Zero-Shot Classification of Biomedical Articles with Emerging MeSH Descriptors

Nikolaos Mylonas
myloniko@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Greece

Stamatis Karlos
stkarlos@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Greece

Grigorios Tsoumakas
greg@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Greece

ABSTRACT

Although numerous applications that have been developed during the last years produce vast amounts of data, the inability to obtain their ground truth target values has triggered the appearance of several new machine learning (ML) variants that tackle such phenomena. The main reasons why this happens are the evolutionary nature that characterizes the majority of real-world problems, highly hindering the conventional approaches to be applied because of incompatibility, as well as the noisy sources of data or even the shortage of available training data to produce robust predictive models. The objective of this work is to provide a new ML approach in the field of zero-shot classification, focused on classifying abstracts that come from PubMed, a well-known resource of publications from the biomedical field. The proposed approach differs in that it uses bioBERT embeddings for transforming the textual data into a new semantic space exploiting them on sentence-level, instead of adopting the usual *n*-grams solution. Moreover, its asset of constructing a learning model without demanding any collected training data leads to an instance-based approach, while at the same time, it can be used as an internal mechanism for assigning labels to collected unlabeled training data, creating appropriate weakly supervised learning batch-based variants. Our evaluations over 3 different MeSH terms highlights the usefulness of these approaches against a state-of-the-art approach and a well-defined baseline, respectively.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; • **Computing methodologies** → *Online learning settings*.

KEYWORDS

Zero-Shot Classification, MeSH indexing, semantic similarity

ACM Reference Format:
Nikolaos Mylonas, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Zero-Shot Classification of Biomedical Articles with Emerging MeSH Descriptors. In *11th Hellenic Conference on Artificial Intelligence (SETN 2020), September 2–4, 2020, Athens, Greece.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3411408.3411414

SETN 2020, September 2–4, 2020, Athens, Greece

ACM ISBN 978-1-4503-8878-8/20/09…$15.00
https://doi.org/10.1145/3411408.3411414

1 INTRODUCTION

Although procedures of obtaining data from several kinds of sources in the field of Machine Learning (ML) have been highly automated, offering an abundance of data, the needs that arise during the learning stages of more complex ecosystems have not been completely covered. Assigning the actual label value of the target variable remains a difficult task in numerous domains, demanding human supervision, which apart from time delays may lead to great economic expenses, since either the amount of data could be too large, or only expert annotators should constitute trustworthy sources of information [8, 38]. More importantly, some outputs of the tackled problem may not have previously been observed, leading to ill-defined behaviors when trying to face them with the typical scenario of supervised learning. These situations call for weakly supervised learning (WSL) [40], as well as dataless or zero-shot learning (ZSL) [6] approaches. Such approaches have been considered recently as some of the most interesting topics of study by the community of ML, since they are closely related to phenomena that occur during real-world applications [33].

The use of ML techniques for analyzing textual data has already been addressed in the last decades through several publications. As part of them, the aforementioned concepts have been employed to address problems that are characterized by inaccurate training data or a shortage of some class labels, respectively. Some seminal approaches were based on the concept of semi-supervised Learning (SSL), trying to assign pseudo-labels to the collected unlabeled examples, mainly through the predictions of probabilistic base classifiers [39]. However, less implications occur during SSL techniques, where a few labeled data are initially provided before applying incremental stages which augment their cardinality based on the decisions of a base learner or the convergence of a learning mechanism [10, 15]. Seed words have been used in [20] under an SSL approach, managing to bridge the gap between these kinds of approaches and dataless ones.

Our goal in this work is to exploit bioBERT embeddings [17], a BERT representation that is specifically tuned towards creating a favorable representation in the field of biomedical text mining, for dealing with the multi-label classification (MLC) of PubMed[1] abstracts with MeSH[2] terms. Our proposed approaches operate without being based on *n*-grams, whose meaning may be heavily affected inside an abstract, but compute similarities that concern whole sentences, assuming that operating as standalone segments could enrich the decisions' quality by capturing the general scope per instance. Because of the underlying dynamic evolution of biomedical knowledge, several parts of MeSH are re-structured,

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] https://www.ncbi.nlm.nih.gov/mesh

gaining more or less importance, or even brand new entities are inserted into the total structure, introducing new labels for PubMed's existing articles. Therefore, a mechanism for distinguishing with high accuracy examples that are suitable for those labels is highly needed, since we deal with a real-world application that constitutes an open issue, and that is one of the main motivations behind our work. From our research we have found several possible ways for new labels to be introduced in the PubMed database from the changes occurring in MeSH (see Section 2).

Furthermore, in contrast with other approaches that are either based on word-level statistics [2], which demand much more computational resources, or which adopt Deep Neural Networks (DNNs), whose tuning of parameters is usually a difficult and time expensive procedure [22, 23], our proposed methods avoid such overheads. As it has already been mentioned, we exploit the semantic relations of embeddings on sentence-level, while, the core of our ZSL model depends only on one parameter: the threshold of similarity between each sentence and the query label (*th*). As it concerns the introduction of the WSL variant, we actually exploit the proposed ZSL mechanism for acquiring suitable weak labels on the collected training instances, assigning the most prevalent of the unseen labels. Then, we fit a typical ML algorithm during the prediction stage on the unknown data. Thus, we investigate further the chances of boosting the classification performance, adopting a batch-based learning scheme without tuning the parameters of the ML classifier, since this task remains out of our interest here. Our results over 3 separate MeSH terms which are characterized by different properties, show the efficacy of the proposed methods against a state-of-the-art approach that works at the word-level [30], as well as a simplified baseline approach that depends on the label's occurrence into the training data for producing its weak label in the case of WSL.

In summary, the main novel contributions of this work are the following:

- This is the first work to bring up a real-world ZSL use case and associated data: that of the evolving MeSH ontology. Past work has focused on simulated data and abstract use cases (news classification, intent recognition of bank consumers).
- This is the first work to examine the utility of contextual word embeddings in ZSL and WSL settings.

The rest of this paper is structured as follows: in Section 2, a brief description of the MeSH structure is provided, facilitating the reader to understand better this specific kind of problem. In Section 3, some related works are discussed further, while the proposed approaches follow in the next Section. Section 5 consists of the description of the constructed datasets, regarding their formulation and the manner that they were mined, as well as the achieved learning behavior against a state-of-the-art approach and a baseline method. Finally, we draw the most important conclusions in Section 6 and provide some future work.

