**PROPOSAL**

Graham Harris, Judy Zhang, Christine Ma, Jong Heon Han

**PROBLEM STATEMENT**

Given a movie review dataset with positive and negative sentiment labels and a chosen binary classification model, our team aims to investigate the effectiveness of various tokenization methods. We explore how these tokenizers differ in parsing and encoding linguistic information from a body of text, and their performances on the specific task of sentiment classification. We measure their effectiveness to one another on the same dataset and classification model. We then provide the best system among those investigated.

**DATASET**

- 50,000 IMDb movie reviews
    - Balanced distribution of positive and negative
    - 10, 000 in validation set, 15,000 in test set, 25,000 in training set
    - 50,000 unlabeled reviews, will not be used during the testing process
    - Web link: https://ai.stanford.edu/~amaas/data/sentiment/.
- Entries in training_set.txt format:
    - Text filenames act as IDs
    - Sentiment labeled as 1 (positive) or 0 (negative)
    - Movie review text
- Entries in test_set.txt format:
    - Same as training set, without labeled sentiment
- Entries in answer_key.txt format:
    - Same as training set, without movie review text
    - *Additional file answer_key_with_review.txt contains the movie reviews*

**MODEL AND DIFFERENT TOKENIZATION METHODS**

- Model: Logistic Regression
    - Input: Tokenized movie reviews

    - Output: Discrete movie sentiment label (0 for negative, 1 for positive)
- Tokenization Methods:
    - Split by space (control)
    - Adjectives as sole tokens
    - Ignoring punctuation
    - TF-IDF weights
    - Stemming included

**EVALUATION METHOD**

Our team will use recall, precision, and f-measure to evaluate token methods. These results will then be compared to answer_key.txt.

**ROLES AND COLLABORATION**

1. Jong Heon Han - Data Analyst, Evaluation
2. Graham Harris - Writer, Editor, Researcher
3. Christine Ma - General Coder, Dataset Implementation, Tokenization
4. Judy Zhang - Researcher, Tokenization

Graham and Judy will work together on the research component. Christine and Judy create the tokenizaters. Jong will perform analysis, the results of which will be communicated to the group. Collectively we will analyze the results and determine conclusions.

**ACADEMIC ARTICLES**

- Breaking Sentiment Analysis of Movie Reviews (link)
    - Movie reviews allow for pragmatic and stylistic manipulations, it is difficult for systems to properly recognize sentiment
- Learning to Generate Reviews and Discovering Sentiment (link)
    - Sensitivity of learned representations of data models on distributions that they are trained on
- Overcoming Language Variation in Sentiment Analysis with Social Attention (link)
    - Group together reviews with similar language patterns

- How to Fine-Tune BERT for Text Classification? (link)
    - Sentiment analysis tasks used to rigorously test different fine-tuning methods of BERT on text classification
    - Yelp reviews proposed for sentiment analysis
- Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings (link)
    - Best results were obtained on untrained data, suggests that there were complementary text characterizations in their model
    - May be beneficial to try to combine multiple tokenization methods for optimal results (if the methods are complementary)

**CITATIONS**


Bonfil, Ben, and Ieva Staliunaite. "Breaking Sentiment Analysis of Movie Reviews." *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, 2017, pp. 61-64, https://www.aclweb.org/anthology/W17-5410.pdf.

Johnson, Rie, and Tong Zhang. "Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings.", 2016, https://arxiv.org/pdf/1602.02373.pdf.

Raford, Alec, Rafal Josefowicz, and Ilya Sutskever. "Learning to Generate Reviews and Discovering Sentiment.", 2017, https://arxiv.org/pdf/1704.01444.pdf.

Maas, Andrew, et al. *Learning Word Vectors for Sentiment Analysis*. www.aclweb.org/anthology/P11-1015.

Sun, Chi, et al. "How to Fine-Tune BERT for Text Classification?", https://arxiv.org/pdf/1905.05583.pdf.

Yang, Yi, and Jacob Eisenstein. "Overcoming Language Variation in Sentiment Analysis with Social Attention." *Transactions of the Association for Computational Linguistics*, vol. 5, 2017, pp. 295–307, https://www.aclweb.org/anthology/Q17-1021.pdf.