Topic Modeling and Multi Output Classification of Pubmed Abstracts

For Cancer Clinical Trials and Randomized Controlled Trials

Michael McManus, Antoine Wermenlinger[t]

University of Michigan, Masters of Applied Data Science, Milestone 2 Project
Submitted 26 March 2021 | https://github.com/awermenlinger/Milestone2

Abstract

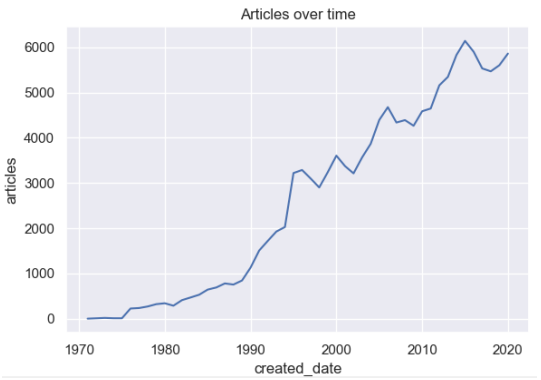
A topic model was built over 132,660 pubmed abstracts on the topic of cancer. From these we were able to demonstrate trends in oncology over the last 35 years. While topics like radiotherapy and hematologic toxicity have reduced in the overall research body; others like safety and efficacy as well as healthy habits and behavioral intervention have increased. In addition to modeling topics, a multi output classification model was trained to predict MeSH term tags associated with PubMed articles pertaining to cancer. We believe this work could provide researchers a way to accurately predict which MeSH terms their articles should be tagged with, thus saving time and increasing accuracy by reducing human error.

Introduction & Motivation

We explored the world of cancer research over the past 35 years in this paper to better understand if topic modeling and supervised learning can add insights into the overall domain. While not researchers in cancer, we believed that we can contribute by providing a different understanding of the domain through these methods.

Data Gathering

Pubmed was the main source of data for this paper. Using the Entrez python library as well as our own API keys (free), we proceeded to extract all the abstracts, mesh terms, publication dates and authors of the papers relating to the keyword cancer and which were in the Clinical Trial and Randomized Control trial subcategories. These categories were selected as they indicate much more funding requirements and are usually more reflective of the changes in a treatment area. Medical science follows the pattern: Theory, Lab, Invivo, Trials (phase 1, phase 2, phase 3), Regulatory Approval, post approval studies (phase 4, phase 5). The most costly of these being the phase 3 clinical trials. We ended up with 132,660 abstracts over the last 35 years.We encountered our first roadblocks in this exercise learning how to deal with pubmed's API and Entrez as well as how to work around the large files limitation of github. In this case we split the data into multiple CSV files. To properly extract the files without exceeding PubMed's call limit and blocking our IP,, we had to specify a delay in retrieval and set up our code to continuously append a csv file.



t : equal contribution, order is alphabetical

Unsupervised Learning - Topic Modeling

Methods

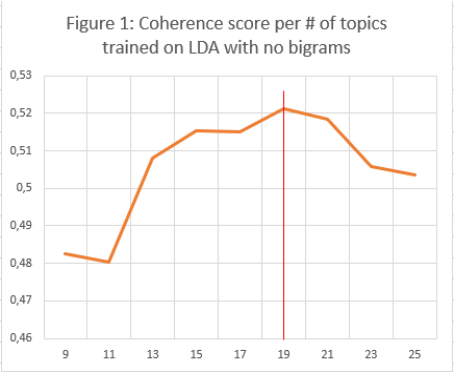
Before diving into the work itself, a high level research of topic modeling methods was performed to identify potential models to use. Three were selected: NMF, LDA and BERTopic. We also explored different parameters for text preprocessing: simple tokenization, lemmatization, stemming, bigrams, stopwords and a deeper stopwords list. The code was set up on a GitHub repository and we used Visual Studio Code and PyCharm as integrated development environments (IDE) and Git for version control. Visualizations were done with pyLDAvis, BERTopic, Seaborn, Matplotlib, Altair, and Excel.