2 MESH EVOLUTION AND GROUND TRUTH

Medical Subject Headings is a hierarchically structured controlled vocabulary created by the National Library of Medicine (NLM)[3] of the United States. MeSH entries are organized in a tree-like structure and are divided into three main categories: *descriptors* also known as main headings, *qualifiers*, commonly called subheadings, and *supplementary records*, which are used for indexing drugs and substances. MeSH descriptors have one or more sets of synonymous *terms* associated with them, which are considered equivalent for semantic search and indexing. These sets of terms are called *concepts*. One of these concepts of each descriptor is called the preferred concept, and usually has the same name as the descriptor, while the rest can be narrower, broader, or simply related to the preferred concept.

Mainly, MeSH is used for indexing and searching biomedical information, while PubMed uses it to facilitate the semantic search of relevant articles. Numerous new articles appear every day in PubMed and as a result manually annotating them can be a very time consuming and expensive task. For this reason, we need tools that automate this indexing procedure. The importance of this attempt could be understood better considering that over the years several researchers have come up with techniques that aim to automate the indexing of biomedical articles with relevant MeSH descriptors [1, 14, 27].

The vocabulary of MeSH is constantly evolving. Each year various changes take place to bring the vocabulary up to date with new biomedical knowledge. This work focuses on the changes that introduce new descriptors to MeSH, since these changes *may* lead to a ZSL situation for developers of supervised MeSH indexing models.

We stress *may*, as the additions of new descriptors can come along with existing ground truth data in the following cases:

- A supplementary record gets promoted to a descriptor. As PubMed articles are indexed with supplementary records, existing ground truth annotations at the supplementary record level are automatically moved at the descriptor level. For example supplementary record *RNA,circular* became a main heading in the 2019 version of MeSH.
- An existing descriptor and qualifier combination known as *entry combination* is replaced by a new single descriptor. Existing annotations of PubMed articles with this combination are automatically moved to the new descriptor. An example of this kind of change is the entry combination *Cornea/Injuries* being replaced by the new descriptor *Corneal Injuries* during the MeSH 2015 revision.
- A descriptor being replaced by two new descriptors. This usually happens when the replaced descriptor has ambiguous meaning and gets split into two different ones with similar but not identical meanings. In this case, NLM manually adjusts existing ground truth annotations. One example of this change is the splitting of descriptor *Gizzard* into *Gizzard, Avian* and *Gizzard, Non-Avian* in the 2017 version of MeSH.

Unfortunately, in several other cases, new descriptors cannot automatically be associated with past annotations:

- One of the entry terms of a concept of an existing descriptor gets promoted to a descriptor. In this case, existing annotations are at the level of the existing descriptor and it is not straightforward to know which of these annotations hold

[3] https://www.nlm.nih.gov/

for the promoted term. One such example is the term *Brain Injuries,Traumatic* who got promoted from being an entry term of *Brain Injuries* to a descriptor in MeSH 2017.
- A brand new descriptor gets introduced. In this case, no past ground truth annotations exist. The descriptors *Biomineralization,Chlorophyceae*, and *Cytoglobin* we use for our experiments in Section 5 are examples of such changes.

The new descriptors that get introduced may have one or more old descriptors in the *previous indexing* field in their record, meaning that articles indexed with that old descriptor(s) in the past, may be related to the new descriptor. The descriptor(s) mentioned in previous indexing can be related to the new one e.g. with a parent-child relation, or simply have a meaning similar or broader to the new descriptor. Previous indexing descriptors, when present, can help us find possible ground truth examples for new descriptors.

We can see that since there are different types of changes happening in MeSH each year, we have to deal with different ways for a new label to be introduced in the PubMed database. As a result we need some techniques that could take into account these kinds of evolutionary changes, dealing with one or more of them at the same time.

3 RELATED WORK

In this Section we briefly summarize research on topics related to ours, having separated them into 3 distinct categories: ZSL, Dataless Classification Protocol, and Emerging labels, although their differences may be subtle.

3.1 Zero-Shot Learning

This kind of learning constitutes a variant of multi-class classification problem, where no training data is available for some of the classes. A zero-shot classifier must be able to correctly classify instances of a given never before seen class and this is usually done by using some kind of description for the unseen class.

An example of zero-shot Learning for image classification is that by [31]. In their research they map the images of the training set into a 50-dimensional semantic word space along with labels in order to capture semantic relationships between them. This is done by computing the distance between their respective representation vectors in that space. Whenever a new test example arrives they use outlier detection to determine the probability of this example belonging to an unseen class.

As far as Zero-Shot Multi-Label Learning is concerned [35] proposed a method for correctly classifying instances of unseen classes in the Multi-Label image classification domain. Their approach projects the instances from the visual feature space and the classes from the word embedding space into a low-dimensional semantic space. With this they aim to exploit the similarity matching score between an instance and its positive labels as well as information about relations between labels.

A simple similarity-based approach for multi-label zero-shot text classification was proposed in [30]. For each unseen label, it computes the semantic similarity between the label and the document, and outputs a positive prediction if this similarity is larger than a threshold parameter t. In particular, it computes the cosine similarity between the word2vec embedding of the label (sum of embeddings if more than one word) and the word2vec embeddings of all n-grams (sum of embeddings when $n > 1$) of the document for $1 \leq n \leq c_{\max}$, where $c_{\max}$ is a second parameter of the method. For the seen labels, the binary relevance approach is used.

If there is some kind of structured relation between the labels (for example if the labels are part of a graph like the labels in MeSH) a technique from [29] can be used for zero-shot classification. In their research they use Convolutional Neural Networks (CNN) with label-wise attention to obtain document feature vectors for each example. The output layer of their CNN tries to match each document with its corresponding label vectors using the label relations for those labels that are not present in the training set.

3.2 Dataless Classification Protocol

Humans unlike machines don't need labeled data in order to correctly classify examples. That is because humans can capture the semantic information from the examples and the labels. Dataless classification is a technique based on the concept of learning the semantic relationships between the features and the labels, making it possible to correctly classify instances without the need of training data.

An algorithm called Seed-guided Topic Model (STM) from [18] is an example of the above technique. Given a collection of unlabeled documents and a few seed words for each category (label) it is able to classify documents based on topic influence. It does that by splitting topics in two sets category-topics and general-topics. Category-topics are associated with one category each and represent the meaning of that category while general-topics cover the general semantics of the documents. The prediction is made by selecting the category-topic with the highest probability based on the distribution of topics in each document. STM assumes that each document is associated with only one category so it's used for multi-class single-label classification (SLC).

