University of Manchester, Manchester, UK
[2]Department of Astronomy, Tsinghua University, Beijing, China
[3]Thüringer Landessternwarte, Tautenburg, Germany
[4]School of Physics and Astronomy, Sun Yat-sen University, Zhuhai, China
[5]Minnesota Institute for Astrophysics, University of Minnesota, Minneapolis, USA
[6]The Alan Turing Institute, London, UK
[7]School of Mathematics and Physics, University of Queensland, Brisbane, QLD, Australia
[8]CSIRO Space and Astronomy, Epping, NSW, Australia

Abstract

We define *deriving semantic class targets* as a novel multi-modal task. By doing so, we aim to improve classification schemes in the physical sciences which can be severely abstracted and obfuscating. We address this task for upcoming radio astronomy surveys and present the derived semantic radio galaxy morphology class targets.

1 Introduction

Language evolves and changes - sometimes quite quickly. When a new idea arises and demands new terminology, terms are commonly invented [e. g. utopia; 9], named after an early adopter or pioneer (e. g. Newtonian physics), or adopted from similar ideas [e. g. 'modern' in philosophy and art; 11, 1]. This is also common across various physical sciences, where use of language is key, as it is even believed to affect how we think [12].

In computer science applications, language used to describe target classes is not a pressing issue at this point. Consider ImageNet [3], whose class targets are built on WordNet [5]. WordNet is explicitly constructed around the conceptual-semantic and lexical relations of the English language. Because of this, models trained on ImageNet carry inductive biases of the semantic (meaningful) terms used as class targets. These inductive biases do not necessarily hold true for data sets in the physical sciences, where in fewer than one hundred years new language has been developed to capture entirely novel and crucial concepts.

The physical sciences are quickly entering large data regimes where automation is essential. This includes classification. Supervised classification approaches usually have definitions for class targets and terms to describe to which class a given data point belongs.

To improve supervised models, it is common in both computer science and applied fields to attempt to optimise approaches and models for a fixed set of class targets. Often, computer science led state of the art methods are implemented in the hope of surpassing previous benchmarks on a specified (fixed) supervised task. However, in dynamically evolving fields, fixing such class targets in place may not

---

[*]micah.bowles@postgrad.manchester.ac.uk; `https://mb010.github.io/`

be ideal or provide meaningful results. For instance, in radio astronomy, the field has developed a detailed understanding of radio galaxies and how they form, yet the same abstract classes defined through the field's understanding in the 1970s [4] still persists. We therefore propose that rather than optimising predictions of ineffective class targets, in certain scenarios it may be more beneficial to *change the target classes* with the aim of developing more robust, generalisable and feature rich models. Consequently, in this work, we propose a task to derive semantic class targets.

Sec. 2 details the proposed task and its potential consequences for the physical sciences. Sec. 3 presents the proposed method. An application of the method to radio astronomy is presented in Sec. 4 before conclusions are drawn in Sec. 5. Code and data used in this work are available at `https://github.com/mb010/Text2Tag`.

2   Task

To improve target classes in labelled data sets, we propose a multi-modal task which can be phrased as:

> Given a set of documents describing labelled data samples, return a set of natural language terms / phrases which capture the semantic features of the labelled data set.

For any task, the derived set of class labels should be able to:

1. Map the science targets,
2. Map the semantic features of the data,
3. Use clear (non-technical) language.

The set of targets must be able to *map to the previous set of classes*, as otherwise a given scientific community will not be able to translate classifications into the historical classes that they are used to. For example, 'fur length' as the class target in the supervised task of classifying cats and dogs; although useful, it does not suffice to classify a given image back into the cat/dog scheme.

Targets which map the *semantic features* of the data are ideal, as populations which contain semantic feature differences may not be captured by abstract classes. For example, a classifier could be trained to predict features of buildings (spires, column designs, materials used, etc.) rather than architectural styles (gothic, baroque, neoclassical, brutalist etc.). This would enable the model to generalise to architectural styles not included in the abstract target classes, and could even be used to highlight designs which include hybrid elements.

The benefit of *clear non-technical language* is the ability it provides experts in a given field to capture, communicate, and collaborate in and around their data. If the terms map the science targets sufficiently well, they could even replace terms reducing that community's dependence on obtuse, and sometimes inconsistent, definitions of technical terminology. It could also lower barriers to entry for inter-disciplinary research, outreach, and citizen science projects.

3   Method

The methods we discuss here are based on annotations and science targets. We use the term *annotations* to describe short documents which each describe a feature of a single data sample using non-technical terminology. We use the term *science targets* to mean the traditional abstract classifications (or engineered features) for each annotated data sample.

