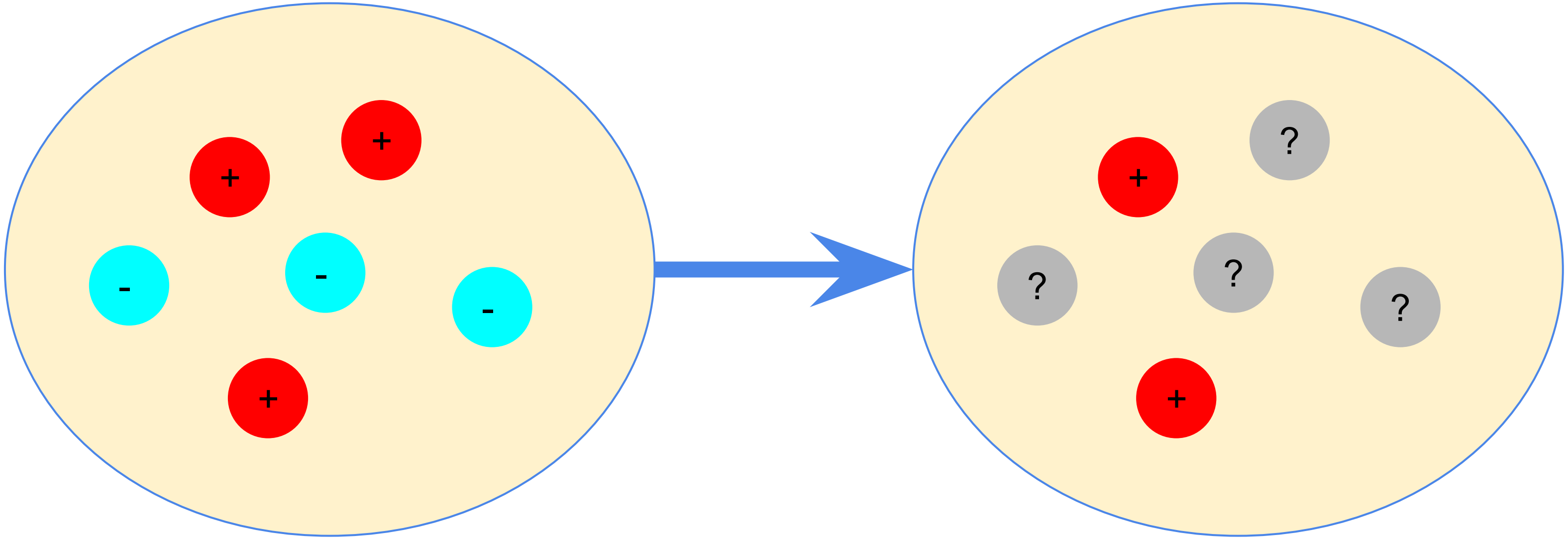
# Scalable Evaluation and Improvement of Document Set Expansion via Neural Positive-Unlabeled Learning

Alon Jacovi, Gang Niu, Yoav Goldberg, Masashi Sugiyama

## Positive Unlabeled (PU) Learning:

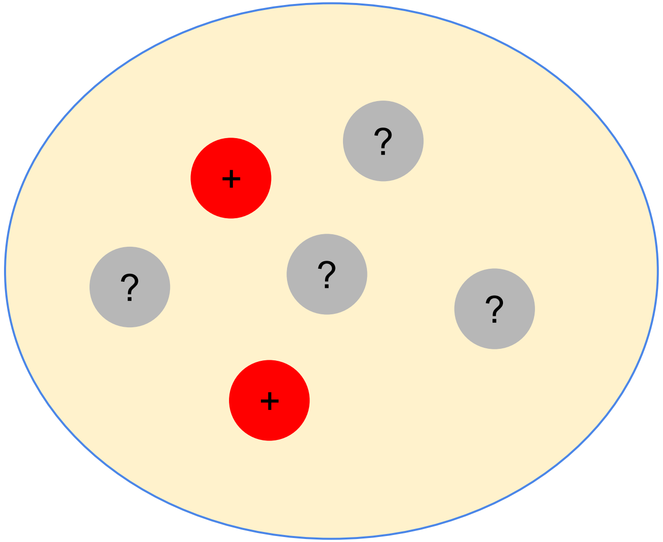
Learning a binary classifier only from:

- examples of one class
- unlabeled examples



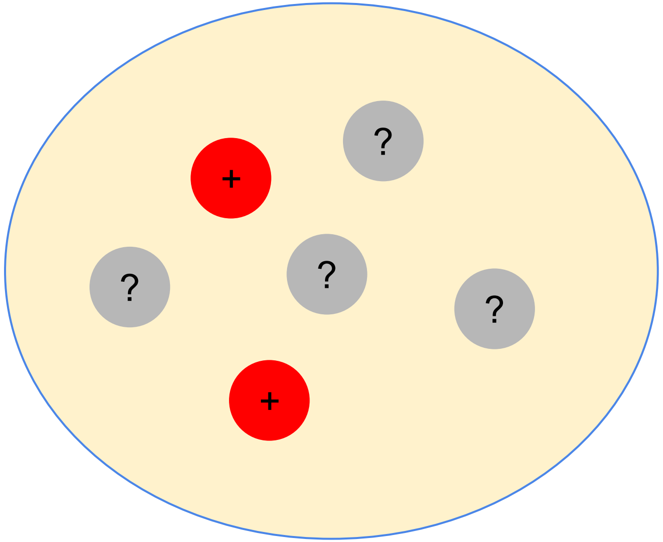
The PU setting **occurs naturally** in many ML and NLP tasks.

