Multi-label Classification of Scientific Research Documents Across Domains and Languages

Autumn Toney-Wails
Georgetown University
autumn.toney@georgetown.edu

James Dunham
Georgetown University
james.dunham@georgetown.edu

Abstract

Automatic organization of scholarly literature is a challenging but essential task. In particular, assigning key concepts to scientific publications allows researchers, policymakers, and the general public to search for and discover relevant research. But any meaningful organization of scientific publications must evolve with new research, requiring up-to-date and scalable text classification models. Additionally, scientific research publications benefit from multi-label classification, particularly with more fine-grained sub-domains. Prior work has focused on classifying scientific publications from one research area (e.g., computer science), referencing static concept descriptions, and implementing English-only classification models. We propose a multi-label classification model that can be implemented in non-English languages, across all scientific literature, with dynamic concepts.

1 Introduction

Maintaining an up-to-date organization of scientific literature in any domain requires an automated approach—a comprehensive and real-time solution for a constant influx of text data. Specifically, research publications require characterization or indexing in order to be searchable and accessible to researchers, policymakers, and the public. Many academic databases and publishers maintain a taxonomy that authors or editors reference in order to manually assign topics, research fields, or concepts to scientific publications. Yet, manual labeling is notoriously laborious and error-prone. Automation is necessary to accurately label documents with taxonomy concepts in a timely manner.

Here, we focus on scientific publication classification based on Microsoft Academic Graph's field of study taxonomy (Shen et al., 2018). This taxonomy contains a hierarchy of scientific concepts (fields of study) to organize scholarly literature. Our objective is to design an **updatable** and **scalable** multi-label classification model that is independent of manual annotation or input language. We experiment with scientific research documents in English and Chinese, as these are by far the two most frequent languages for publications in our database.

Our work leverages a multi-lingual knowledge base, Wikipedia, in order to obtain up-to-date concept descriptions in English and other languages. Using MediaWiki's API, we first locate an English concept's Wikipedia page and are then able to find the corresponding page in other languages (MediaWiki, 2022). Hence, a multi-lingual knowledge base provides multi-lingual concept descriptions without requiring any direct translating of the concept taxonomy or concept descriptions.

We represent both the concept descriptions and research publications text data in embedding form. By using vector space representations of text (word embeddings) we can compute the cosine similarities between concept embeddings and publication embeddings, with the cosine similarity score indicating the relevance of a concept to a publication. In this way, we are able to compute either one top field (most similar) or multiple fields of study that are relevant (determined by a similarity score threshold for the task at hand) to a given publication. A multi-label classification model is a practical approach to scientific publication classification, as most scientific research publications are relevant to more than one field of study, particularly at the more granular level of fields. For example, a publication can be relevant to *natural language processing* and *machine learning*.

We implement our multi-label classification model in English and Chinese, generating field descriptions, embeddings, and field-to-publication similarity scores in each language. Our database of scholarly literature contains more than 184 million documents in English and more than 44 million

Proceedings of the Third Workshop on Scholarly Document Processing, pages 105–114
October 12–17, 2022.

documents in Chinese, which serve both as input text for word embeddings and as target publications for classification. Applying our scientific publication word embeddings and field of study descriptions from Wikipedia, we compute field embeddings for 313 different fields of study, and publication embeddings for the scientific research publications in English and Chinese.

Because we do not have a manually annotated, ground-truth dataset with field labels assigned to publications, we provide extensive evaluations of our results and include a case study on artificial intelligence and machine learning publications.

The contributions of the paper are summarized as follows: 1) word embeddings in English and Chinese, trained on a comprehensive set of scholarly literature, 2) a scientific text classification model not restricted to the English language, and 3) a Python library for updating field embeddings and models in sync with changes to underlying field definitions (from Wikipedia articles and the sources they cite), to address conceptual drift. All results and code will be made public in our GitHub repository[1].

2 Related Work

Classifying text according to a defined taxonomy is applied across a wide range of domains, such as patents, news articles, and scientific literature, using numerous machine learning approaches. Text classification for scientific literature typically involves text extraction, topic modeling, or citation graphs to cluster related documents (Aljaber et al., 2010; Tsai et al., 2013; Yau et al., 2014; Kim and Gil, 2019). Prior research that uses a predefined taxonomy for multi-label classification is generally limited to one broad area of research, and selecting a dataset with annotated publication data (i.e., a dataset limited to a classification scheme).

