Convolutional Neural Networks for American Accented English Region Localization and Analysis

Andea J. Scott
Department of Computer Science
Department of Energy Resources Engineering
Stanford University
andea98@stanford.edu

James Thieu
Department of Computer Science
Department of Electrical Engineering
Stanford University
jthieu23@stanford.edu

John M. Wall
Department of Electrical Engineering
Stanford University
johnwall@stanford.edu

Abstract

Accent classification has important applications in the development of voice-activated devices, as well as in ethnolinguistic research. This project is aimed at developing a model that can successfully classify accented varieties of North American English. Through our project, we applied a 2-layer Convolution Neural Network (CNN) to Mel-frequency cepstral coefficient (MFCC) graphs of 1,368 recordings of a standardized English text being read. We predicted that the models would be able to perform better on datasets that had been constrained to have equitable class distributions, given the same data augmentation preprocessing. Overall, we found that models trained on class-constrained datasets performed better than non-constrained counterparts, with a observed difference of 10-15% for both accuracy and micro F1 scores. Additionally, the confusion matrices of the models corroborate demographic realities for early 20th century America.

1 Introduction

Because of the significantly increased rate of movement with rapid economic development in the past century, we have observed that accents are, on average, converging and merging. In cosmopolitan settings such as Stanford, it is not uncommon to hear accented English spoken by individuals originating from various countries around the world. Two people from the same region of origin can develop incredibly different accents over their lives. Similarly, people from different regions can converge to have very similar accents. Ethnolinguistic researchers will want to record and catalogue regional accents and dialects in order to preserve regional heritage.

This is especially true for the "American" accent, which evolved out of many different demographic groups, such that many Americans have difficulty identifying distinct regional accents today. Thus, it is of great interest to see if accents drawn from the early 20th century have clearer differences.

The input to our algorithm is a MFCC graph of speech represented as a 2D matrix. We feed this into a Convolutional Neural Network to output a predicted regional accent as a one-hot vector of labels. Using this and the indices of the labels, the predicted regional accent can then be found.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

There has been some work into accent identification from various institutions, though the methods and particular foci are widely varied.

Chu et al. provides a broad overview of different methods for accent recognition, namely KNNs with support vectors, LSTMs, and CNNs, given MFCCs as data input [1]. Interestingly, this project finds that the optimal MFCC length is 200 as opposed to 50, which contradicts our finding that 50 is generally optimal. However, this may be due to differences in sampling rate and dataset size.

Wu et al. and Jiao et al. both focus on using LSTM (Long Short Term Memory) networks, with Jiao et al. achieving similar accuracies as our project [2]. Wu et al., on the other hand, focuses on applications of LSTMs for accent identification via syllable detection, achieving 89% accuracy [3].

Another broad grouping of projects is those that focus on using convolutional neural networks. In this group, we have Zhong et al., which focuses on demonstrating that using MFCCs for CNNs leads to a significant increase in performance when used with appropriate speech segmentation methods [4]. There is also Sheng and Edmund, which focused on non-native English speakers, similar to our original focus [5]. Sheng and Edmund were able to achieve accuracy far beyond ours, which may be due to the addition of redundant clips with injected Gaussian noise.

By far, the project most similar to ours was Chionh et al. from Carnegie Mellon, which used a 2-layer CNN structure very close to ours with a different institutional dataset [6]. With this, they were able to achieve 77.9% test accuracy. The principal difference between our projects was that Chionh et al. focused on filters of size (1,50) with graphs of size (13,469) whereas we used filters of size (3,3) with graphs of (13, COL_SIZE), where COL_SIZE is a tuned hyperparameter. While size (1,50) may have a theoretical grounding in MFCC graph structure, we were unable to replicate the findings in their work, which may be due to the drastic difference in the audio qualities of our datasets.