As we stated before semantic relationships between documents and labels are really important for Dataless classification, this can be seen in the work from [6]. In their paper they research how different vector representations of text data help with the above problem. Specifically, they use the Nearest Neighbors (NN) algorithm for on the fly classification in conjunction with two different representations for the documents and labels, Bag of Words (BOW) which represents the text as a vector in the space of words and Explicit Semantic Analysis (ESA) which represents the text as a vector in the space of concepts. The results of their research show that using the ESA representation gives better scores overall because it is able to catch the semantic relationships between the words. They also tested an extension of their approach with the use of unlabeled data achieving results comparable to those of fully supervised algorithms.

An extension of the Latent Dirichlet Allocation (LDA) commonly used in Natural Language Processing (NLP) called Descriptive LDA (DescLDA) from [7] can be used for Dataless Text Classification. Their technique uses a describing device to infer priors from category descriptions, and then use these priors in the standard LDA algorithm. The describing device used to obtain those priors is also an LDA model. The goal of this approach is to drive the LDA model to induce category aware topics thus decreasing classification error. The results of this research are comparable to those of fully supervised algorithms.

A technique for dealing with datasets where there are no labeled data was proposed by [21]. In their research they tried a novel approach of labeling words instead of labeling entire documents. Specifically they find the words with the most information gain from clusters of unlabeled documents and rank them based on that criterion. Then, they ask human experts to label these words with the class that seems more representative. Using these labeled words they build an initial dataset from the unlabeled documents which is then used to train a Naive-Bayes based classifier. This technique, unlike most Dataless classification approaches, requires knowledge injection by human experts in order to label a few words, operating under an Active Learning fashion, violating thus the initially strict Dataless Protocol.

3.3 Emerging labels

Two relevant works are recorded in this category. The first regards MUENL [41], an approach which deals with a similar problem to ours. They propose an algorithm for dealing with emerging new labels in the multi-label data-stream setting. This algorithm has three key components (i) a multi-label classifier for the known labels, (ii) a detector for the new labels and (iii) a procedure to update the previous two components. The detector in this case, similarly to [31], considers the new labels as outliers, and the algorithm proposed for the detection is called MuENLForest. Each tree in the forest has a ball constructed around its leaf nodes. This ball is constructed during training based on the data that fall on that leaf. If during testing a new instance falls outside of that ball then it is predicted to have a new label. For the final prediction, majority voting is used. The results of this approach seem very promising and it is worth looking into further. The Second is the approach of Y. Zhu et al. [37] which was recently demonstrated has been actually inspired by the previous work. A Gaussian distribution was used for detecting instances that could be theorized that belong to a new class, whose features are first exported by a sparse Auto-Encoder algorithm. The main defect of their approach is the fact that only one new label is considered each time, but its innovation could be really helpful as a method of detection.

4 PROPOSED METHODS

In order to deal with the constantly evolving MeSH data as well as the new labels added each year, we propose two methods: an instance-based that needs no training stage covering the more complex case of online learning [25, 32], thus eliminating the time-consuming task of finding ground truth data for these new labels, and a batch-based framework that applies the same mechanism over the collected training data for obtaining class predictions and assigning the most confident label, before providing them to the selected base classifier. We use the term framework, because the manipulation of the training data remains on our choice, thus we have applied two different approaches: i) use of the tf-idf model, ii) use of bioBERT embeddings. The instance-based method is called *ZSLbioSentMax*, and is actually based on the fact that the embeddings of abstracts related to a specific label will have a high similarity matching relationship with the embedding of the later. For that reason, we use the well-known metric of *Cosine Similarity* as it is defined in Eq. (1) in order to calculate the similarity scores between labels and the sentences per instance's abstract, assuming that each instance is expressed as $x_i = \{sent^1, sent^2, ..., sent^{\mathrm{r}}\}, 0 \leq i \leq n_{\text{test}}$, where the n_{test} denotes the number of test instances. The final similarity matching score is the maximum one from the similarities of each sentence. If that score is higher than a pre-defined threshold (*th*) for a label-abstract pair then the abstract is considered as relevant to that label.

$$Cosine\ Similarity(A, B) = \frac{A * B}{||A|| * ||B||} \tag{1}$$

The use of embeddings under the scenario of ZSL for predicting never-before-seen labels is a typical approach. The main difference of our proposed method is the treatment of each abstract as a bag of sentences, calculating the similarity between the label embeddings and each one of these entities. The final similarity per instance is the maximum score among the computed. The embeddings for the abstracts and labels are obtained using the BioBert pre-trained model [17] which is a biomedical language representation model fine-tuned using data from the PubMed database. The main idea behind our method is that if an abstract is indeed related to a query label, then we can find at least a sentence in it that is semantically close to that label. We decided to use the maximum similarity between the sentences and the query label to trigger our labeling stage, instead of a more common measure like the average of all the *r* similarity scores per instance. In that scenario, the chance of miss-labeling an instance $\mathrm{x_i}$ increases, because the average similarity gets lowered by one or more completely unrelated sentences. For the sake of completeness, we present visually through favorable density plots the distinguishing ability that is acquired through the strategy of keeping the maximum similarity score per instance (abstract), instead of keeping all of them in case of *Biomineralization* MeSH term (Figure 1). This is one of the three MeSH terms that are described better into the next Section, while the rest cases are also included in our repository. In Algorithm 1 we present *ZSLbioSentMax* in a proper pseudo-code format as it acts per test instance.

The previous method completely bypasses the time consuming and difficult task of training a classifier in order to obtain suitable decisions per test instance. Instead, it executes an on-the-fly classification of any provided text segment that consists of at least one sentence. As a result, *ZSLbioSentMax* can be quite time efficient making it suitable for problems where the amount of data is large and manually annotating them can result in major time loss. It also depends only on the choice of the *th* variable.

As far as the WSL framework is concerned, we propose the use of the same procedure with the presented one, to obtain weakly labeled examples $\mathrm{y_{weakly}}$ for the unseen labels, this time into the collected training set. Assuming that its decisions are trustworthy, we proceed towards training a typical ML classifier through them, after having transformed the raw data from their textual-structure into any selected representation, and finally produce the necessary predictions for the corresponding query labels into new unknown instances. Therefore, a variant of Support Vector Machines (SVMs)

