# BioMedical NLP - project topics

1. Survey of research on a specific sub-topic in bio-medical NLP + some demo implementation
2. Health related fake news: detect topics in existing datasets / collect new dataset
3. Semantic change in the medical domain: compare embeddings
4. Collect and annotate (semi-automatic) mental health dataset for Romanian / other low-resource language
5. Implement NER on existing dataset (see shared tasks below)
6. Categorize medical transcripts (example targets: medical specialties); data: <https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions>
7. Compare performance of medical and non-medical embeddings on a given task
8. Topics in health related scientific papers over time
9. Detecting symptoms in social media data
10. Reproduce methods described in a paper discussed during the seminar
11. Collect and annotate dataset from social media on different disorder than depression: anxiety, …
12. Participate in a shared task (current): eRisk, CLPsych, [n2c2 Challenge: Contextualized Medication Event Extraction](https://portal.dbmi.hms.harvard.edu/projects/n2c2-2022-t1/),
13. Solve a past shared task:
    1. <https://n2c2.dbmi.hms.harvard.edu/challenges> with associated datasets here: <https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/>
    2. <https://temu.bsc.es/meddoprof/> (Spanish)
    3. <https://ehealthkd.github.io/2021> (Spanish + English)
    4. CLEF eHealth (<https://sites.google.com/site/clefehealth/tasks>): <https://sites.google.com/site/shareclefehealth/> , <https://sites.google.com/site/clefehealth2015/> , <https://sites.google.com/site/clefehealth2016/>
    5. SemEval (Temporal): <https://alt.qcri.org/semeval2016/task12/> , <https://alt.qcri.org/semeval2017/task12/> ,
14. Other potential clinical datasets to analyze using NLP and ML: <https://physionet.org/about/database/> , <https://clinical-nlp.github.io/2020/resources.html> , <https://sites.google.com/site/clefehealth/datasets>

Synthetic data: <https://synthea.mitre.org/fhir-api>

1. Compare natural and synthetic data, identify discriminating patterns (to explore)

### Requirements

Deliverables include:

* implementation of a solution to a problem in bio-medical NLP
* document describing the methods used (2-4 pages)
* presentation

The document/paper will follow the classical structure of a research article:

* short summary (abstract)
* analysis of main idea
* related work: state of the art (SOTA) where it exists, short history, recent and/or related results
* **methods**
* conclusions and future work, directions for further improvement
* references

In case you’re presenting a survey: explain main methodologies and selection process (i.e. you are surveying either chronologically, or in order of SOTA achievements), discuss advantages and disadvantages to the methods used.

In the case of presenting specific applications: describe the method, compare it with other results in the field

Teams are 1, 2, or max 3 students. Add your choice next to your name on the google sheets document with your name.

## Resources

for researching your topic and writing the documentation

* [Google Scholar](https://scholar.google.com/): [search for author](https://scholar.google.gr/citations?hl=en&view_op=search_authors), paper name, topic ...; look at year of publication (more recent - better), number of citations (more - better), publication venue (preferably a top conference or journal)
* [ACL Anthology](https://aclanthology.org/)
* LaTeX / Word doc; tools: [Overleaf](https://www.overleaf.com/project) / [Google Docs](https://docs.google.com/) - for collaborative editing of papers and presentations