**Data 620 - Final Project Proposal**

*Members/Researchers: Jenkins, Tamiko; Marks, Juanelle; Wong, Calvin*

## *Proposed Title: An Exploration of Gender Classification Techniques for Blogs*

## Project Background

Blogging has become a common way of communication on the web. Blogs are written by persons of all ages and gender. Gender classification of blogs can have commercial applications. For example, when integrated with recommender systems, it can help service providers to push relative gender related products/articles/materials to their customers. Conversely, gender classification and deriving other demographic features derived from stylistics can be used to, for example, help identify online predators.

In recent years, text analysis research has explored the use of various techniques to classify a blog as being written by a male or a female. An example of one such research is conducted by Liu and Mukherjee, titled, Improving Gender Classification of Blog Authors(2010). In their research, they proposed two novel techniques for gender classification of blogs namely:

1. POS sequence patterns (for updated POS methods, see <http://nlpprogress.com/english/part-of-speech_tagging.html>) that are able to capture complex stylistic regularities of male and female authors and
2. Ensemble feature selection method which can take advantage of many different types of feature selection criteria. (see: <https://scikit-learn.org/stable/modules/ensemble.html>)

A more recent survey of online stylistics in Online Social Networks and Writing Styles–A Review of the Multidisciplinary Literature by Tai, K. Y., Dhaliwal, J., & Shariff, S. M. (2020) proposes extracting a more specific tense morphology, “syntactic patterns of “to be” verbs in simple past, simple present, present participle and past participle tenses, with age and gender” with the following methods:

1. Classification Algorithm: Extreme gradient boosting (XGBoost) (see: <https://xgboost.readthedocs.io/en/latest/>)
2. Dependency Parsing: spaCy (see: <https://spacy.io/usage/linguistic-features#dependency-parse>)
3. Cross-categorization: Stratified K-Folds (see: <https://scikit-learn.org/stable/modules/cross_validation.html>)

## Research Question

Which document classification techniques provide a high level of accuracy when conducting gender classification of blogs?

NB: The specific techniques to be tested and compared have to be finalized by group members.

## Data source

The data to be used will be taken form one or both of the following sources:

1. Blog Author Gender Classification (<https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets>) This data set is a spreadsheet with 3226 rows and 2 columns. Column one contains rows of blog posts and column two contains the accurate gender of the author. This might most likely be the dataset the team will work with.
2. The Blog Authorship Corpus (<http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm>)

The Blog Authorship Corpus consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. The corpus incorporates a total of 681,288 posts and over 140 million words - or approximately 35 posts and 7250 words per person.

Each blog is presented as a separate file, the name of which indicates a blogger id# and the blogger’s self-provided gender, age, industry and astrological sign. (All are labeled for gender and age but for many, industry and/or sign is marked as unknown.)

All bloggers included in the corpus fall into one of three age groups:

* 8240 "10s" blogs (ages 13-17),
* 8086 "20s" blogs(ages 23-27)
* 2994 "30s" blogs (ages 33-47).

For each age group there are an equal number of male and female bloggers.

This may serve as our alternative source.

## Proposed Methodology and Concerns (High Level)

**Data:** Thehe blogger.com dataset has been successfully employed in several separate studies over the past decade. However, we are concerned that grammatical trends may change over time both among age groups, and by online communities.

**Pre-processing:** With this caveat, we will explore some of the more recent work measuring POS-based tense morphology (using spaCy or a similarly well-performing dependency parser) using best practice data cleaning guidelines for that specific library. Our concern in pre-processing is the potential that some spaCy models are paid, in which case we will use the most accurate (based on benchmarks), free POS dependency parser available to us given the processing power available on <https://colab.research.google.com/>. We also question whether dependency parsing is as accurate for online writing versus benchmark corpora. One the other hand, most online writing benchmarks focus on twitter which has very specific grammatical trends we feel may not be useful for blog analysis.

**Classification and Training:** We will attempt to reproduce the study by Tai et al. (2020) by running our POS sequence features through a XGB Ensemble method with 10-fold K Fold Cross Validation to tune our XGD hyperparameters. Our concern here is once again POS tagging accuracy.

**Result Analysis:** We will next measure POS pattern frequency, information gain, and accuracy, and will consider analyzing our results in the context of the work of Mukherjee, A & Liu, B. (2010) who provide an excellent description of common and more traditional feature selection criteria such as F-measure and chi-statistic. Tai et al. (2020) achieved a 97.4% accuracy on gender identification which they compare to similar approaches by Schler et al. (80.1%) and Goswami et al. (89.3%).

**Comments:** While online platforms, stylistics and morphological analysis, and thus gender classification, may change over time, the development of Machine Learning models which can be trained on new data is increasingly important. The standardization through widespread scientific extraction of demographic features from online text can one day be used to anonymize advertising information, as well as to quickly train classifiers on new or unusual online platforms, such as processing niche forums to identify online predators.

## Group Members Roles and Responsibilities

1. Jenkins, Michelle(Team Lead, Lead Analyst)
   1. Data preprocessing
   2. Application classification techniques to dataset
   3. Discussion on accuracy of technique
2. Marks, Juanelle(Support Analyst)
   1. Application of classification technique to dataset
   2. Discussion on accuracy of technique
   3. Consolidation of project parts and materials
3. Wong, Calvin (Support Analyst)
   1. Application of classification techniques to dataset
   2. Discussion on accuracy of technique
   3. Consolidation of project parts and materials

## Reference

Tai, K. Y., Dhaliwal, J., & Shariff, S. M. (2020). Online Social Networks and Writing Styles–A Review of the Multidisciplinary Literature. IEEE Access, 8, 67024–67046. Retrieved April 27, 2020, from https://doi.org/10.1109/access.2020.2985916

‌Koppel, M. (2020). The Blog Authorship Corpus. Retrieved April 27, 2020, from Bar-Ilan University website: http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm

‌Liu, B. and Hu, M. (2015). Opinion Mining, Sentiment Analysis, Opinion Extraction. Retrieved April 27, 2020, from University of Illinois at Chicago website: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets

Liu, B. and Mukherjee, A. (2010). Improving gender classification of blog authors. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP ’10). Association for Computational Linguistics, USA, 207–217. Retrieved April 27, 2020, from https://dl.acm.org/doi/10.5555/1870658.1870679

‌