Project Proposal

For the structure of the proposal, we expect you can follow a research paper format, which consists of introduction, related work, datasets/environments, methodology, and evaluation description. In general, we want to see that you have a clear goal in the project. The technical details can be described in a rough manner, but in principle, you need to show what problem you want to study.

* Introduction: This part includes the background and motivation of your project. The problem has to be well-defined. What are the input and output. Why this is an important problem to study.
* Related work: Every project must include a description of the current state of the problem. In this part describe what has been done in previous work on the same or related subject. We don’t want you to just download their source code and run it on the released data. We want you to do something interesting. So here describe why what you propose to do here is useful, any aspect is novel and different from prior work. Doing comparison work can also be treated as novel if nobody has reached the same level of depth or scale.
* Datasets/Environments: For problems that work on a dataset, the following information is required. What data do you want to use? What is the size of it? What information is contained? Why is it suitable for your task?
* Methodology: What models do you want to use? You may change the model as the project goes, but you may want to indicate some type of models that might be suitable for your problem. For example, is it a supervised learning problem or unsupervised? What classifiers can you start with? Are you making improvements? Are you going to implement any baseline methods? You don't have to be crystal clear on this section, but it can be used to indicate the direction that your project goes.
* Evaluation: What metrics or methods do you want to use for evaluating your models?

[1] <https://aclanthology.org/S19-2145.pdf>

[2] <https://towardsdatascience.com/sentiment-analysis-using-logistic-regression-and-naive-bayes-16b806eb4c4b>

[3] <https://aclanthology.org/P14-1105.pdf>

[4] <https://www.kaggle.com/lingshuhu/political-partisanship-tweets?select=cong_politician_tweets_2020-3-12-2021-5-28_text_party_balanced_anonymous.csv>

Progress Report

**Data preprocessing**. If you are working on a dataset, how did you preprocess the dataset? Which features do you select, and for which purposes? What is the resulting dataset (size, splits, any characteristics that are relevant to the project)? Make sure you motivate the decisions made along the way.

Ideas from [2]

1. Eliminate handles and URLs
2. Tokenize the string into words.
3. Remove stop words like “and, is, a, on, etc.”
4. Stemming- or convert every word to its stem. Like a dancer, dancing, danced, becomes ‘danc’. You can use porter stemmer to take care of this.
5. Convert all your words to lowercase.
6. Mask rare words as “UNK” tokens in frequency counts

Then:

1. Create a frequency dictionary

**Methodology**. Report what are your current models/methods in solving the problem. Including: model architecture, parameter spaces, how to train the model, any model selection and hyperparameter tuning. Do you use any comparisons? What are the baselines (if applicable)? What is the rationale for designing such a model (your proposal)?

Ideas from Professor meeting

1. Try out LSTM and Logistic Regression (starting with latter)

Baselines from [1]

* The RANDOM baseline chooses a label at random from {liberal, conservative}.
* LR1, our most basic logistic regression baseline, uses only bag of words (BoW) features.
* LR2 uses only BoW features. However, LR2 also includes phrase-level annotations as separate training instances
* LR3 uses BoW features as well as syntactic pseudo-word features from Greene & Resnik (2009). These features from dependency relations specify properties of verbs (e.g., transitivity or nominalization)
* LR-(W2V) is a logistic regression model trained on the average of the pretrained word embeddings for each sentence (Section 2.2).

Hyperparameter tuning similar to homeworks (cross validation, grid search, random search, etc.)

**Current result**. Report the results you’ve obtained thus far. Describe the evaluation metrics or objectives/goals. Including: what are the performances of your model on the problem and the baseline methods if applicable.

Ideas from Professor meeting

1. Use both accuracy and f1-score (to cover the case where the classes are imbalanced)

**Future plan for improvement**. Report what you are anticipating to improve and achieve in the finalized version.

Enhancements

1. Checkpoints

Final Report

**To Do**

* Try out RNN, LSTM, or Naive Bayes for comparison
* Implement validation set!!!!
* More hyperparameter tuning
* Use non-politician tweets
  + (currently only using: cong\_politician\_tweets…)
* ~~Try n-grams as feature (in frequency dictionary)~~
* Presentation slides
* Project write-up
  + Add references (Annika’s in code)

**Problem Description**

* What is the task?
  + A description of the particular problem(s) you are addressing
  + The motivation behind the project and why this is an important problem to study (e.g. you can discuss its social impact or other impact).
* System input and output (Examples are encouraged)
  + What is the input that the model receives?
  + What is the output that the model generates?
* What are the challenges of the task?
* What are your contributions in the project? Give a brief overview of what you have done and achieved. Report the task allocations among your team members

**Reference/Related work**

* Put your work in context: What has been done before related to your topic?
  + Please make sure that you cite relevant and (if possible) relatively new research.
* What’s new in your work? E.g. Do you propose a new model or model variant? Do you provide new analysis or insights which can be used for future model development?

**Methodology**

* What are your datasets?
  + What data do you use?
  + What is the size of it?
  + What information is contained? (Examples are encouraged)
  + Why is it suitable for your task?
  + Provide some details about the dataset (e.g. If it is a classification task, is the dataset biased? If so, what is the percentage of each class? Are there any characteristics of the data?)
* Preprocessing of the data (if needed)
  + How do you collect and clean the data?
  + How do you process the data to fit into the structure that the model expects?
* What methods do you experiment with? And why do you think they’re reasonable and suitable for the task? Please include the details of your model architecture (e.g. for a neural model, number of layers, loss functions, parameters, etc.) A diagram for your model can be helpful here.

**Experiments**

* How do you split the dataset into training/validation/test sets? Please also include the implementation details of the training process of your model (learning rate, batch size, number of epochs, parameter tuning, etc.)
* What evaluation metrics do you use? Are they proper to calibrate system performance? Have a discussion on pros and cons if needed.
* What are your models and what are their performances?
  + What are the baselines and what are their performances? How do they compare with your model?
  + Discuss your results. If it is not satisfying, what might be the reasons? Have you tried other variants?
  + Summarize the insights you have gained through this exercise. You can do an error analysis or other types of analysis here.

Note: Please use tables, figures, etc. to support your claims and observations when it’s proper (e.g. displaying examples, results, or system outputs).

Project Presentation

The presentation content guideline is the same as the report. However, given the time limit (each team has 7 minutes), the team should use their own judgment on what are the most important elements to be covered. The principle will be that each aspect will need some discussion to allow the audience to understand the problem to be solved, models/datasets used, and the main results and observations. In other words, the presentation should be self-explanatory.

NOTE: everyone turns in slides by end-of-day 4/13