There are many possible approaches to address this task. Two simple approaches include manual selection of plain English terms by a panel of experts, or using a large language model (LLM) for a zero-shot approach. We expect, given appropriate experts, that manual selection via an expert panel would be acceptable to a given community. However, expecting a panel of experts to agree on a set of plain English class targets may not be realistic depending on the background of each expert and/or their ability to distil abstract concepts into simple terms. Manual selection may also lack the reproducibility and tractability that the physical sciences should demand. Using a LLM in a zero-shot approach may work; however, it is not clear how prompts should be engineered in order to extract

semantic class names at scale, or if the scientific community could be convinced to accept the results of such an opaque process.

Here we propose a third method that derives class names in a bag-of-words paradigm using a pre-trained language model with embedded tokens, and a random forest (RF). *Conceptually*, we assign each annotation the token which most closely resembles the encoding of all 'similar' annotations. A random forest is then trained to predict the abstract science targets from the derived tokens. The newly derived class targets are determined by calculating the Shapley values (estimating feature importance) of each input token and selecting the most impactful tokens as the targets. This method only results in individual tokens for each annotation. These can be manually adjusted if their original annotations clearly refer to a more specific idea which may require multiple tokens / specific grammar to capture.

The *details* of our approach are as follows. We define $\mathbf{a}_j$ as the jth annotation of the M annotations in our corpus. Using a pre-trained model, f_{emb}, the annotations are embedded into a k-dimensional vector through $f_{\text{emb}}(\mathbf{a}_j) = \mathbf{v}_j \in \mathbb{R}^k$. This is currently implemented in a bag-of-words paradigm, where the order of the words does not affect the encoding. More explicitly, in the case of annotations longer than one word, $\mathbf{v}_j$ is the average vector of each embedded word within the annotations.

A similarity value is calculated for each pair $(i, j) \in [1, M]^2$ of the M entries. The similarity between two vectors is calculated using the cosine similarity, $g_{\text{cs sim}} : \mathbb{R}^k \to [-1, 1]$. M averaged vectors are then calculated according to a similarity threshold σ as

$$\mathbf{v}'_i = \frac{1}{M'} \sum_{j=1}^{M} \begin{cases} \mathbf{v}_j & \text{if} \quad g_{\text{sim}}(\mathbf{v}_i, \mathbf{v}_j) \geq \sigma \\ 0 & \text{else,} \end{cases}$$

where M' is the number of non-zero elements in the summation. The model originally used to embed the annotations, f_{emb}, is then used to produce the token closest to $\mathbf{v}'_i$: $f_{\text{emb}}^{-1}(\mathbf{v}'_j) \approx s_q$, where s_q is the qth entry of all Q unique derived s. Note that $Q \leq M$ as tags are derived from the M annotations.

We define $\mathbf{t} = (t_1, ..., t_Q)$ as a vectorised encoding of the assigned derived class targets and train a model, f_y, to predict values of each science class $y \in Y$, where Y is the set of considered science classes. For each model, f_y, and tag representation, t_q, we calculate an importance

$$f_{\text{Importance}}(t_q, f_y) = I_{(q,y)} \in \mathbb{R},$$

where larger values of $I_{(q,y)}$ mean that s_q is is more important to the classification output. We use Shapley values as a proxy for importance.

To recover the importance of the qth tag, we take the support weighted average across $I_{(q,y)} \forall y \in Y$:

$$I_q = \frac{1}{Q} \sum_{y \in Y} n_y I_{(q,y)},$$

where n_y is the number of positive occurrences of class y. We normalise the importance values, I_q, such that

$$\sum_{q=1}^{Q} I'_q = 1, \quad \text{and} \quad I'_q \in [0, 1].$$

Finally, we sort the tags, S_q, according to their respective I'_q. While most tags will have non-zero I'_q in this ranking, the majority of the information is seen to be contained within the highest ranked tags. Furthermore, the lowest ranked tags are expected to contain significant noise. We recommend making a selection cut on the minimum importance a tag can have to be included in a final $Q' < Q$ tags. These most important tags, consisting of Q' strings $s \in S_{\text{targets}}$, are the derived semantic class targets.

Some of the tags may require clarification to improve their clarity. If a tag is not immediately clear, we suggest considering the raw annotations from which that tag was derived in order to verify what it represents and make adjustments as, or if, necessary.

