VICTORIAN AUTHORSHIP CLASSIFICATION

GIORGIO MONTENEGRO, NICHOLAS SYNOVIC, STEFAN NELSON

COMP 338B

DR. DLIGACH

2 INTRODUCTION

- Over 3.5 million written books from the 1800s to 2015 have been digitized
  - Fraction of the ~130 million books that have been published since the invention of the printing press
- Authorship Attribution (AA) tooling to assist in the discovery and classification of texts by author
  - Three different AA models using different architectures to evaluate their performance at classifying a subset of documents written by 19th century (1800 – 1900) authors
    - Multinomial Naïve Bayes, RNN, and CNN

3 DATASET DESCRIPTION: VICTORIAN ERA AUTHORSHIP ATTRIBUTION DATA SET

A collection of literary works from various authors from the Victorian Era (1800s – 1900s)

Extracted from the GDELT database, an open-source research dataset project

Dataset consists of a matrix of (53,680 x 1,000) word text fragments, for 50 different authors represented

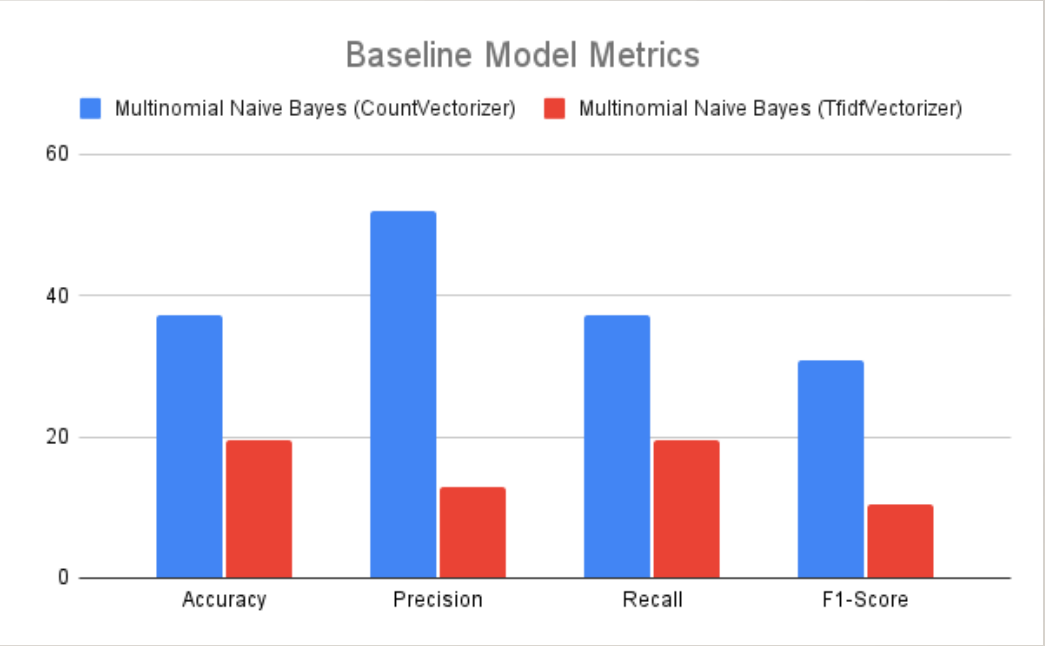
Original authors of the dataset cleaned the data by removing author, title names, the first and last 500 words per document, and only keeping the 10,000 most common words amongst all included texts

We further split the dataset via a random stratified sample were 80% was kept for training purposes (~42,000 documents). And 20% was used for testing (~10,000 documents).

