What makes TED talks worth spreading?

Exploration of the linguistic and text features of TED video transcripts

Ozge Yasar

Goals of the project

- Using linguistic & non-linguistic features such as film date, inspiring and persuasive votes etc. to:
  - Build a classifier in order to predict the gender of the presenters
  - Predict the number of views for each video
- Create a classification model to accurately classify the multiple topic tags of each video from text (bag of words)

Data Set

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

Initial Size:

- 2406 TED Talk scripts scraped from TED website.
- 187 features

Video features included:

- Number of views, comments, votes (e.i., Persuasive, Unconvincing),
- Video transcripts
- Tags, occupations of the speakers

*https://data.world/owentemple/text-and-content-features-of-most-persuasive-ted-talks/workspace/file?filename=all_with_liwc_segmented.xls

Linguistic Inquiry Word Count (LIWC**) features:

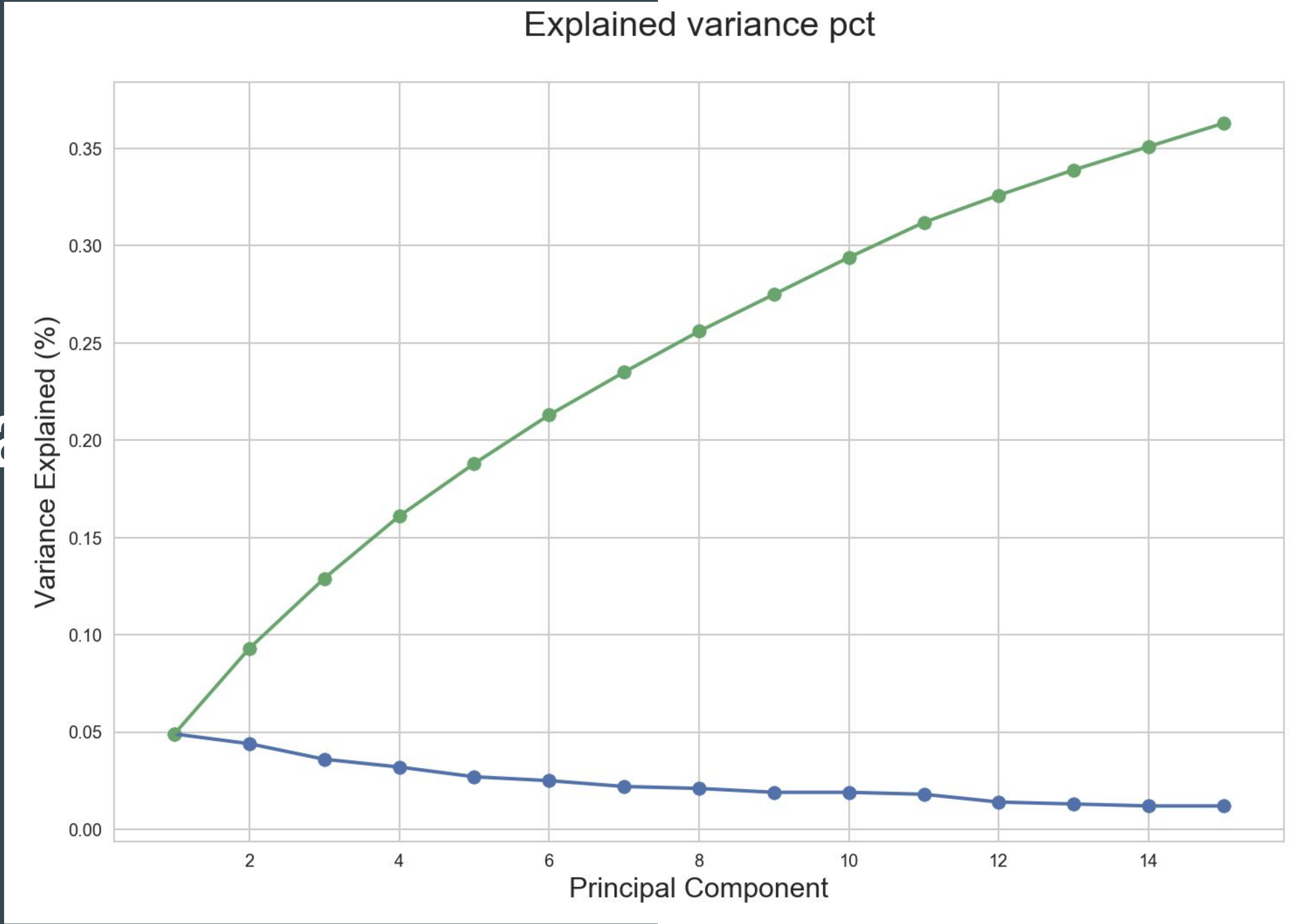
- Linguistic Dimensions (use of pronouns, punctuation, etc.)
- Other Grammar (verbs, adjectives, etc.)
- Psychological Processes (Affective processes, Cognitive processes, Time orientations , etc.)

