Beyond Annual Revisions: A Multi-Label Concept Drift Analysis of MeSH*

1st Nikolaos Mylonas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
myloniko@csd.auth.gr

2nd Ioannis Mollas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
iamollas@csd.auth.gr

3rd Grigorios Tsoumakas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
greg@csd.auth.gr

***Abstract*—MeSH (Medical Subject Headings) is a hierarchically structured thesaurus used for indexing biomedical information. This vocabulary contains most of the biomedical knowledge available to date. To keep up with the continuous evolution and expanding of our understanding on the medical field, yearly revisions take place in MeSH. These revisions introduce new descriptors in the thesaurus, in addition to changes in already existing ones, either directly or indirectly. This constant evolution of the thesaurus causes many older descriptors to exhibit some form of drift in their meaning, which in turn affects the performance of Machine Learning models trained on an older version of the thesaurus when used to predict data obtained from more recent versions. In this paper, we study the phenomenon of concept drift in MeSH, through evaluating the performance of a state-of-the-art text classification algorithm in articles from different years. We also investigate how changes in descriptors indirectly affect different ones that are related to them by studying the shifts in their co-occurrence, using this shift as a measure of concept drift.**

***Index Terms*—Concept Drift, Multi-Label, Medical Subject Headings, Outlier Detection, Label Co-occurrence**

I. Introduction

New methods for indexing biomedical articles using Medical Subject Headings (MeSH) descriptors are constantly being introduced, in large thanks to the BioASQ challenge [1]. However, the volatile nature of MeSH, with its constant revisions, makes indexing models trained on previous versions of the vocabulary less effective on newer ones. The cause of this accuracy decrease is two-fold, with the most obvious one being the changes in the vocabulary, and specifically the introduction of new descriptors [2], [3], which the model has never seen before. The second, more inconspicuous reason is the drift in meaning of older descriptors, which is introduced during MeSH revisions or due to the passing of time.

The latter is addressed by techniques from the concept drift domain, which detect and adapt to drift events [4]. Concept drift finds application in a variety of domains, including hate speech detection [5], where temporal changes in vocabulary occur, hospital discharge records classification [6], in which changes occur monthly, network intrusion detection [7], due to changes in network characteristics, and energy load forecasting [8], due to climate or equipment changes.

This work explores the concept drift phenomenon in models indexing biomedical articles with MeSH from two different perspectives. The first one is an investigation of the performance of a state-of-the-art text classification algorithm trained on MeSH articles of a particular past year and tested on subsequent ones. We also study how the inherent multi-label nature of the thesaurus and the relationships between MeSH descriptors may indirectly cause their meaning to drift, affecting their performance. Secondly, we study the shifts in co-occurrence between descriptors and use them as an indicator of a possible drift.

The majority of drift detection methods are used on data sets with artificially created drift. In contrast, we contribute real-world data sets and use them for our analysis. Particularly, we modify the BioASQ[1] challenge data sets for the MeSH versions 2013-2019, to only include articles from the respective year, covered by a common set of descriptors. The modified data sets and the code for our work is available at our GitHub[2].

II. Related Work

Concept drift refers to the phenomenon, where the relationship between input and output values of the examined data changes over time. This phenomenon has been given different names, such as concept shift [9] or data set shift [10]. Concept drift is generally categorized in six types, namely *sudden*, *incremental*, *gradual*, *recurring*, *blip* and *noise* drifts [11], which differ in the relation between the input and target data of the drifting concept. Three approaches for dealing with concept drift are identified: weight-based, window-based, and ensemble-based. The latter two are the most popular ones, while drift detection is usually performed by observing the error of the model on fresh instances.

