Multiclass Tweet Classification

Dominnic Chant, Isaac Lo

Project Goal

- Predict domain/category of a tweet given its text
  - 1M tweets (80-20 split)
- Architectures:
  - RNN
  - LSTM
- Hyperparameters:
  - Learning Rate
  - Epochs
  - Word Embedding Size, Hidden Layer Size
  - Number of linear layers, LSTM layers
- NLTK TweetTokenizer

3 - TV Shows
4 - TV Episodes
6 - Sports Events
10 - Person
11 - Sport
12 - Sports Team
26 - Sports League
27- American Football Game
28 - NFL Football Game
35 - Politicians
38 - Political Race
39 - Basketball Game
40 - Sports Series
45 - Brand Vertical
46 - Brand Category
47 - Brand
48 - Product
49 - Product Version
54 - Musician
55 - 56 - Actor
58 - Entertainment Personality
60 - Athlete
65 - Interests and Hobbies Vertical
66 - Interests and Hobbies Category
67 - Interests and Hobbies
68 - Hockey Game
71 - Video Game
78 - Video Game Publisher
79 - Video Game Hardware
84 - BookMusic Genre
85 - Book Genre
86 - Movie
87 - Movie Genre
88 - Political Body
89 - Music Album
90 - Radio Station
91 - Podcast
92 - Sports Personality
93 - Coach
94 - Journalist
110 - Viral Accounts
114 - Concert
115 - Video Game Conference
116 - Video Game Tournament
117 - Movie Festival
118 - Award Show
119 - Holiday
120 - Digital Creator
122 - Fictional Character
130 - Multimedia Franchise
132 - Song
136 - Video Game Personality
137 - eSports Team
138 - eSports Player
139 - Fan Community

RNN

- Fixed hyper-parameters
  - Number of epochs = 5
  - ReLU layer
  - Number of linear layers = 2
- Variables hyper-parameters
  - Learning Rate [0.03, 0.01, 0.001, 0.0001]
  - Word embedding size [50, 100, 150]
  - Hidden layer size [50, 100, 150]

RNN - Results

Learning Rate	Emb Size	Hidden Size	Accuracy
0.03	150	100	0.29
0.01	50	50	0.24
0.01	**150**	**100**	**0.32**
0.001	50	50	0.24
0.001	150	100	0.30

LSTM

- Fixed hyper-parameters
  - Dropout = 0.3
  - Learning rate = 0.01
  - ReLU
- Variables hyper-parameters
  - Word embedding size
  - Hidden layer size
  - Number of LSTM layers
  - Number of linear layers

LSTM

Emb Size	Hidden Size	LSTM Layers	Linear Layers	Accuracy
100	50	1	2	0.73
100	50	2	2	0.72
200	**100**	**2**	**2**	**0.74**
200	100	1	4	0.52
200	100	2	4	0.55

Project Outcome

- Model performance was decent
- Training time was very long (3+ min per epoch)
- Best Model: LSTM (0.74 accuracy)

Emb Size	Hidden Size	LSTM Layers	Linear Layers	Accuracy
200	**100**	**2**	**2**	**0.74**