WHAT MAKES TED TALKS WORTH SPREADING?

Exploration of the linguistic and text features of TED video transcripts

Ozge Yasar

GOALS

- Predict the number of views for each video.
- Build a classifier in order to predict the gender of the presenters.
- Create a classification model to accurately classify the multiple topic tags of each video from text (bag of words).

DATASET

Shared on data.world - Words of Persuasion: Text Predictors of Persuasive TED Talks*

2406 TED Video Transcripts & Other Video Features

Total 187 Features including:

Linguistic Inquiry Word Count (LIWC)

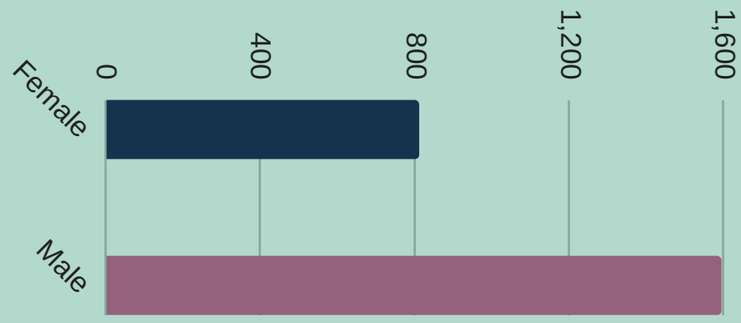
Morality Word Counts

VIDEO TOPICS

423 Unique tags in the original dataset

Reduced to 22 for multi label classification

GENDER VARIABLE



gender-guesser library

genderize.io

manually looking up the names

PREDICTION MODELS

VIEW PREDICTION

Best Model

R-Squared

0.55

RMSE

0.68

GENDER PREDICTION

Best Model: Logistic Regression

Baseline Accuracy

0.67

Model Accuracy

0.72

	Precision	Recall
Female	0.61	0.39
Male	0.75	0.88

TOPIC CLASSIFICATION

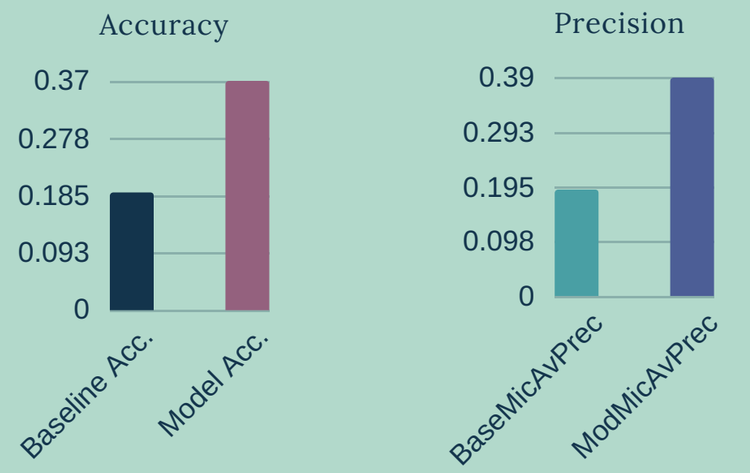
TextBlob

Lemmatized

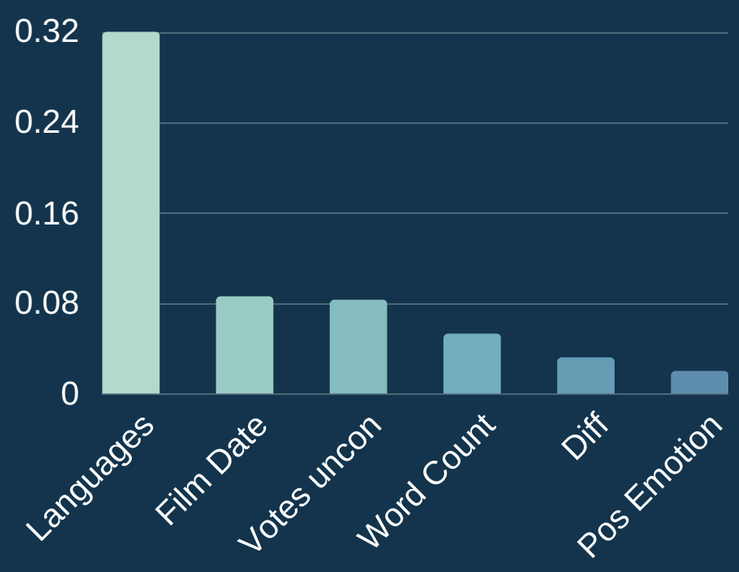
CountVectorizer

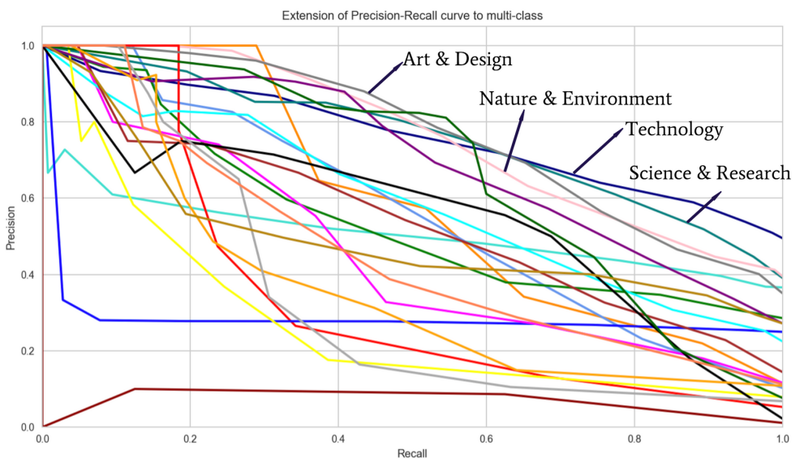
Tf-idf transformer

OneVSRestClassifier(KNN)



FEATURE IMPORTANCES



ROC CURVES

PRECISION - RECALL CURVES