The most common concept drift approach is the Drift Detection Method (DDM) [12]. DDM monitors the error rate p_t of a learning model, as well as its standard deviation s_t during a specific time step t, constantly updating these values with each step. When the sum of these values passes a specific

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)

[1]http://bioasq.org/
[2]https://tinyurl.com/2sa9sbue

TABLE I
DATA SET SIZES FOR EACH EXAMINED MESH YEAR WHEN INCLUDING ALL DESCRIPTORS, THE TOP 300 MOST FREQUENT EACH YEAR, THE 198 MOST FREQUENT ACROSS ALL EXAMINED YEARS

	Year	2013	2014	2015	2016	2017	2018	2019
	# Desc.	24,933	24,620	23,662	23,319	25,094	25,361	25,471
# Articles	**All Desc.**	541,024	368,527	243,035	129,998	224,424	224,608	237,079
	Top 300	317,446	241,628	147,611	119,896	206,901	205,897	215,195
	Final 198	304,116	231,231	141,980	113,155	196,326	191,580	197,180

threshold, a warning state is signaled. If they pass a second threshold, then concept drift is detected, and the learning model is retrained.

There has been a lot of research dealing with single label concept drift, particularly in data streams. One such work uses k-means to detect and address possible concept drifts [13], by creating clusters for the instances belonging to the normal concept and derive a global boundary, which is used to identify instances belonging to unknown ones. Another work proposes the use of ensembles with dynamic integration of classifiers to handle local concept drift in the domain-specific problem of *antibiotic resistance in nosocomial infections* [14]. Local concept drift refers to the phenomenon of changes occurring only in specific parts of the instance space. A performance-based concept drift detection method in evolving data streams, called Accurate Concept Drift Detection Method, was also introduced [15], using Hoeffding's Inequality to observe the changes in the error rate of the learner. A concept drift detection method based on contrastive auto-encoders, named CADE, detects underlying concept drift and explains how the drift occurred [16]. This is accomplished by identifying a feature space's subset that differentiates the drifting examples from the normal ones present in the training data set.

Concerning multi-label concept drift, a study introduces a sliding window-based approach for drift detection in stream classification, making use of two different windows for each label, for positive and negative examples, respectively [17]. One similar approach uses two windows to simulate short-term and long-term memory (STM and LTM) [18]. STM stores the m most recent instances, where m is the window size, while LTM stores older instances. STM, LTM, and the combined union of the two, along with a multi-label kNN classifier, make decisions for each label independently. An ensemble-based method employs a weighted ensemble of classifiers, each trained in a different chunk of the examined data stream, to address concept drift [19]. The weight of each classifier is adjusted based on its performance in the current data.

Analysing the temporal shift in MeSH data sets, three different types of data distribution changes can be identified [20]: a) the interest of researchers changes with the passing of time, b) the National Library of Medicine of USA changes the thesaurus each year, and c) the annotation rules change each year, possibly resulting in the same article being indexed with different descriptors during different years. To address this, a method called Time-aware Concept Embedding Learning (TaCEL) is employed, offering time-aware embeddings for any data set that follows the MeSH annotation rules.

Concept drift has been explored in several domains, including autonomous vehicle control systems and toxicology. In the former, a recent study proposed a framework that employs an online sliding window-based drift detection algorithm utilizing a distance metric, to effectively capture traffic pattern changes and prevent unnecessary model updates in time-varying traffic environments [21]. In the latter, a drift detection algorithm on graphs generated by a chemical compound described in multiple toxicology datasets to detect the impact of the compound on biochemical pathways is introduced [22].

III. DATA SET CREATION PROCESS

At the beginning of each year, a new BioASQ data set is introduced containing articles from MEDLINE, annotated with MeSH descriptors. The articles in these data sets are the ones introduced in MEDLINE up to that year and are indexed using the MeSH version of that year. For example, the BioASQ data set of year 2015 contains articles with dates up to 2015 and follows the indexing policy of MeSH 2015. The focus of this work is to analyze the concept drift of MeSH throughout the years. Therefore, the data set for each year should contain articles introduced during that year. To do so, we mined each BioASQ data set, looking only for the articles that specifically mention the year we are interested in. However, as each BioASQ data set is introduced at the start of the corresponding year of the challenge, the number of articles that refer to that year is very limited, in some cases less than 1000. Therefore, we use the BioASQ data set of the version that corresponds to the year right after the one we want to create the data set for. This way, the number of articles that refer to that year is much higher.