Santos and Rodrigues reference the Association for Computing Machinery (ACM) Concept Classification System (CCS) to assign multiple concept labels to computer science papers (Santos and Rodrigues, 2009). The authors crawl relevant web pages to identify concept-related descriptive text and implement three different classification models: Binary Relevance, Label Powerset, and Multi-Label k-Nearest Neighbors (Santos and Rodrigues, 2009). Similarly, Mustafa et al. reference the ACM CCS, but use Word2Vec embeddings to represent scientific research publication text and cosine similarity to compute a similarity score and determine concept assignment (Mustafa et al., 2021).

Shen et al. generate a six-level scientific document taxonomy for all of science. Using Word2Vec and term frequency-inverse document frequency (TF-IDF) embeddings trained on scientific publication titles and abstracts, Shen et al. generate field of study embeddings and publication embeddings. Each scientific publication is assigned multiple field labels using cosine similarity between the publication embedding and the field embeddings (Shen et al., 2018).

3 Data

We use three datasets in our model: 1) scientific research documents, 2) a scientific research field of study taxonomy, and 3) a knowledge base.

3.1 Scientific Research Documents

In this work, we use a comprehensive set of scientific research documents that we compiled from six scholarly literature databases: Clarivate's Web of Science (WOS), Digital Science's Dimensions[2] (DS), Microsoft Academic Graph (MAG), arXiv, Papers with Code (PWC) and the Chinese National Knowledge Infrastructure[3] (CNKI). There is no common publication identifier across these six datasets, so we deduplicate publications to generate a merged corpus of scholarly literature.

We deduplicate documents in a two-step process illustrated in Figure 1. In step one, we extract six document identifiers (DOI, citations, normalized abstract, normalized author names, normalized title, and publication year) for each document. To normalize the document abstracts, author names, and titles, we implement the Normalization Form Compatibility Composition standard, which decomposes Unicode characters by compatibility and recomposes them by canonical equivalence. We de-accent letters, strip copyright signs, HTML tags, punctuation, non-alphanumeric characters, and numbers, and remove white space from the strings. If any three identifiers between documents are equal, we assign those documents a

[1] https://github.com/georgetown-cset/scientific-field-classification

[2] Data sourced from Dimensions, an inter-linked research information system provided by Digital Science http://www.dimensions.ai

[3] All China National Knowledge Infrastructure content is furnished for use in the United States by East View Information Services, Minneapolis, MN, USA

unique merged ID.

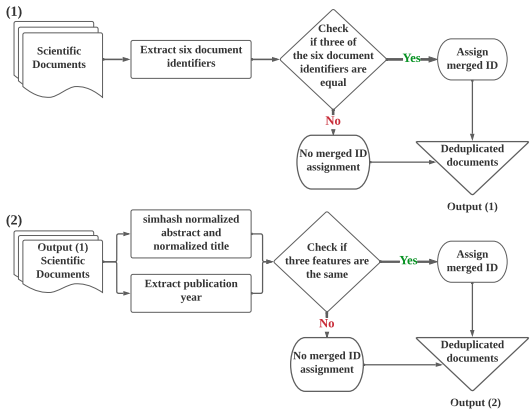

Figure 1: Scientific document de-duplication process.

In step two, we use the SimHash fuzzy matching algorithm with a rolling window of three characters in order to match articles that were published in the same year and have similar abstracts and titles (Manku et al., 2007). Articles matched in step two are also assigned a merged ID. Articles that do not have a distinct merged ID assigned in either deduplication step are included in the final corpus as unique documents.

From the deduplicated set of scientific research documents, we generate a set of English documents, EN-PUBLICATIONS (184,381,319 publications), and a set of Chinese documents, ZH-PUBLICATIONS (44,166,696 publications), using Chromium Compact Language Detector 2 (CLD2). Each document is represented by the text available from the title and abstract; if both title and abstract are present then the text is concatenated.

3.2 Field of Study Taxonomy

We use MAG's Field of Study (FoS) taxonomy, which contains six levels (0 through 5) of fields. Level 0 ("L0") represents the most broad fields, such as *computer science* and *medicine*, and Level 5 ("L5") represents the most granular fields, such as *key clustering* and *gene density*. FoS L0 and L1 were derived from Science-Metrix classification scheme[4] and refined manually by the authors, whereas and L2-L5 were automatically identified (Shen et al., 2018).

In this study, we select the 19 L0 and the 294 L1 FoS as our target classification scheme; L1 FoS are sub-domains of L0. In Table 1 we display all 19 L0 FoS with several examples of their L1 child FoS. We denote the total number of L1 FoS under each L0 in parentheses next to their label. *Medicine* has the most L1 child FoS, with 45, followed by *engineering* with 44 and *economics* with 40.