4 Application to Radio Galaxy Morphology

Radio galaxy classification schemes are complex, require explicit definition as their use can vary from person to person, and rely on a shared set of features. However, these classifications are essential for
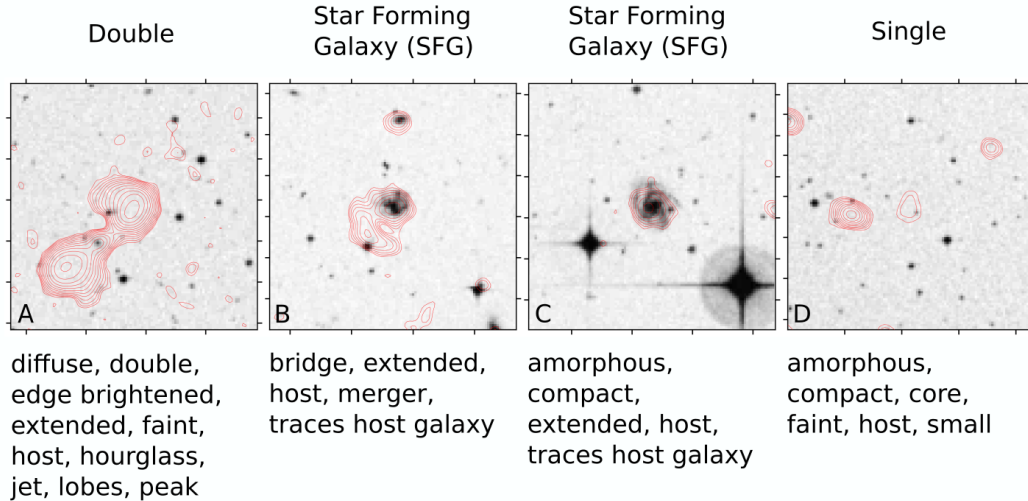
Double	Star Forming Galaxy (SFG)	Star Forming Galaxy (SFG)	Single
diffuse, double, edge brightened, extended, faint, host, hourglass, jet, lobes, peak	bridge, extended, host, merger, traces host galaxy	amorphous, compact, extended, host, traces host galaxy	amorphous, compact, core, faint, host, small

Figure 1: Example of science (above) and semantic (below) classifications in radio galaxy morphology. The classes all describe radio morphology of the source associated with the central emission in each frame. Radio EMU pilot data (red contours) overlayed on optical DSS data. We make no statements on the accuracy of these classes, but simply present them

studying populations of radio galaxies and their properties in order to probe physical environments and conditions which cannot be recreated on earth. Working towards the Radio Galaxy Zoo EMU citizen science project (RGZ EMU; Tang & Vardoulaki et al. in prep), the RGZ EMU team collected annotations and science classifications for 299 data samples. These samples consist of radio maps of galaxies as well as optical [DSS; 6] and infra-red [WISE; 13] backgrounds to help users understand which radio components belong together. The radio maps are cutouts from the Evolutionary Map of the Universe [EMU; 7, 8] pilot survey and were subject to selection effects designed to sample galaxies with interesting morphologies from simple catalogues.

In total, 5 experts made 1,257 multi-label classifications using provided science targets. The presented science targets were selected from a compiled list of radio galaxy morphology classes presented in [10]. These are: Single, Double, Classical double, Triple, Narrow-angle tail (NAT), Wide-angle tail (WAT), Bent tail, Fanaroff & Riley Class 1 (FR I), Fanaroff & Riley Class 2 (FR II), Fanaroff & Riley Class 0 (FR 0*), Hybrid, X-shaped, S-shaped, C-shaped, Diffuse, Double-double (DDRG), Core-dominant, Core-jet, Compact Symmetric Object (CSO), 1-sided, Odd Radio Circle (ORC), and Star Forming Galaxy (SFG). The plain English annotations were made by 19 users, who collectively entered 2,920 descriptions of the presented samples for a total of 8,486 annotations.

By applying the approach presented in Sec. 3 and making hyperparameter selections based on the performance of the random forests through cross validation, we derive 22 terms. In alphabetical order, these are: amorphous, asymmetric brightness, asymmetric structure, bent, bridge, compact, core, diffuse, double, edge brightened, extended, faint, host, hourglass, jet, lobe, merger, peak, plume, small, tail, and traces host galaxy. These will be used in the RGZ EMU project and in the value added EMU catalogues. We present four sources in Fig. 1, highlighting the stark difference between the science classes and the newly assigned semantic classes.

5   Conclusion

We propose and define the novel task of deriving semantic class targets. The task can be approached in many different domains, and could have far reaching effects across the physical sciences when addressed.

We present a tractable and reproducible method using annotations of, and respective science targets for, a given data set. Our method's implementation is domain-agnostic and publicly available: `https://github.com/mb010/Text2Tag`.

4

We present a first application of this method to radio galaxy morphology classification, where highly technical and disputed science classes are common. We derive 22 semantic plain English class targets. These will be implemented in the upcoming RGZ EMU citizen science project.

Impact Statement

We hope that addressing this task will have far reaching impacts in the physical sciences. In data-driven scientific domains, the science targets may be a severe limitation in deep learning, where representations are learned to map targets, even if they are abstracted classes [2]. Altering the class targets to be more semantic may be a significant advantage as models are required to fit the *semantic features* which define the abstract. We expect this to result in more robust mappings of real data, where the representations which the model maps are clear features, even if in some regions the respective abstract classes might vary.