Morality features***

- Harm. Fairness, Purity & sub categories

**https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf
***http://www.moralfoundations.org/sites/default/files/files/downloads/moral%20foundations%20dictionary.dic

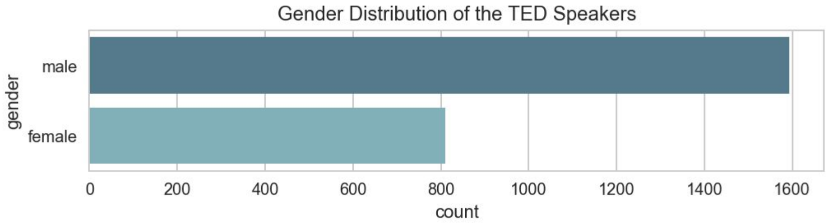
Fea

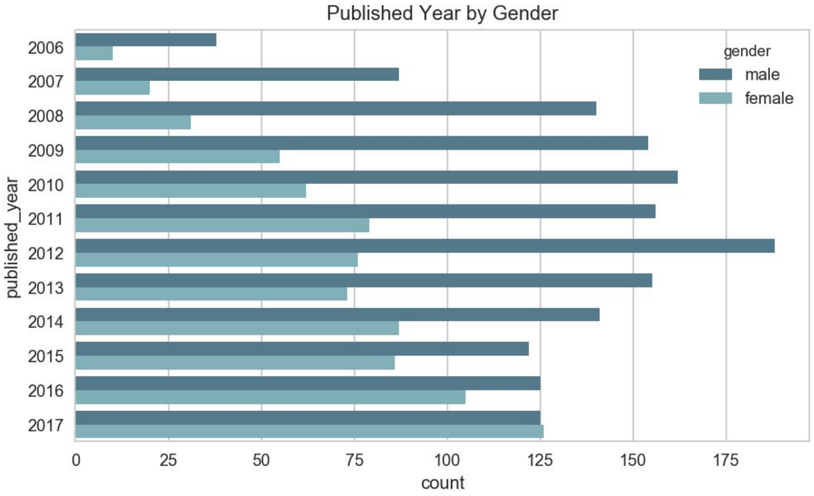


Explained variance pct
Variance Explained (%)
Principal Component
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2
4
6
8
10
12
14