The FoS taxonomy we reference in this study defines the fields: their names and parent/child relations. All FoS in this taxonomy are provided in English only.

3.3 Knowledge Base

For our knowledge base we use Wikipedia, an open-collaboration online encyclopedia accessible for free, with articles published in 327 languages (Wikipedia, 2022). We access Wikipedia articles through MediaWiki's API (MediaWiki, 2022). Given the English Wikipedia page title for a field (if known) or otherwise the field name in English, we query the Mediawiki API for metadata on any such page in English Wikipedia. Specifically, we request its `langlinks` property, which describes corresponding pages in other languages/Wikipedias. In this way, the English FoS can be linked to any language of interest without manual translation, making Wikipedia an ideal knowledge base for our multilingual classification model.

In Figure 2 we display a portion of the Wikipedia articles for *natural language processing*, in English and Chinese. We use the full-body text in the article, as well as the publication titles and abstracts listed in the "References" section.

4 Field of Study Classification Model

Our field of study multi-label classification model is adapted from MAG's scientific publication classification scheme, with key design modifications. In Shen et al.'s model, the descriptive text used to generate L0 and L1 FoS embeddings are titles and abstracts from sets of scientific publications for each field, in which the publications are selected from a sample of unknown journals and conferences (Shen et al., 2018). For one of their embeddings set the authors generate Word2Vec vectors.

We use Wikipedia article text and reference publications for L0 and L1 FoS descriptive text. In this way, field descriptions can be replicated, extended to languages other than English, and updated as the fields evolve. We describe in this section the project workflow to process our data and design our field of study classification model. Figure 3 shows the high-level pipeline to produce the field

[4] http://science-metrix.com/en/classification

Art (displaying 6 of 6 L1) *Aesthetics*, *Art History*, *Classics*, *Humanities*, *Literature*, *Visual Arts*	***History*** (displaying 6 of 7 L1) *Ancient History*, *Archaeology*, *Classics*, *Economic History*, *Ethnology*, *Genealogy*
Biology (displaying 7 of 32 L1) *Anatomy*, *Animal Science*, *Bioinformatics*, *Botany*, *Genetics*, *Immunology*, *Zoology*	***Materials Science*** (displaying 5 of 7 L1) *Ceramic Materials*, *Composite Material*, *Metallurgy*, *Nanotechnology*, *Optoelectronics*
Business (displaying 6 of 13 L1) *Accounting*, *Actuarial Science*, *Commerce*, *Finance*, *International Trade*, *Marketing*	***Mathematics*** (displaying 6 of 20 L1) *Algebra*, *Combinatorics*, *Geometry*, *Mathematical Optimization*, *Statistics*, *Topology*
Chemistry (displaying 5 of 21 L1) *Biochemistry*, *Food Science*, *Mineralogy*, *Organic Chemistry*, *Radiochemistry*	***Medicine*** (displaying 7 of 45 L1) *Audiology*, *Cancer Research*, *Nursing*, *Orthodontics*, *Pediatrics*, *Surgery*, *Virology*
Computer Science (displaying 5 of 34 L1) *Algorithm*, *Artificial Intelligence*, *Database*, *Internet Privacy*, *Parallel Computing*	***Philosophy*** (displaying 6 of 7 L1 *Aesthetics*, *Epistemology*, *Humanities*, *Linguistics*, *Religious Studies*, *Theology*
Economics (displaying 5 of 40 L1) *Accounting*, *International Trade*, *Management*, *Political Economy*, *Socioeconomics*	***Physics*** (displaying 5 of 27 L1) *Astronomy*, *Geophysics*, *Nuclear Physics*, *Quantum Mechanics*, *Thermodynamics*
Engineering (displaying 5 of 44 L1) *Aeronautics*, *Control Theory*, *Nuclear Engineering*, *Simulation*, *Systems-Engineering*	***Political Science*** (displaying 3 of 3 L1) *Law*, *Public Administration*, *Public Relations*
Environmental Science (displaying 4 of 8 L1) *Agricultural Science*, *Agroforestry*, *Environmental Planning*, *Environmental Protection*	***Psychology*** (displaying 5 of 14 L1) *Cognitive Science*, *Criminology*, *Neuroscience*, *Psychiatry*, *Social Psychology*
Geography (displaying 6 of 11 L1) *Archaeology*, *Cartography*, *Forestry*, *Geodesy*, *Meteorology*, *Regional Science*	***Sociology*** (displaying 5 of 13 L1) *Anthropology*, *Demography*, *Ethnology*, *Gender Studies*, *Media Studies*, *Political Economy*
Geology (displaying 6 of 18 L1) *Climatology*, *Earth Science*, *Geophysics*, *Hydrology*, *Oceanography*, *Petrology*	

Table 1: The 19 L0 Fields of Study and a sample of their child fields (L1). Next to each field is the number of L1 FoS displayed and the total number of child fields.

and document embedding outputs necessary for our classification model.