3 Dataset and Features

Originally, our dataset was comprised of 2,930 audio samples from George Mason University's Speech Accent Archive (SAA) [7]. The SAA dataset consisted of 30 second samples of English speakers from around the world reading a standard paragraph with personal metadata (e.g. birthplace, native language). Unfortunately, the majority of the SAA samples are of speakers of United States origin, which means there is a very unequal class distribution [INSERT into appendix graph from milestone: Appendix 2]. This unevenness leads to an inflated accuracy as the probability that a particular English speaker originated from the US is much higher.

Thus, we shifted the direction of our project to utilize the University of Wisconsin - Madison's DARE dataset which focused on regional accents of the United States [8]. Although the DARE dataset is older, from 1969, the 1,368 standardized audio samples are on average 7x longer. Additionally, since fewer people moved cross-country at that time period, these accents could be considered to be more representative of one of the 8 major North American accents. This dataset distribution was also more balanced; we further limited and equalized the distributions in our testing.

In order to pre-process the data, we used BeautifulSoup [9], a web scraping tool, to access the SAA and DARE websites. This allowed us to fetch the metadata and the URLs of all data points and export them into a CSV file. For the SAA data, we initially improved upon Akshansh Chaudhry's code base in order to correct labelling and metadata scraping [10]. For example, countries such as 'South Korea' were divided into 'South' 'Korea', throwing off labelling. For the DARE samples, the code base had to be almost completely rewritten. After metadata collection, the output CSV containing URLs and metadata was fed into another Python script to pull the audio files from the website. Likewise, for DARE, the script had to be rewritten to satisfy our needs.

When switching to the DARE dataset, we found that it did not have labels according to regional accent, so we hand-labelled the samples according to the 8 major dialects founds in Aschmann's map of regional American English dialects [11]. In the process of labelling this dataset, we made the conscious choice to split the Southern dialect into Inland and Lowland Southern dialects, as this led to a more even distribution of classes; since the DARE dataset did not include Canadian English, we maintained 8 classes of American English accents: North, New England, North Central, Midland, Inland South, Lowland South, New York City, and West. After this, a preprocessing script allowed us to select which feature(s) to use as labels. Then, we loaded the WAV files and converted them into graphs of 13 MFCCs. These graphs are then sliced into units of COL_SIZE length to produce more data for the model.

We did identify other suitable datasets, such as the IDEA dataset [12] and the Wildcat Corpus [13], though lack of sufficient metadata and/or funding prevented us from acquiring these.

4 Methods

We began our testing model on top of Chaudhry's Speech Accent Recognition GitHub repository [10]. As we are attempting to predict a class given several potential classes, we used categorical cross-entropy loss.

$$\mathcal{L} = -\sum_{c=1}^{Classes} y_{o,c} \log p_{o,c}$$

where $y_{o,c}$ is the prediction for class c at observation o and $p_{o,c}$ is the probability of class c at observation o.