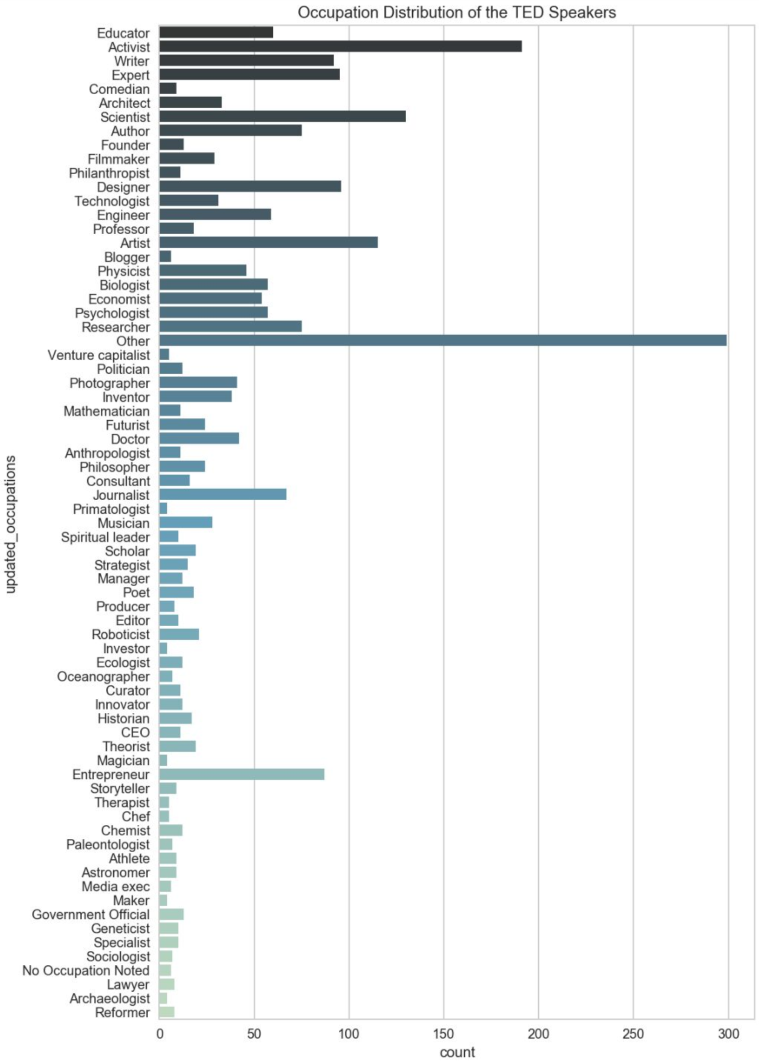
Feature Distributions

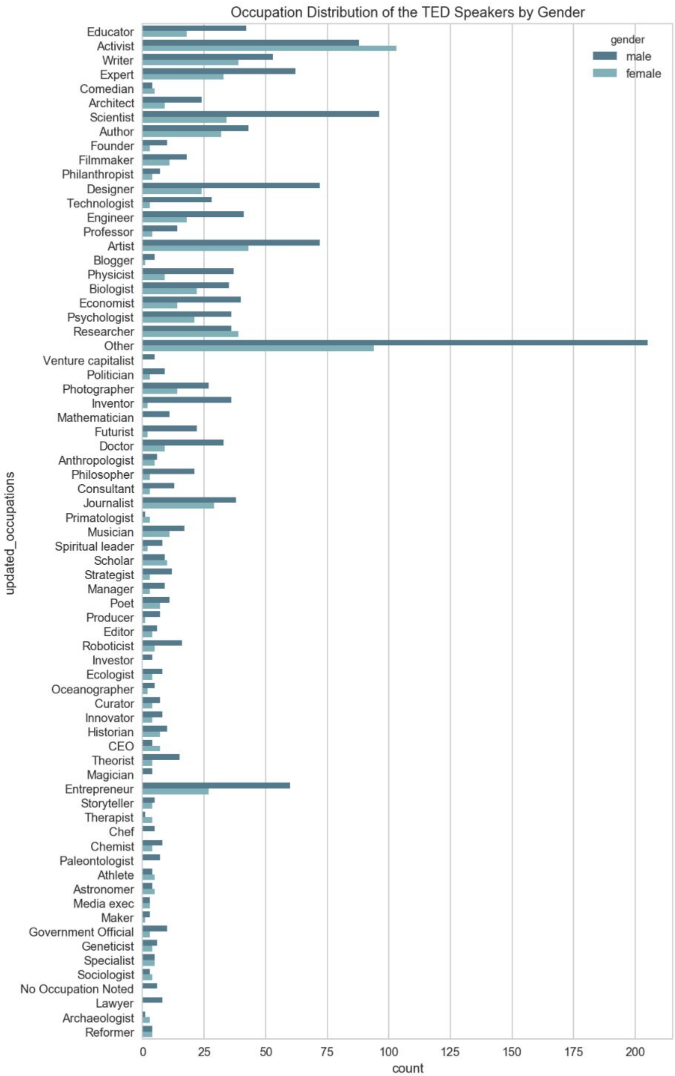
Gender Distributions



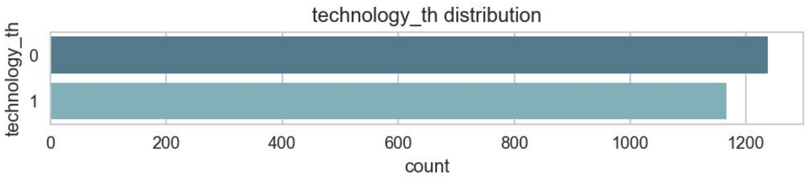


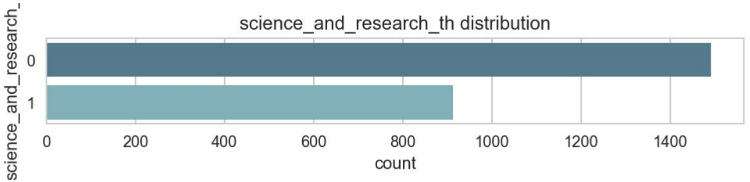
0

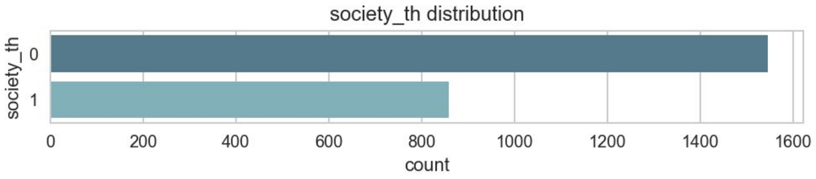

Occupation Distribution of the TED Speakers
updated_occupations
count
Educator
Activist
Writer
Expert
Comedian
Architect
Scientist
Author
Founder
Filmmaker
Philanthropist
Designer
Technologist
Engineer
Professor
Artist
Blogger
Physicist
Biologist
Economist
Psychologist
Researcher
Other
Venture capitalist
Politician
Photographer
Inventor
Mathematician
Futurist
Doctor
Anthropologist
Philosopher
Consultant
Journalist
Primatologist
Musician
Spiritual leader
Scholar
Strategist
Manager
Poet
Producer
Editor
Roboticist
Investor
Ecologist
Oceanographer
Curator
Innovator
Historian
CEO
Theorist
Magician
Entrepreneur
Storyteller
Therapist
Chef
Chemist
Paleontologist
Athlete
Astronomer
Media exec
Maker
Government Official
Geneticist
Specialist
Sociologist
No Occupation Noted
Lawyer
Archaeologist
Reformer
0
50
100
150
200
250
300