We describe each step in our classification model pipeline as follows:

Step 1: To normalize the scientific publication text, we remove all punctuation and numeric tokens. For languages that are case-sensitive, we set all text to lowercase. For example "COVID-19" is transformed to "covid19" in English. The normalized texts are used as inputs for the TF-IDF and fastText embeddings in Step 2.

Step 2: With the normalized scientific publication text (from Step 1) as input, we produce TF-IDF embeddings using `TfIdfTransformer` from gensim and 250-dimensional fastText word embeddings using the skipgram model (Rehurek and Sojka, 2011; Bojanowski et al., 2017). TF-IDF provides a measurement of how important a word is to a document based on the word's occurrences in the entire document. FastText word embeddings encode n-grams in a vector space that represents semantics. Since both vector representations of words (TF-IDF and fastText) are determined by the input corpus, it is necessary to use a representative corpus for the task at hand.

Step 3: For each of the 19 L0 and 294 L1 FoS, we retrieve the corresponding associated text (page content and reference publications) in Wikipedia, which we refer to as descriptive field text. Combining the Wikipedia page text and scientific publication text we aim to capture both definitions and exemplar research for a given field.

Step 4: We compute field TF-IDF and fastText embeddings using the embedding sets from Step

Figure 2: Sample Wikipedia article on Natural Language Processing

2 and the descriptive field text from Step 3. We follow the procedure in "Algorithm 1" to generate TF-IDF and fastText embeddings for each FoS.

Algorithm 1 Full Text to Single Embedding

Input: Word embedding dictionary, E
Text, t
Output: Single text vector, $\vec{t}$

procedure EMBED_TEXT(t, E)
  V = [] ▷ Empty array to store word vectors
  for $word$ in t **do**
    if $word$ in E`.keys()` **then**
      $\vec{w} = E[word]$
      V.append($\vec{w}$)
    end if
  end for
  $\vec{t}$ = `sum(`V`, axis=0)`
  $l2$ = `linalg.norm(`$\vec{t}$`, 2, axis=0)`
  if $l2 == 0$ **then return** $\vec{t}$
  else $\vec{t} = \frac{\vec{t}}{l2}$
  end if
  return $\vec{t}$ ▷ The text vector is $\vec{t}$
end procedure

Step 5: Separate from FoS embeddings in Step 4, we compute entity embeddings. We generate these for a FoS or publication as the average over the embeddings of each FoS mention in its text.

Step 6: Using "Algorithm 1", we compute document embeddings for each scientific research publication in our corpus.

Step 7: We use cosine similarity to compute a similarity score for each document compared to each FoS, for each embedding set (TF-IDF and fastText). Our similarity score is the average of the two cosine similarities. The cosine similarity between two vectors is defined as:

$$\cos(\vec{f}, \vec{d}) = \frac{\mathbf{f} \cdot \mathbf{d}}{\|\mathbf{f}\|\|\mathbf{d}\|} \quad (1)$$

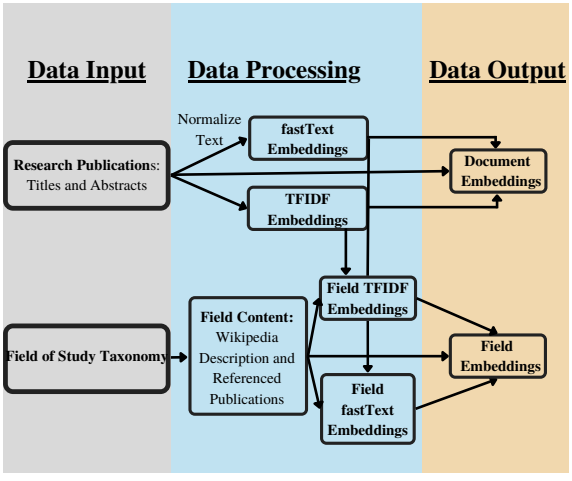

Figure 3: Process to generate document embeddings and three sets of FoS embeddings

Here, $\vec{f}$ represents a FoS embedding and $\vec{d}$ represents a document embedding. Cosine similarity returns a value between 0 and 1, with 0 indicating no similarity and 1 indicating perfect similarity. By computing cosine similarity for all FoS and document pairs, we can choose if we want to label a document with only one field (the most similar FoS), or set a similarity score threshold and assign multiple fields. This is particularly useful with

more granular fields. For example, a publication can be relevant to *computer vision* and *machine learning* L1 FoS.