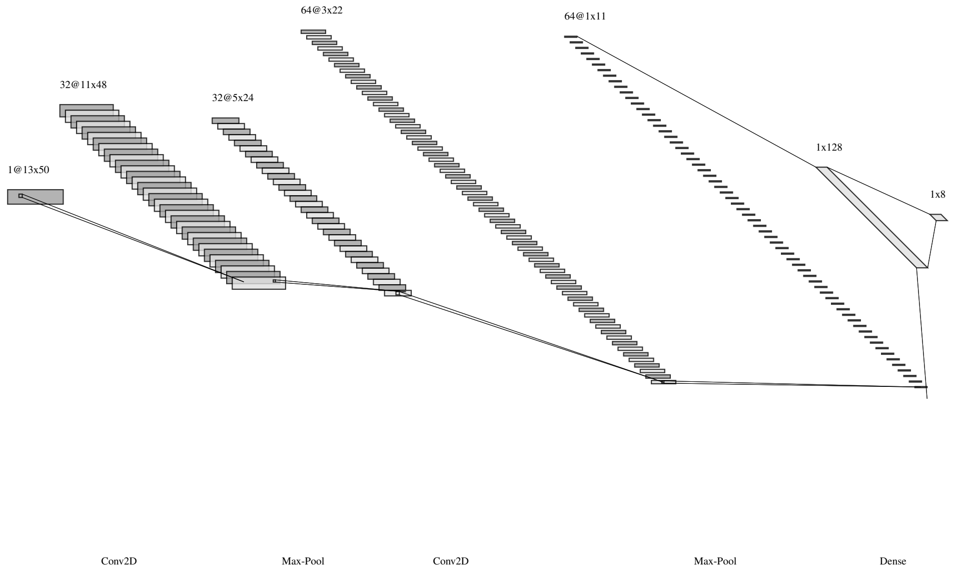


Figure 1: CNN Architecture

The optimal structure of our CNN after our experiments was two sets of 2D Convolutional layers with kernel size (3,3) and Max-Pooling layers with window size (2,2); the first Conv2D layer had 64 units while the second layer had 32 units. Following this, there is a 128-unit Dense layer and a 8-unit Dense layer for prediction. All layers used a relu activation except for the final layer, which used softmax for categorical classification. The number of units in the last layer is variable, as some experiments constrain the dataset and as such constrain the number of classes. We used Adadelta as the optimization function [14]. This model is implemented in Keras with a Tensorflow backend.

A major part of our methodology was making the size of each MFCC fed into the model a tunable hyperparameter, such that each individual MFCC was of size (13, COL_SIZE). This was achieved by taking each MFCC graph and cutting it into segments of the aforementioned size. This allowed for data augmentation and corrected for non-uniform data sizes by avoiding processing samples that had significant variation in length. This was done after datapoints were split into train and test sets.

At first, our model was trained with a 80%/20% train/test split, since SAA provided a limited amount of data. However, after switching to the DARE dataset, we changed the ratio to a 90%/10% train/test split, as the amount of data available increased by over an order of magnitude. This more generous split led to higher accuracy and better training as we had more samples to work with.

5 Results

Our two primary metrics are accuracy and F1 score. Initially, we were only concerned with the accuracy of the algorithm in classification. However, several of our experiments were able to achieve high accuracy by merely classifying everything as one category; this was because the data was distributed such that one class formed a majority of the labels. This is why we looked for a new, more evenly distributed dataset and further distributed the categories to avoid this bias.

F1 score is defined as the harmonic mean of the precision and recall. In our situation, we defined the micro F1 score as the sum of the F1 scores of each class, divided by the total number of classes.

$$Micro\ F1\ Score = \frac{\sum_{i \in Classes} \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i}}{Number\ of\ Classes}$$

$$Precision_i = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall_i = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

For our iterations, we chose to use a mini-batch size of 64, as we experimentally determined that was resulted in the highest performance metrics in a test of 100 samples from each of 5 classes.

Batch Size	32	50	64	256
Accuracy	46%	40%	58%	52%
Micro F1	45.15%	39.72%	56.60%	50.58%

Figure 2: Confusion Matrix (Test Set = 50 samples)

For each model, we tested it on a 10% test set and generated a confusion matrix between the class. An ideal confusion matrix would have all of the elements along the main diagonal. For this particular confusion matrix, we have an accuracy of 52% and a micro F1 score of 52.34%.

Predicted/Actual	West	Inland South	North	Lowland South	Midland
West	6	0	1	1	2
Inland South	0	5	0	3	2
North	0	1	5	0	4
Lowland South	1	0	2	3	4
Midland	1	1	1	0	7

Figure 3: Confusion matrix for $COL\_SIZE = 50$, on 100 samples from each of 5 classes.