Training and Learnings

The first portion of the process was transforming the large corpus of 132,660 pubmed abstracts on the topic of cancer. To increase relevance, the filters in pubmed were set to randomized control trials and clinical trials. These were selected as they are the main goal of this exercise to understand where on the ground cancer research is being performed. From a text transformation perspective we wanted to see if different transformations would have a positive or negative impact on the quality of the topics extracted. The code was set up with pickle to export the different steps of text processing. In the end we settled on 4 texts corpus: [no treatment], [stopwords, lemmatized and stemmed], [stopwords, lemmatized and stemmed with tf-idf], [stopwords, lemmatized and stemmed, tf-idf and bigrams]. This was in part done to avoid the many iterations of all the different possible combinations. The “no-treatment” category was only applicable to Bertopic. As an interesting note, when testing the visualization with PyLDAvis we noticed that the most common words were patient, group, trial, placebo, etc. so these were added to the stop words list to improve the output of the results.

LDA Model

We started with the Latent Dirichlet Allocation model (LDA) as it was the first we came across in our research. This model has become the de-facto approach for topic modeling. We needed to find a way to target the right amount of topics for the corpus. Here the first approach taken was to use the [stopwords, lemmatized and stemmed] text corpus and run it through many iterations of the LDA model to calculate the coherence value of the resulting topics. This gave us a target number of topics of 19 topics. We then used hierarchical dirichlet process (HDP) to infer the # of topics[1] as it was designed for this purpose. The results were the same: 19 topics. Further research led us to the optimization parameters of LDA alpha: α and eta: η these determine the prior distribution over topic weights in each document and the prior distribution over word weights in each topic respectively. To tune these hyperparameters



was very computationally expensive, to save time we used only nearby quantity of topics [15, 17, 19, 21, 23] and restrained the search for both parameters to [0.01, 0.5, 1, 'symmetric', 'asymmetric'] (without asymmetric for eta as it is not available). We also used the multicore version of gensim's LDA model to speed things up. With these reductions, it still took almost 3 days to get the results. The top 5 optimal parameters when measuring coherence (Table 1) were interesting as they showed better results than at 19 topics. We then re-ran the LDA for tf-idf and bigram models with these parameters (α = symmetric, η = 1). This gave us a result of 0.5842 for standard text preprocessing, 0.5811 with tf-idf and 0.4326 with bigrams added in. We therefore picked the first model.

Table 1

# Topics	Alpha	Beta	Coherence
21	symmetric	1	0,5842
21	0,01	1	0,5624
17	0,5	1	0,5617
21	asymmetric	1	0,5569
21	0,5	0,01	0,5562

NMF Model

The non-negative matrix factorization (NMF) model was then trained using 21 topics, initialization with Nonnegative Double Singular Value Decomposition (NNDSVD) since sklearn recommends it for sparseness. Sklearn was used since we learned through trial and error that the sklearn was able to compute on our text corpus while the gensim model failed to execute or even give out relevant errors. Due to time and processing constraints, we used the default parameters for NMF. Since there are no coherence measures integrated into sklearn for NMF, hyperparameter tuning would have been very difficult.

BERTopic

We made an attempt at a constrained BERTopic model to target 21 topics, but the algorithm was not able to reduce dimensionality with UMAP (this was tested for 48h+). So with an unconstrained BERTopic model, we arrived at 60 topics, however, 50,030 documents were "topic-less" [-1]. Trying to generate a coherence score for BERTopic was a complex process that we should probably have stopped earlier. There is no direct way in the model to calculate topic coherence, so there is a need to patch gensim and BERTopic together, this led to obscure error and references and checking the source code for BERTopic. The learning is that: sometimes one needs to stop and consider the amount of time involved vs the reward.

So the decision was to compare the outputs of the NMF model and the selected LDA topic models. To do this, once the models were selected and topics created, we picked the most relevant abstract for each topic (the centroid) and applied sentence summarization through transformers[9] for text summarization to help us name the topics. This gave us the following results:

<u>NMF Topics:</u>

Topic # 0	Topic # 1	Topic # 2
toxic	surgeri	tumor
grade	postop	recurr
toler	oper	bladder
phase	resect	resect
cycl	complic	hcc
activ	group	tissu
combin	undergo	brain
infus	surgic	carcinoma
partial	laparoscop	malign
respons	preoper	size
administ	gastric	local

```
0: our report describes the safety and toxicity profile of irinotecan , an oral small - molecule
inhibitor of phosphatidylinositol 3-kinase ( pi3k ) , in combination with gemcitabine in a phase iia
safety trial , a phase iib safety trial and a phase iiia safety and efficacy trial in advanced non -
small cell lung cancer ( nsclc ) , and a phase iiia safety and efficacy trial in metastatic nsclc

1: the efficacy of postoperative prophylactic defibrotide and heparin was evaluated in group of 47
patients undergoing general surgery .<n> the first defibrotide dose was 400 mg i.m .<n> b.i.d .<n>
heparin was administered on postoperative day 1 .<n> the group of 47 patients was divided into two
subgroups .<n> the first group was treated with postoperative prophylactic defibrotide and heparin .<n>
the second group was treated with postoperative prophylactic defibrotide and calcium

2: findings in 180 patients with cancer of the larynx and hypopharynx , including supraglottic , glottic
, pyriform fossa , posterior pharyngeal wall and posterior wall tumors , have been reported .<n>
inadequate margins were found in 12.2% of all tumors , 12.5% of supraglottic tumors , 16% of
transglottic tumors , 11.1% of pyriform fossa tumors , and 5.3% of posterior pharyngeal
```