Following this process, we used the BioASQ data sets corresponding to years 2014-2020 to create the ones we used for our experiments that refer to 2013-2019. The number of descriptors for each year in these data sets surpasses 20,000. For computational reasons, we decided to focus on a smaller subset of them. We first removed the top 10 most frequent descriptors of each year, such as *Humans*, *Male*, and *Female*, as these descriptors typically have a very general meaning and are not particularly interesting. Then, we kept the 300 most frequent remaining descriptors of each year. To arrive at a common set of descriptors across the data sets of the 7 years of our study, we took the intersection of their descriptor sets, leaving us with 198 descriptors. The size of each year's data set can be found in Table I, where we show the number of articles available each year along with the number of descriptors found

among them. The number of articles corresponding to the top 300 and final 198 descriptors are also presented.

Interestingly, we observe that the number of articles shows significant differences from year to year. A possible cause for this is that MeSH indexers may fall behind schedule, which can affect the number of articles for a particular year. We can notice such an occurrence for 2016, where the number of articles corresponding to it in the BioASQ 2017 data set is rather small. Furthermore, by mining the 2016 articles from BioASQ 2018, we can see that this number increases to 496,445, indicating a delay in indexing during these years.

IV. Concept Drift Analysis in MeSH

With MeSH constantly evolving, new descriptors as well as changes to already existing ones are introduced. These changes, along with the passing of time, result in the meaning of certain descriptors drifting. This drift is easy to pinpoint when it is the consequence of a direct change to the descriptor, i.e. changes in its indexing policy. However, it can be hard when the reason is not as apparent, for example, changes in the context in which it is used. Since MeSH is inherently multi-label, with the terms present inside the vocabulary being part of a hierarchical structure and as such related to each other, shifts in one descriptor's meaning may affect the other ones as well. In this section, we will study the phenomenon of concept drift in MeSH from two different perspectives to determine if it occurs, as well as the reasons behind its occurrence.

A. Performance-based Semantic Shift Detection

For this analysis, we decided to evaluate the performance of a state-of-the-art text classifier on data annotated using different versions of the thesaurus. BERT [23] was selected for this procedure, which has competitive results on text classification tasks. We fine-tuned BERT on the task of text classification for 10 epochs and a batch size of 16, with data corresponding to the first available MeSH year in the BioASQ challenges, namely 2013. The classifier was then used to predict the data for the rest of the years. This way we can evaluate the performance of the model each year and catch drifts corresponding to changes in meaning. To do so, we compute the quantitative differences in F_1-score between consecutive years for each descriptor and examine these differences to find the descriptors that act as outliers. These descriptors are the ones we consider as having drift in meaning for that year pair. It is worth noting that we only keep the differences for descriptors that the model has an F_1-score of at least 0.1. This is done to avoid large fluctuations in the quantitative difference due to the very small values of F_1.

To better understand when these changes in performance between the examined years signify a possible drift, we calculate the percentage change in performance for each descriptor for a year pair $((F_1^{latter} - F_1^{former})/F_1^{former})$. Then, averaging these changes across all descriptors, we show the mean change of F_1-scores between each year pair, *2014-2015*: -0.39%, *2015-2016*: 0.53%, *2016-2017*: -0.93%, *2017-2018*: 1.92%, *2018-2019*: 1.34%. These values are signed numbers, with a positive sign denoting that the mean F_1 has increased between these years, while a negative one indicates a decrease. The mean quantitative difference between F_1 scores for each year pair is rather small, meaning that the performance for the majority of descriptors remains stable. Consequently, descriptors who exhibit much larger fluctuations during these year pairs can be considered as outliers and as a result candidates for having drift in their meaning. We can also see that for some year pairs, this mean difference is positive, meaning that there are descriptors who exhibit an increase in their performance.