5 Experiments

We perform Steps 1-7 on EN-PUBLICATIONS and ZH-PUBLICATIONS. Text normalization and embedding generation (Steps 1-2) may require different tools and packages depending on the choice of non-English languages; we use `jieba` for Chinese text processing.

For knowledge base information retrieval (Step 3), we reference MAG's FoS metadata for field ID, field name, field level, and field Wikipedia page. The field of study attributes metadata includes English Wikipedia URLs for all fields. We query MediaWiki with the assigned Wikipedia pages for each FoS in English to store the descriptive text and search for the corresponding page in Chinese. This results in several outcomes that we detail below for non-English implementations of our model:

1. **The Wikipedia page does not exist** (maybe it once did; maybe not). We fall back to searching Wikipedia for this term (in a second API request), in case there exists a near match. We store these "near-match" results for manual review to ensure they are accurate.

2. **The desired English Wikipedia page exists but the `langlinks` property does not include a link to a corresponding page on Chinese Wikipedia.** We store the English page name and page ID, and leave the Chinese page fields blank to flag for manual review.

3. **We find the desired English page and a linked Chinese page**. We store each page name and page ID, for the English and Chinese results.

With the completed links between FoS and Wikipedia pages, we are able to retrieve the descriptive text from Wikipedia pages and the text from referenced publications. At this stage in the process, the Chinese implementation is self-contained and no longer relies on any data linkages in English, which would be the case for any non-English language implementation.

We generate document embeddings for each scientific document in EN-PUBLICATIONS and ZH-PUBLICATIONS, and we generate FoS embeddings and entity embeddings for our English and Chinese

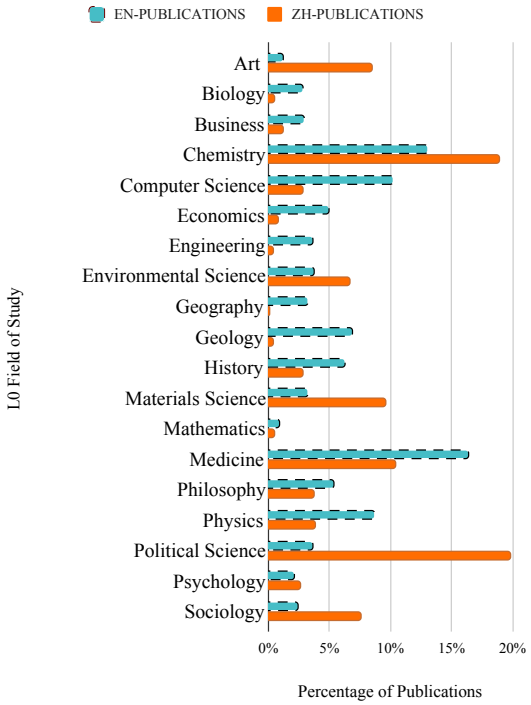

Figure 4: Percentage of papers in EN-PUBLICATIONS and ZH-PUBLICATIONS by the top L0 FoS label

results, respectively (Steps 4-6). We then compute the cosine similarity between every document and FoS embedding pair in both languages (Step 7).

6 Results and Evaluation

Evaluating our results is particularly challenging without a ground-truth dataset that contains publications and their corresponding field of study labels. Because of this limitation, we offer several methods of evaluation that do not require annotation (to limit human bias and error) and can be replicated. Our evaluation methods compare results at the FoS level and the publication level in order to measure our taxonomy representation results (FoS embeddings) and our publication classification results.

6.1 Top Field of Study Labels

With each publication in EN-PUBLICATIONS and ZH-PUBLICATIONS having cosine similarity scores for the L0 and L1 FoS, we first analyze the top L0 field assignments (i.e., the L0 field with the highest cosine similarity score). Figure 4 displays the percentage of papers from EN-PUBLICATIONS and ZH-PUBLICATIONS with each top L0 field label. In EN-PUBLICATIONS, *medicine*, *chemistry*, and *computer science* have the most top field labels,

