DATA MINING – PROJECT PROPOSAL

ORIN BROWN

PROBLEM STATEMENT

- Text classification
  - The goal of this data mining project is to develop a classification model that accurately categorizes new articles into predefined topic categories. Given a vast amount of unlabeled news articles, the objective is to create a reliable system that can automatically assign appropriate categories to incoming articles based on their content.

RELATED WORK

- Traditional models
  - Bag-Of-Words (BOW), N-gram, Term Frequency-Inverse Document Frequency (TF-IDF), word2vec, and Global Vectors for word representation (GloVe).
- Deep learning models
  - Sentiment Analysis (SA), Topic Labeling (TL), News Classification (NC), Question Answering (QA), Dialog Act Classification (DAC), Natural Language Inference (NLI), and Relation Classification (RC).
- Excellent and current overview on text classification work:
  - ACM Transactions on Intelligent Systems and Technology Volume 13Issue 2 Article No.: 31pp 1–41
  - https://doi.org/10.1145/3495162

PROPOSED WORK

- Use one traditional model and one deep learning model.
- Find most “diagnostic” words / phrases / word pairs, see if they make sense to a human.
- Find distinguishing patterns for different article types.
- Focus on “problem areas” for text classification models:
    - Stability
    - Interpretability

DATA

- Source
  - I’ve identified a website with a large number of different data sets. Many have already been processed and prepared for analysis.
  - https://ana.cachopo.org/datasets-for-single-label-text-categorization
- Warehousing
  - All the data sets are in the range of 10’s of MB, so I will have no problem storing them locally or on github while developing the project.

EVALUATION

- Accuracy
  - The primary metric I will use to judge model performance will be the accuracy of the text classification.
- Runtime
  - Speed of training and prediction.
- Interpretability
  - A model will be rated higher if it produces some kind of intermediate output that is interpretable by humans. For instance, a list of the most “diagnostic” words or phrases.

TIMELINE

- There are roughly six weeks left in the term at time of writing, so I propose to divide the tasks as follows:
  - Data gathering – 1 week
  - Research on models – 1 week
  - EDA and data cleaning – 1 week
  - Modeling and model tuning – 2 weeks
  - Analysis and report writing – 1 week
- This is not a hard and fast division of tasks, as I'm sure they will blend into each other, but more of a general prediction as to how I will spend the remaining time.