For example, the interpretation for the first 3 topics: topic 0 centroid is non-small-cell lung cancer, this does not match very well with the key words; topic1: postoperative surgery - very clear from keywords; topic 2: head and neck surgery, but the key words do not match.

LDA Topics:

Topic 0	Topic 1	Topic 2
placebo	pet	progress
infect	fdg	phase
receiv	uptak	grade
random	thyroid	diseas
incid	valu	overal
risk	tomograp	safeti
control	scan	event
oral	therapi	pfs
compar	suv	advers
group	evalu	efficaci

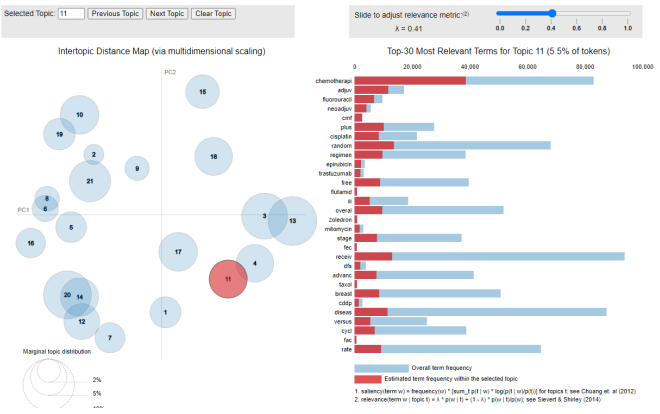
```
0: fungal infections account for more than a third of all cancer - related deaths1 .<n> the efficacy and
tolerability of the systemic anti - fungal agent fluconazole , which was developed in the late 1980s to
treat skin and soft - tissue infections , has been questioned because of side effects .<n> a recent us
food and drug administration report on the effectiveness of fluconazole in neutropenic leukemia patients
suggested otherwise .<n> here , we review the existing literature on the effects of

1: objective : to determine if there is an inverse relationship between thyroid stimulating hormone (
tsh ) and glucose utilization in patients with metastatic differentiated thyroid
carcinoma.methods:seventy-four patients with metastatic lesions underwent whole - body mri and were
divided into two groups : ( 1 ) control ( n = 32 ) and ( 2 ) group ( n = 64 ) , which were compared in
terms of uptake of tsh and glucose .<n> the

2: brigatinib is an oral inhibitor of the serine / threonine kinase 4,nine that is being investigated
for the treatment of advanced nonsmall cell lung cancer as well as for the treatment of other types of
cancer , because it is active against several kinases that are highly resistant to conventional
inhibitors ( e.g. , crizotinib ) .<n> we conducted a phase 1/2 trial in nine academic hospitals to
assess the safety and efficacy of brigatinib in patients with
```

Repeating the above exercise: Topic 0 infection risks, which matches well with the keywords; Topic 1 thyroïd cancer, which also matches well with the keywords; topic 2 safety and efficacy which also matches well.

Based on these results we decided to proceed with the LDA Model for the next steps (visualization and trending over time).

One important point to note is that even though we selected the LDA topics for understandability, the speed of the NMF model outperformed significantly the LDA model (reduction by a factor of 10 to 20). This is due to the NMF uses matrix factorization for multivariate analysis while LDA is a probabilistic model.

Refining Topics



To refine the topics pyLDAvis (right) was used to visualize which keywords were more salient in the corpus for each topic. Also, we consulted some Subject Matter Experts to better understand the interplay between some of the terms and the overarching theme in some of the topics. This helped refine how the model built itself. For example, topic 11 [chemotherapi, random, receiv, adjuv, diseas, plus, therapi, regimen, overal, compar] was very difficult to label using only the summarized text and the keywords. However, searching with pyLDAvis, some google-fu and some quick texts with SME colleagues (Antoine works in Biotech): the following words highlighted the central theme [adjuv, trastuzum, fluorouracil, flutamid]. They first led to first antimetabolites (molecules similar to normal body molecules but they have a slightly different structure that stop cancer cells working properly) and more broadly to adjuvant therapies as we kept digging at the core common concept.