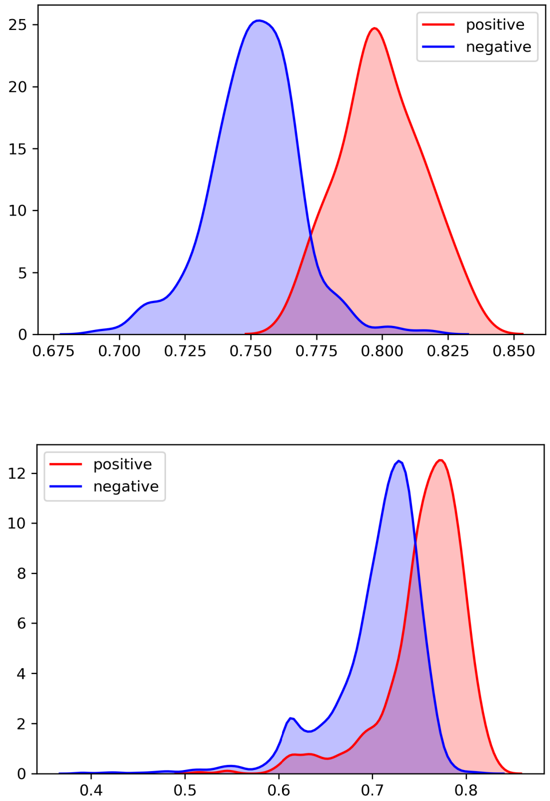

Figure 1: Density plots of similarity scores in case of Biomineralization. (Upper) The maximum sentence similarity value per instance. (Lower) All the sentences's similarities per instance.

```
Algorithm 1: ZSLbioSentMax
  Input:
    x_test: Text instance,
    L: query label
    th: threshold (0,1)
  Main Body
  Split $x_{test}$ to sentences: $x_{test} = \{sent^1, sent^2, ..., sent^r\}$
  $label\_embed \leftarrow bioBert(L)$
  $max \leftarrow 0$
  $j \leftarrow 0$
  while $j \leq r$ do
      $sen\_embed \leftarrow bioBert(x_{test}(j))$
      $sim \leftarrow cosine\_similarity(label\_embed, sen\_embed)$
      if $sim \geq max$ then
          $max \leftarrow sim$
      end
      $j \leftarrow j + 1$
  end
  Output:
    if (max > th) then export 1, otherwise 0
```

algorithm that exploits linear kernel has been combined with the proposed WSL algorithm, hoping to obtain accurate predictions without consuming many computational resources and based on the wide applicability that SVMs have presented. In general, this classifier is a popular algorithm of parametric discriminant approaches that optimizes its assigned loss function trying to separate the instances from binary problems into a new projected space that is formatted based on the original feature space [5]. Its successful behavior over several ML problems, especially text-classification ones, enforces its application here as a trustworthy enough algorithm [3, 28].

This approach even though it requires training, acting as a batch-based approach, thus making on-the-fly classification impossible, could provide robust predictions over detecting those unseen labels in several cases making it more suitable for occasions when time expenditure is not the main concern but the quality of predictions actually is. Its interaction also with the final transformation mode of the textual data might highlight some valuable insights about the compatibility of each such mode per occasion. In contrast with the instance-based approach, here, each selected classifier inserts its own hyper-parameters that need to be tuned, apart from the similarity threshold which is necessary into *ZSLbioSentMax*. The choice of these additional parameters as well as their fine-tuning is an optimization problem separate from ours, so we will not provide further details about this stage here. Adoption of suitable tuning stages could be examined as a future direction [16]. Algorithm 2 describes the above approach in pseudo-code format exploiting internally the Algorithm 1 for acquiring the labels of the training data.

```
Algorithm 2: WSLbioSentMax(transform_mode)
  Input:
    X_train: Text Database of $n_{train}$ instances,
    L: query label
    th: threshold (0,1)
    transform_mode: Representation Model
    Base_Learner: Linear_SVM
  Main Body
  $j \leftarrow 0$
  $y_{weakly} \leftarrow zeros(n_{train})$
  while $j \leq n_{train}$ do
      $y_{weakly}(j) \leftarrow ZSLbioSentMax(X_{train}(j)\ L, th)$
      $j \leftarrow j + 1$
  end
  Transform $X_{train}$ based on transform_mode
  Fit Base_Learner($X_{train}, y_{weakly}$)
  Output:
  For any provided test instance, export decisions of trained
  Base_Learner.
```

5 EXPERIMENTS AND RESULTS

In this Section, we first provide a more detailed description of our data and then describe the executed experiments, to support the posed assumptions about the efficacy of retrieving valuable information through the embeddings applied on sentence-level. Appropriate comments follow later, we also highlight two specific examples where the sentence-based and word-based ZSL approaches differ on their decisions, trying to understand how they are led towards them. It is worth mentioning that since *ZSLbioSentMax* deals with each label separately, this flexibility enables its application over both SLC and MLC problems adding another advantage to it. Here, we have applied the first mode, but assuming that the existence of each label is an independent binary problem, the application of our methods can easily be generalized into the second mode also.

5.1 Dataset Description

Our source of data coincides with those provided by the BioASQ[4] challenge page, facilitating the reproducibility of the experiments, and at the same time contributing to tasks that are under research from the corresponding scientific community. The training data refer to abstracts that have been recorded until 2018, while the test data come solely from 2019. Since the goal of our research was to predict emerging new labels, we tried to emulate this by training on older examples than those we are trying to predict.

We focus on three descriptors that were added in the 2019 version of MeSH: *Biomineralization*, *Cytoglobin* and *Chlorophyceae*. The first two had a MeSH term as previous indexing, meaning that we can narrow our search for train data on those, while the third one did not. For the third one, due to the absence of previous indexing we tried to find suitable training data by exploiting its relations in the MeSH tree-like architecture. Specifically, we found training data indexed with the MeSH term that is its parent node and used them as a base to find some related examples for it. The above method results in weaker training examples than the former, and as such, lower classification scores are expected. For *Biomineralization* that term was *Minerals*, for *Cytoglobin* it was *Globins* and for *Chlorophyceae* who has no previous indexing we used the term *Chlorophyta* which is actually its parent node. Figure 2 presents the frequency of the above terms in the MeSH-2018 dataset per year. When the number of appearances is positive but very small to be legible as bar height, we write it as a number with the color of the corresponding MeSH term according to the corresponding legend. We initially tried to find 1000 of those examples when possible, as was the case with *Biomineralization* and *Chlorophyceae*, while this number was restricted to its half in case of *Cytoglobin* due to its more scarce appearance.

Regarding the test set of both ZSL and WSL experiments, we joined some negative instances per MeSH term mining them from other unrelated labels randomly in order to reach a ratio equal to 1:3 into them. This means that per each positive instance, three new negative instances have been selected additionally. However, due to some technical issues that emerged when trying to compute the corresponding embeddings (e.g. phrases that contain many consecutive special characters), we finally rejected a small portion of the candidate negative instances per MeSH term, biasing slightly the desired *positive:negative* ratio.