For example: Document Set Expansion, Relation Extraction...



## Document Set Expansion (DSE):

Given *n* documents, retrieve more documents similar to them.



# What is our paper about?

# What is our paper about?

Problem 1:

**Evaluation and benchmarking**

# What is our paper about?

Problem 1:

**Evaluation and benchmarking**

Current DSE and PU benchmarks are relatively *small* and *simple*.

(repurposed 20News and CIFAR10, typically)

# What is our paper about?

Problem 1:

**Evaluation and benchmarking**

Problem 2:

**Training**

Current DSE and PU benchmarks are relatively *small* and *simple*.

(repurposed 20News and CIFAR10, typically)

## What is our paper about?

Problem 1:

**Evaluation and benchmarking**

Current DSE and PU benchmarks are relatively *small* and *simple*.

(repurposed 20News and CIFAR10, typically)

Problem 2:

**Training**

Current DSE and PU solutions *do not scale* with larger and harder settings.

# What is our paper about?

Problem 1:

**Evaluation and benchmarking**

We design a new DSE/PU benchmark by using PubMed labels.

Problem 2:

**Training**

We test existing algorithms on the benchmark (they fail), diagnose the issues, and propose solutions.

# Problem 1:
# Evaluation and Benchmarking

## What do we need from a scalable DSE benchmark?

- Small prior
  - small quantity of *positive* documents (vs negative)
- Unbalanced labels
  - small quantity of *labeled positive* documents (vs unlabeled)
- A lot of topics
  - Able to generate many DSE tasks for various cohesive topics
- Large scale data
  - Ideally, millions of documents
- Fully labeled!
  - Positive/negative ground truth for ***all*** documents.

Problem 1:
**Evaluation and benchmarking**

We are going to repurpose **PubMed** to create a DSE benchmark for PU learning.

We are going to repurpose **PubMed** to create a DSE benchmark for PU learning.

PubMed is a collection of millions of biomedical articles.

Problem 1:
**Evaluation and benchmarking**

We are going to repurpose **PubMed** to create a DSE benchmark for PU learning.

PubMed is a collection of millions of biomedical articles.

Each PubMed instance has (among other things):

- Paper text / abstract
- MeSH (Medical Subject Headings) tags

Problem 1:
**Evaluation and benchmarking**

We are going to repurpose **PubMed** to create a DSE benchmark for PU learning.

PubMed is a collection of millions of biomedical articles.

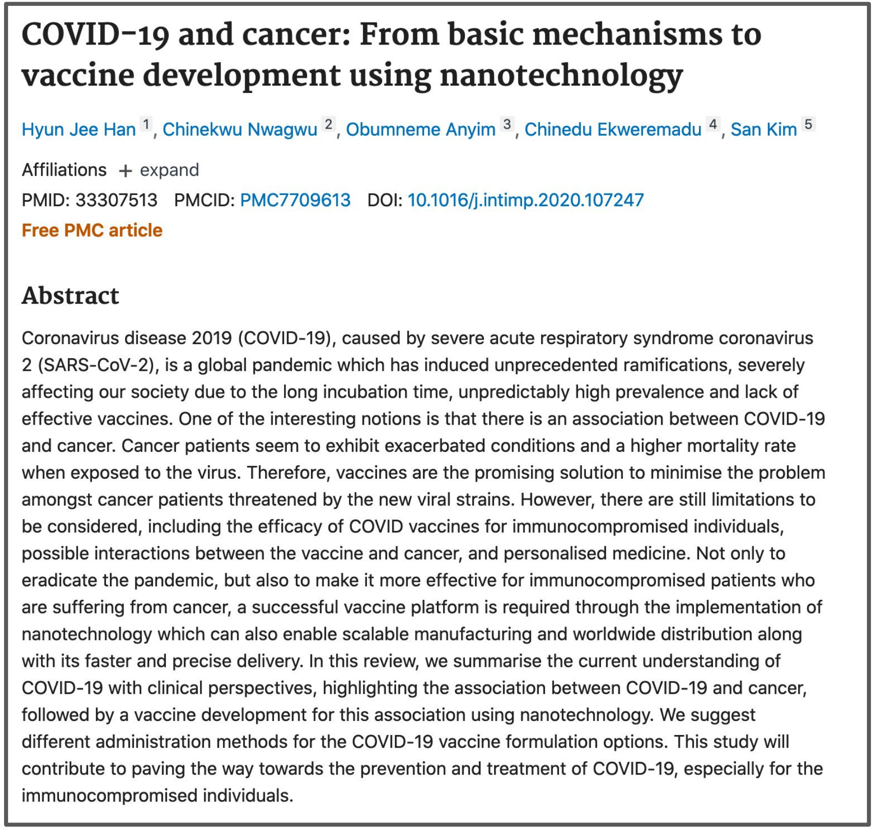
Each PubMed instance has (among other things):

- Paper text / abstract
- MeSH (Medical Subject Headings) tags

Let’s use this!