We hope that models trained in this way, will provide improvements to transfer learning tasks, robustness in representation learning, and domain specific foundation models across the physical sciences. Furthermore, in scenarios where language is a barrier to future research, domains within the physical sciences may benefit from partially freeing themselves from overly constrained and abstracted language to improve the accessibility and utility of complex scientific concepts for all parties including domain experts, inter-disciplinary collaborators, and the general public. Negative impacts of this work may be assumed effectiveness leading to the loss of valuable scientific information through the oversimplification of classification targets.

References

[1] T. Adajian. The Definition of Art. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2022 edition, 2022.

[2] M. Bowles, M. Bromley, M. Allen, and A. Scaife. E(2) Equivariant Self-Attention for Radio Astronomy. *NeurIPS 2022: Machine Learning and the Physical Sciences Workshop*, Nov. 2021.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] B. L. Fanaroff and J. M. Riley. The morphology of extragalactic radio sources of high and low luminosity. *Monthly Notices of the Royal Astronomical Society*, 167:31P–36P, May 1974.

[5] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, 1998.

[6] B. M. Lasker, C. R. Sturch, B. J. McLean, J. L. Russell, H. Jenkner, and M. M. Shara. The guide star catalog. i-astronomical foundations and image processing. *The Astronomical Journal*, 99:2019–2058, 1990.

[7] R. P. Norris, H. T. Intema, A. D. Kapińska, B. S. Koribalski, E. Lenc, L. Rudnick, R. Z. E. Alsaberi, C. Anderson, G. E. Anderson, E. Crawford, R. Crocker, J. English, M. D. Filipović, T. J. Galvin, A. M. Hopkins, N. Hurley-Walker, S. Inoue, K. Luken, P. J. Macgregor, P. Manojlović, J. Marvil, A. N. O'Brien, L. Park, W. Raja, D. Shobhana, T. Venturi, J. D. Collier, C. Hale, A. Hotan, V. Moss, and M. Whiting. Unexpected circular radio objects at high Galactic latitude. *Publications of the Astronomical Society of Australia*, 38:e003, Jan. 2021.

[8] R. P. Norris, J. Marvil, J. D. Collier, A. D. Kapińska, A. N. O'Brien, L. Rudnick, H. Andernach, J. Asorey, M. J. I. Brown, M. Brüggen, E. Crawford, J. English, S. F. u. Rahman, M. D. Filipović, Y. Gordon, G. Gürkan, C. Hale, A. M. Hopkins, M. T. Huynh, K. HyeongHan, M. James Jee, B. S. Koribalski, E. Lenc, K. Luken, D. Parkinson, I. Prandoni, W. Raja, T. H. Reiprich, C. J. Riseley, S. S. Shabala, J. R. Sheil, T. Vernstrom, M. T. Whiting, J. R. Allison, C. S. Anderson, L. Ball, M. Bell, J. Bunton, T. J. Galvin, N. Gupta, A. Hotan, C. Jacka, P. J. Macgregor, E. K. Mahony, U. Maio, V. Moss, M. Pandey-Pommier, and M. A. Voronkov. The Evolutionary Map of the Universe pilot survey. *Publications of the Astronomical Society of Australia*, 38:e046, Sept. 2021.

[9] J. Romm. More's strategy of naming in the utopia. *The Sixteenth Century Journal*, 22(2):173–183, 1991.

[10] L. Rudnick. Radio Galaxy Classification: #Tags, Not Boxes. *Galaxies*, 9(4):85, Oct. 2021.

[11] F. von Schelling, F. von Schelling, and F. W. J. S. Schelling. *On the history of modern philosophy*. Cambridge University Press, 1994.

[12] P. Wolff and K. J. Holmes. Linguistic relativity. *Ltd. WIREs Cogn Sci*, 2:253–265, 2010.

[13] E. L. Wright, P. R. M. Eisenhardt, A. K. Mainzer, M. E. Ressler, R. M. Cutri, T. Jarrett, J. D. Kirkpatrick, D. Padgett, R. S. McMillan, M. Skrutskie, S. A. Stanford, M. Cohen, R. G. Walker, J. C. Mather, D. Leisawitz, I. Gautier, Thomas N., I. McLean, D. Benford, C. J. Lonsdale, A. Blain, B. Mendez, W. R. Irace, V. Duval, F. Liu, D. Royer, I. Heinrichsen, J. Howard, M. Shannon, M. Kendall, A. L. Walsh, M. Larsen, J. G. Cardon, S. Schick, M. Schwalm, M. Abid, B. Fabinsky, L. Naes, and C.-W. Tsai. The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. *The Astronomical Journal*, 140(6):1868–1881, Dec. 2010.