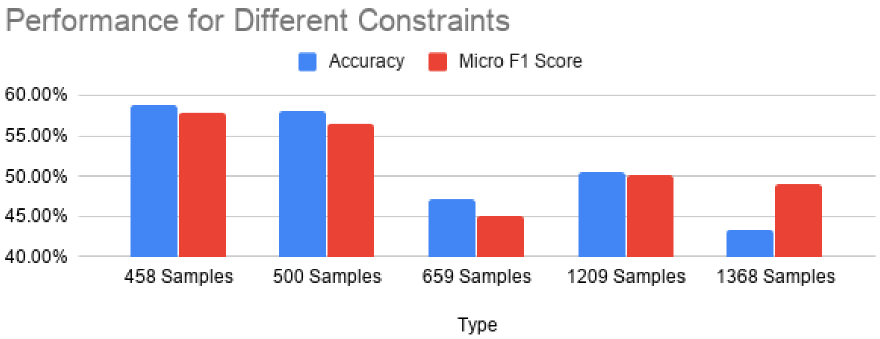


Figure 4: Graph of performance for the different constraints

We had give major groups of tests, in which we constrained the dataset to different degrees:

1. Restrict the dataset to 60 samples from each class. In this case, NYC only had 38 samples (458 samples)
2. Restrict the dataset to 100 samples each from North, Midland, Inland South, Lowland South, and West (500 samples)
3. Restrict the dataset to 100 samples from each class. In this case, NYC, New England, and North Central all had fewer than 100 samples. (659 samples)
4. Restrict the dataset to the top 5 classes (1209 samples)
5. Do not restrict the datasets (1368 samples)

6 Analysis

Overall, we found that models trained on datasets that were intentionally constrained so that classes had an equal number of data points performed, on average, 10%-15% better than models trained on unconstrained datasets. Interestingly, the 500 sample model could be considered the most "constrained", as it has exactly the same number of samples per class, yet the 458 sample model has 1% greater performance in both metrics. However, this could easily be due to randomness. Models with very large imbalances in classes could achieve high accuracies by simply classifying all samples as the majority class(es), but would have extremely low micro F1 scores.

In 3/5 cases, the models performed best with $COL\_SIZE = 50$. $COL\_SIZE = 30$ was best for the 458-sample case and $COL\_SIZE = 75$ was best for the 1209-sample case.

Additionally, analyzing the confusion matrices allowed us to see that

- 30-50% of Inland Southern dialect samples are classified as Midland or Lowland Southern
- Western dialect samples had the lowest average precision (i.e. most likely to be misclassified), generally as Midland or Lowland South

After looking at the Aschmann map, we see that Inland Southern for much of its extent in a narrow strip bordered by Midland and Lowland Southern dialects [11]. Thus, it would be easy to misclassify these accents due to their high degree of contact. For the West, we must realize that over 60% of the recorded speakers with West dialect are over 60 years old, so would have been 1st or 2nd generation settlers of the West, due to the Homestead Act of 1862 [15]. As such, they would for the large part still speak like people from their areas of ancestral origin; many immigrated from the Eastern seaboard or the deep South, which correlates to Midland and Lowland South dialects, respectively. As such, the "mistakes" in the confusion matrix actually tell us about the ethnolinguistic history of the area.

7 Conclusion/Future Work

Over the course of this project, we found that deep imbalances in underlying datasets and errors in data collection could significantly hamper the accuracy of trained models. By constraining our datasets such that class distributions were more balanced, we could achieve a 10-15% increase in both accuracy and micro F1 score. By splitting up the data handling into discrete, one-time preprocessing steps, we could rapidly iterate our experiments and learn more about our model trends.

Generally, a column size of 50 was best, which is likely due to this low column size leading to more data per class, allowing the model to train more thoroughly.

Even without perfect prediction accuracy, the confusion matrices uncovered similarities between regions that experienced large population shifts in the past (Lowland South to West, for example) which could prove very useful in studying underlying migration and cultural histories of an area.

Continuing onwards, we could further preprocess the data by removing periods of silence and augmenting the dataset by applying Gaussian noise to the background. Additionally, we could attempt to further split the regional dialects into subdialects in order to explore both how finely the model can identify local accents and how robust data constraints are to increased numbers of classes, as well as less data per class.

8 Contributions

Andea rewrote the webscraping and audio download code to work with the DARE dataset as the requisite information and html code was vastly different. Additionally, she fine tuned the SAA webscraping code and ran experiments on the SAA dataset to tune hyperparametes. She also worked writing the report and co-presented for the final video.