METHOD DESCRIPTION

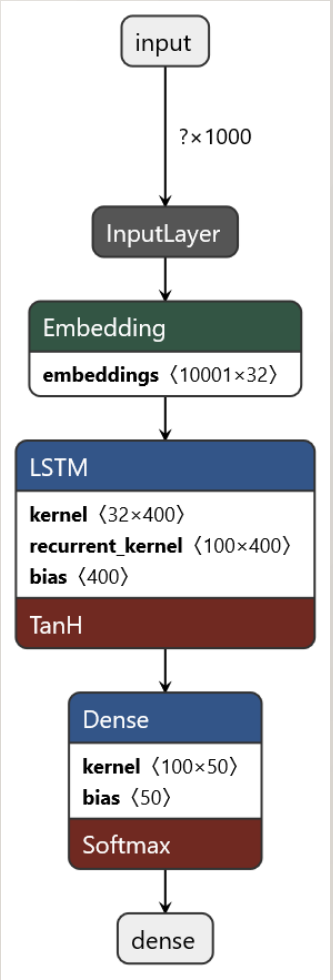
5 BASELINE APPROACH: MULTINOMIAL NAÏVE BAYES

- Multinomial Naïve Bayes model
  - 20% accuracy achieved using *TfidfVectorizer*
  - Best accuracy of ~37% achieved using *CountVectorizor*
- Since accuracy achieved < 50% we decided to try to perform this task using neural networks
- Evaluation on Slide 9



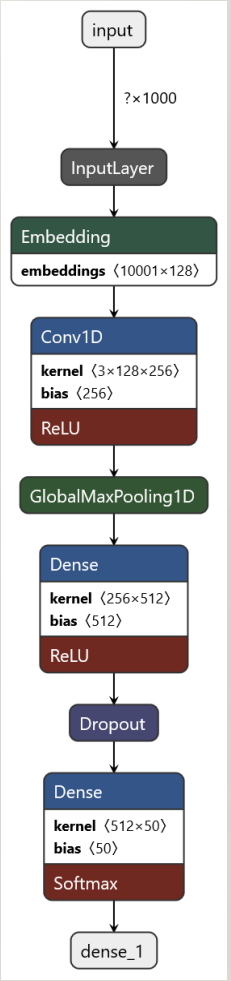
6 METHOD 1: LONG-SHORT TERM MEMORY (LSTM) RNN

- RNN model continually updates learning weights based on a calculated gradient – difference between expected label and return label following a run through of the data
- Where training weights are less than one, LSTM limits the data that the model learns from to only the most significant to avoid gradient becoming too small for training weights to substantially adapt
- Embedding layer includes replacing words in training data with numbers representing ordered frequency of each word in the dataset
- Following embedding layer is the LSTM layer and then the SoftMax layer



METHOD II: CNN

- More commonly used to handle multidimensional data, such as images
  - Possible to use CNNs on one dimensional data, such as text
  - Convolutions work by sliding over the input data with different sized kernels (e.g. 3x3, 5x5)



CLASSIFIER DEMO

EVALUATION

BASELINE APPROACH: MULTINOMIAL NAÏVE BAYES EVALUATION

Model Type	Accuracy	Precision	Recall	F1-Score
CountVectorizer	37.1647%	51.8899%	37.1647%	30.8241%
TfidfVectorizer	19.5417%	12.9272%	19.5417%	10.5286%

METHOD 1: LSTM RNN EVALUATION

- Model returned an 75.90% accuracy amongst sequences in the testing data and the following author specific accuracy:
- Variation of accuracy amongst authors could be explained by consistency in writing style for each author

Most Accurate Classes		Least Accurate Classes	
Anne Manning	97.5225%	George Eliot	39.5062%
James Payn	93.6275%	Robert Louis Stevenson	39.3939%
Mark Twain	91.2801%	Charles Darwin	31.5789%
Horace Greeley	89.6602%	Charles Dickens	27.9070%
Thomas Hardy	86.5510%	Ralph Emerson	27.0270%

METHOD 1: LSTM RNN EVALUATION (CONTINUED)