1) Concept Drift identification based on outlier detection: In this section, we present the top 10 descriptors found as having a possible drift in their meaning for each year pair. The identification of these descriptors was done using Isolation Forest (IF) [24], on the computed F_1 differences for each year pair, keeping the 10 most drifting ones according to IF. Based on this procedure, Table II presents the descriptors found as most drifting (outliers) based on IF between these year pairs along with the change in F_1.

We can see from Table II that the most drifting descriptors are different for each year pair, with a few of them being present in multiple pairs. These abnormal changes in performance for those descriptors can be a sign of drift in their meaning, but can also be caused by other changes, for example, the frequency of that descriptor in the examined data set suddenly increasing or decreasing. Additionally, since the examined data are multi-label, changes in the descriptors relations with the other labels, can also affect the model's performance on that descriptor. Finally, yearly MeSH revisions that do not necessarily change the descriptors' meaning can also be the cause of these changes in performance, for example changes in their indexing policy.

What stands out the most is the 129.12% increase in performance of the model on the *Computational Biology* descriptor between years 2018 and 2019. This descriptor shows a 17.07% decrease in performance for years 2015-2016 and a 21.32% decrease for 2017-2018, while showing a smaller increase of 18.75% for 2016-2017. This volatile behavior is a sign of the descriptor's usage changing from year to year.

This becomes more apparent if we check the other descriptors appearing together for each year pair. During 2016, the descriptor *Models, Biological* started appearing more frequently alongside *Computational Biology*. Specifically, it emerges as its 5th most frequent descriptor, while in previous years it wasn't even in its top 10. This behavior changes again in 2017, where the most common descriptors used with it are mostly the same as in 2013, which was the year used to train the model, explaining the increase in performance. A similar pattern is observed during 2018, where *MicroRNAs* appears as its 5th most frequent descriptor, while not being used as frequently in previous years. Finally, in 2019, where the top 3 most frequent co-occurrent descriptors are the same as those in 2013, the performance of the model increases.

2) MeSH Descriptor relations and concept drift: Usually, concept drift appears as a consequence of changes happening directly on the examined label. In MeSH, these changes can be

TABLE II
TOP 10 MOST DRIFTING DESCRIPTORS PER YEAR PAIR BASED ON QUANTITATIVE DIFFERENCE

Year	Descriptor	F_1-diff	New
2014-2015	Survival Rate	11.42%	1
	Mice, Nude	11.74%	2
	Polymerase Chain Reaction	13.06%	2
	Chromatography, High Pressure Liquid	14.84%	2
	Species Specificity	-28.69%	1
	Recombinant Proteins	-21.68%	0
	Sex Factors	-15.74%	1
	Real-Time Polymerase Chain Reaction	-14.43%	0
	Image Processing, Computer-Assisted	-14.32%	1
	Rats	-10.49%	2
2015-2016	RNA, Small Interfering	17.08%	1
	Recombinant Proteins	21.27%	3
	Image Processing, Computer-Assisted	23.13%	1
	Real-Time Polymerase Chain Reaction	30.84%	2
	Base Sequence	49.24%	3
	DNA, Bacterial	-18.46%	0
	HeLa Cells	-18.08%	2
	Computational Biology	-17.07%	4
	RNA, Ribosomal, 16S	-16.83%	1
	Software	-13.14%	1
2016-2017	NF-kappa B	11.32%	2
	RNA, Ribosomal, 16S	14.18%	1
	Transcription Factors	14.93%	2
	Computational Biology	18.75%	3
	DNA, Bacterial	28.55%	0
	Base Sequence	-40.86%	0
	RNA, Small Interfering	-24.00%	0
	Image Processing, Computer-Assisted	-16.85%	1
	Mass Spectrometry	-16.79%	2
	Models, Biological	-15.55%	3
2017-2018	Phenotype	16.47%	4
	DNA	21.00%	1
	Polymerase Chain Reaction	22.95%	5
	Enzyme-Linked Immunosorbent Assay	24.46%	5
	Databases, Factual	28.58%	3
	Real-Time Polymerase Chain Reaction	37.76%	3
	Computational Biology	-21.32%	3
	Models, Theoretical	-19.81%	0
	Proportional Hazards Models	-16.45%	2
	Disease Progression	-14.58%	1
2018-2019	Databases, Factual	23.48%	2
	Body Weight	31.12%	2
	Dose-Response Relationship, Drug	35.64%	2
	Gene Expression Regulation	39.72%	0
	Computational Biology	129.12%	1
	Disease-Free Survival	-34.4%	2
	Mice, Nude	-26.78%	1
	Kaplan-Meier Estimate	-25.91%	2
	Species Specificity	-21.04%	4
	Protein Conformation	-16.20%	1