James worked on adjusting the Chaudhry repository model to work with the updated SAA and DARE datasets. He also researched different available datasets and existing accent recognition research, as well as did a large part of the model tests and writing of the report.

John set up and managed the group's AWS, colab, and github Deep Learning environments and ran experiments tuning hyperparameters and model structure to arrive at the final model. He also assisted with manual labelling and web scrape cleaning, report writing, some research to inform experiments, and the final video.

References

[1] Chu, A., Lai, P. & Le, Diana. (n.d.) Accent Classification of Non-Native English Speakers.

[2] Wu, Y., Li, S. & Li, H. (2019) Automatic Pitch Accent Detection Using Long Short-Term Memory Neural Networks. *Proceedings of the 2019 International Symposium on Signal Processing Systems*. New York, NY: Association for Computing Machinery.

[3] Jiao, Y., Tu, M., Berisha, V. & Li, J. (2016) Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features. Interspeech.

[4] Zhong, J., Zhang, P. & Li, X. (2018) Adaptive recognition of different accents conversations based on convolutional neural network. Multimedia Tools and Applications.

[5] Shang, L.M.A. & Xiong, E.M.W. (2017) Deep Learning Approach to Accent Classification. Stanford University.

[6] Chionh, K., Song, R. & Yin, Y. (n.d.) Application of Convolutional Neural Networks in Accent Identification. Carnegie Mellon University.

[7] Speech Accent Archive. George Mason University. https://accent.gmu.edu/

[8] Dictionary of American Regional English. University of Wisconsin - Madison. https://uwdc.library.wisc.edu/collections/amerlangs/

[9] Richardson, L. Beautiful Soup. https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[10] Chaudhry, A. Speech Accent Recognition. https://github.com/akshanshchaudhry/Speech-Accent-Recognition

[11] Aschmann, R. (2018) North American English Dialects, Based on Pronunciation Patterns. https://aschmann.net/AmEng/

[12] International Dialects of English Archive. (2020) https://www.dialectsarchive.com/