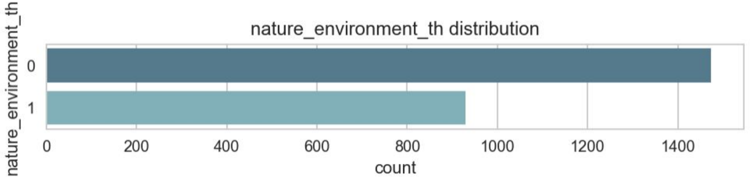

Occupation Distribution of the TED Speakers by Gender
gender
male
female
updated_occupations
count
Educator
Activist
Writer
Expert
Comedian
Architect
Scientist
Author
Founder
Filmmaker
Philanthropist
Designer
Technologist
Engineer
Professor
Artist
Blogger
Physicist
Biologist
Economist
Psychologist
Researcher
Other
Venture capitalist
Politician
Photographer
Inventor
Mathematician
Futurist
Doctor
Anthropologist
Philosopher
Consultant
Journalist
Primatologist
Musician
Spiritual leader
Scholar
Strategist
Manager
Poet
Producer
Editor
Roboticist
Investor
Ecologist
Oceanographer
Curator
Innovator
Historian
CEO
Theorist
Magician
Entrepreneur
Storyteller
Therapist
Chef
Chemist
Paleontologist
Athlete
Astronomer
Media exec
Maker
Government Official
Geneticist
Specialist
Sociologist
No Occupation Noted
Lawyer
Archaeologist
Reformer
0
25
50
75
100
125
150
175
200

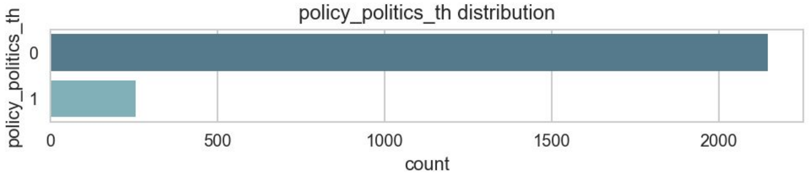
Video Tags - I

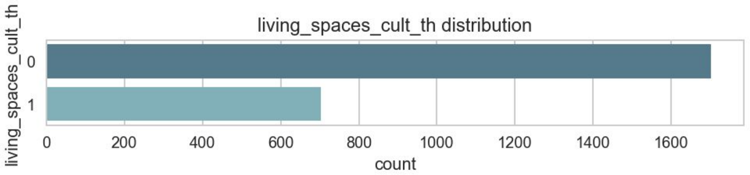

technology_th distribution
technology_th
0
1
0 200 400 600 800 1000 1200
count

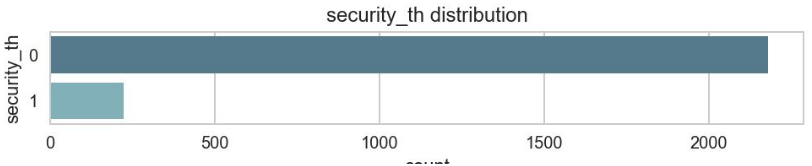

science_and_research_th distribution
science_and_research_
0
1
0 200 400 600 800 1000 1200 1400
count

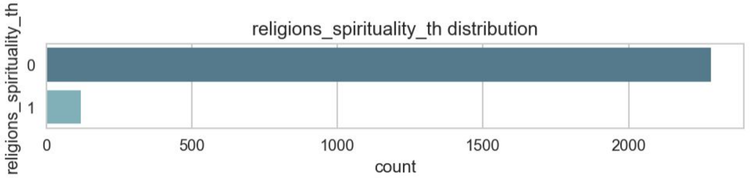

society_th distribution
society_th
0
1
0 200 400 600 800 1000 1200 1400 1600
count


nature_environment_th distribution
nature_environment_th
0
1
0 200 400 600 800 1000 1200 1400
count

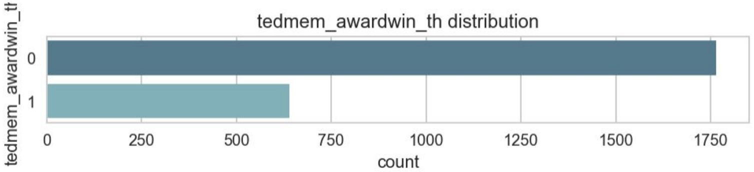

policy_politics_th distribution
policy_politics_th
0
1
0 500 1000 1500 2000
count

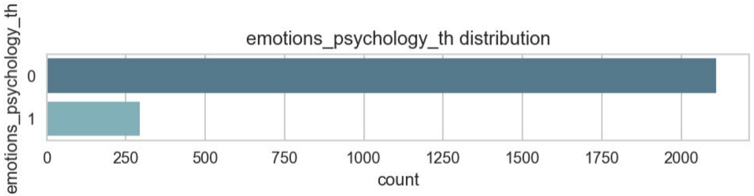

living_spaces_cult_th distribution
living_spaces_cult_th
0
1
0 200 400 600 800 1000 1200 1400 1600
count