Problem 1:
**Evaluation and benchmarking**

# For example...



## COVID-19 and cancer: From basic mechanisms to vaccine development using nanotechnology

Hyun Jee Han [1], Chinekwu Nwagwu [2], Obumneme Anyim [3], Chinedu Ekweremadu [4], San Kim [5]

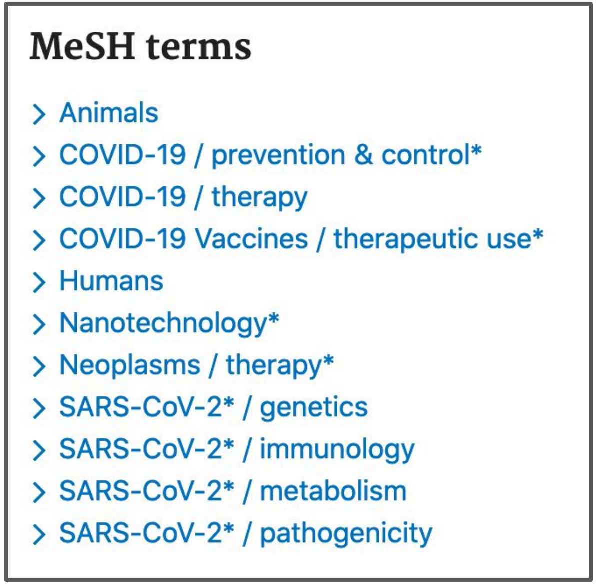
Affiliations + expand

PMID: 33307513 PMCID: PMC7709613 DOI: 10.1016/j.intimp.2020.107247

Free PMC article

### Abstract

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a global pandemic which has induced unprecedented ramifications, severely affecting our society due to the long incubation time, unpredictably high prevalence and lack of effective vaccines. One of the interesting notions is that there is an association between COVID-19 and cancer. Cancer patients seem to exhibit exacerbated conditions and a higher mortality rate when exposed to the virus. Therefore, vaccines are the promising solution to minimise the problem amongst cancer patients threatened by the new viral strains. However, there are still limitations to be considered, including the efficacy of COVID vaccines for immunocompromised individuals, possible interactions between the vaccine and cancer, and personalised medicine. Not only to eradicate the pandemic, but also to make it more effective for immunocompromised patients who are suffering from cancer, a successful vaccine platform is required through the implementation of nanotechnology which can also enable scalable manufacturing and worldwide distribution along with its faster and precise delivery. In this review, we summarise the current understanding of COVID-19 with clinical perspectives, highlighting the association between COVID-19 and cancer, followed by a vaccine development for this association using nanotechnology. We suggest different administration methods for the COVID-19 vaccine formulation options. This study will contribute to paving the way towards the prevention and treatment of COVID-19, especially for the immunocompromised individuals.



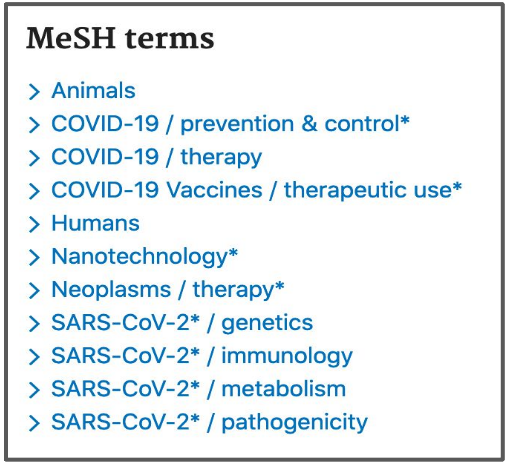
## MeSH terms

- Animals
- COVID-19 / prevention & control*
- COVID-19 / therapy
- COVID-19 Vaccines / therapeutic use*
- Humans
- Nanotechnology*
- Neoplasms / therapy*
- SARS-CoV-2* / genetics
- SARS-CoV-2* / immunology
- SARS-CoV-2* / metabolism
- SARS-CoV-2* / pathogenicity

Problem 1:
**Evaluation and benchmarking**

Given a **set of MeSH terms** (= our topic),

1. Randomly choose *n* PubMed articles with these these terms
2. Retrieve more documents with these MeSH terms from PubMed



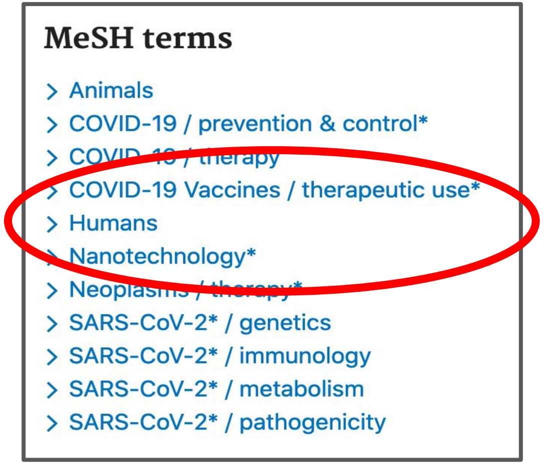
**MeSH terms**

- Animals
- COVID-19 / prevention & control*
- COVID-19 / therapy
- COVID-19 Vaccines / therapeutic use*
- Humans
- Nanotechnology*
- Neoplasms / therapy*
- SARS-CoV-2* / genetics
- SARS-CoV-2* / immunology
- SARS-CoV-2* / metabolism
- SARS-CoV-2* / pathogenicity