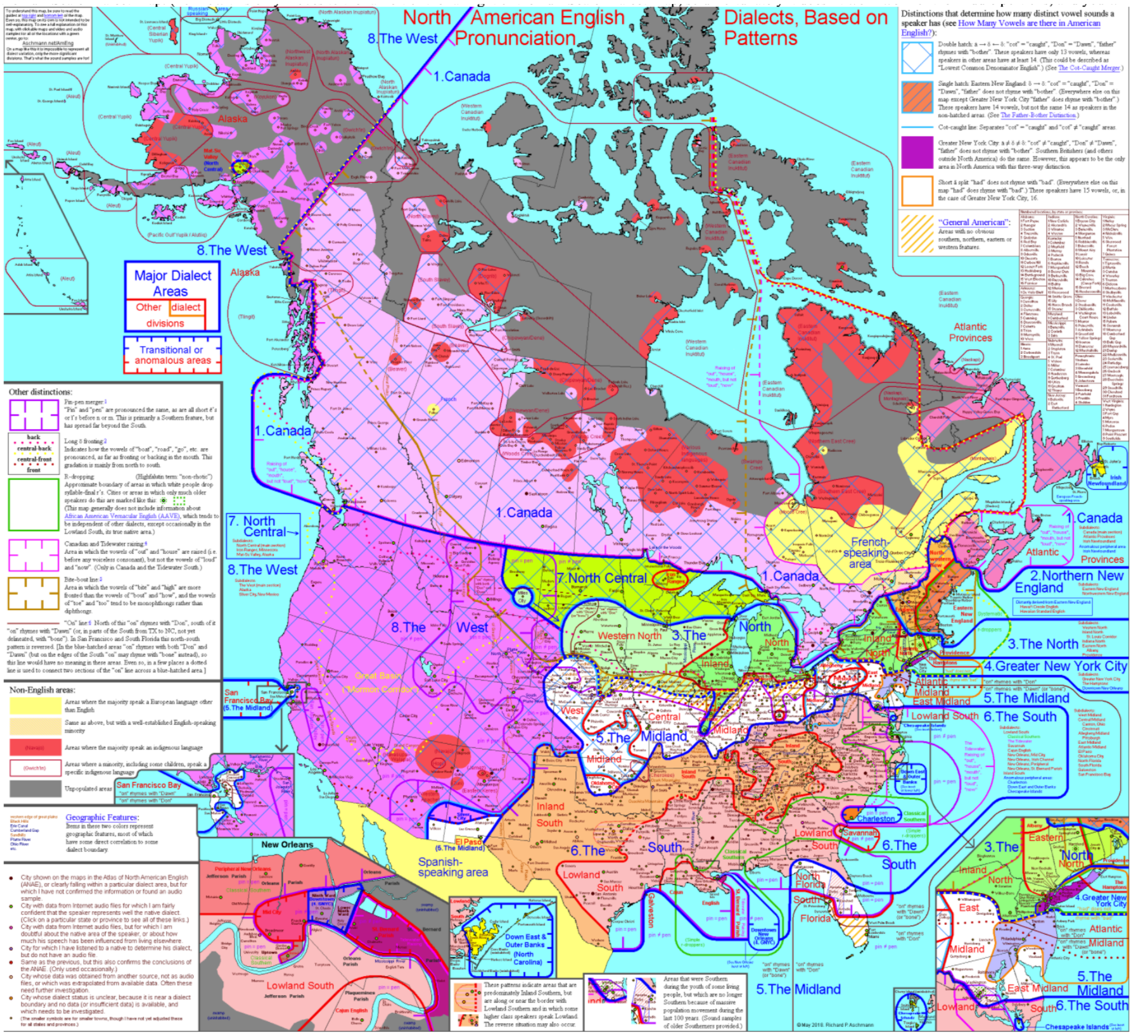
[13] Bradlow, A. Wildcat Corpus of Native- and Foreign-Accented English. Northwestern University. http://groups.linguistics.northwestern.edu/speech_comm_group/wildcat/content.html

[14] Zeiler, M. (2012) ADADELTA: An Adaptive Learning Rate Method. Cornell University. https://arxiv.org/abs/1212.5701

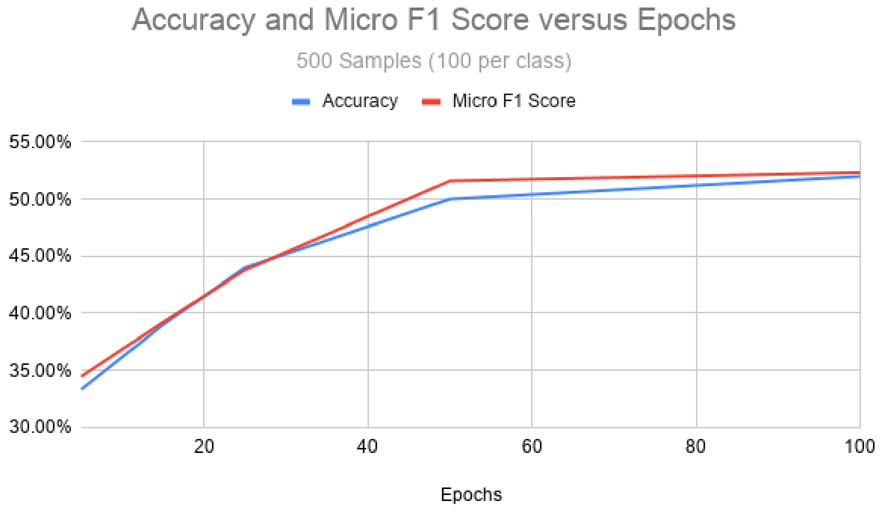
[15] Pessen, E., et al. *Encyclopaedia Brittanica.* "United States." https://www.britannica.com/place/United-States