The naming exercise was repeated for all 21 topics in the LDA model giving us the following list of topics: infection risk, thyroid cancer, safety and efficacy, leukemia chemotherapy, surgical intervention, detection with

lymph nodes, pain management, cervical cancer, bladder cancer, risk prediction, adjuvant therapy, healthy habits, hematologic toxicity, surgical complications, tumor angiogenesis, intraoperative radiation therapy, external radiotherapy, stem cell transplantation, glioma, behavioral intervention, prostate cancer. One may ask themselves why there is no specific breast cancer topic. The LDA model tries to find words that are more uncommon, and since "breast" was a keyword in many topics, this may be that it got redistributed. Inspecting the model for "breast" the returned topics which are all extremely relevant:

- detection with lymph nodes - lymph nodes are one of the primary detection techniques for breast cancer[2]
- adjuvant therapy - originated from the National Surgical Adjuvant Breast and Bowel Project (NSABP) in the 60s
- healthy habits - would include frequent self examination
- intraoperative radiation therapy - frequently used in early stage breast cancer

Results & Visualization

Now the question arises, how has the field of cancer research evolved in the last 35 years. As mentioned in the Data Gathering section, the amount of published articles has grown significantly year on year. Therefore, to properly visualize how the field has moved we calculated the proportion of the topic weights. For example, if we resampled by month (t_r), we added all the weights of a specific topic (w_i), then divided by the overall sum of the weights of all topics at that resampled time (W) to obtain the proportion of topics in the sampled timescale. This gave us the following visualizations:

$$w_{tr,i} = \frac{\sum_{tr} w_i}{\sum_{tr} W}$$

Fig 1. Topics over time with their own Y-axis:

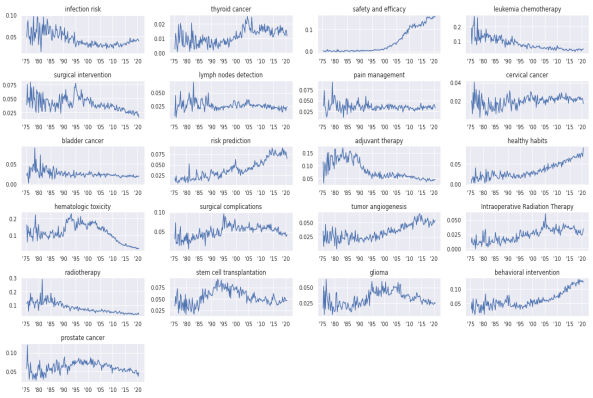
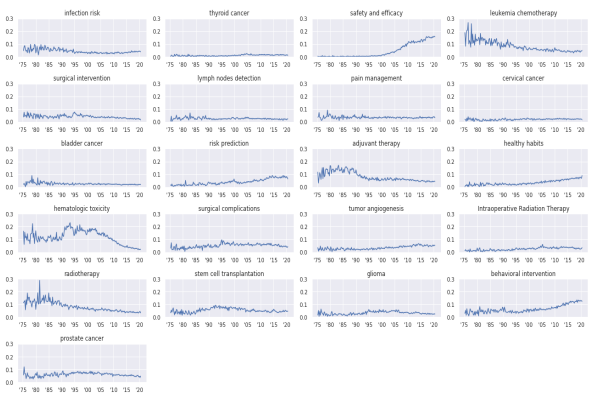


Fig 2. Topics over time with same Y-axis:



The full graphs can be found here: https://github.com/awermenlinger/Milestone2/tree/main/results

Some interesting trends can be identified:

- We notice in the same Y axis plot an increase in the topic of healthy habits and behavioral intervention which points to a shift from treatment only mentality to more proactive prevention measures. Which is consistent with medicine shifting towards a more proactive/preventative approach over the last decades. For example the Affordable Care Act in the US mandates preventative care[3]
- We also notice a decrease in the proportion of topics for leukemia chemotherapy, hematologic toxicity radiotherapy and adjuvant therapy which makes sense since these are older cancer treatment approaches and may elicit less research.

- Finally, some topics seem to remain constant and at a smaller proportion of the overall corpus: glioma (brain cancers), bladder cancer, cervical cancer, lymph node detection and pain management. This may or may not indicate progress on these topics. What it does indicate is that less clinical trials or RCTs have been run on them.