a result of the yearly revisions of the vocabulary, modifications of the indexing policy, or simply the meaning of the descriptor shifting to something else over the years. Examining descriptor *Rats* in Table II, we notice a decrease in F_1 between 2014 and 2015. This decline is a result of the indexing policy for that specific descriptor changing during 2015. In particular, the descriptor *Rats*, which has a very general meaning, was decided as of 2015 to no longer be used on articles where one of its children terms is selected by the indexer[3].

Even though descriptors may drift in meaning due to direct changes, in a hierarchically structured thesaurus such as MeSH, concept drift can be caused indirectly as well, through changes in related concepts. By studying the changes in the vocabulary in conjunction to the descriptors with the biggest shift in their performance, we found that a descriptor can exhibit a drift in its concept indirectly, due to changes in its relations with other descriptors. MeSH descriptors are closely related to each other. Hence, changes in one descriptor may indirectly affect others that are correlated with it. These changes may cause that descriptor to be indexed differently, causing an indirect concept drift.

An example of such a scenario concerns descriptor *Base Sequence*, that shows a decrease in performance of 40.86% in 2016-2017. During the MeSH revisions of the latter year, descriptor *Molecular Sequence Data*, which is parent-child related to our examined one, had its indexing policy changed[4]. Between 1988-2016 this term was indexed for articles that contained: i) accession numbers for sequences deposited in a molecular sequence databank, such as Genbank, ii) base sequences of 50 or more bases, iii) amino acid sequences of 15 or more amino acids, iv) carbohydrate sequences of 3 or more carbohydrate units. During the 2017 revisions, this indexing policy changed so that the descriptor in question will only be indexed for general articles about sequence data.

B. Co-Occurrence Based Semantic Shift Detection

MeSH indexing is a multi-label text classification task. As such, another technique to determine if a descriptor exhibits concept drift is to examine the other descriptors commonly indexed with it. Changes in co-occurrence may indicate that the descriptor's meaning is drifting. The main motivation behind this analysis is that in a multi-label learning task, drift in the meaning of a descriptor is expected to affect the descriptors it frequently appears alongside with. To capture the changes between the most popular descriptors used in tandem each year, we chose to conduct an analysis of the top 10 most common descriptors used in conjunction to our investigated ones. For each one of the 198 descriptors, we obtain 7 co-occurrence sets of 10 descriptors (one per year from 2014 to 2019) that denote the ones it was most commonly used with in that year. The difference between two such consecutive sets indicates the different descriptors between these years, and can be used as a measure of change in that descriptor's meaning.

In Figure 1, the number of different co-occurrent descriptors for each year pair, along with the percentage of the total descriptors that exhibit these differences, are presented. The x-axis of the plot denotes the year pair, with each number representing the new descriptors appearing in the co-occurrence set of the former year when compared to the latter one. Finally, the y-axis indicates the percentage of descriptors that showcase the difference.

