Don’t Trust The Data

Classification of Spotify Podcast Transcripts and Metadata

Team 9: Chris Cappiello, Keith McBride, Andrés Medina, Bryan Smith, Joseph Smith, Jessica Wheeler

The Problem and Goals

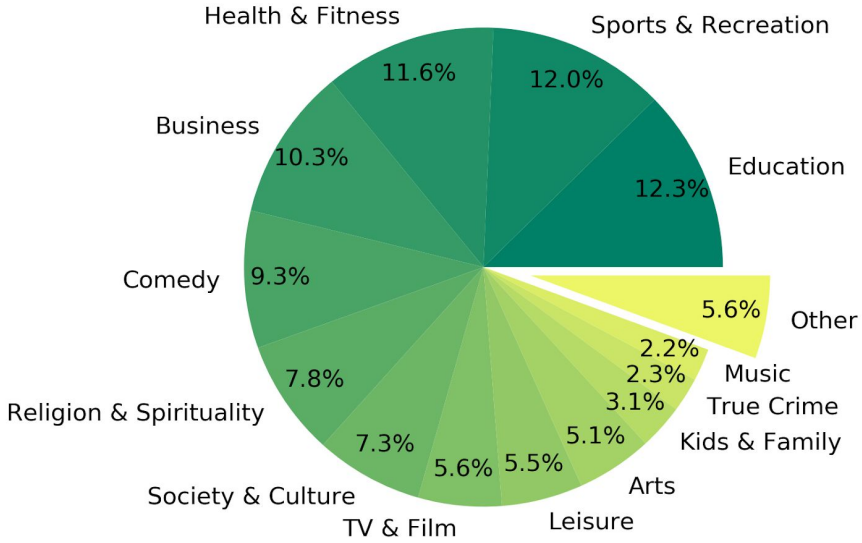
Facts: There are more than **800,000 active podcasts** [Forbes 2019] and more than **50%** of consumers in the US listen to podcasts [Statistica, 2019]

Problem: Content creators need to find their audience and connect with their listeners

Goal: Use podcast transcripts and metadata to classify podcasts based on their categories (e.g. True Crime vs. Kids & Family)

Data Cleaning

- Cleaned and vectorized **88 GB** of transcript data from over **100,000 podcasts** on Spotify (Bryan, Chris) [Clifton, 2020 arXiv:2004.04270]
  - Natural Language Toolkit (NLTK)
- Extracted show information from iTunes XML files (Andrés, Keith)
  - XML parsing, Spotify api (Spotipy)
- Performed cleaning and analysis on metadata (Jessica, Andrés, and Joey)
  - Emoji cleaning, pattern recognition



Modeling

- Preprocessing
  - Hash vectorizer (Bryan)
  - PCA (Joey)
  - Lexical Complexity (Chris)
- sklearn classification algorithms:
  - SVC/LinearSVC (Bryan)
  - MultiNB/GaussNB (Keith, Bryan)
  - Random Forest (Jessica/Keith)
  - Decision Trees (Jessica/Keith)
  - SGDC (Bryan, Keith)
  - KNN (Joey)
- Keras/TensorFlow
  - Neural Networks (Simple and LSTM) (Andrés)

Features
Duration
Explicit
Episode description lexical complexity
Transcript lexical complexity
Episode description contains emojis
Professional
Show frequency
Hash vectorized transcripts

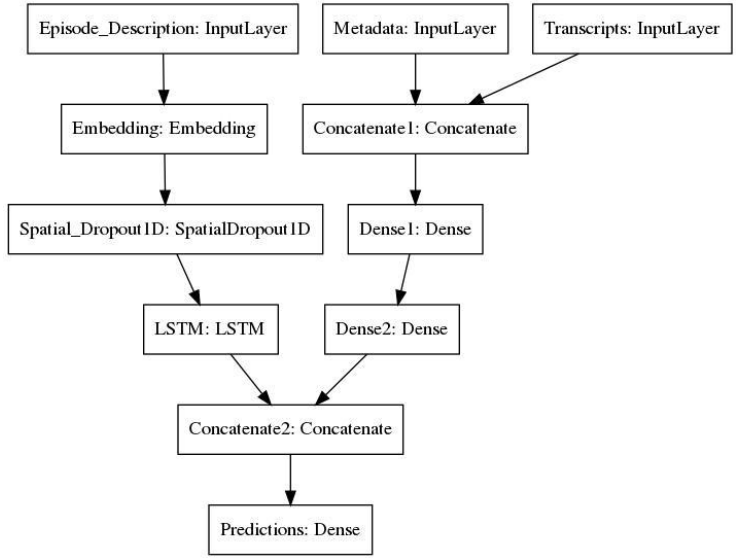
Accuracy from models in sklearn:

MultiNB	Dec.Tree	GaussNB	Forest
20%	42%	44%	60%
SGDC	**KNN**	**LinSVC**	**SVC**
62%	66%	67%	70%

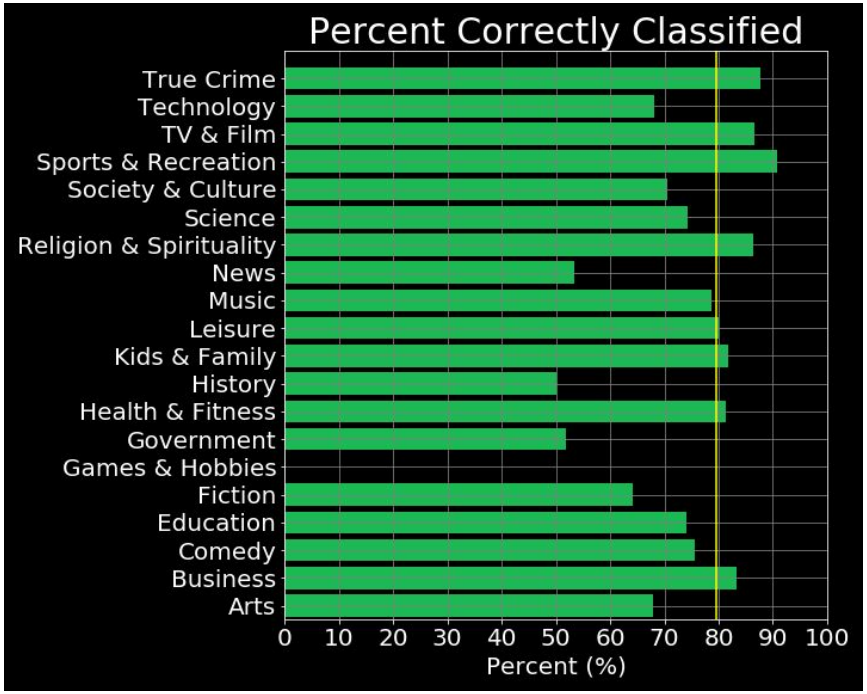
Results

Our models in sklearn achieved ~40-70% accuracy

By merging two neural networks (LSTM+DNN) we achieved **~80% category classification accuracy**

