Final Project Proposal

Christine Shen zs1534@nyu.edu
Aysja Johnson aj2745@nyu.edu
Xintong Li xl3269@nyu.edu
Andrew Yeh ay1626@nyu.edu

Project Description

Our project's aim is to use machine learning and neural network techniques to determine authorship of English poems from their textual content. Although authorship identification is not a novel task and has been applied to various genres of text, it is rarely tested on poems. Poetic language differs drastically from most other texts (fiction and non-fiction), both in its linguistic structures, innovative and creative use of language, and goals. This will provide interesting challenges for feature extraction and engineering, especially for our machine learning approaches. In addition, English poetry tends to group by era, location and movement–making specific author identification an unique task to attempt. Machine learning techniques with heavy feature engineering will be more interpretable and provide insights into what differentiates authors. On the other hand, neural network techniques have proven to be extremely effective in natural language processing; we are curious to see how it performs on authorship attribution in English poetry.

Literature Review

There has been increased research of applying machine learning models to text classification. And most of these models follow two procedures: generate word presentation using either traditional bag of words or pre-trained word embeddings and then feed those features into classifiers (Minaee et al., 2020). In Can et al.'s study, they referred to text features such as most frequent words, word length and two-word collections as style markers and employed SVM and Naive Bayes on these text features (Can et al., 2013). In their experiment, they managed to achieve 90% accuracy on authorship classification for Ottoman poems (Can et al., 2013). Deep learning based methods have also drawn a lot of attention in recent years. Qian et al. used the GloVe word vectors of size 50 as the pre-trained word embeddings and trained a Gated Recurrent Unit (GRU) network and Long Short Term Memory (LSTM) network on both a news dataset (Reuters 50-50) and story dataset (Gutenberg) at both sentence level and article level and achieved 69.1% and 89.2% accuracy respectively (Qian et al., 2017). Besides the classification methods, the pre-trained deep bi-directional language representation model BERT obtained new state-of-art results for a wide range of tasks (Devlin et al., 2018). We would like to test its performance on this authorship attribution task.

Plan of Work

We plan to look at several different models ranging in complexity. The first models will be based on previous work on author attribution in poetry (discussed below) and will leverage techniques like bag of words, SVM, and Naive Bayes classifier. We'll extract features mentioned in the literature, such as lexical, character, structural, poetic (meter and rhyme), syntactic, semantic, and specific words features. We will try different feature engineering iterations like bag of words versus TF-IDF to see the lower level impact on classification accuracy but do not expect high accuracy at this stage. Our next models will be more complex, e.g., neural network classification using LSTM units with word2vec embedding, or pre-trained neural networks like BERT. We expect the accuracy of these models to be much higher, and most of the work will be learning how to implement them successfully and efficiently (most likely on a cluster) as well as interpreting the results and lessons. For example, we may look at how accurate these models are based on genre or length, and we can also look at specific classifications and see if the authors do in fact write in similar styles.

Dataset and Tools

The dataset we are using is scraped from poetryfoundation.org. The dataset contains approximately 15.7K poems and 3,310 unique authors. We will require an author to have at least a certain threshold of poems in the dataset to be included in the training process and will experiment with different thresholds to determine an appropriate one with enough samples in the training, validation, and test set.

Threshold	Number of Authors
10	451
20	103
30	40
50	6

There are some cases in the data where the poems are extremely long. In those cases, we can truncate it, e.g. to the first 1000 words, for computational efficiency and to avoid length being used to determine authorship as our objective is to uncover stylistic distinctions.

Tools we will use include sklearn for machine learning models, pytorch for deep learning models and pre-trained models, various word embedding packages, as well as the NYU cluster and Google Colab to run computationally expensive models.

We will evaluate our models using cross entropy loss and tune based on accuracy. We are as interested in interpreting how the model classifies poems as the accuracy of the models.

Collaboration Statement

Andrew and Christine are planning to work on the data cleaning while Aysja and Xintong research possible pre-trained models that are effective for authorship problems. We will then work on the modelling in parallel, splitting model runs between us to save time and allow hyperparameter tuning.

Lastly, we will read the literature and test alternative formulations of the models when it comes time to interpret our results and what the model is looking at in its distinctions.

References

Fazli Can, Ethem Can, Pinar Duygulu Sahin, and Mehmet Kalpakli. 2013. Automatic categorization of ottoman poems. *Glottotheory*, 4(2):40–57.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep learning based text classification: A comprehensive review. *arXiv preprint arXiv:2004.03705*.

Chen Qian, Tianchang He, and Rao Zhang. 2017. Deep learning based authorship identification. *Department of Electrical Engineering, Stanford, CA*.