Final Project Proposal

Liangpeng Zhuang, Xiaofan Jin, Sizhe Wen

1. **Method:**
   - For our project, we will employ a supervised classification method which uses Natural Language Processing (NLP) features which measure different aspects of a user comment. To be more specific, our feature can be divided to three classes: N-grams features, embedding-derived features and syntactic features
     i. N-gram features (TF-IDF weighted vectors): word and char level
     ii. Embedding-derived features: word2vec, doc2vec
     iii. Syntactic features: natural language parsing
   - Models we might use to detect personal-attack comments
     i. Support vector machine (SVM)
     ii. Naive Bayes
     iii. Random Forest Classifier
     iv. CNN (Multi-layer Perceptron (MLP)) / RNN

2. **Related topics that are covered in class:**
   - We learned about text classification on class using Naïve Bayes, which is under conditional independence assumption. Here, we want to explore more ways to first extract feature vector from a given text, and then do text classification, which in our case is to identify personal attacks in among all comments.
3. **What we are specifically interested in studying.**
   - Many of the platforms are trying to curb the users to use damage personal attacks for online discourse. However, the difficult part is to figure out the prevalence since there is a large scale of personal comments in online platforms. The goal of this project is to develop a machine learning method to analyze personal attacks. Also, by using approximation of crowd-workers, we are going to develop an evaluation method for a classifier. We feel very excited to use the course knowledge to solve a meaningful practical problem like this.
4. **Why we think this is an interesting problem**
   - We will follow the complete industrial steps, text cleanup, matrix generation, feature extraction, data modeling, and hyper-parameters tuning. The potential future job interviewers might find it professional, and also interesting.
     In this project, we are trying to solve a practical problem. It will be a very useful tool to detect personal attacks in online platforms. This dataset we collected from Wikimedia Foundation. It includes all 95 million users and article talks between 2001 – 2015. Not only the depth, but also the breadth of the data makes it more precious.
5. **Reference to datasets that we plan on using.**
   - **Wikipedia Talk labels: personal Attacks**
     This data set includes over 100k labeled discussion comments from English Wikipedia. Each comment was labeled by multiple annotators via Crowdflower on

whether it contains a personal attack. Demographic data for each crowdworker is also included.

Wulczyn, Ellery; Thain, Nithum; Dixon, Lucas (2017): Wikipedia Talk Labels: Personal Attacks. figshare.**https://doi.org/10.6084/m9.figshare.4054689.v6**

6. **Related work**
    - ○ Ellery Wulczyn, Nithum Thain and Lucas Dixon. Ex Machina: Personal Attacks Seen at Scale. *Computation and Language.* In arXiv:1610.08914v2
    - ○ C.Nobata, J.Tetreault, A.Thomas, Y.Mehdad, and Y.chang. Abusive language detection in online user content. In WWW, 2016.

We find this *Ex Machina: Personal Attacks Seen at Scale* paper and the dataset from the final project of CS5100-02 Seattle: Foundations of Artificial Intelligence that is taught this semester. In the requirement from AI's final project, we have to try at least 3 modeling methods other than the Logistic Regression module given by authors of the paper. There are two of us(Sizhe Wen and Xiaofan Jin) are taking the AI course this semester. Therefore when we are doing this project for the ML course, we will not use any modeling methods that we would use for AI's project. That means we have to employ at least four different machine learning algorithms for the same paper and datasets and our goal is to go beyond that number. We really hope dive deeper and explore broader with all the knowledge that we have been taught throughout the semester, so that we could have a comprehensive understanding of all the algorithms from the lectures.