In Table III, we show the descriptors who have the most differences in their co-occurrence sets for each year pair. Based on Figure 1, for each year pair the vast majority of descriptors (around 90%) had between 0, 1, and 2 different descriptors in

[3]https://tinyurl.com/8kez6yed

[4]https://tinyurl.com/3m5h2w5d

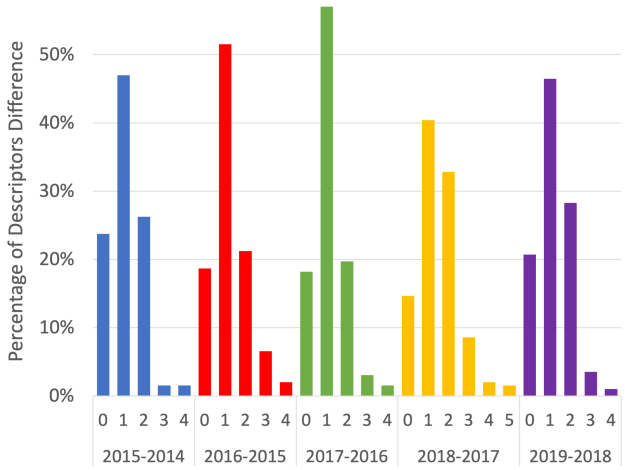

Fig. 1. Co-occurrence difference plot for each year-pair.

each year pair, while the rest had either 3, 4, or 5. Therefore, the descriptors shown in the table belong to the second group, with most of them having 3 new co-occurrent descriptors between their respective year pairs. This means that in the latter year of the year pair, the descriptor started to appear more frequently with 3, 4, or 5 new descriptors that it was less commonly used with in the past, signifying a possible drift in meaning for that descriptor.

The descriptor *Dogs* appears in almost every year pair, being frequently used alongside different descriptors each successive year. This is most likely due to the descriptor having a very broad meaning and hence being used in every article related to dogs. Similarly, descriptor *Species Specificity* is used to denote the differences between organisms of different species, essentially being used in articles that reference characteristics that separate one species from others.

C. Performance and Co-occurrence based comparisons

Going one step further, we combine the results of the two methods. Thus, the decisions made by the co-occurrence based method for the performance-based descriptors can be found in Table II while those made by the latter for the descriptors of the former can be found in Table III. We can see from the two tables that there are common descriptors between these two methods. Specifically, there are 10 common descriptors, which indicates a link between descriptor co-occurrence and model performance in the vocabulary of MeSH. A descriptor, being present in both tables during the same year-pair, can signify that the changes in its co-occurrences are the cause of the abnormal fluctuations on the model's performance for that descriptor in that year-pair.

However, that may not always be the case, since there are descriptors with many co-occurrence changes that do not exhibit any significant difference in performance. The latter can be observed for descriptors with broad meanings that are used in various topics, and as such the descriptors they frequently appear with can change yearly without significantly impacting the performance of the model. These descriptors appear as most drifting by the co-occurrence based method during multiple year-pairs, as discussed in Section IV-B.

In addition, the same phenomenon may be caused due to these shifts in co-occurrences, resulting in the most common descriptors used alongside each other being closer to the train set used during the training of the model. Such an example is for descriptor *Calcium* who has 4 different co-occurrent descriptors between 2014-2015 while having only 2 different ones between 2013-2015, meaning that it was used in a context more similar to the train set during 2015.