Corpus	Computer Science	Economics	Medicine	Sociology
EN-PUBLICATIONS	1. **Data Science**	1. Economic Growth	1. Cancer Research	1. Media Studies
	2. Machine Learning	2. **Economy**	2. Surgery	2. Socioeconomics
	3. Internet Privacy	3. Microeconomics	3. Cardiology	3. **Gender Studies**
	4. Computer Network	4. International Econ.	4. Virology	4. Communication
	5. Computer Security	5. Economic Policy	5. Medical Physics	5. Criminology
ZH-PUBLICATIONS	1. Algorithm	1. Commerce	1. Pharmacology	1. Regional Science
	2. **Data Science**	2. **Economy**	2. Immunology	2. **Gender Studies**
	3. Simulation	3. Monetary Econ.	3. Audiology	3. Law & Economics
	4. Real-time Computing	4. Macroeconomics	4. Oncology	4. Social Science
	5. Software Engineering	5. Financial System	5. Family Medicine	5. Anthropology

Table 2: Top five L1 fields of study for computer science, economics, medicine, and sociology L0 fields. L1 fields in bold font indicate that they appear in both the English and Chinese top five results for the same L0 field.

whereas in ZH-PUBLICATIONS *political science*, *medicine*, and *chemistry* have the most.

Next, we analyze the top L1 FoS (child) for each L0 FoS (parent). In Table 2, we present results from four representative L0 FoS (*computer science*, *economics*, *medicine*, and *sociology*) and list the top five L1 FoS from EN-PUBLICATIONS and ZH-PUBLICATIONS. We bold the fields that appear in both the English and Chinese top five L1 results; medicine has no overlapping top five L1 fields.

6.2 L0-to-L0 Similarities

Each FoS has a unique vector representation, calculated in Step 4; thus we can evaluate how similar FoS are to each other using cosine similarity. In Figure 5, we compare all L0 FoS embeddings using their cosine similarity scores; we present the results for English (left) and Chinese (right).

The diagonal represents the cosine similarity score for each L0 FoS to itself, which is 1. We find that the results in English are stronger than the results in Chinese. For example, in English, we see high similarities between L0 FoS we know are related: [*computer science*, *engineering*]; [*political science*, *sociology*]. Additionally, we see low similarities between L0 FoS that are unrelated: [*biology*, *political science*], [*chemistry*, *political science*], [*materials science*, *philosophy*]. In Chinese, we find L0 FoS pairs with high similarities that we would expect, such as [*political science*, *economics*] and [*mathematics*, *physics*]. However, we also find L0 pairs with high similarities that do not align with field relatedness, such as [*chemistry*, *economics*] and [*history*, *physics*].

6.3 L0-to-L1 Field Similarities

We evaluate the parent-child relationship between L0 and L1 FoS. For each L0 FoS, we generate a t-Distributed Stochastic Neighbor Embedding (t-SNE) plot with its corresponding L1 FoS. Using t-SNE, we implement dimensionality reduction on our 250-dimensional embeddings and plot the FoS embeddings in a 2-D space. In this way, we can visualize the organization of the parent FoS to its children. Figure 6 shows our results in both languages; the L0 FoS (parent) is highlighted in yellow.

We display the same four representative FoS (economics, computer science, medicine, and sociology) from Section 6.1 in Figure 6, but all L0 FoS graphs will be available in our GitHub repository. The t-SNE plots allow us to see how the L1 FoS are represented in the embedding space, and they highlight similarities and differences between the results in English and Chinese. For example, in computer science the L1 FoS have different groupings, such as data science and data mining in English, and pattern recognition and computer vision in Chinese. Alternatively, in economics, both languages have strong similarities between finance and actuarial science.

The t-SNE plots also help us compare the L1 field embeddings to their L0 (parent) field embeddings. We find that the English results for *economics* and *medicine* show the L0 fields as more central, with the L1 fields tightly clustered, as opposed to the *computer science* and *sociology* results. The Chinese graphs highlight that the L1 fields are not as tightly clustered as the English L1 fields.