Problem 1:
**Evaluation and benchmarking**

Given a **set of MeSH terms** (= our topic),

1. Randomly choose *n* PubMed articles with these these terms
2. Retrieve more documents with these MeSH terms from PubMed



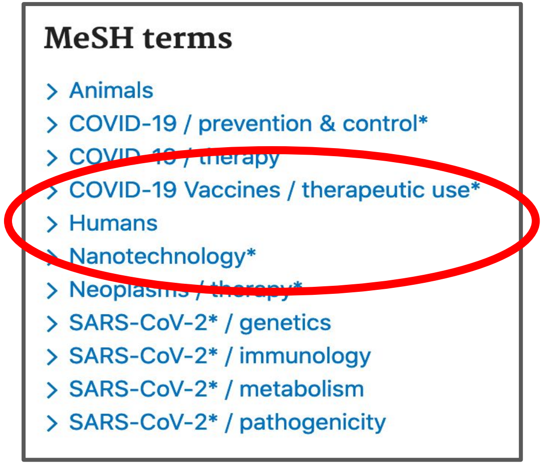
**MeSH terms**

- Animals
- COVID-19 / prevention & control*
- COVID-19 / therapy
- COVID-19 Vaccines / therapeutic use*
- Humans
- Nanotechnology*
- Neoplasms / therapy*
- SARS-CoV-2* / genetics
- SARS-CoV-2* / immunology
- SARS-CoV-2* / metabolism
- SARS-CoV-2* / pathogenicity

Problem 1:
**Evaluation and benchmarking**

Given a **set of MeSH terms** (= our topic),

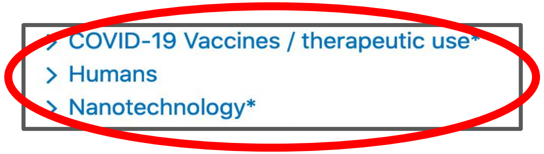
1. Randomly choose *n* PubMed articles with these these terms
2. Retrieve more documents with these MeSH terms from PubMed



**MeSH terms**

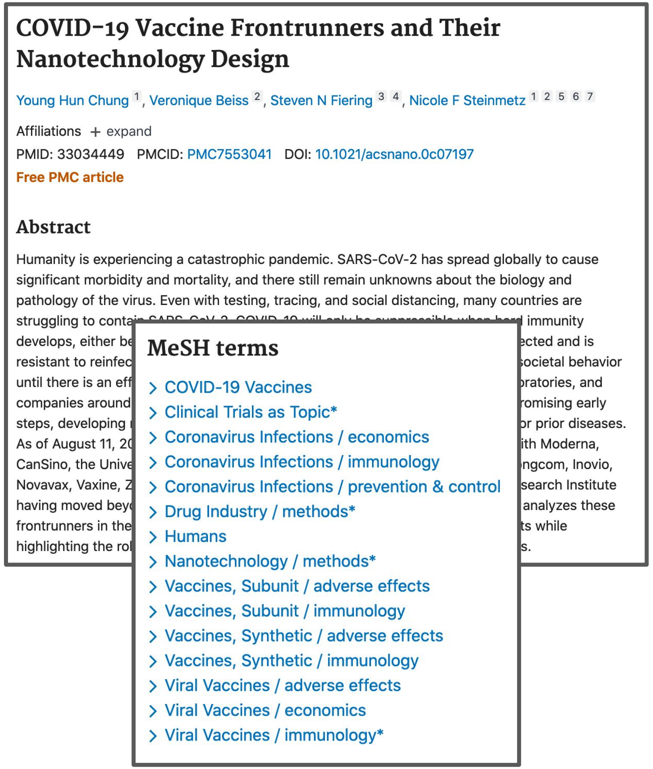
- Animals
- COVID-19 / prevention & control*
- COVID-19 / therapy
- COVID-19 Vaccines / therapeutic use*
- Humans
- Nanotechnology*
- Neoplasms / therapy*
- SARS-CoV-2* / genetics
- SARS-CoV-2* / immunology
- SARS-CoV-2* / metabolism
- SARS-CoV-2* / pathogenicity

note: in the paper, we retrieve the unlabeled documents using Okapi BM25 from general PubMed, and then use PU. Check the paper for details/motivation.