TABLE III
MOST DRIFTING DESCRIPTORS PER YEAR PAIR BASED ON CO-OCCURRENCE

Year	Descriptor	New	F_1-diff
2014-2015	Fibroblasts	3	-0.32%
	Liver Neoplasms	3	2.75%
	Water Pollutants, Chemical	3	-2.91%
	Calcium	4	7.55%
	Dogs	4	-1.71%
	Transcriptome	4	3.32%
2015-2016	Antioxidants	3	4.87%
	Base Sequence	3	49.24%
	Calcium	3	0.40%
	Fibroblasts	3	2.37%
	Oxidative Stress	3	2.33%
	Recombinant Proteins	3	21.27%
	Computational Biology	4	-17.07%
	Dogs	4	5.08%
	Polymerase Chain Reaction	4	-8.46%
	Transcriptome	4	10.05%
2016-2017	Cognition	3	-2.47%
	Computational Biology	3	18.75%
	Models, Biological	3	-15.55%
	Software	3	9.01%
	Swine	3	3.81%
	Tandem Mass Spectrometry	3	0.65%
	Binding Sites	4	-4.50%
	Dogs	4	3.75%
	Species Specificity	4	-1.46%
2017-2018	Cell Differentiation	3	2.97%
	Cell Movement	3	8.57%
	Comorbidity	3	13.03%
	Fibroblasts	3	1.76%
	Hydrogen-Ion Concentration	3	2.04%
	Models, Biological	3	6.71%
	Computational Biology	3	-21.32%
	Databases, Factual	3	28.58%
	Real-Time Polymerase Chain Reaction	3	37.76%
	Base Sequence	4	7.79%
	Diet	4	-0.08%
	Immunohistochemistry	4	7.05%
	Phenotype	4	16.47%
	Enzyme-Linked Immunosorbent Assay	5	24.46%
	Polymerase Chain Reaction	5	22.95%
	Species Specificity	5	-9.25%
2018-2019	Cognition	3	-11.28%
	Logistic Models	3	-3.95%
	Mice, Knockout	3	-1.25%
	Neoplasm Staging	3	5.15%
	Odds Ratio	3	$< 0.1\%$
	Real-Time Polymerase Chain Reaction	3	12.57%
	Tumor Necrosis Factor-alpha	3	7.15%
	Dogs	4	0.56%
	Species Specificity	4	-21.04%

V. Conclusions

In this work, we studied the phenomenon of concept drift in MeSH from two different viewpoints: i) the performance of a state-of-the-art model for text classification on each descriptor during different years, ii) shifts in the co-occurrences of descriptors from year to year. Concept drift in MeSH can be a direct result of the yearly changes taking part in the vocabulary, or a product of the meaning of descriptors changing through time.

Through our analysis, we found that due to the complex hierarchical structure of the thesaurus, changes to one descriptor may also indirectly influence other related ones, causing models trained on earlier iterations of the thesaurus to miss-classify these descriptors more frequently. Moreover, co-occurrence shifts between descriptors seem to be linked to model performance, but this is not always the case. Descriptors whose performance was lowered between two consecutive years, which also exhibit changes in their co-occurrent descriptors in the same years, are the most prone to have drift in their meaning. This highlights the importance of regularly updating the models used to predict descriptors from the MeSH thesaurus, to accurately reflect the evolution of our biomedical knowledge

A few limitations of this work include that several newly emerging biomedical concepts might not be yet well-represented in MeSH, and thus cannot be captured by our analysis. Furthermore, our analysis primarily relied on quantitative measures, and further qualitative analysis could provide deeper insights into the reasons behind concept drift. In the future, we aim to perform our analysis using multi-label models or strategies that consider label dependencies. This way, we can get a better insight on how the shifts in co-occurrences affect the performance, since the aforementioned models are more sensitive to these kinds of changes. Additionally, collaboration with medical experts would allow us to identify biomedical topics that are more prone to concept drift based on our findings. Finally, future studies should consider examining concept drift over longer time frames to gain a more comprehensive understanding.

References

[1] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, L. Gascó, M. Krallinger, and G. Paliouras, "Overview of bioasq 2021: The ninth bioasq challenge on large-scale biomedical semantic indexing and question answering," *CoRR*, vol. abs/2106.14885, 2021. [Online]. Available: https://arxiv.org/abs/2106.14885

[2] N. Mylonas, S. Karlos, and G. Tsoumakas, "A multi-instance multi-label weakly supervised approach for dealing with emerging mesh descriptors," in *Artificial Intelligence in Medicine*, A. Tucker, P. Henriques Abreu, J. Cardoso, P. Pereira Rodrigues, and D. Riaño, Eds. Cham: Springer International Publishing, 2021, pp. 397–407.

[3] A. Nentidis, A. Krithara, G. Tsoumakas, and G. Paliouras, "What is all this new mesh about? exploring the semantic provenance of new descriptors in the mesh thesaurus," *CoRR*, vol. abs/2101.08293, 2021. [Online]. Available: https://arxiv.org/abs/2101.08293

[4] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2019.