As it concerns the latter format of experiments, we constructed two different training sets per label, discriminating them based on the cardinality of the instances that regard each specific query label and the corresponding quantity of the non-related. The selected ratios are *1:1* and *1:3*, while the test data remain the same. The formulation of each case is presented in Table 1, where the column names of "actual" and "others", correspond to positive and negative instances, respectively, while Table 2 includes the corresponding information about the transformed datasets for the WSL experiments regarding the training set.

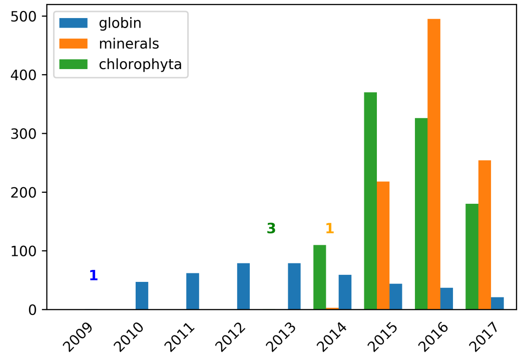

Figure 2: Frequency of the examined MeSH terms per year in MeSH-2018 dataset

Table 1: Size of examined Datasets

Label	training set		test set	
	ratio 1:1	*ratio 1:3*	*pos.*	*neg.*
Biomineralization	2000	4000	86	252
Chlorophyceae	2000	4000	93	273
Cytoglobin	1000	2000	55	164

Table 2: Data Dimensions for each examined dataset

Label	tfidf		bioBert	
	ratio 1:1	*ratio 1:3*	*ratio 1:1*	*ratio 1:3*
Biomineralization	26930	40271	768	768
Chlorophyceae	26692	40207	768	768
Cytoglobin	16744	26138	768	768

5.2 Experiment Setup

In this paragraph we describe the rest of the algorithms that have been implemented in order to perform our comparisons, as well as provide the appropriate information on our proposed approaches. First of all, in order to verify the efficacy of the *ZSLbioSentMax* approach, we implemented the algorithm in [30], which is referred to as Label Word Similarity (LWS). It is mainly based on the application of a series of sliding windows on words-level and the computation of the underlying similarities with each demanded query label. Therefore, its operation depends on two parameters: a threshold value, similar to our ZSL approach, and the definition of the windows' lengths (c_{max}). The latter one was set equal to 3,

[4] http://bioasq.org/

scanning for all possible unigrams, bigrams, and trigrams inside any asked abstract before it exports its decision, as was proposed in the original work. The final decision is of course related to the overpass of the corresponding threshold value.

Intending to conduct a more fair comparison, we applied bioBERT embeddings instead of Word2Vec. Otherwise, there would be no consistency between the default LWS approach and the context of the Pubmed segments. Furthermore, in maintaining the fairness of our experiments, we tuned the *th* value over both the proposed method and the LWS, making a grid search over the total range of threshold values, since there is still not any mechanism embedded into none of these approaches so as to adaptively select the most suitable threshold value. The criterion for implementing our model selection choice based on the most proper *th* value is depicted in Equation 2, as follows:

$$Model\ Selection : th^{*} = \underset{th}{\operatorname{argmax}} F_1(y, \hat{y}(th)) \qquad (2)$$

where y refers to the ground truth vector of the test data, while $\hat{y}(th)$ denotes the estimated labels on the same subset depending on the the *th* parameter. To continue with the experimental procedure, two distinct variants of the proposed WSL framework were implemented for reaching to safe conclusions about the usefulness of its internal labeling mechanism when it acts as a mechanism for acquiring weak labels. Both variants exploit SVMs with linear kernel over the transformed collected training data. According to the choice of applied transformation, *WSLbioSentMax(bioBERT)* and *WSLbioSentMax(tfidf)* have been produced, while our selected baseline approach operates by assigning a true label on train instances where the query label is present at least once inside their abstract. This last approach does not differ from other baseline methods stated at the literature for this kind of problems (label occurrence), wasting less resources, either computational or human-based, since this concept is really straightforward, and sets a typical performance threshold that should be surpassed by any more sophisticated method.

Our implementation along with the appropriate raw data and the corresponding results could be found in the next link: *https://github.com/intelligence-csd-auth-gr/zsl-wsl-sentence-mesh.git*. Since we later report the time response of the examined algorithms, we have to mention the technical specifications of our computing machine: 64-bit OS (Windows10) embedded with Intel Core i7-9700 (3GHz) processor and 32GB RAM.

5.3 Results

The produced results between the proposed ZSL approach and this variant of LWS are presented in Table 3. There, the metrics of precision (*Pr*), recall (*Re*), the binary f1 score related with the positive labels (F_1) and the classification time, measured in seconds, produced by the best threshold value according to the maximum F_1 metric are placed, respectively. The best performance value per query label is highlighted in bold format. The results for the range [0.65, 0.85] with a fixed step equal to 0.01 are included in our repository, providing a more compact aspect of the total performance of these approaches. Values outside of that range produced trivial results. The best specific threshold for each label is also shown in Table 3. Although this kind of tuning stage cannot be applied in practice, it is interesting that for all the examined MeSH terms the best threshold values are really close (0.770 for our approach and around 0.805 for [30]). Table 3 shows that both methods have similar performance as far as predictive ability goes with our method being ahead in 6 out of 9 cases.

Table 3: Comparison of Zero-Shot Learning Approaches

Algorithm	Pr	Re	F_1	th	time
Biomineralization					
Ours	0.824	**0.977**	**0.894**	0.77	**168**
[30]	**0.946**	0.814	0.875	0.81	6175
Chlorophyceae					
Ours	**0.683**	**0.882**	**0.77**	0.77	**183**
[30]	0.675	0.785	0.726	0.80	6154
Cytoglobin					
Ours	**1**	0.891	0.942	0.77	**114**
[30]	0.982	**0.982**	**0.982**	0.80	3866

We also note that our approach is much more time-efficient than [30]. This is due to the total embedding computations that are executed per approach. For example, in case of *Biomineralization* as query label, the test data contains 4327 sentences, out of which 238,542 *n*-grams are produced for $n \leq 3$ (or $c_{max} = 3$). To be more precise, during the splitting of abstracts into sentences we ignored any resulting sentence with up to 10 characters, such as 'OBJECTIVE:', 'METHODS:', 'RESULTS:', because they typically carry minimal information. Thus, the final executed computations for the former case is 3975. Additionally, our method depends only on the threshold parameter, while the LWS demands the definition of the maximum window size.

It is also worth noting that our method achieves its peak performance for lower similarity values while LWS for higher ones. This is due to the fact that whole sentences lead to smoother similarities than *n*-grams, and as a result, if a sentence has achieved high similarity with one label, we can be almost certain that they are related. This property is not present in *n*-grams because due to their smaller size the similarity value may be higher even for two unrelated words.