Supervised Learning - Multi Output Classification

Methods

The goal of our supervised learning approach was to create a classifier that could predict MeSH Terms associated with PubMed article abstracts. MeSH Terms are "the NLM controlled vocabulary thesaurus used for indexing articles for PubMed"[4]. In short, each article is tagged with roughly 5-10 MeSH terms to aid in indexing the article. Our goal was to create a multi output classifier which could predict MeSH terms associated with each PubMed article based on its abstract.

Our first step toward realizing this goal was to use MultiLabelBinarizer to create one column for each unique mesh term. To begin, we used a multi labeled binarizer to create columns based on each unique MeSH term. Binarizing in this way meant we created one column for each unique MeSH term; 0 indicating the MeSH term was not associated with the pub Med article and 1 indicating it was associated. Upon binarizing in this way, we realized quickly that the MeSH terms created a sparse matrix and thus, creating a multi output classifier would be rather challenging. In addition, many of the MeSH terms had slight variations in their spelling, case, or plurality which led to duplication of MeSH terms. To compensate for this, while loading the dataset we converted all MeSH terms to lower casing, lemmatized, and removed punctuation including hyphens and slashes from all MeSH terms thereby reducing duplicate MeSH terms throughout the data set from 72,168 to 54,471 thereby reducing the total number of duplicate MeSH terms by 17,697.

Once we ensured that all MeSH terms were unique, we began the process of cleaning and identifying appropriate representations of the abstract text. Several methods were attempted including TFIDF vectorizer and BERT word embeddings. The first method we attempted was to use TFIDF vectorization to represent the abstract text. Running a simple K nearest neighbors classifier using the TFIDF vectorized text, it became clear that accuracy score was not an appropriate way to measure model performance. This again is due to the sparse nature of the matrix being that MeSH terms are rarely duplicated throughout the entire dataset. Instead, to measure model performance we used F1 score which is the harmonic mean of precision and recall. We also used micro F1 score to see model performance at the micro level. Simply put, micro averaged F1 score measures the F1 score of the aggregated contribution of all the classes. This method is commonly used in multi output classification problems as labels that are very rare may not be intended to influence the overall F1 score heavily if the model is performing well on the other more common genres[5]. In other words, mesh terms which are very rare should have less of an influence on the model than those which appear more commonly throughout. Thus, we began training our models by tuning to increase micro F1 score.

Having identified the appropriate performance metric, we began again attempting to identify the appropriate model with which to train our classifier. Models attempted included support vector machines, K nearest neighbors, and random forest classifiers. Training on each of these classifiers led to varying results and it became evident that training time presented a major barrier, especially when training the random forest classifier. To compensate, we determined to train models on articles which appeared in the dataset from the year 2015 forward, and dropped any MeSH terms which appeared in less than 1% of the abstracts. This reduced the dimensions of our dataset from roughly 130,000 rows to 30,328 rows and from 54,471 MeSH terms to 107 MeSH terms. Reducing MeSH term inclusion to those which appeared in greater than 1% of the articles, greatly increased our overall micro F1 score because these outlier MeSH terms had less of an effect on model training. Training time was also greatly reduced, from days to hours, as there were fewer rows of data to classify making it possible to complete the analysis within the time constraints.

As a next step in determining the best model with which to create our classifier, we compared two models on a subset of the data. Due to imposed time constraints and extensive training time of the Random Forest Classifier, we decided to compare the model performance using Support Vector Machine and K Nearest Neighbors multi output classifiers. We ran comparisons between SVM and KNN utilizing both TFIDF and BERT representations of the abstract text feature on a subsample of the data which included 1327 articles and 63 MeSH terms. Our results on this subset of the data pointed to better performance overall from the KNN classifier utilizing BERT sentence embeddings pretrained model Allenai-Specter which was trained on scientific articles[6]. As can be seen in the figure below, the nearest neighbors model outperforms the SVC in F1-Score at the micro level which is our target indicator, as well as in other performance metrics such as precision.