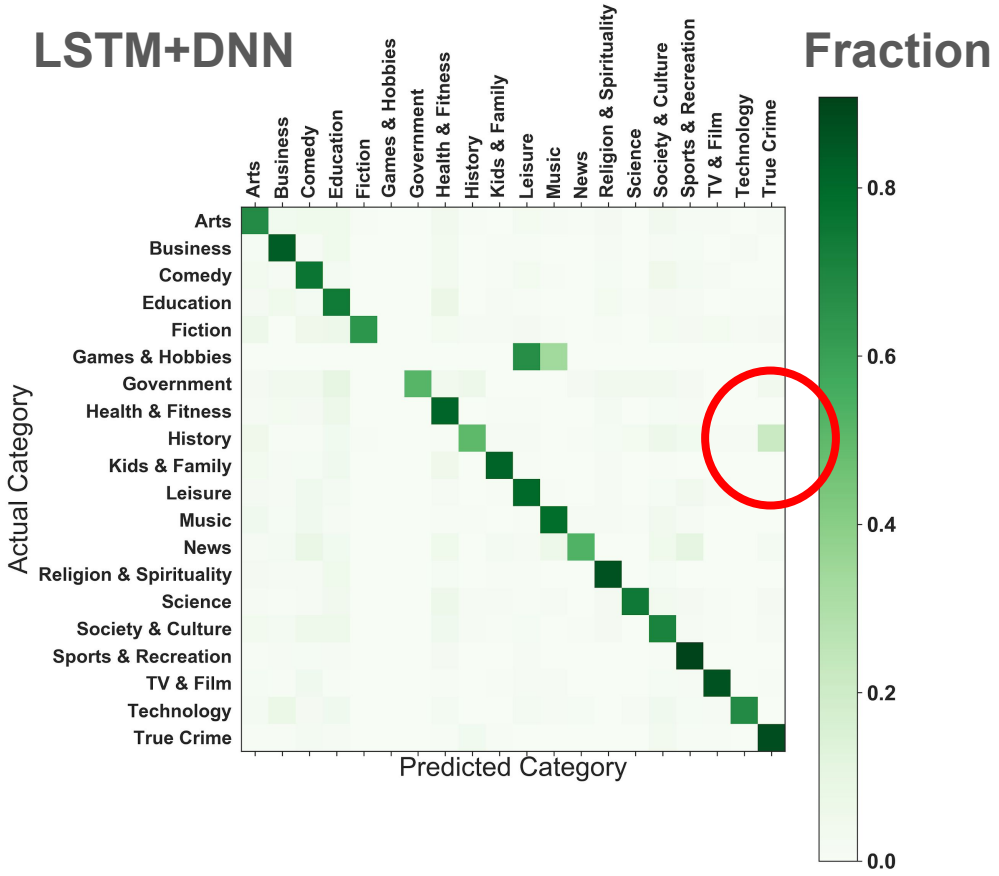


LSTM+DNN Results



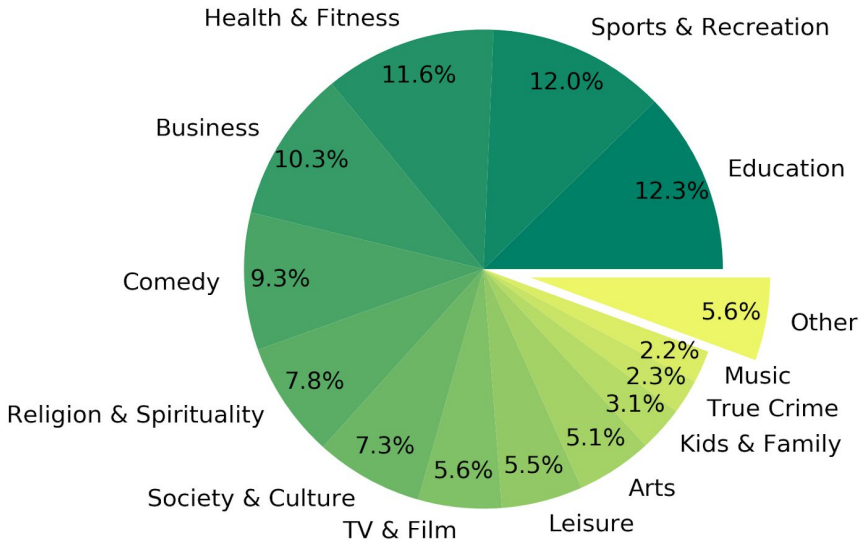
Challenges

- Classifying “Games & Hobbies” due to small number of episodes
- Episodes sometimes misclassified as similar categories
- More computational time and resources for better model
- Actually getting the data



Next Steps

- Reduce or join similar categories
  - Additional data cleaning
- Test best performing model with larger hash vector size
- Explore ensemble methods for improved performance
- Unsupervised learning



Thank You!

TEAM 9: Don't Trust The Data

Chris Cappiello
cappiello.7@osu.edu
The Ohio State University

Bryan Smith
smith.10851@osu.edu
The Ohio State University

Keith McBride
mcbride.342@osu.edu
The Ohio State University

Joseph Smith
smith.10838@osu.edu
The Ohio State University

Andrés Medina
medina.101@osu.edu
The Ohio State University

Jessica Wheeler
wheeler.1011@osu.edu
The Ohio State University

*https://github.com/TheEssentials/TheEssentials

* Access on request