Projects

Arjun Mukherjee†

Course webpage:
http://www.cs.uh.edu/~arjun/courses/ml

Grading Scheme

- Form groups of two student ~ 15 groups (OK to do independently but work should tantamount groups effort) :
- Project proposal (10%) – This is a free if you happen to choose one of the projects I have developed for this course. If you choose to work on a project of your own choice – please develop problem statement and find your own dataset and have it approved by me.
- (Original/Novel) Idea (25%) [This will invariably require reading a few relevant papers]
- Implementation (35%) [research quality code with train/dev/test splits, 5-fold cross validation etc.]
- Report/Presentation (30%) [professional quality report e.g., drafted with LaTeX]
- **Ideally, if you are a PhD student you may want use this opportunity to learn to write good papers.**

Gender Classification from Text

- **Problem Statement:** Given a set of labeled blogs written by males and females, predict the gender of the author of a new blog.

- **Dataset:** Sample blog author dataset used in [Mukherjee and Liu, EMNLP 2010] available from: http://www.cs.uic.edu/~liub/FBS/blog-gender-dataset.rar

- **Some Ideas (you are required to develop your own ideas that beat the baseline):**
  - Baselines: Standard Word, POS n-grams as features with SVM, NB, DT, etc.
  - Baselines and other approaches reported in [Mukherjee and Liu, 2010]
  - Employ (regularized) regression and model classification as regression [can be made to work better if tweaked well]
  - Find the most effective features via PCA and feature selection. Ensemble learning [?]
  - Cluster relevant features to feature groups [?] combine with L1/L2 regularized regression (rationale is that probably features of the same discriminative strength should have the same weights – so you constrain them via your cost function in the regression model rather than standard regression]

Opinion Spam Detection

- **Problem Statement:** Given a set of labeled reviews [Filtered i.e., fake vs non-filtered non-fake], figure out whether a new (unseen) review is fake or not.

- **Dataset:** Yelp filtered review dataset used in [Mukherjee et al., ICWSM 2013; Kc and Mukherjee WWW 2016 available from: http://arjun5.rcc.uh.edu/dumps/yelp_hot_res.rar

- **Some Ideas (you are required to develop your own ideas that beat the baseline):**
  - Baselines: Standard Word, POS n-grams as features with SVM, NB, DT, etc.
  - Baselines and other approaches reported in [Mukherjee et al., ICWSM 2013]
  - Employ time series (regularized) regression models (e.g., see Yogatama et al., 2011) on top of ideas presented in [Kc and Mukherjee, WWW 2016]
  - Novel ideas using Markov models, PCA and Ensemble learning [?]
  - Time-series clustering of different restaurants to understand the spamming policies and using them for future prediction of fake reviews [?] (grow on top of [Kc and Mukherjee, WWW 2016])

Fine-Grained Sentiment Analysis

- **Problem Statement:** Given a set of labeled aspect specific sentiment expressions (e.g., "I experience *recurrent signal drops* on this wireless router…" for the aspect signal) across variety of review sentences, find new sentiment expression spans on unseen sentences of the same aspect.

- **Dataset:** Amazon review issue sentence dataset used in [Mukherjee CICLING 2016] and is available from: http://arjun5.rcc.uh.edu/dumps/issue_sent_labeled_data.zip
- **Some Ideas (you are required to develop your own ideas that beat the baseline):**
  - Baselines: HMMs, Standard Word, POS n-grams as features with CRF, etc.
  - Baselines and other approaches reported in [Mukherjee CICLING 2016]
  - Convert the sequence extraction as a windows span regression problem with spans as response variables around the aspect. [think how ?]
  - Novel ideas using clustering of features around the aspect using PCA, Markov models (e.g., see approaches in Li et al., AAAI 2015),
  - Use an ensemble of different classifiers on different feature spaces to yield a more effective boundary around the aspect [?]

Authorship Attribution

- **Problem Statement:** Given a set (e.g., 25 or more) of "known" documents by a single person and a "questioned" document, the task is to determine whether the questioned document was written by the same person who wrote the known document set.

- **Dataset:** PAN benchmark datasets from 2011, 2013, 2015 years or Review dataset used in [Shrestha et al., CICLING 2016] and is available from: http://arjun5.rcc.uh.edu/dumps/AA.7z
- **Some Ideas (you are required to develop your own ideas that beat the baseline):**
  - Baselines: Standard Word, POS n-grams as features with multi-class SVMs.
  - Baselines and other approaches reported in [Shrestha et al., CICLING 2016]
  - Convert multi-class classification as a regression [?] this can often help determine the edge cases of authors and allow you to develop a post-hoc function on using the output of regression to actual classification (instead of hard multi-class where you don't have this flexibility)
  - Novel ideas using Markov models, Use an ensemble of classifiers on different feature spaces [?]

Debate Mining

- **Problem Statement:** Given a set debate discussions across politics and religion domains by various users on various topics (threads), find agreement ("point taken", "I think you're right") and disagreement expressions ("I disagree", "you have no clue", etc.) mentioned in the given discussions.

- **Dataset:** Volconvo.com debate forum dataset used in [Mukherjee and Liu, KDD 2012] and is available from: http://arjun5.rcc.uh.edu/dumps/DebateData.zip
- **Ideas (you are required to develop your own ideas that beat the baseline):**
  - Baselines and other approaches reported in [Mukherjee and Liu, ACL 2013], [Mukherjee and Liu, KDD 2012]
  - Soft/noisy labeling of debate expressions (e.g., manually or semi-automatically using heuristic rules) and then employing ideas for Fine-Grained Senitmetn Analysis project
  - Novel ideas using Markov models for sequences (e.g., see approaches in Li et al., AAAI 2015)
  - Sentence clustering combined with expression span regression around select "arguing" keywords [?]