> COVID-19 Vaccines / therapeutic use*
> Humans
> Nanotechnology*

Problem 1:
**Evaluation and benchmarking**



## COVID-19 Vaccine Frontrunners and Their Nanotechnology Design

Young Hun Chung [1], Veronique Beiss [2], Steven N Fiering [3] [4], Nicole F Steinmetz [1] [2] [5] [6] [7]

Affiliations + expand

PMID: 33034449 PMCID: PMC7553041 DOI: 10.1021/acsnano.0c07197

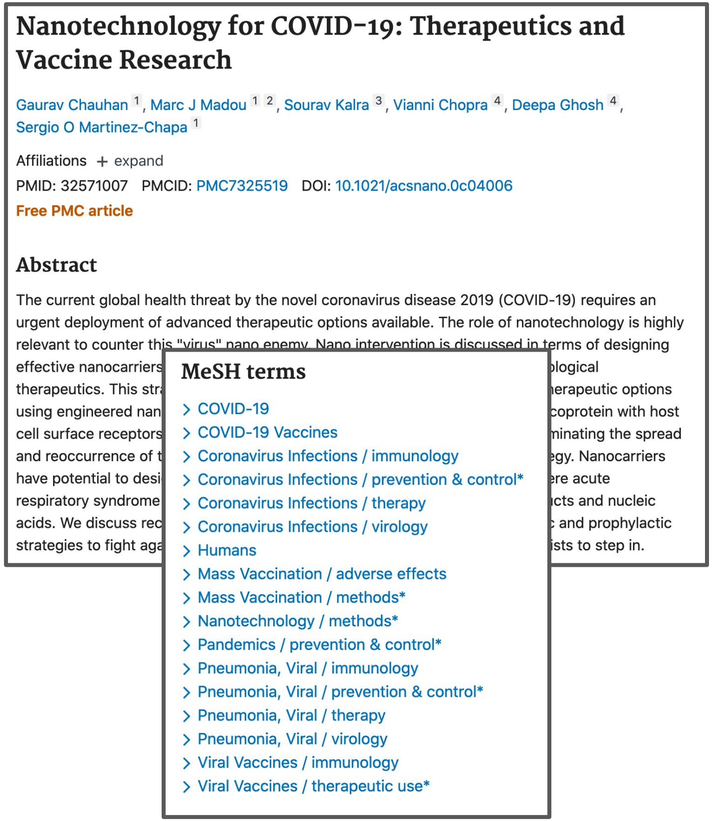
Free PMC article

### Abstract

Humanity is experiencing a catastrophic pandemic. SARS-CoV-2 has spread globally to cause significant morbidity and mortality, and there still remain unknowns about the biology and pathology of the virus. Even with testing, tracing, and social distancing, many countries are struggling to contain SARS-CoV-2. COVID-19 will only be suppressible when herd immunity
develops, either be ... ected and is
resistant to reinfec ... societal behavior
until there is an eff ... oratories, and
companies around ... romising early
steps, developing ... or prior diseases.
As of August 11, 20 ... ith Moderna,
CanSino, the Unive ... ngcom, Inovio,
Novavax, Vaxine, Z ... search Institute
having moved bey ... analyzes these
frontrunners in the ... ts while
highlighting the rol ... s.

### MeSH terms

- COVID-19 Vaccines
- Clinical Trials as Topic*
- Coronavirus Infections / economics
- Coronavirus Infections / immunology
- Coronavirus Infections / prevention & control
- Drug Industry / methods*
- Humans
- Nanotechnology / methods*
- Vaccines, Subunit / adverse effects
- Vaccines, Subunit / immunology
- Vaccines, Synthetic / adverse effects
- Vaccines, Synthetic / immunology
- Viral Vaccines / adverse effects
- Viral Vaccines / economics
- Viral Vaccines / immunology*



## Nanotechnology for COVID-19: Therapeutics and Vaccine Research

Gaurav Chauhan [1], Marc J Madou [1] [2], Sourav Kalra [3], Vianni Chopra [4], Deepa Ghosh [4], Sergio O Martinez-Chapa [1]

Affiliations + expand

PMID: 32571007 PMCID: PMC7325519 DOI: 10.1021/acsnano.0c04006

Free PMC article

### Abstract

The current global health threat by the novel coronavirus disease 2019 (COVID-19) requires an urgent deployment of advanced therapeutic options available. The role of nanotechnology is highly relevant to counter this "virus" nano enemy. Nano intervention is discussed in terms of designing
effective nanocarriers ... ological
therapeutics. This stra ... herapeutic options
using engineered nan ... coprotein with host
cell surface receptors ... minating the spread
and reoccurrence of t ... egy. Nanocarriers
have potential to desi ... ere acute
respiratory syndrome ... ucts and nucleic
acids. We discuss rec ... and prophylactic
strategies to fight aga ... ists to step in.

### MeSH terms

- COVID-19
- COVID-19 Vaccines
- Coronavirus Infections / immunology
- Coronavirus Infections / prevention & control*
- Coronavirus Infections / therapy
- Coronavirus Infections / virology
- Humans
- Mass Vaccination / adverse effects
- Mass Vaccination / methods*
- Nanotechnology / methods*
- Pandemics / prevention & control*
- Pneumonia, Viral / immunology
- Pneumonia, Viral / prevention & control*
- Pneumonia, Viral / therapy
- Pneumonia, Viral / virology
- Viral Vaccines / immunology
- Viral Vaccines / therapeutic use*

Problem 1:
**Evaluation and benchmarking**

For every set of MeSH terms, we can easily generate a new DSE task.

These tasks are large, realistic, **fully labeled**, and difficult with small priors.