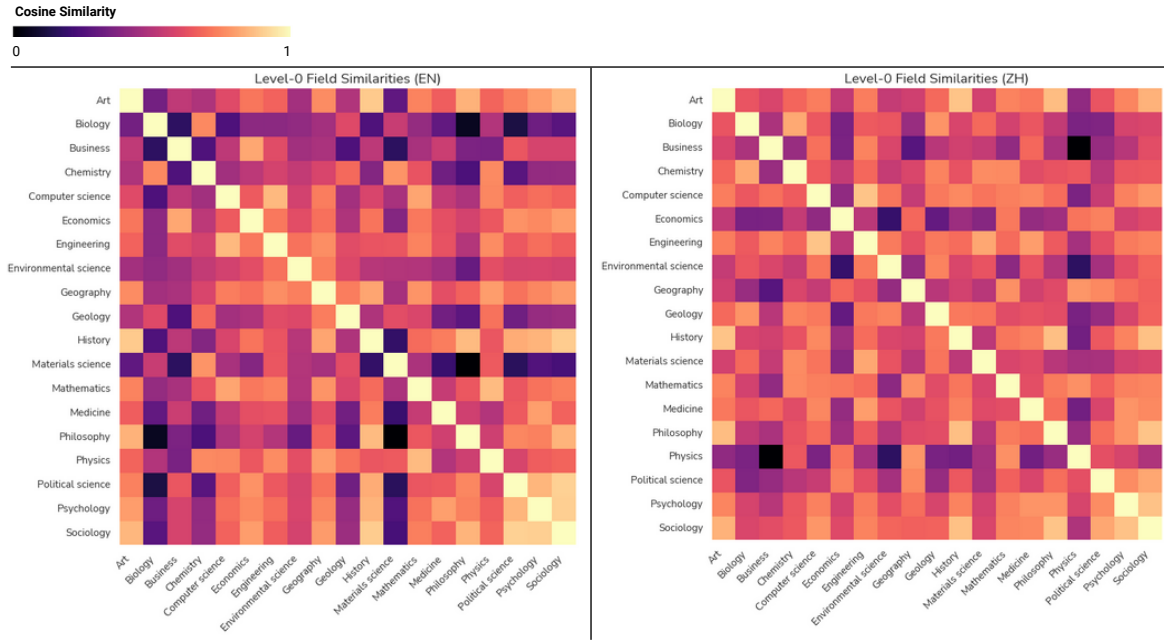

Figure 5: L0 Fields of Study cosine similarity heatmaps.

6.4 Case Study: Publication Field of Study Labels in Artificial Intelligence and Machine Learning

In order to evaluate how well our model assigns field labels to scientific research publications, we select publications from 13 top artificial intelligence (AI) and machine learning (ML) conferences identified by CSRankings[5]:

1. AAAI Conference on Artificial Intelligence
2. International Joint Conference on Artificial Intelligence
3. IEEE Conference on Computer Vision and Pattern Recognition
4. European Conference on Computer Vision
5. IEEE International Conference on Computer Vision
6. International Conference on Machine Learning
7. International Conference on Knowledge Discovery and Data Mining
8. Neural Information Processing Systems
9. Annual Meeting of the Association for Computational Linguistics
10. North American Chapter of the Association for Computational Linguistics
11. Conference on Empirical Methods in Natural Language Processing
12. International Conference on Research and Development in Information Retrieval
13. International Conference on World Wide Web.

There are 127,257 publications in EN-PUBLICATIONS that were published in a top AI/ML conference; this evaluation is limited to EN-PUBLICATIONS. We find that 57% of these publications have *computer science* as the top L0 FoS, with *physics* coming in second with 27%. Additionally, we check for the number of L0 FoS that are children of *computer science* and find that 59% of the publications have a top L1 FoS that is a child of *computer science*.

7 Conclusion and Future Work

Organizing scholarly literature is necessary for accessibility and usefulness of scientific research publications. Prior work has focused on a few broad areas of research, English-only research publications and taxonomies, and static taxonomy descriptions. In this paper, we implement a multi-label classification model that encompasses research fields from all of science, can be updated using a comprehen-