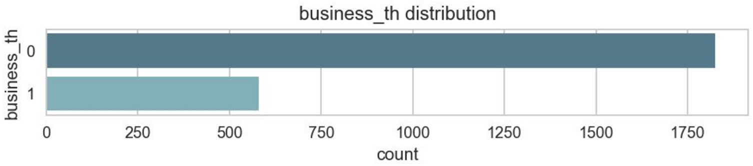

security_th distribution
security_th
0
1
0 500 1000 1500 2000

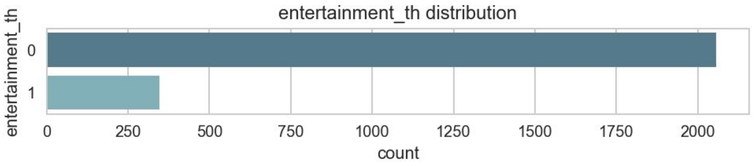

religions_spirituality_th distribution
religions_spirituality_th
0
1
0 500 1000 1500 2000
count

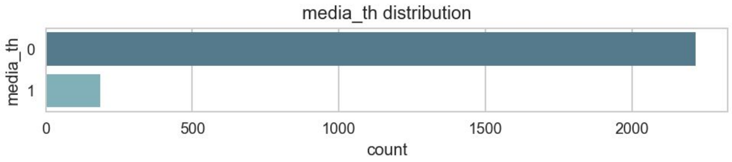
Video Tags - II

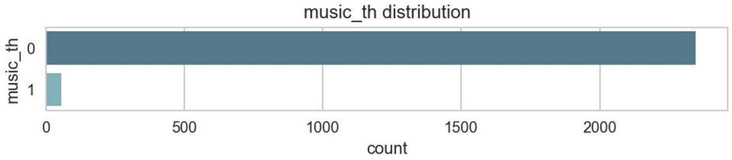


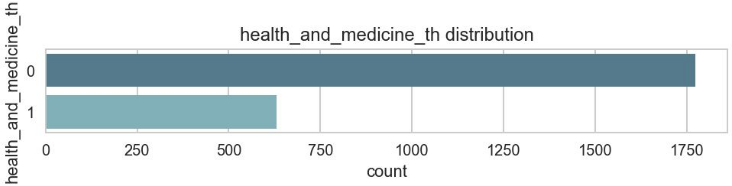


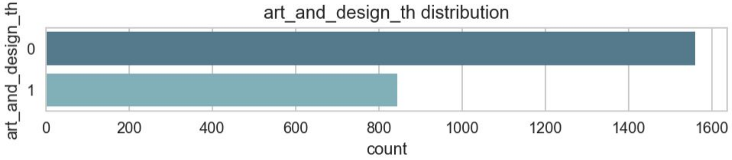




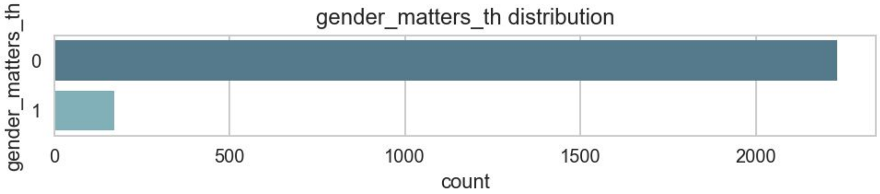


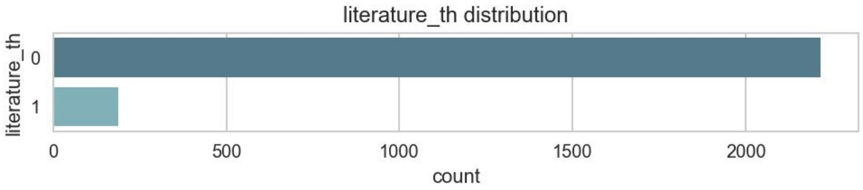


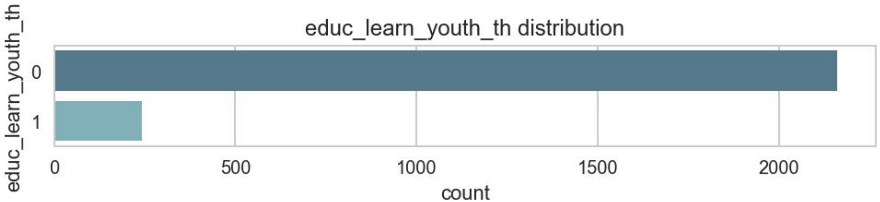