Table 4 presents two examples of disagreement between these two ZSL methods, along with the maximum similarity score they computed for each one of those two abstracts, depicting the sentence or words, respectively, that were responsible for these values. Note that positive predictions are obtained when the threshold is surpassed, otherwise a negative decision is drawn. These examples highlight the differences that exist between sentence-based and word-based approaches: our approach, by checking whole sentences instead of smaller *n*-grams, is able to better capture the similarity between label and abstract when the former is not present in the latter. This can be seen in the first case, where the combination of words "pancrustaceans", "chitinous exoskeleton" and "cuticular proteins" is what makes *Biomineralization* relevant to that abstract. The smaller *n*-grams are not able to completely capture this combination thus misclassifying that instance. On the other hand, when the query label is present inside the abstract, the similarity may be lowered when using sentences due to the presence of many unrelated words to the label. In that case, *n*-grams are better since they

Table 4: Presentation of the maximum similarity scores achieved per online method for two meaningful instances.

Mesh Term	Method	Prediction - Actual	Raw text	Similarity Value
Biomineralization	ZSLbioSentMax	Positive - Positive	A key feature common to all pancrustaceans is their chitinous exoskeleton with a major contribution by cuticular proteins.	0.771
	LWS	Negative - Positive	chitin–binding chitinous exoskeleton chitinous exoskeleton with	0.769 0.794 0.802
Cytoglobin	ZSLbioSentMax	Negative - Positive	We demonstrate that p63-iRHOM2 regulates cell survival and response to oxidative stress via modulation of SURVIVIN and Cytoglobin respectively.	0.768
	LWS	Positive - Positive	Cytoglobin and Cytoglobin SURVIVIN and Cytoglobin and Cytoglobin respectively Cytoglobin respectively. Furthermore	0.869 0.865 0.843 0.810 0.826

only focus on small parts of the abstract eliminating that concern. This can be seen in the second example, where our method fails to correctly classify the abstract even though the label exists into the context. However, this fact could operate inversely in examples where the query label(s) appear in abstracts for other reasons - such as a counterexample of a medical description - while the smoother decision profile and contextual representation of a sentence-based approach would probably not suffer much on that cases.

The total results regarding the WSL experiments are presented in Table 5, presenting again the same metrics as in the case of ZSL, including also the selected baseline apart from our proposed approaches and examining both train-test spit scenarios: balanced and non-balanced. In order to facilitate the presentation of Table 5, we mention only the distinct *transform_mode* that is applied each time along with the prefix of WSL. We reach to the point that both of the proposed methods perform much better than the baseline approach, consuming of course much more computational resources, a fact that is not as concerning, since the time response is not the ultimate priority during the off-line methods. A strong proof of the need of exploiting more sophisticated approaches than the baseline, is its biased performance in the case of *Chlorophyceae*, where it has predicted all the test instances as negative, presenting too weak discriminating ability. This is likely due to the fact that this term is a scientific name for a class of 'green algae' and therefore is not always present inside the relevant abstracts.

As far as the proposed methods are concerned, they do not seem to deteriorate much when more negative training instances are provided, distorting the ideal balanced scenario, as this behavior holds on the 3 out of 6 cases. This reveals a robust behavior, since the larger the size of the training data, the higher the possibility of noisy examples to be considered as their annotation is based on its internal mechanism rather than a laborious process of asking one or more human oracles. Their prevalence against each other is also not stable, since the dataset of *Biomineralization* term has better results using WSLbioSentMax(bioBert), while WSLbioSentMax(tfidf) achieved better learning rates over the case of *Chlorophyceae*, presenting similar time complexity. However, it has to be mentioned that the dataset related to the *Cytoglobin* term has not favored the use of training data, since the performance of the instance-based *ZSLbioSentMax* approach recorded the best results. The performance of the baseline method in this dataset is also pretty high, this is possibly because the term appears in almost every abstract related to the query label, favoring the simple and straightforward approach of abstract occurrence. However, this scenario is very rarely met.

5.4 Investigation of threshold tuning

Since the definition of the *th* parameter constitutes a serious issue for generalizing our approach, especially when more labels have to be adressed, we discuss briefly some ideas over alleviating this selection stage. First, we could exploit the seed words (similar words with each MeSH term) or the corresponding description phrase that are provided by PubMed database and obtain some initial thresholds which could be tuned periodically after a number of incoming test instances have arrived, either by human supervision or via an adaptive weighting process. Hopefully, the majority of

Table 5: Comparison of WSL Approaches

Algorithm	Pr	Re	F_1	th	time(sec)
		1:1			
Biomineralization					
baseline	**1.0**	0.1627	0.28	–	1.338
WSL(bioBert)	0.907	**0.907**	**0.907**	0.78	1563
WSL(tfidf)	0.930	0.767	0.841	0.78	1072
Chlorophyceae					
baseline	0.0	0.0	0.0	–	1.16
WSL(bioBert)	0.632	0.850	0.725	0.77	1585
WSL(tfidf)	**0.670**	**0.957**	**0.788**	0.77	1072
Cytoglobin					
baseline	**1.0**	0.764	0.865	–	0.756
WSL(bioBert)	0.960	0.854	**0.904**	0.77	846
WSL(tfidf)	0.923	**0.873**	0.897	0.76	575
		1:3			
Biomineralization					
baseline	**1.0**	0.174	0.297	–	2.813
WSL(bioBert)	0.938	**0.884**	**0.910**	0.78	2927
WSL(tfidf)	0.777	0.849	0.811	0.77	1998
Chlorophyceae					
baseline	0.0	0.0	0.0	–	2.263
WSL(bioBert)	0.588	**0.968**	0.731	0.76	2969
WSL(tfidf)	**0.654**	0.914	**0.762**	0.77	2025
Cytoglobin					
baseline	**1.0**	0.764	0.865	–	1.54
WSL(bioBert)	0.957	0.820	0.882	0.77	1547
WSL(tfidf)	0.925	**0.891**	**0.907**	0.76	1068

the MeSH terms are accompanied by such information facilitating this postulation. Second, we may assume that the distribution of the sentence similarities could reveal useful insights about the optimum threshold value. To investigate further, we applied an unsupervised threshold selection process assuming that all the test instances per examined MeSH term are provided into a batch. Then, a Gaussian mixture model was fitted, searching for two distinct components into the original population of sentence similarities, while the method of Expectation-Maximization takes place for estimating the parameters of the corresponding density functions [4]. Adopting the midpoint value of the mean per component as the output of this stage, we record the next values: *Biomineralization*: 0.776, *Chlorophyceae*: 0.766 and *Cytoglobin*: 0.767. The fact that in all cases these values are really close to the optimum threshold is a promissing indication.