Appendix

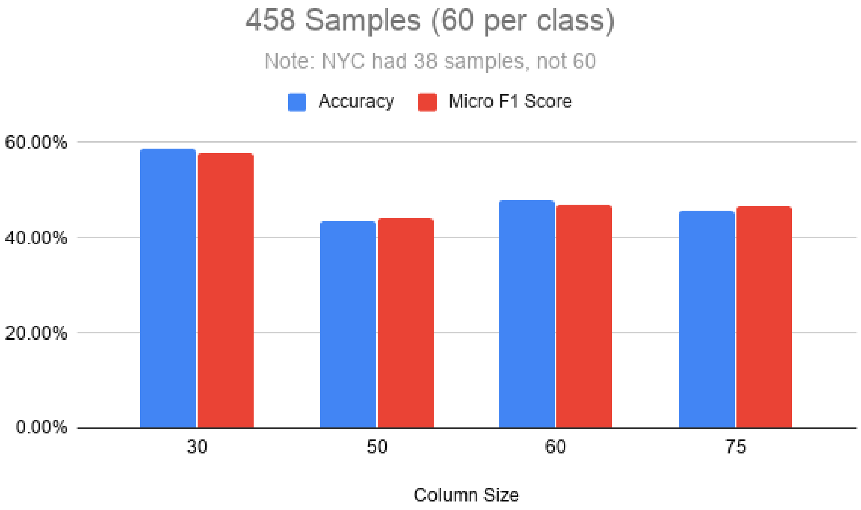
A Map of American Regional Accents

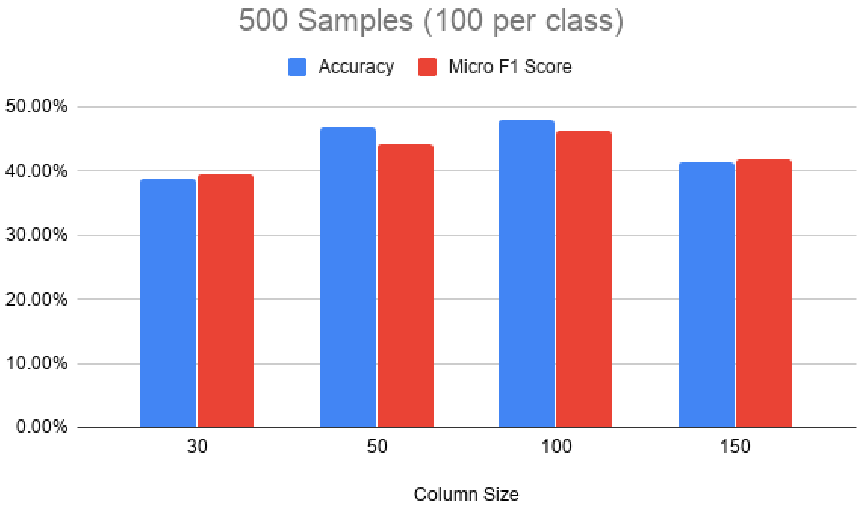


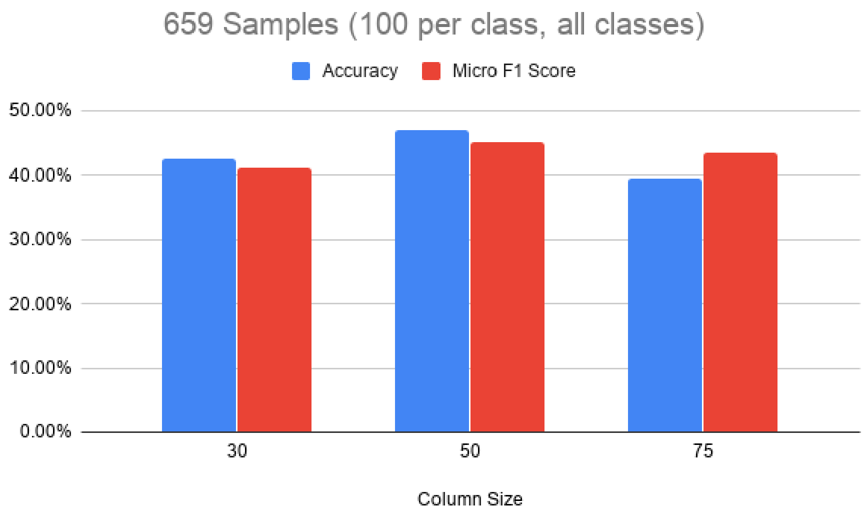
B Accuracy and Micro F1 Score versus Epochs