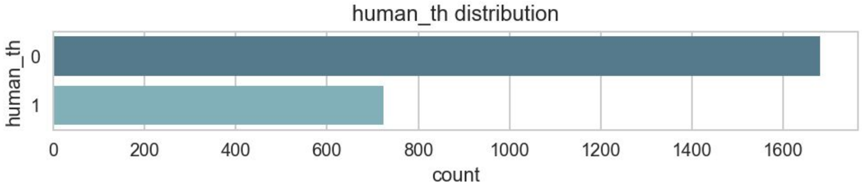


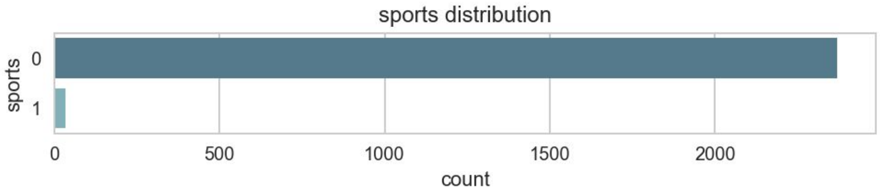
Video Tags - III

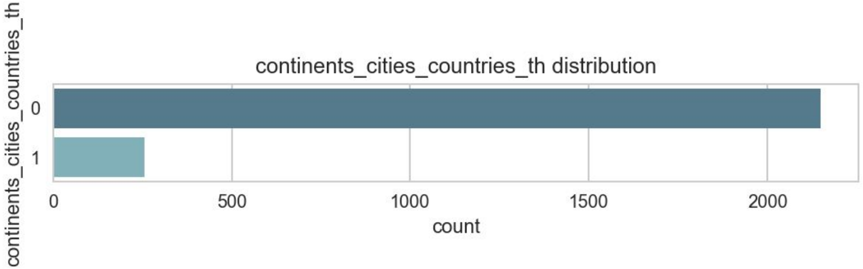




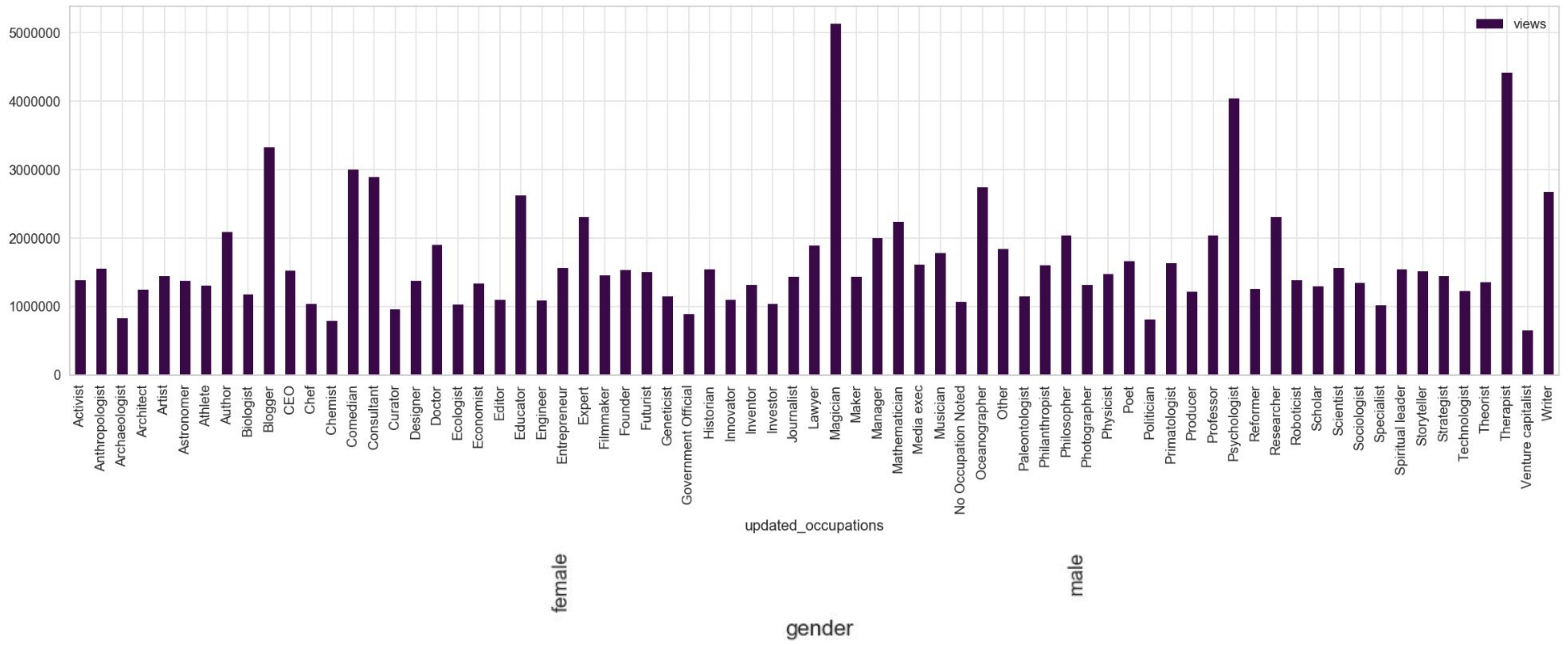








Distribution of Views



Cleaning Process

Highly correlated continuous variables were removed. For example:

- **Fairness (general)**
- Fairness virtue, Fairness Vice

Log transformed outcome variable: **views**

Outliers were removed

Outside +/- 10 Std range

Models

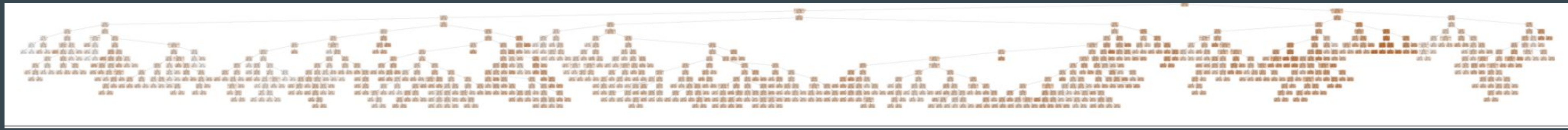
View Prediction

Predicting Number of Views

Models used:

- $\boxed{\text{Random Forest Regressor}}$
- LinearSVR
- Linear Regression
- $\boxed{\text{LassoCV}}$
- Ridge CV
- MLPRegressor

Importance	Variable
0.32	Languages
0.086	Film date
0.083	Normalized unconvincing votes
0.053	Word count
0.032	Words that indicates differentiation
0.02	Number of positive emotion words



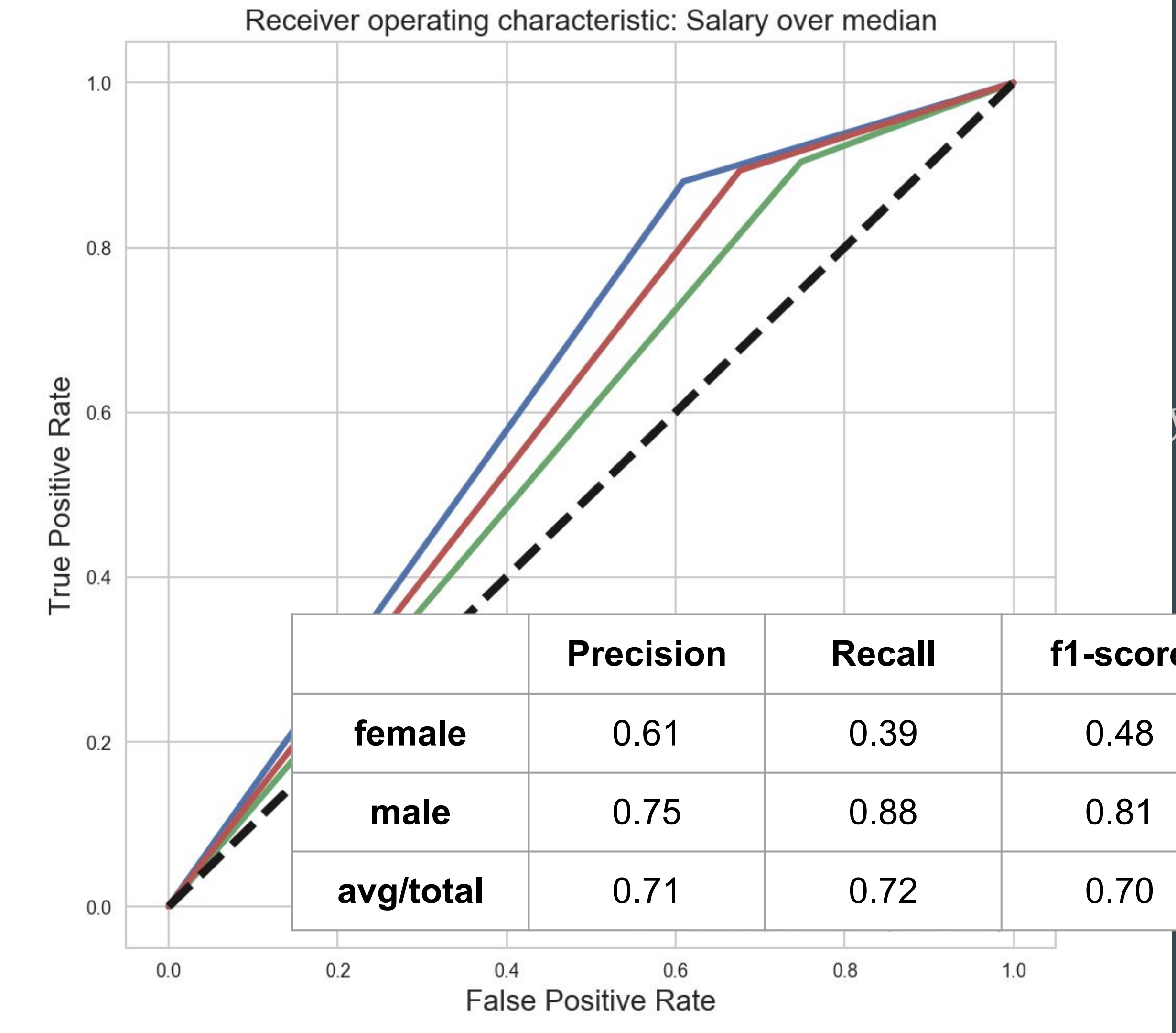
Gender Prediction

Predicting

Models Used:

- Randon
- KNN C
- Logistic

y = 0.72



	Precision	Recall	f1-score	Support
female	0.61	0.39	0.48	179
male	0.75	0.88	0.81	375
avg/total	0.71	0.72	0.70	554