[5] L. Justen, K. Müller, M. Niemann, and J. Becker, "No time like the present: Effects of language change on automated comment moderation," in *2022 IEEE 24th International Conference on Business Informatics (CBI)*, 2022.

[6] G. Stiglic and P. Kokol, "Interpretability of sudden concept drift in medical informatics domain," in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 609–613.

[7] G. Andresini, F. Pendlebury, F. Pierazzi, C. Loglisci, A. Appice, and L. Cavallaro, "Insomnia: Towards concept-drift robustness in network intrusion detection," in *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, ser. AISec '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 111–122.

[8] R. K. Jagait, M. N. Fekri, K. Grolinger, and S. Mir, "Load forecasting under concept drift: Online ensemble learning with recurrent neural network and arima," *IEEE Access*, vol. 9, pp. 98 992–99 008, 2021.

[9] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine Learning*, vol. 23, no. 1, pp. 69–101, Apr 1996.

[10] S. Amos, "When training and test sets are different: Characterizing learning transfer," in *Dataset Shift in Machine Learning*. The MIT Press, Dec. 2008, pp. 2–28.

[11] K. Wadewale and S. Desai, "Survey on method of drift detection and classification for time varying data set," 2015.

[12] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *Advances in Artificial Intelligence – SBIA 2004*, A. L. C. Bazzan and S. Labidi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 286–295.

[13] E. Spinosa and J. Gama, "Olindda: A cluster-based approach for detecting novelty and concept drift in data streams," pp. 448–452, 01 2007.

[14] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen, "Handling local concept drift with dynamic integration of classifiers: Domain of antibiotic resistance in nosocomial infections," in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, 2006, pp. 679–684.

[15] M. M. W. Yan, "Accurate detecting concept drift in evolving data streams," *ICT Express*, vol. 6, no. 4, pp. 332–338, 2020.

[16] L. Yang, W. Guo, Q. Hao, A. Ciptadi, A. Ahmadzadeh, X. Xing, and G. Wang, "CADE: Detecting and explaining concept drift samples for security applications," in *30th USENIX Security Symposium*. USENIX Association, Aug. 2021. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/yang-limin

[17] E. Spyromitros-Xioufis, M. Spiliopoulou, G. Tsoumakas, and I. Vlahavas, "Dealing with concept drift and class imbalance in multi-label stream classification," 01 2011, pp. 1583–1588.

[18] M. Roseberry and A. Cano, "Multi-label knn classifier with self adjusting memory for drifting data streams," in *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, ser. Proceedings of Machine Learning Research, L. Torgo, S. Matwin, N. Japkowicz, B. Krawczyk, N. Moniz, and P. Branco, Eds., vol. 94. ECML-PKDD, Dublin, Ireland: PMLR, 10 Sep 2018, pp. 23–37. [Online]. Available: http://proceedings.mlr.press/v94/roseberry18a.html

[19] Y. Sun, H. Shao, and S. Wang, "Efficient ensemble classification for multi-label data streams with concept drift," *Information*, vol. 10, no. 5, 2019.

[20] Q. Jin, H. Ding, L. Li, H. Huang, L. Wang, and J. Yan, "Tackling mesh indexing dataset shift with time-aware concept embedding learning," in *Database Systems for Advanced Applications*, Y. Nah, B. Cui, S.-W. Lee, J. X. Yu, Y.-S. Moon, and S. E. Whang, Eds. Cham: Springer International Publishing, 2020, pp. 474–488.

[21] S. Lee and S. H. Park, "Concept drift modeling for robust autonomous vehicle control systems in time-varying traffic environments," *Expert Systems with Applications*, vol. 190, p. 116206, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421015207

[22] V. Bharti, S. S. Nair, A. Jain, K. Kumar Shukla, and B. Biswas, "Concept drift detection in toxicology datasets using discriminative subgraph-based drift detector," *Briefings in Bioinformatics*, vol. 24, no. 1, 12 2022, bbac506. [Online]. Available: https://doi.org/10.1093/bib/bbac506

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[24] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *8th IEEE International Conference on Data Mining*, 2008, pp. 413–422.