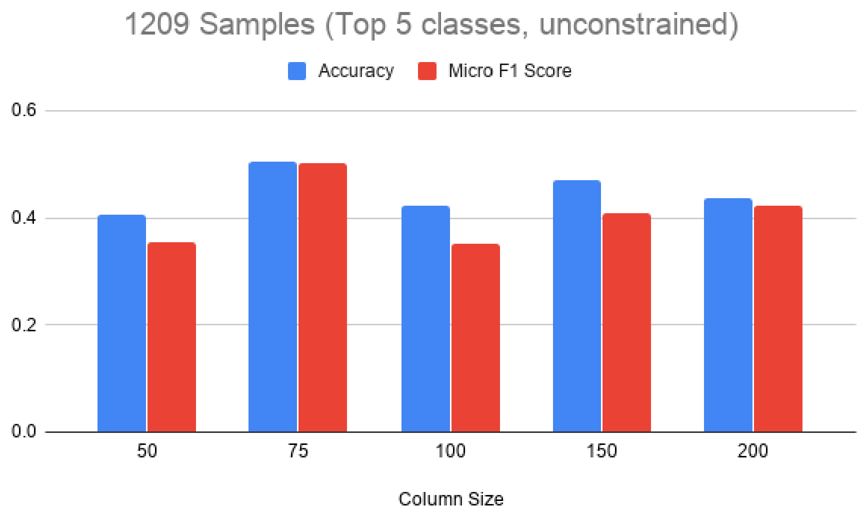


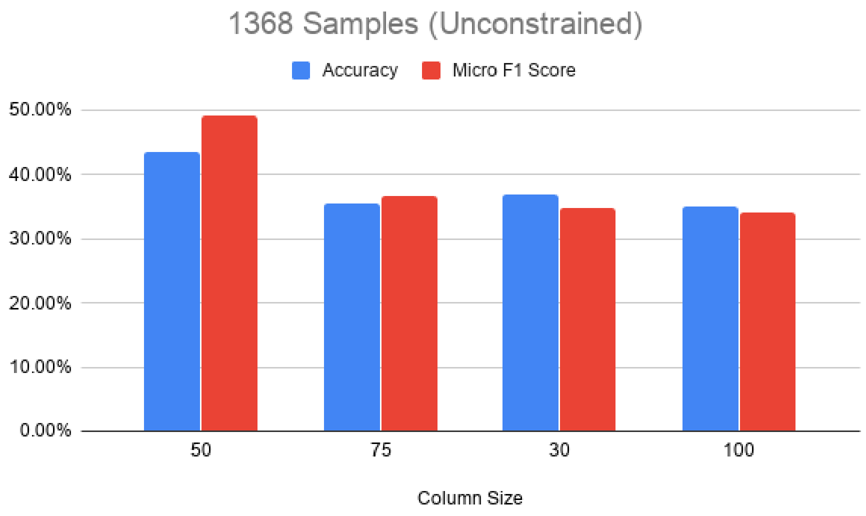
C Performance Graphs for Different Constraint Models






659 Samples (100 per class, all classes)
🟦 Accuracy
🟥 Micro F1 Score
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
30
50
75
Column Size


1209 Samples (Top 5 classes, unconstrained)
🟦 Accuracy
🟥 Micro F1 Score
0.6
0.4
0.2
0.0
50
75
100
150
200
Column Size


1368 Samples (Unconstrained)
🟦 Accuracy
🟥 Micro F1 Score
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
50
75
30
100
Column Size