These tasks can serve as a benchmark for DSE (or any PU!) algorithms.

# Problem 2:
# Training

Problem 2:
**Training**

The current state-of-the-art for training PU classifiers is the **nnPU loss** (details in paper).

Problem 2:
**Training**

The current state-of-the-art for training PU classifiers is the **nnPU loss** (details in paper).

I’ll go over two key limitations of nnPU which prevent it from scaling well to our new PubMed DSE task:

Problem 2:
**Training**

The current state-of-the-art for training PU classifiers is the **nnPU loss** (details in paper).

I’ll go over two key limitations of nnPU which prevent it from scaling well to our new PubMed DSE task:

1. It assumes a known true prior.

Problem 2:
**Training**

The current state-of-the-art for training PU classifiers is the **nnPU loss** (details in paper).

I’ll go over two key limitations of nnPU which prevent it from scaling well to our new PubMed DSE task:

1. It assumes a known true prior.
2. It assumes balanced supervision, and large batch size.

Problem 2:
**Training**

## True prior is unknown

In the DSE task (and often in any PU tasks), we don’t know the true prior of the positive and negative classes.

## True prior is unknown

In the DSE task (and often in any PU tasks), we don’t know the true prior of the positive and negative classes.

nnPU relies on this assumption to remain unbiased.

Problem 2:
**Training**

## True prior is unknown

In the DSE task (and often in any PU tasks), we don’t know the true prior of the positive and negative classes.

nnPU relies on this assumption to remain unbiased.

Unfortunately, this prior is very difficult to estimate.

Problem 2:
**Training**

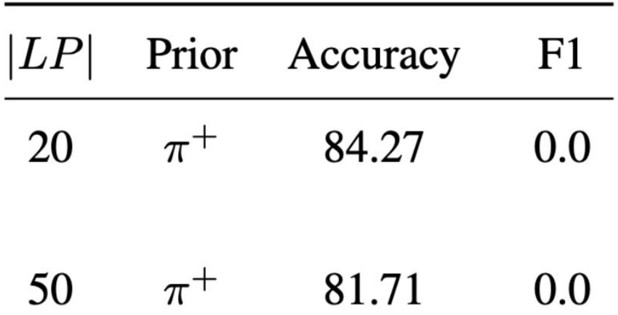
## True prior is unknown

Furthermore - even if we knew the true prior, the PU loss optimizes for *accuracy*, often at the cost of F1 (the favorable metric for DSE).

Problem 2:
**Training**

## True prior is unknown

Furthermore - even if we knew the true prior, the PU loss optimizes for *accuracy*, often at the cost of F1 (the favorable metric for DSE).



| $\|LP\|$ | Prior | Accuracy | F1 |
|---|---|---|---|
| 20 | $\pi^+$ | 84.27 | 0.0 |
| 50 | $\pi^+$ | 81.71 | 0.0 |

Problem 2:
**Training**

## True prior is unknown

## Solution: Balanced Error (BER) optimization

Problem 2:
**Training**

## True prior is unknown

## Solution: Balanced Error (BER) optimization

Instead of optimizing for accuracy, we can optimize for BER by assuming that the prior is 0.5 (derivation in paper).

$$BER(g) = \frac{1}{2}(\frac{FP}{TN + FP} + \frac{FN}{FN + TP})$$

Problem 2:
**Training**

**True prior is unknown**

**Solution: Balanced Error (BER) optimization**

Instead of optimizing for accuracy, we can optimize for BER by assuming that the prior is 0.5 (derivation in paper).

BER is correlated with AUC - accomplishing both goals with one trick.

$$AUC = \frac{3}{2} - 2BER$$

Problem 2:
**Training**

## True prior is unknown

## Solution: Balanced Error (BER) optimization

Instead of optimizing for accuracy, we can optimize for BER by assuming that the prior is 0.5 (derivation in paper).

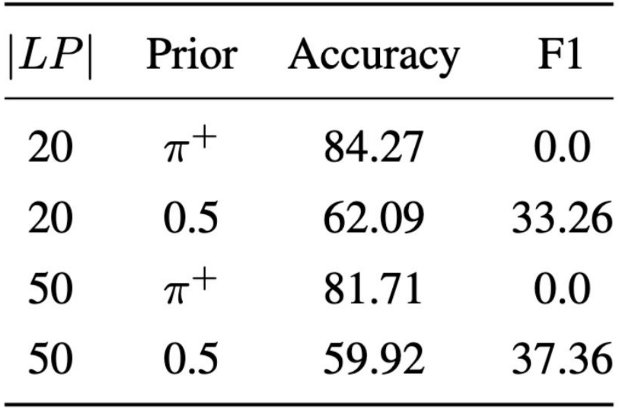
BER is correlated with AUC - accomplishing both goals with one trick.

$$\begin{aligned} R_{PU}(g) = \\ &\frac{1}{2}\mathbb{E}_{x\sim p^+(x)}[\ell(g(x),+1) - \ell(g(x),-1)] \\ &\qquad + \mathbb{E}_{x\sim p(x)}[\ell(g(x),-1)]. \quad (3) \end{aligned}$$