		Precision	Recall	F1-Score	Support
KNN	micro avg	0.73	0.59	0.65	3087
	macro avg	0.33	0.16	0.19	3087
	weighted avg	0.61	0.59	0.57	3087
	samples avg	0.73	0.62	0.64	3087
SVC	micro avg	0.59	0.6	0.6	3087
	macro avg	0.29	0.29	0.29	3087
	weighted avg	0.6	0.6	0.6	3087
	samples avg	0.6	0.63	0.59	3087

Due to the success of the KNN to outperform the SVC on the data, we pursued further tuning of the KNN model to derive the best results. Looking further at the abstract representation, we were able to increase micro F1 score from 0.6167 utilizing TFIDF vectorizer to 0.6485 utilizing BERT sentence embeddings further confirming our choice to utilize BERT as our feature representation on the PubMed abstract text. Utilizing GridSearch, we were able to determine optimal parameters on a subsample of the data. Our results on this subsample can be seen in the table below.

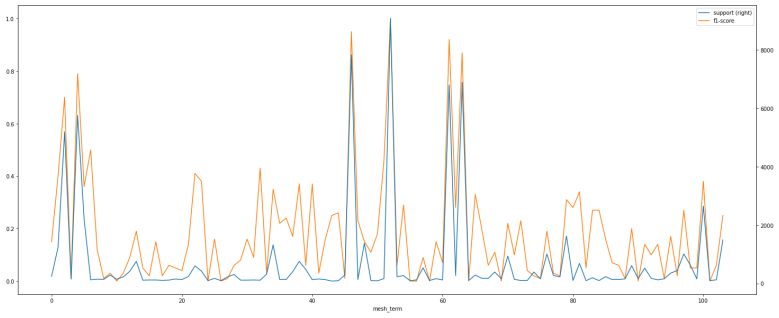
	Precision	Recall	F1-score	Support
micro avg	0.81	0.86	0.84	22683
macro avg	0.73	0.76	0.74	22683
weighted avg	0.81	0.86	0.83	22683
samples avg	0.82	0.88	0.82	22683

Upon training our model on the complete dataset for articles published from 2015 and beyond, and removing those MeSH terms which appeared in less than 1% of the articles, we trained our multi output KNN classifier utilizing BERT sentence transformers on 30,328 articles containing 107 MeSH terms were able to achieve a micro F1 score of 0.6558, a precision (average by samples) of 0.768, and a hamming loss of 0.048.

	Precision	Recall	F1-Score	Support
micro avg	0.75	0.58	0.66	75921
macro avg	0.35	0.15	0.19	75921
weighted avg	0.64	0.58	0.58	75921
samples avg	0.77	0.61	0.65	75921

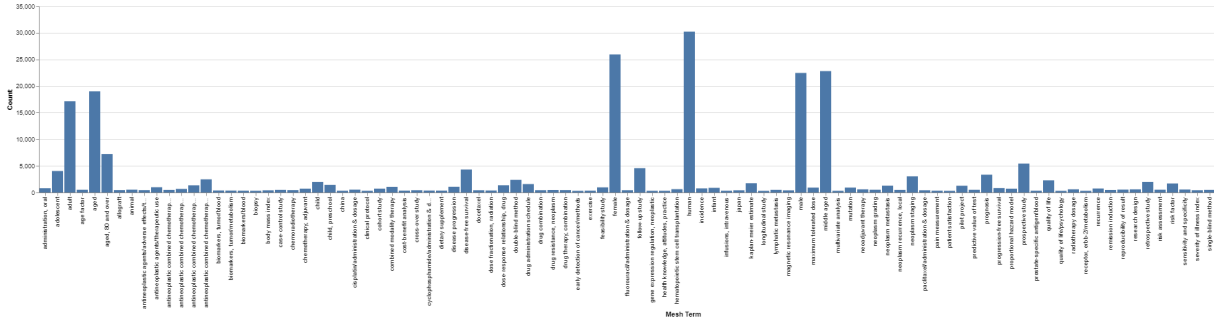
Performance Evaluation

Based on the results of our model fully trained on 30,328 articles containing 107 MeSH terms, we were able to achieve some decent results. Looking at where the model failed to predict multiple MeSH terms accurately, we can see that the model completely failed to predict MeSH terms at positions 10, 24, 26, 55, 56, 58, 64, 69, 90, 101 where precision, recall, f1-score were all 0.0. The MeSH terms at these positions were 'antineoplastic combined chemotherapy protocols/administration & dosage', 'china', 'clinical protocol', 'infusions, intravenous', 'japan', 'longitudinal study', 'multivariate analysis', 'neoplasm recurrence, local', 'risk assessment', and 'tumor burden'.



