IU ADFN: Project 2

Choose from one of the following four tasks:

Fake News Challenge

Implement a system to take on the Fake News Challenge. It's described at http://www.fakenewschallenge.org/ .

- Stance classification
- Four target classes
- Data is supplied already

Goal: try to detect suspicious articles. Evaluate how well your system does using F1 score and Accuracy.

Many examples on the web

Stance detection in social media comments

Use an existing stance dataset as training data. They're available in lots of languages - ask me if you can't find them! English data is available from the PHEME dataset on Figshare.

Try machine learning / NLP approaches to get a good F1 score. Accuracy is not so useful because there tends to be a lot of class imbalance, It's tough, so getting over 40% is good, over 60% is excellent.

E.g.: https://www.aclweb.org/anthology/D16-1084/

Stance detection for a new language

The goal is to label some new stance data. To do this, first gather social media data; include a claim, and also the response to the claim.

Then, annotate the data for stance, using e.g. Snorkel (https://www.snorkel.org/). Finally, build a baseline classifier for predicting stance on this data.

E.g.: https://www.aclweb.org/anthology/W19-6122/

Veracity prediction from social media comments

Use the HMM system - or a similar one - to develop a veracity predictor. You can use the PHEME data for training. Try to get the best model you can.

Paper: https://www.aclweb.org/anthology/C18-1284/

Project format

- You can work in groups of 1 or 2

- Write as an academic paper
  - Use the LREC 2020 style files
  - https://lrec2020.lrec-conf.org/en/submission2020/authors-kit/
  - Results will be published informally
  - You're welcome to submit to LREC with my help

- Main sections:
  - Introduction
    - Discuss the general problem and your understanding of it
  - Background (literature, similar previous work)
  - Method
    - Baseline
    - Dataset (where you get the text from; characteristics of the data)
    - Your approach
  - Analysis
    - What worked, what didn't work, and why; what was difficult (e.g. in the text)
  - Conclusion

- Project assignment due **November 18**
- Mail me for any corpora/annotations, there are many

- Submission:
  - Code or a link to a colab
  - Documentation (4-page short paper, or max. 8-page long paper)