Problem 2:
**Training**

## True prior is unknown

## Solution: Balanced Error (BER) optimization



| $\|LP\|$ | Prior | Accuracy | F1 |
|---|---|---|---|
| 20 | $\pi^+$ | 84.27 | 0.0 |
| 20 | 0.5 | 62.09 | 33.26 |
| 50 | $\pi^+$ | 81.71 | 0.0 |
| 50 | 0.5 | 59.92 | 37.36 |

Problem 2:
**Training**

## Extreme imbalance and small batch size

Realistic DSE data has a very small quantity of positive docs, and very large quantity of unlabeled docs. (can be 1:10,000 or more)

## Extreme imbalance and small batch size

Realistic DSE data has a very small quantity of positive docs, and very large quantity of unlabeled docs. (can be 1:10,000 or more)

Literature in imbalanced classification typically goes up to 1:50 imbalance. Our setting is what we call “extreme imbalance”.

Problem 2:
**Training**

## Extreme imbalance and small batch size

Realistic DSE data has a very small quantity of positive docs, and very large quantity of unlabeled docs. (can be 1:10,000 or more)

Literature in imbalanced classification typically goes up to 1:50 imbalance. Our setting is what we call “extreme imbalance”.

This is problematic in stochastic gradient descent.

## Extreme imbalance and small batch size

This is especially limiting for modern NLP models, which are so large that it’s computationally prohibitive to use a large batch size.

Problem 2:
**Training**

## Extreme imbalance and small batch size

This is especially limiting for modern NLP models, which are so large that it’s computationally prohibitive to use a large batch size.

→ Most batches during learning are homogeneous (no positive docs).

Problem 2:
**Training**

## Extreme imbalance and small batch size

This is especially limiting for modern NLP models, which are so large that it’s computationally prohibitive to use a large batch size.

→ Most batches during learning are homogeneous (no positive docs).

→ The model gets a very sparse training signal, complicating training.

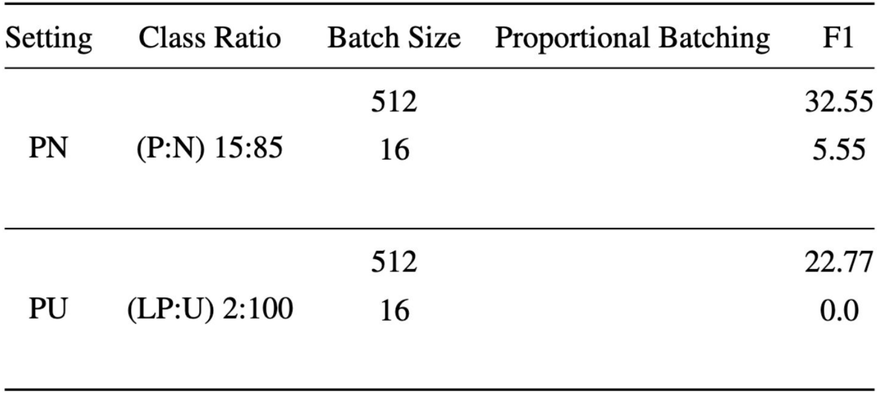
Problem 2:
**Training**

## Extreme imbalance and small batch size

This is especially limiting for modern NLP models, which are so large that it’s computationally prohibitive to use a large batch size.

→ Most batches during learning are homogeneous (no positive docs).

→ The model gets a very sparse training signal, complicating training.



| Setting | Class Ratio | Batch Size | Proportional Batching | F1 |
|---|---|---|---|---|
| PN | (P:N) 15:85 | 512 | | 32.55 |
| | | 16 | | 5.55 |
| PU | (LP:U) 2:100 | 512 | | 22.77 |
| | | 16 | | 0.0 |

Problem 2:
**Training**

## Extreme imbalance and small batch size

## Solution: Proportional Batching

## Extreme imbalance and small batch size

## Solution: Proportional Batching

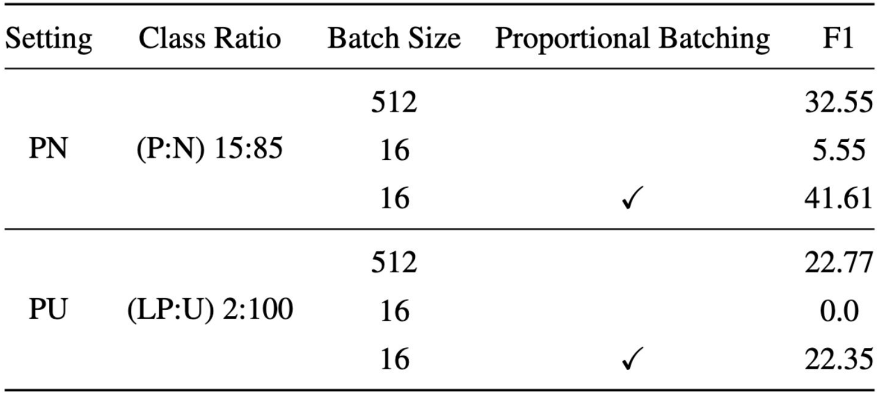
We propose a simple trick: enforce heterogeneous batches by forcing each batch to contain at least one positive example.