Topic Classification

TED Words

see
one
think
say
take
want
world
good
lot
really
give
make
go
year
thing
time
know
little
kind
find
back
right
call
way
now
people
look
well
use
come
work
talk
show
new
actually
first
start
two
try
three
live
still
need
much
life
mean
thank
will
something
happen
even
around
story
tell
day
problem
change
child
many
country
every
percent
different
great
ask
number
woman
part
put
sort
love
ago
let
place
technology
question
system

Text Processing

- TextBlob
- Words Lemmatized
- CountVectorizer
- Tf-idf transformer

Multilabel Classification

Models used:

- OneVSRestClassifier(KNN)

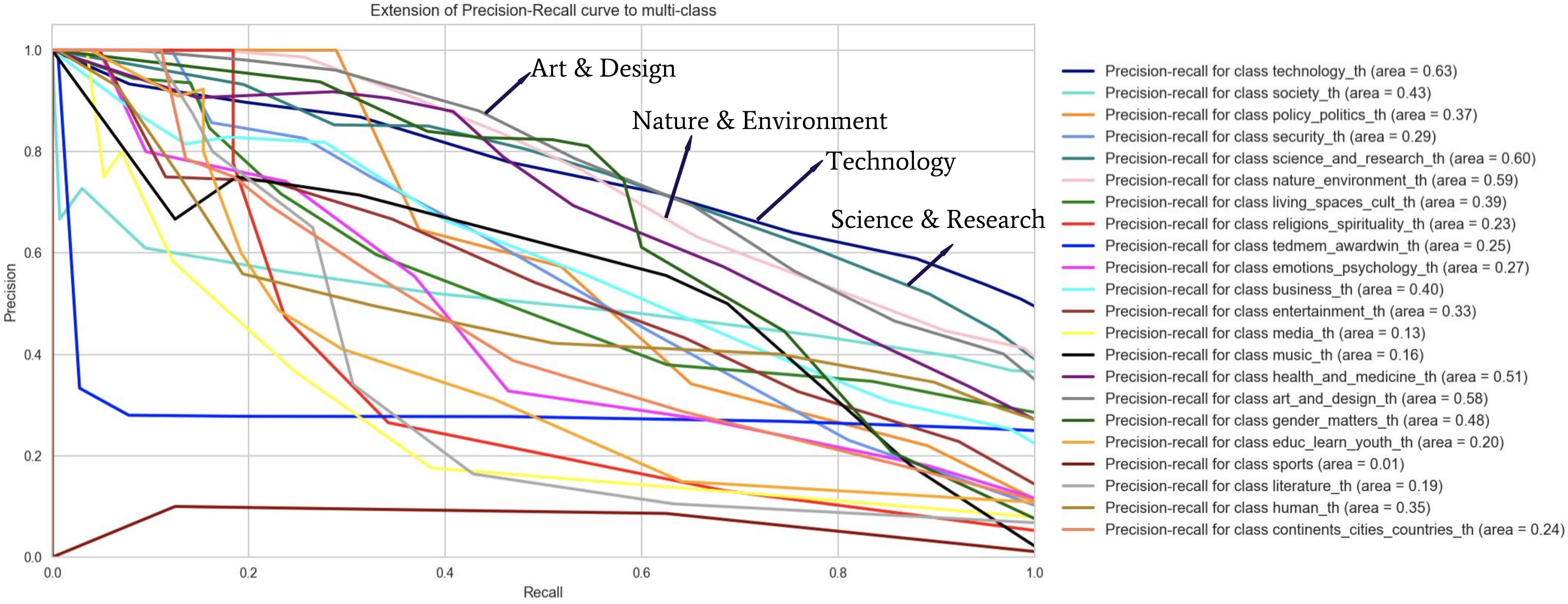
- Jaccard_similarity_score (accuracy) = 0.37
- Baseline (mean of y_train matrix) = 0.19

- Model Micro-Average Precision Score = 0.39
- Baseline Micro-Average Precision Score = 0.19

Individual Label Accuracy

	Base Line	Model
art_and_design_th	0.351367	0.785021
business_th	0.247919	0.822469
continents_cities_countries_th	0.103448	0.897365
educ_learn_youth_th	0.097503	0.898752
emotions_psychology_th	0.124257	0.901526
entertainment_th	0.144471	0.886269
gender_matters_th	0.068966	0.955617
health_and_medicine_th	0.258026	0.819695
human_th	0.313317	0.719834
literature_th	0.082640	0.938974
living_spaces_cult_th	0.294887	0.744799
media_th	0.077289	0.925104
music_th	0.024376	0.980583
nature_environment_th	0.383472	0.753121
policy_politics_th	0.102854	0.918169
religions_spirituality_th	0.048751	0.955617
science_and_research_th	0.374554	0.757282
security_th	0.088585	0.918169
society_th	0.353151	0.643551
sports	0.015458	0.988904
technology_th	0.480975	0.679612
tedmem_awardwin_th	0.272889	0.721221

Precision-Recall Curves



Next Steps

- Processing the text with Gensim Doc2Vec, Word2Vec

- Other algorithms to tackle multi-classification problem: Classifier Chain, Label Powerset from skmultilearn