COMS4054A / COMS7066A

Natural Language Processing

Project

Semester 2, 2021

Description

Your assignment for this course is to perform NLP research (accompanied by a research report). The NLP research will require you to train one or more NLP models, to which you should write and submit model cards for each of your models.

You've been provided a number of datasets with potential accompanying research questions or project ideas. You can select your own dataset and research question if you'd like, but be conscious of training times. You're allowed to make use of any pre-trained models which you can then fine-tune for your problem set.

You may do this assignment in groups. Groups of up to 4 people are allowed.

You will also need to submit a video that explains your system or research, including details about methodology. The video should be similar to research videos that are often produced alongside publications. The video should clearly explain and demonstrate your approach. The video should be uploaded onto YouTube and the link to it shared at submission

You've been provided a list of datasets below that you can select from, along with suggested research questions. The datasets are organised into 3 themes:

- Local language datasets
- COVID-19
- Cyber Safety

Your report should include

- Problem Definition / Introduction
- Background [a short literature review as seen in papers]
- Methodology
- Results
- Analysis
- An Impact Statement
- Conclusion
- Bibliography

Submission

You will need to submit:

- Your code
- Your trained models (if feasible) and their respective model cards
- Your report
- A link to your YouTube video

Please zip all files and submit via moodle.
Deadline: 23rd November 23h00

Evaluation

You will be evaluated on

- The quality of your methodology and analysis
- Originality of your research question or approach
- Presentation quality of your research video
- Your execution of the scientific method.
- The thoughtfulness of the impact statement
- The quality of your model cards
- Proper Referencing

Format

Research Report

- The papers must be no less than 4 pages, and no more than 6 pages, plus space for unlimited references.
- Please use the main ACL-IJCNLP 2021 paper style files: Overleaf template or direct download: Latex and Word

Bonus marks will be given to projects who open source their code, with respective licenses, and/or submit their model(s) onto the HuggingFace model hub or to Zenodo.

Datasets

We list a number of datasets and accompanying ideas for tasks can be used in this project. If you think up a more interesting research question or come across an interesting dataset, you may choose that instead. Please be cognisant of training times - where possible, use transfer learning and fine tuning of pre-trained models to limit time-consuming training.

Local Language NLP Datasets

- **Umsuka English - isiZulu Parallel Corpus**
  Ideas:
  - How do tokenization strategies affect machine translation performance?
  - What machine translation model works best for English - to - isiZulu?

- **South African Disinformation [Fake News] Website Data - 2020**
  Ideas:
    - What features get used by different classification models to determine if a model is fake news or not? (Note: you will need to scrape some actual news. This dataset only reports fake news)
- **masakhaner · Datasets at Hugging Face**
  Accompanying paper
  Ideas:
    - What features are the models using from Swahili which are transferring to other African languages?
    - [This paper has trained models which you can re-use and do interpretability studies on, if you wish]
- LAFAND-MT - Document level translation. This is a Masakhane project. Join **#lafand-mt**
  Ideas:
    - In the document, you'll see a number of short term research questions they'd like to look into. You can join the slack group. We have some compute credits on google cloud for this project which can be provided on request in the group.

COVID-19 Tasks

- COVID-19 FAKE NEWS INFODEMIC RESEARCH DATASET (COVID19-FNIR DATASET).
  Ideas:
    - What classification models perform best on classifying COVID 19 fake news?
    - What features are the classification models using?
- AYLIEN COVID-19 News Dataset
  Ideas:
    - What classification models perform best on classifying COVID 19 sentiment?
    - What classification models perform best on entity recognition?
    - What features are the classification models using?

Cyber Safety

CW: Please note that some of the datasets on cyber safety contain sensitive and offensive material. When processing this data, it is quite likely that you'll encounter hateful and upsetting content.

Hate Speech

- Multilingual detection of hate speech against immigrants and women in Twitter (hatEval) - Competition
- OffensEval: Identifying and Categorizing Offensive Language in Social Media - Competition
- HateSpeech data
- **Ideas:**
    - Which classification models transfer well to hate speech classification?

-

Misinformation/Disinformation

- South African Website Fake News [South African Disinformation [Fake News] Website Data - 2020]
- Hyperpartisan News Detection - [PAN @ SemEval 2019 - Hyperpartisan News Detection]
- Fake News Challenge - [FakeNewsChallenge/fnc-1]
- Fake News - [Fake News]
- RumourEval - [Competition]
- **Ideas:**
    - What features get used by different classification models to determine if a model is fake news or not?
    - Which classification models work best for this dataset?

Helpful Libraries

- [JoeyNMT]
- [HuggingFace]
- [NLTK]