6 CONCLUSIONS

To sum up, in this study, we have tried to apply a variant of the ZSL protocol in order to predict unseen labels in cases where finding reliable training data is difficult or even impossible due to their sparsity. To that end, we proposed a method that needs no training in order to make predictions as well as a framework that uses the aforementioned method to create a weakly-labeled dataset for training a linear variant of SVMs, assuming that the overall predictive ability may be boosted further with the addition of this extra training step. The fact that the examined data indeed describe a problem that suffers from the shortage of collected training data for MeSH terms that appear on upcoming years, mainly because of the evolutionary nature of the biomedical domain, denotes an interesting field of study for approaches that operate under such kind of protocols, avoiding the inaccurate process of being based solely on artificially generated datasets or simulating similar behaviors into datasets that come from non-suffering fields.

Our research has shown that splitting the textual data into sentences and then calculating the cosine similarity between the labels and each one of those sentences, generally outperforms similar methods that use directions mainly based on n-grams, not only in performance but in time efficiency as well. The property of exporting smoother decisions according to the underlying similarity values in case of the proposed sentence-based approach proved more successful, avoiding misclassifications that the existence of a familiar with the query label term could potentially trigger on the word-based approaches. This fact combined with the strategy to retain only the largest similarity value per instance, seems to achieve an overall better distinguishing ability, especially on MeSH terms without previous indexing. Also the use of methods that make the assumption about the absence of seed or related words to the query label should be investigated [19].

In the Weak Classification part, the results vary between datasets, with them improving as the quality of data increases. This leads us to investigate ways for creating an approach that is more stable in the future as far as performance is concerned. Possible improvements could be the proper manipulation of produced features [11, 26] or the choice of the most consistent training samples, avoiding thus both redundancy and reducing the chance of over-fitting [9]. Cost-sensitive methods could offer even more robust learning behaviors, as it has been demonstrated in related data [13].

In general, some of the next steps that should be considered during our future research are the use of lightweight models [24], which may benefit mainly the WSL approaches, as well as applying techniques that guarantee an improvement of learning rates, by avoiding the degenerating effects of noisy labels [33]. Moreover, further examination of mechanisms that could benefit the automated tuning of the threshold per different MeSH term should take place for presenting a more generic approach that does not demand exhaustive grid searches in the most realistic scenario of MLC with several labels [12, 36]. Adoption of an Active Learning strategy has also been proven a quite promising method for acquiring further information through an oracle towards tackling multi-label problems with large number of unseen labels, as well as considering the existing dependencies between different labels [34].

ACKNOWLEDGMENTS

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: 514)

REFERENCES

[1] Joel Robert Adams and Steven Bedrick. 2014. Automatic classification of PubMed abstracts with Latent semantic indexing: Working notes. In *CLEF 2014 - Working Notes for CLEF 2014 Conference*, Vol. 1180. CEUR-WS, 1275–1282.

[2] Berna Altnel, Murat Can Ganiz, and Banu Diri. 2017. Instance Labeling in Semi-Supervised Learning with Meaning Values of Words. *Eng. Appl. Artif. Intell.* 62, C (June 2017), 152–163. https://doi.org/10.1016/j.engappai.2017.04.003

[3] George Apostolopoulos, Athanasios Koutras, Ioanna Christoyianni, and Evaggelos Dermatas. 2014. Computer Aided Classification of Mammographic Tissue Using Shapelets and Support Vector Machines. In *Artificial Intelligence: Methods and Applications - 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15-17, 2014. Proceedings.* 510–520. https://doi.org/10.1007/978-3-319-07064-3_44

[4] Gerasimos Arvanitis, Otilia Kocsis, Aris S. Lalos, Stavros Nousias, Konstantinos Moustakas, and Nikos Fakotakis. 2018. 3-Class Prediction of Asthma Control Status Using a Gaussian Mixture Model Approach. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence, SETN 2018, Patras, Greece, July 09-12, 2018.* ACM, 53:1–53:2. https://doi.org/10.1145/3200947.3201056

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM TIST* 2, 3 (2011), 27:1–27:27. https://doi.org/10.1145/1961189.1961199

[6] Ming Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI-08/IAAI-08 Proceedings*, Vol. 2. 830–835.

[7] Xingyuan Chen, Yunqing Xia, Peng Jin, and John Carroll. 2015. Dataless Text Classification with Descriptive LDA. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas) *(AAAI'15)*. AAAI Press, 2224–2231.

[8] Diego Fernandes de Araújo, Carlos Eduardo Santos Pires, and Dimas Cassimiro do Nascimento. 2020. Leveraging active learning to reduce human effort in the generation of ground-truth for entity resolution. *Comput. Intell.* 36, 2 (2020), 743–772. https://doi.org/10.1111/coin.12268

[9] Debashree Devi, Saroj K. Biswas, and Biswajit Purkayastha. 2017. Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance. *Pattern Recognit. Lett.* 93 (2017), 3–12. https://doi.org/10.1016/j.patrec.2016.10.006

[10] Vijay Garla, Caroline Taylor, and Cynthia Brandt. 2013. Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management. *J. Biomed. Informatics* 46, 5 (2013), 869–875. https://doi.org/10.1016/j.jbi.2013.06.014

[11] Yeming Hu, Evangelos E. Milios, and James Blustein. 2016. Document Clustering With Dual Supervision Through Feature Reweighting. *Comput. Intell.* 32, 3 (2016), 480–513. https://doi.org/10.1111/coin.12064

[12] Hsin-Hsiung Huang, Zijing Wang, and Wingyan Chung. 2019. Efficient parameter selection for support vector machines. *Enterprise IS* 13, 6 (2019), 916–932. https://www.tandfonline.com/doi/full/10.1080/17517575.2019.1592233

[13] Suvir Jain, Kashyap R., Tsung-Ting Kuo, Shitij Bhargava, Gordon Lin, and Chun-Nan Hsu. 2016. Weakly supervised learning of biomedical information extraction from curated data. *BMC Bioinform.* 17, S-1 (2016), 1. https://doi.org/10.1186/s12859-015-0844-1

[14] Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2018. AttentionMeSH: Simple, Effective and Interpretable Automatic MeSH Indexer. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*. ACM, Brussels, Belgium, 47–56. https://doi.org/10.18653/v1/W18-5306

[15] Mohammad Reza Keyvanpour and Maryam Bahojb Imani. 2013. Semi-Supervised Text Categorization: Exploiting Unlabeled Data Using Ensemble Learning Algorithms. *Intell. Data Anal.* 17, 3 (May 2013), 367–385.