Problem 2:
**Training**

## Extreme imbalance and small batch size

## Solution: Proportional Batching

We propose a simple trick: enforce heterogeneous batches by forcing each batch to contain at least one positive example.

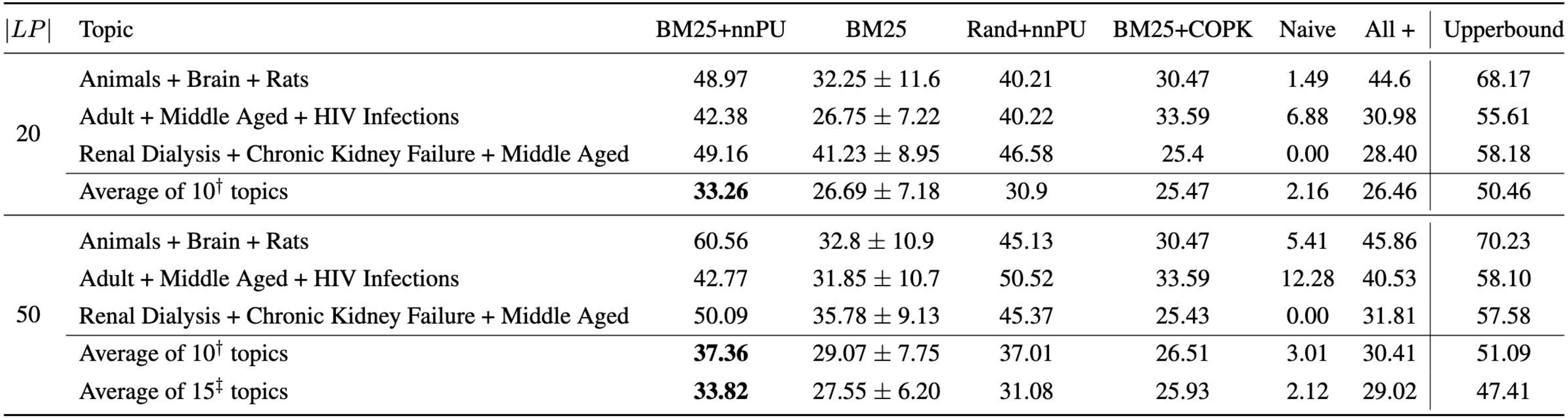


| Setting | Class Ratio | Batch Size | Proportional Batching | F1 |
|---|---|---|---|---|
| PN | (P:N) 15:85 | 512 | | 32.55 |
| | | 16 | | 5.55 |
| | | 16 | ✓ | 41.61 |
| PU | (LP:U) 2:100 | 512 | | 22.77 |
| | | 16 | | 0.0 |
| | | 16 | ✓ | 22.35 |

Problem 2:
**Training**

To recap, we talked about two modifications to allow nnPU to overcome its limitations:

**BER optimization** and **proportional batching**



| $\lvert LP \rvert$ | Topic | BM25+nnPU | BM25 | Rand+nnPU | BM25+COPK | Naive | All + | Upperbound |
|---|---|---|---|---|---|---|---|---|
| 20 | Animals + Brain + Rats | 48.97 | 32.25 ± 11.6 | 40.21 | 30.47 | 1.49 | 44.6 | 68.17 |
| | Adult + Middle Aged + HIV Infections | 42.38 | 26.75 ± 7.22 | 40.22 | 33.59 | 6.88 | 30.98 | 55.61 |
| | Renal Dialysis + Chronic Kidney Failure + Middle Aged | 49.16 | 41.23 ± 8.95 | 46.58 | 25.4 | 0.00 | 28.40 | 58.18 |
| | Average of 10[†] topics | **33.26** | 26.69 ± 7.18 | 30.9 | 25.47 | 2.16 | 26.46 | 50.46 |
| 50 | Animals + Brain + Rats | 60.56 | 32.8 ± 10.9 | 45.13 | 30.47 | 5.41 | 45.86 | 70.23 |
| | Adult + Middle Aged + HIV Infections | 42.77 | 31.85 ± 10.7 | 50.52 | 33.59 | 12.28 | 40.53 | 58.10 |
| | Renal Dialysis + Chronic Kidney Failure + Middle Aged | 50.09 | 35.78 ± 9.13 | 45.37 | 25.43 | 0.00 | 31.81 | 57.58 |
| | Average of 10[†] topics | **37.36** | 29.07 ± 7.75 | 37.01 | 26.51 | 3.01 | 30.41 | 51.09 |
| | Average of 15[‡] topics | **33.82** | 27.55 ± 6.20 | 31.08 | 25.93 | 2.12 | 29.02 | 47.41 |

# Conclusion

Problem 1:

**Evaluation and benchmarking**

Problem 2:

**Training**

We design a new DSE/PU benchmark by using PubMed labels

We test existing algorithms on the benchmark (they fail), diagnose the issues, and propose solutions.

PubMed DSE

BER optimization
Proportional Batching

## Conclusion

These solutions are more generally useful than the specific context we presented them in:

- PubMed DSE is useful for benchmarking any PU system
- BER optimization is useful for PU losses generally
- Proportional Batching is useful for general imbalanced classification

**Thanks for watching!**

PubMed DSE

BER optimization
Proportional Batching