Journal of Physics Special Topics
An undergraduate physics journal

A2 1 Physics Special Topics Predictions

R. McElroy, A. Hawkins, F. Stacey, M. Robertson, S. Dias

Department of Physics and Astronomy, University of Leicester, Leicester, LE1 7RH

October 28, 2026

Abstract

A topic modelling technique is used to identify 77 topics across 1,192 physics papers published in Physics Special Topics. These are grouped into 13 broad themes and historical trends are used to forecast the expected thematic composition of 2026 papers. Astronomy (18.9%) and Mechanics (16.43%) are expected to be the most popular.

Introduction

The subject matter of *Physics Special Topics* varies highly. Using data [1] taken from volumes between 2009 and 2025, we attempt to predict the topics that will emerge in the 2026 edition. This is done using an advanced topic modeling tool, BERTopic [2]. It uses transformer based language embedding and density clustering to identify topics across the existing papers. We aggregate these into broad themes and give an estimate for the proportion of articles that will appear in each theme in 2026.

Methodology

To collate the data, we used an automated browser-based retrieval script to systematically access and download each article from the PST archive. The `.pdf` files were reduced to a plain text dataset. Formatting and linguistic artefacts like headers, bibliographies and page numbers were removed. Key scientific content from the title, abstract, body, figure captions, and data was retained. The cleaned text of each paper was converted into a dense vector representation using the default BERTopic sentence-transformer model, all-MiniLM-L6-v2 [3]. This embedding is dimensionally reduced using UMAP [4], which preserves useful structures within the data. A density based clustering algorithm, HDBSCAN [5], is applied to identify common themes among the papers. The model identified 77 topics using words that are representative of a cluster relative to the other data. A further 218 papers were assigned an outlier class. The 77 *fine* topics were manually mapped into 13 broad themes given in Table 1. Without this grouping, the diversity of papers made it difficult to provide a stable set of themes for forecasting. Topics were grouped based on their dominant physical subject matter, for example Astronomy & planetary science includes black holes, exoplanets, lunar dynamics, and planetary orbits. Mechanics & motion includes vehicles, impacts, rotational motion, projectiles, and forces.

Modelling and Backtesting

To predict the thematic distribution of papers in 2026, several models were evaluated by doing a rolling origin backtest. Forecasting each year using only historical data and computing the RMS error using the actual data allowed us to test several models:

- historical mean (theme averaged across all preceding years)
- three year mean (theme averaged across the last three years)
- five year mean (theme averaged across the last five years)
- linear regression model (least squares regression model extrapolated one year ahead)

The rolling backtests accounted for *observed* years. The previous year was used in 2015, as the journal was not published so there was no data. Although other models fitted some individual themes better, the historical mean model produced the lowest overall RMS error in the backtest. The annual proportions each theme takes are displayed in Figure 1. Data leakage from topics assigned using future papers was considered insignificant for this test.

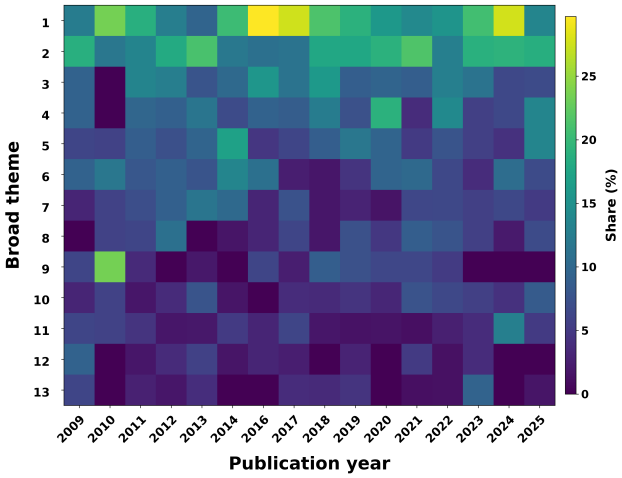


Figure 1: Heatmap showing the proportion of papers assigned to each broad physics theme for each publication year. Lighter cells indicate a higher proportion of papers in that theme. Theme names are matched to numbers in Table 1.

Results

Table 1 shows the expected distribution of papers for 2026. This is very similar to the historical values. The highest RMS error across these was 6.5% (Astronomy & Planetary). Note that this excludes the 218 out of 1192 total papers that were assigned an outlier class.

Table 1: Predicted proportion of papers in each broad theme for 2026.

Broad theme	**Forecast (%)**
1 Astronomy & planetary science	18.90
2 Mechanics & motion	16.43
3 Flight & aerodynamics	10.05
4 Waves, sound & acoustics	9.20
5 Magnetism & electromagnetism	8.27
6 Climate & Earth systems	7.90
7 Fluids & hydrodynamics	5.86
8 Relativity & spacetime	4.95
9 Other/applied physics	4.88
10 Thermal & statistical physics	4.44
11 Nuclear & radiation physics	4.07
12 Medical & biological physics	2.56
13 Energy & power	2.50
Total *(excluding outliers)*	**100.00**

Discussion and Conclusion

This analysis shows that BERTopic [2] can identify the broad structure of the *Physics Special Topics* corpus. The 77 fine-grained topics capture specific subjects, while the 13 broad themes provide a simpler representation for studying changes over time. The backtesting results showed that the historical mean gave the lowest overall RMS error of the forecasting methods tested. This suggests that, for this dataset, a simple estimate is appropriate for predicting how many papers will fit each theme. There are several limitations to the analysis. In particular, 218 of the 1,192 papers were classified as BERTopic outliers and were excluded from the broad-theme proportions. The manual mapping of fine-grained topics to broad themes also introduces some subjectivity. The values given in Table 1 should be interpreted as estimates of the expected thematic distribution of papers classified by the model. Future work might involve testing more sophisticated models, modeling individual themes independently or applying tuning to BERTopic to better classify the papers into initial topics.

Supporting Material

BERTopic [2] was used for this project. Documentation is available on GitHub at `https://maartengr.github.io/BERTopic/index.html`. Claude Sonnet 5 and OpenAI GPT-5.6 Luna were used as python debugging aids.

References

[1] University of Leicester Open Journals. *Physics Special Topics*. 2008-2025.

[2] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.057s*, 2022.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[4] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[5] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer, 2013.