[5] www.csrankings.org

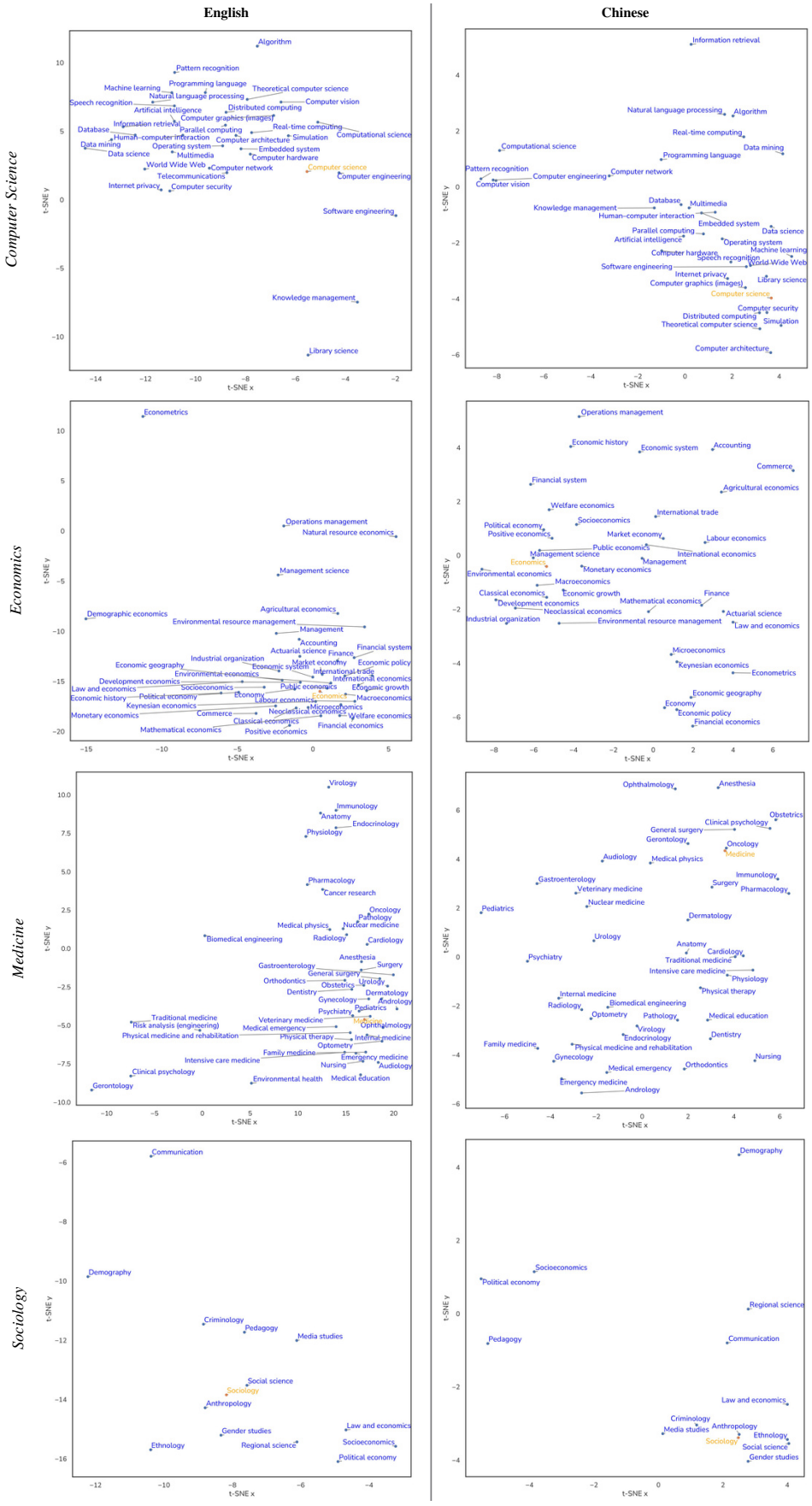

Figure 6: English and Chinese L1 embedding t-SNE plots for Economics, Computer Science, Medicine, and Sociology L0 fields of study

sive, online knowledge base, and is not restricted to the English language.

In future work, we plan to expand to additional languages and explore the longitudinal dynamics of fields: how their relative positions have shifted, within and between languages, as Wikipedia article text and references have changed.

References

Bader Aljaber, Nicola Stokes, James Bailey, and Jian Pei. 2010. Document clustering of scientific texts using citation contexts. *Information Retrieval*, 13(2):101–131.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Sang-Woon Kim and Joon-Min Gil. 2019. Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences*, 9(1):1–21.

Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150.

MediaWiki. 2022. Api:main page — mediawiki,. [Online; accessed 1-July-2022].

Ghulam Mustafa, Muhammad Usman, Lisu Yu, Muhammad Sulaiman, Abdul Shahid, et al. 2021. Multi-label classification of research articles using word2vec and identification of similarity threshold. *Scientific Reports*, 11(1):1–20.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

António Paulo Santos and Fátima Rodrigues. 2009. Multi-label hierarchical text classification using the acm taxonomy. In *14th Portuguese Conference on Artificial Intelligence (EPIA)*, volume 5, pages 553–564. Springer Berlin.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations*, pages 87–92.

Chen-Tse Tsai, Gourab Kundu, and Dan Roth. 2013. Concept-based analysis of scientific literature. In *Proceedings of the 22nd ACM international conference on information & knowledge management*, pages 1733–1738.

Wikipedia. 2022. Wikipedia:about. Online; accessed 22-June-2022.

Chyi-Kwei Yau, Alan Porter, Nils Newman, and Arho Suominen. 2014. Clustering scientific documents with topic modeling. *Scientometrics*, 100(3):767–786.