[16] Brent Komer, James Bergstra, and Chris Eliasmith. 2019. Hyperopt-Sklearn. In *Automated Machine Learning - Methods, Systems, Challenges*. Springer, 97–111. https://doi.org/10.1007/978-3-030-05318-5_5

[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (09 2019). https://doi.org/10.1093/bioinformatics/btz682

[18] Chenliang Li, Jian Xing, Aixin Sun, and Zongyang Ma. 2016. Effective Document Labeling with Very Few Seed Words: A Topic Model Approach. In *Proceedings of the 25th ACM International on CIKM* (Indianapolis, Indiana, USA) *(CIKM '16)*. ACM, New York, NY, USA, 85–94. https://doi.org/10.1145/2983323.2983721

[19] Ximing Li, Changchun Li, Jinjin Chi, Jihong Ouyang, and Chenliang Li. 2018. Dataless Text Classification: A Topic Modeling Approach with Document Manifold. In *Proceedings of the 27th ACM International CIKM* (Torino, Italy) *(CIKM '18)*. ACM, New York, NY, USA, 973–982. https://doi.org/10.1145/3269206.3271671

[20] Ximing Li and Bo Yang. 2018. A Pseudo Label based Dataless Naive Bayes Algorithm for Text Classification with Seed Words. In *COLING*.

[21] Bing Liu, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. 2004. Text Classification by Labeling Words. In *Proceedings of the 19th National Conference on Artifical Intelligence* (San Jose, California) *(AAAI'04)*. AAAI Press, 425–430.

[22] Zhaoying Liu, Haipeng Kan, Ting Zhang, and Yujian Li. 2020. DUKMSVM: A Framework of Deep Uniform Kernel Mapping Support Vector Machine for Short Text Classification. *Applied Sciences* 10, 7 (Mar 2020), 2348. https://doi.org/10.3390/app10072348

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.* 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality

[24] Chongyu Pan, Jian Huang, Jianxing Gong, and Xingsheng Yuan. 2019. Few-Shot Transfer Learning for Text Classification With Lightweight Word Embedding Based Models. *IEEE Access* 7 (2019), 53296–53304. https://doi.org/10.1109/ACCESS.2019.2911850

[25] Mahardhika Pratama, Sreenatha G. Anavatti, and Edwin Lughofer. 2014. An Incremental Classifier from Data Streams. In *Artificial Intelligence: Methods and Applications - 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15-17, 2014. Proceedings (Lecture Notes in Computer Science)*, Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles (Eds.), Vol. 8445. Springer, 15–28. https://doi.org/10.1007/978-3-319-07064-3_2

[26] Asriyanti Indah Pratiwi and Adiwijaya. 2018. On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis. *Applied Comp. Int. Soft Computing* 2018 (2018), 1407817:1–1407817:5. https://doi.org/10.1155/2018/1407817

[27] Alastair R. Rae, James G. Mork, and Dina Demner-Fushman. 2020. Convolutional Neural Network for Automatic MeSH Indexing. In *Machine Learning and Knowledge Discovery in Databases*, Peggy Cellier and Kurt Driessens (Eds.). Springer International Publishing, Cham, 581–594.

[28] Sandeep Rathor and R. S. Jadon. 2019. The art of domain classification and recognition for text conversation using support vector classifier. *IJART* 11, 3 (2019), 309–324. https://doi.org/10.1504/IJART.2019.10020168

[29] Anthony Rios and Ramakanth Kavuluru. 2018. Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces. In *Proceedings of the 2018 Conference on Empirical Methods in NLP.* Assoc. for Comp. Ling., Brussels, Belgium, 3132–3142. https://doi.org/10.18653/v1/D18-1352

[30] Prateek Veeranna Sappadla, Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. Using Semantic Similarity for Multi-Label Zero-Shot Classification of Text Documents. In *Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN-16)*. d-side publications, Bruges, Belgium.

[31] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero-shot learning through cross-modal transfer. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. Lake Tahoe, Nevada, US, 935–943.

[32] Rajasekar Venkatesan, Meng Joo Er, Mihika Dave, Mahardhika Pratama, and Shiqian Wu. 2017. A novel online multi-label classifier for high-speed streaming data applications. *Evolving Systems* 8, 4 (2017), 303–315. https://doi.org/10.1007/s12530-016-9162-8

[33] Tong Wei, Lan-Zhe Guo, Yu-Feng Li, and Wei Gao. 2018. Learning Safe Multi-Label Prediction for Weakly Labeled Data. *Mach. Learn.* 107, 4 (April 2018), 703–725. https://doi.org/10.1007/s10994-017-5675-z

[34] Sihong Xie and Philip S. Yu. 2017. Active zero-shot learning: a novel approach to extreme multi-labeled classification. *Int. J. Data Sci. Anal.* 3, 3 (2017), 151–160. https://doi.org/10.1007/s41060-017-0042-5

[35] Meng Ye and Yuhong Guo. 2018. Multi-Label Zero-Shot Learning with Transfer-Aware Label Embedding Projection. arXiv:cs.CV/1808.02474

[36] Daochen Zha and Chenliang Li. 2019. Multi-label dataless text classification with topic modeling. *Knowl. Inf. Syst.* 61, 1 (2019), 137–160. https://doi.org/10.1007/s10115-018-1280-0

[37] Shaohui Zhang, Man Wang, Weihua Li, Jiesi Luo, and Zusheng Lin. 2019. Deep Learning With Emerging New Labels for Fault Diagnosis. *IEEE Access* 7 (2019), 6279–6287. https://doi.org/10.1109/ACCESS.2018.2886078

[38] Xiaoya Zhang, Lianjie Wang, Jin Xie, Pengfei Zhu, Citation X Zhang Y, and Wang L J. 2020. Human-in-the-loop image segmentation and annotation. *SCIS* 63 (2020), 3. https://doi.org/10.1007/s11432-019-2759-y

[39] Li Zhao, Minlie Huang, Ziyu Yao, Rongwei Su, Yingying Jiang, and Xiaoyan Zhu. 2016. Semi-Supervised Multinomial Naive Bayes for Text Classification by Leveraging Word-Level Statistical Constraint. In *AAAI*.

[40] Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review* 5, 1 (08 2017), 44–53. https://doi.org/10.1093/nsr/nwx106 arXiv:https://academic.oup.com/nsr/article-pdf/5/1/44/31567770/nwx106.pdf

[41] Y. Zhu, K. M. Ting, and Z. Zhou. 2018. Multi-Label Learning with Emerging New Labels. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (Oct 2018), 1901–1914. https://doi.org/10.1109/TKDE.2018.2810872