Project Ideas for NLP with PyTorch

A. Identify Sarcasm in social media text

Identification of sarcasm is a difficult task due to the gap between literal and intended meaning. Sarcasm recognition is a task that can potentially provide a lot of benefits to the area of sentiment analysis.

Take a look at this paper from 2016:
https://www.aclweb.org/anthology/C16-1231.pdf

B. Determine if one text contradicts or entails the other text

Textual entailment attempts to capture whether one sentence can be inferred from another. TE is a useful component in several applications, including finding whether a text, speech or manifesto contradicts itself.

Use textual entailment to find inconsistencies in political positions.

As a starting point, this dataset is annotated for entailment:
http://marcobaroni.org/composes/sick.html

C. Extract a timeline of events from historical texts.

A 2015 shared task at SemEval asked participants to build timelines from written news in English. The goal was to order all the events of a target entity on a timeline. This application can be extended to historical texts, news texts, and political addresses.

You can take a look at the full task description, along with some test and training data, here:
http://alt.qcri.org/semeval2015/task4/

D. Classifying the Manifesto Project

The Manifesto Project collects and annotates manifestos (electoral projects). The dataset contains over 114K English labelled text units.
https://manifesto-project.wzb.eu/tutorials/primer

While hand-labelling remains the standard in the political field, automating this process could potentially allow us to label much larger collections of texts for much less time and effort. Supervised machine learning can be used to classify documents or even shorter texts, like those we classified in class. Classifying shorter texts like those labelled in the Manifesto Project remains a more difficult task. Using previous attempts at classification as a jumping off point, you can try to classify short texts and phrases in the Manifesto project, using the labels, and see if you can improve on the current methods.

(You can find the data here: https://www.dropbox.com/sh/ttafzi26ul55pkz/AABH4gDCRjnPB9sk4rzcKVRNa?dl=0)

E. Style Transfer of Text

Style transfer is a well established deep learning technique for visual data, and recent research has applied it to NLP. Different researchers have proposed different algorithms, and you can read about some of them here: https://www.groundai.com/project/what-is-wrong-with-style-transfer-for-texts/1

These methods have some interesting implications in the political science domain. An interesting project could involve applying style transfer to detecting polarisation, fake news, or populism, using contextual knowledge to capture this information.

F. Automated Hate Speech Detection

The impressively named WOAH workshop provides a venue for research into detecting online abuse, in particular hate speech in online forums and social media. As a result, the workshop has put together a number of datasets annotated for hate speech (available at the link below). A useful contribution could be to train a model that detects hate speech, using deep learning NLP techniques. A successful model could potentially participate in the workshop at WOAH in 2020.

If you're interested, the data is available here: http://hatespeechdata.com/
You can also take a look at last year's workshop proceedings here:
https://www.aclweb.org/anthology/volumes/W19-35/