Accuracy	Precision	Recall	F1-Score
75.9%	78.7046%	74.7019%	76.6511%

METHOD II: CNN EVALUATION

- Model returned an 86.75% accuracy amongst sequences in the testing data and the following author specific accuracy

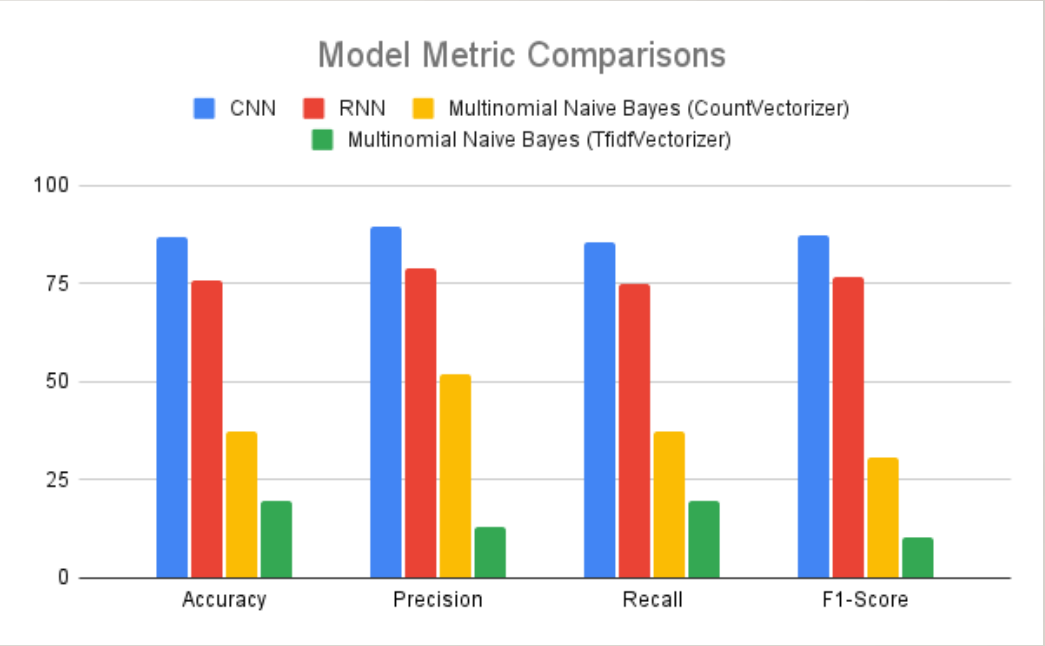
Most Accurate Classes		Least Accurate Classes	
Anne Manning	98.9865%	Charles Dickens	62.7907%
James Payn	98.0392%	Washington Irving	61.5385%
Thomas Hardy	96.0954%	William Carleton	56.5789%
Horace Greeley	94.4324%	John Pendleton Kennedy	50%
Isabella Lucy Bird	94.2605%	Ralph Emerson	45.9459%

METHOD II: CNN EVALUATION (CONTINUED)

Accuracy	Precision	Recall	F1-Score
86.75%	89.4798%	85.5626%	87.477%

EVALUATION OF METHODS BY METRIC COMPARISON

- The CNN method performs better than any other method presented with respect to:
  - Accuracy
  - Precision
  - Recall
  - F1-Score



16 DISCUSSION

- Because the model learns an author's writing style, authors with high accuracy scores could have a distinct or unique writing style.
- Adjusting embeddings for different results
  - Hard to pick up writing style without accounting for all words?
- How could different variables affect contents of author's text?
  - Authors from similar time periods use similar vocab / grammar
  - Authors with different genres could be very disconnected

17 CONCLUSION

- NLP techniques can classify authors based on their writing style
  - LSTM performed strongly
    - ~ 75% accuracy
  - CNN performed even better
    - ~ 86 % accuracy
    - Convolutional layer picks up style subtleties more effectively
- Further steps to increase accuracy
  - Increase dataset range
  - Change vectorization technique