Initially the concern with these failures was that these were MeSH terms which were underrepresented in the dataset. Upon further inspection we determined that these MeSH terms had ample support compared to other mesh terms and were represented in the test set at a rate of 31.07% which aligns with our train/test split of 70/30 and is slightly above the overall average representation of all MeSH terms in the training set at 30.21%. Looking further at the model predictions we see as an example the MeSH term at position 10 occurred in the full dataset 485 times, in the test dataset 155 times, and was only predicted as a positive label for any abstract in the test dataset 9 times. As we know the label was evenly represented in the training data, at least proportionally, our only conclusion is that the label occurred less overall than other MeSH terms. Examining this hypothesis, we determined the average MeSH term use throughout the dataset was 2407.51 times, and the median MeSH term

usage was 551.5. Comparing this to the underperforming MeSH terms which occurred on average 379.4 times in the dataset and had a median representation of 325.5 occurrences.



As can be seen in the figure above, there are several outliers which are over represented in the dataset. Our second hypothesis is that the existence of outliers in the dataset contributed to poor performance in the under-represented MeSH terms. The over-represented MeSH terms in the dataset were 'adult', 'aged', 'female', 'human', 'male', and 'middle aged' and were predicted positive labels for abstracts 9.82% on average above their actual occurrence in the dataset. We believe that while applying uniform weights to each class yielded the best results in the GridSearchCV on the subsample of the dataset, adding custom weights to the classes may have been necessary to improve the performance of the model.

Discussion

Overall we are pleased with our results and believe we have provided some valuable insight:

- The unsupervised learning sometimes felt like alchemy, but the more we explored the corpus, understood the prevalence of some words, used the relevance metric in pyLDAvis and discussed with SMEs the more it came together. It was a great educational experience and perseverance seems to be a key quality for Datascientists.
- The positive results in topic modeling gave interesting insights, while not groundbreaking, indicate an interesting approach where non-specialists can help a very specialized field.
- SMEs found the views interesting and discussed with colleagues
- While BERT did not work well for usable topic models, it provided the best features for the supervised learning model. This is most likely due to the bidirectional nature of BERT and its ability to derive semantic meaning. Having used the Allenai-Specter BERT model pretrained on scientific articles to train our classifier, BERT outperformed TFIDF token count based vectorization in representing the key feature in our model training, i.e. article abstracts.
- While we were able to achieve decent results with our multi output classifier, we believe that outside of time and resource constraints, better results may be achieved through the use of deep learning. Our model does achieve an F1 score of 0.6581 with a precision (average by samples) score of 0.7655, it is unclear whether this outperforms human tagging. Further exploration and research should be done to determine the model's performance against human centered MeSH term tagging.

- While scanning all of these topics gives us an interesting overview of the research, the algorithm does not differentiate between positive and negative outcome clinical trials. So the trend on where the positive research is going can remain hidden. However since scientists tend to follow and learn from each other, it is more than likely that over time, the shifts identified are due to scientists following in the footsteps of transformative papers.

Conclusion

This has been a tremendously educational experience, being able to dive so deeply into specific topics helped us both learn more about the Data Science process "in real life". The movements of the topics over time are useful insights that could be used by the research management community to target areas that are lagging behind or identify where to invest.

The team is interested in doing further work on this project:

- Explore lda2vec (deep learning) for topic modeling.
- Hyperparameter tuning for NMF expanding on the proposed solution in derekgreene's Parameter Selection for NMF github page[7] the approach would be to use the word2vec embeddings for coherence measures and tune on alpha, beta loss, l1 ration, solver and init.
- The LDA model result is significantly weaker using bigrams and would be an interesting future topic of research as to why this happens with this dataset.
- Outlier detection to see where we have very niche research papers.
- The dataset could also be used to map the networks of authors and the dissemination of topics.
- Although here a more targeted corpus would most likely yield better results. We could potentially use a technique similar to Analyzing Evolving Stories in News Articles[8] to understand which research papers trigger waves of new work.
- Utilizing deep learning to more accurately predict MeSH term labels is another area of research which could be explored with this dataset which was impossible given the scope of the project. We believe an RNN LSTM ensemble would likely outperform our current KNN supervised learning approach.

Statement of work

The data gathering, redaction were done jointly, the rest of the work was divided to allow us to go more in depth. Michael focused on the supervised learning section, while Antoine focused on the unsupervised. However, we helped and supported each other during the debugging and difficult phases, ie. Michael's computer overheating and crashing.

Acknowledgments

A very big thanks to Antoine's colleagues who provided valuable insights during the project: Matej Horvat, the Director of Data Science for Novartis technical drug development; Marie-Eve Alary a pharmacologist and now clinical quality assurance manager and Sreekanth Gattu a clinical trial strategist.

