Ling 506: Affective Computing

Term Project Proposal – Jacob Fuehne

Due date: Oct. 15th, 2020 (by midnight)

You are required to submit a one- or two-page written project proposal. The proposal should include enough detail to make sure you've found a good problem, you understand how hard it is, you've thought a bit about a general plan for how to attack it, and you have an idea about what kind of experiments you might run to test the success of your implementation.  Please do not be vague in your written descriptions. You are given the following outline (questions to answer/address) to help you in the process:

* Goal (Problem definition and importance)
  + What am I going to do?
    - My objective in this project is to build upon existing research in using topic modelling, multi-domain text, transfer learning, combining corpora, etc. by incorporating ideas from sociolinguistics. Building on the social roles theory and how social context affects the way someone speaks (formality/friendliness/emotional intensity/etc.), I intend to use this to determine if a meaningful trend can be established and leveraged when deciding which corpora can be adapted into a unified corpus. Ideally, I will also be implementing methods of topic modelling prediction, domain transferring, and more, and comparing different combinations of techniques and different combinations of corpora. In establishing trends, a variety of stats and performance metrics will be measured. If possible, I would also like to attempt a classifier to determine the some of these social context aspects.
  + Who would benefit?
    - The people who stand to benefit the most from this research are the people who have always benefitted from better methodologies in multi-domain sentiment analysis, the people who work in fields with limited data sizes. Arguably just as important as positivity and negativity, if one can draw connections in language with computational models, customer service models, digital assistants, and more will be able to classify how someone perceives their brand (ie, a loyal customer will have different aspects from someone who is simply upset about their product.)
* Problem Difficulty
  + Why is it hard?
    - While there is more research on doing topic modelling, domain transferring, and combining corpuses that I thought, even after going many papers deep in research, I have not found anybody specifically addressing this aspect of linguistics when it comes to combining text data into a corpus, or when it comes to combining corpora together. It is difficult because
  + How hard is it?
    - This research problem has many different routes to go with it, and it could range from impossible for my abilities in terms of resources and abilities, to doable and very useful, to very achievable and useful, but not as impressive as it could be. Thus, I intend to take this project in milestones with realistic goals and be careful with time management.
* Previous Work [this bullet point is NOT mandatory now, but recommended]
  + What have others tried?
    - An Analysis of Annotated Corpora for Emotion Classification in Text
    - Assessing the Corpus Size vs. Similarity Trade-off for Word Embeddings in Clinical NLP
    - In Pursuit of an Efficient Multi-Domain Text Classification Algorithm
    - Frustratingly Easy Domain Adaptation
    - Efficient Topic Level Opinion Mining and Sentiment Analysis Algorithm using Latent Dirichlet Allocation Model
    - LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis
    - Enriching Topic Coherence on Reviews for Cross-Domain Recommendation
    - Challenges and Recommended Solutions in Multi-Source and Multi-Domain Sentiment Analysis
    - Automatically Identifying Complaints in Social Media
    - And more, but none have done research specifically with what I am doing
* Approach
  + What approach am I going to try?
    - I intend to refer to corpora with consistency in the social context among people (ie, newspeak, people in certain communities of reddit, professionals and politicians on twitter [but filtered to the ones that act formally]), but not necessarily a consistent domain. Pre-built corpora may be combined with corpora formed from an API. I intend to show trends among these sets, and look for trends in them. The data will then be separated and tested with various models, including the best SOTA multi-domain and topic modelling algorithms that I can implement. Combining corpora within a domain, but of varying social roles will be evaluated against combining corpora within a domain with consistency in social roles. This can also be tested against randomized corpora
  + Why do I think it will work well?
    - Hypothetically, certain models should perform better as the training data will have controlled for one extra variable, social aspect, thus emotional intensity, friendliness, and formality should lead to more clear trends in the training data.
* Methodology
  + What resources (e.g.: data, annotations, etc.) are required and how am I going to get them (unless I already have them)?
    - As previously mentioned, there is a significant amount of data in the Reddit Corpus, enough to even train GPT2 on individual subreddits. Subreddits tend to focus (mostly) on a singular domain and social context. Also, useful could be the Complaint twitter corpus. I will also use different twitter accounts if needed. In addition, if I can find any consistent corpora, I will use these as well. I will also likely cross reference and cite sociolinguistics papers for suggestions on trends to look for.
  + What computational models do I have in mind to solve the problem?
    - I will use models implemented in the previous works above, but I will also use models of varying structures, (ie, some that scale poorly, some that require lots of data but scale well, some that leverage context better than others, etc.)
  + Which aspects of your model(s) are particularly hard?
    - The hardest part will be my target goal, of performing some level of classification on aspects of social context, such as formality, friendliness, emotional intensity, etc. Also, likely it will be difficult to implement the LDA models mentioned in papers.
  + What to do if the hard steps don't work out (meaning, what is my plan B)?
    - If necessary, I will focus my efforts more towards annotation and gathering of data, and producing more definitive results on the effectiveness of considering these aspects, rather than on complex models from 2019 and 2020, and a classifier for these aspects.
* Metrics
  + How will I measure success?
    - I plan to use a variety of metrics including accuracy, f1, recall, and precision. As I am approaching relatively uncharted territory, I intend to make my data clear, such that, if I were to miss any possible conclusions, or if I am unsuccessful, someone could still gather useful conclusions and inspiration from my paper, just as many publications we have seen in class and as I’ve seen on this topic.
    - I will be measuring inter and intra-domain variance across corpora and areas of Reddit, as well as other metrics
* Summary
  + Why is this project of interest to me and what will I learn by doing this project?
    - I’m interested in this because I am personally a big fan of digital assistants and AI companions, as well as working with NLP analysis in social contexts. Time and time again, we’ve borrowed concepts from our knowledge of the brain, as it is even the foundation of artificial intelligence, and so I think that we could use more techniques that leverage the mechanisms of the brain and mind, not just on the neurological level, but also more abstracted such as in sociolinguistics.