And to both our families, especially our partners, for bearing with us during this milestone project that tested the limits of our sleep deprivation and work/school/life balance.

References

1. Teh, Y. W.; Jordan, M. I.; Beal, M. J.; Blei, D. M. (2006). "*Hierarchical Dirichlet Processes*" (PDF). Journal of the American Statistical Association. **101** (476): *pp.* 1566–1581. CiteSeerX 10.1.1.5.9094. doi:10.1198/016214506000000302.
2. Pietrangelo, A. (2021, March 30). *What Does It Mean If Breast Cancer Spreads to Your Lymph Nodes?* Healthline. https://www.healthline.com/health/breast-cancer/breast-cancer-lymph-nodes
3. Rosenbaum, Sara. '*The Patient Protection and Affordable Care Act: Implications for Public Health Policy and Practice*'. Public Health Reports (Washington, D.C.: 1974), vol. 126, no. 1, Feb. 2011, pp. 130–35. PubMed, https://doi.org/10.1177/003335491112600118
4. Home - MeSH - NCBI. (n.d.). NCBI. Retrieved September 26, 2021, from https://www.ncbi.nlm.nih.gov/mesh/
5. *Micro F1-score. (n.d.)*. Peltarion. Retrieved September 26, 2021, from https://peltarion.com/knowledge-center/documentation/evaluation-view/classification-loss-metrics/micro-f1-score
6. Arman Cohan and Sergey Feldman and Iz Beltagy and Doug Downey and Daniel S. Weld: *SPECTER: Document-level Representation Learning using Citation-informed Transformers*. ACL, 2020. https://huggingface.co/sentence-transformers/allenai-specter)
7. Derekgreene, D. (2021, July 6). *3 - Parameter Selection for NMF*. GitHub. https://github.com/derekgreene/topic-model-tutorial/blob/master/3%20-%20Parameter%20Selection%20for%20NMF.ipynb
8. Barranco, Roberto Camacho, et al. '*Analyzing Evolving Stories in News Articles*'. International Journal of Data Science and Analytics, vol. 8, no. 3, Oct. 2019, pp. 241–56. arXiv.org, https://doi.org/10.1007/s41060-017-0091-9

Code References

Sources of code and inspiration used in the project:

- Kapadia, S. (2019, April 29). *Topic Modeling in Python: Latent Dirichlet Allocation (LDA)*. Medium. https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0
- Li, S. (2018, May 31). *Topic Modeling and Latent Dirichlet Allocation (LDA) in Python*. Medium. https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24
- Kapadia, S. (2019, August 19). *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*. Medium. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
- Norton, S. (2020, June 5). *Gensim: extract 100 most representative documents for each topic.* Stack Overflow. https://stackoverflow.com/questions/62174945/gensim-extract-100-most-representative-documents-for-each-topic/62222475#62222475
- Grootendorst, M. P. (n.d.). *FAQ - BERTopic. BERTopic FAQ*. Retrieved September 26, 2021, from https://maartengr.github.io/BERTopic/faq.html#i-have-only-a-few-topics-how-do-i-increase-them
- Rehurek, R. (n.d.). *Gensim: topic modelling for humans. Gensim Documentation*. Retrieved September 26, 2021, from https://radimrehurek.com/gensim/auto_examples/core/run_topics_and_transformations.html#sphx-glr-auto-examples-core-run-topics-and-transformations-py
- Grootendorst, M. (2020, October 5). *Topic Modeling with BERT*. | Towards Data Science. Medium. https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6
- Gu, Yu, et al. '*Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing*'. ArXiv:2007.15779 [Cs], Sept. 2021. arXiv.org, https://doi.org/10.1145/3458754
- Ravishchawla. (n.d.). *Topic Modeling with LDA and NMF algorithms.* Github. Retrieved September 26, 2021, from https://gist.github.com/ravishchawla/3f346318b85fa07196b761443b123bba
- Deutsch, D. (2020, April 16). *Dynamic subplot layout in Seaborn* - Towards Data Science. Medium. https://towardsdatascience.com/dynamic-subplot-layout-in-seaborn-e777500c7386
- Zaheer, Manzil, et al. '*Big Bird: Transformers for Longer Sequences*'. ArXiv:2007.14062 [Cs, Stat], Jan. 2021. arXiv.org, http://arxiv.org/abs/2007.14062