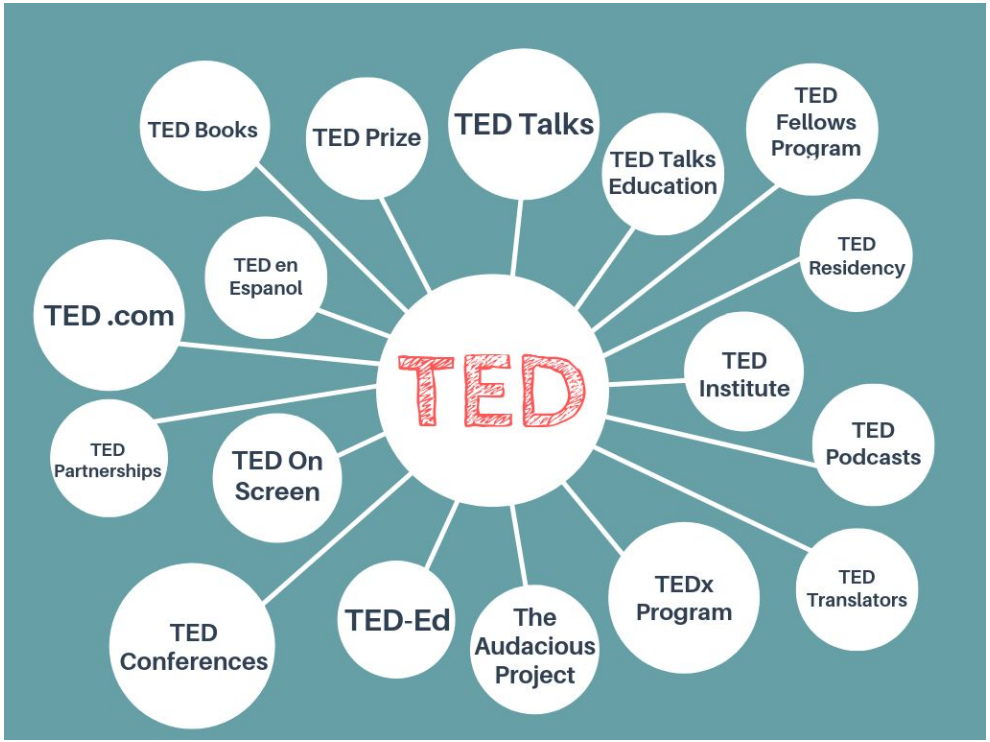
Text Mining on TED Talks

By Jean, Kelly, Sulav, and Ram

Introduction

TED Talks, which operates under the slogan 'Ideas worth spreading' hosts conferences where experts, enthusiasts, and talents from various walks of life share their invaluable insights on the Internet for free.

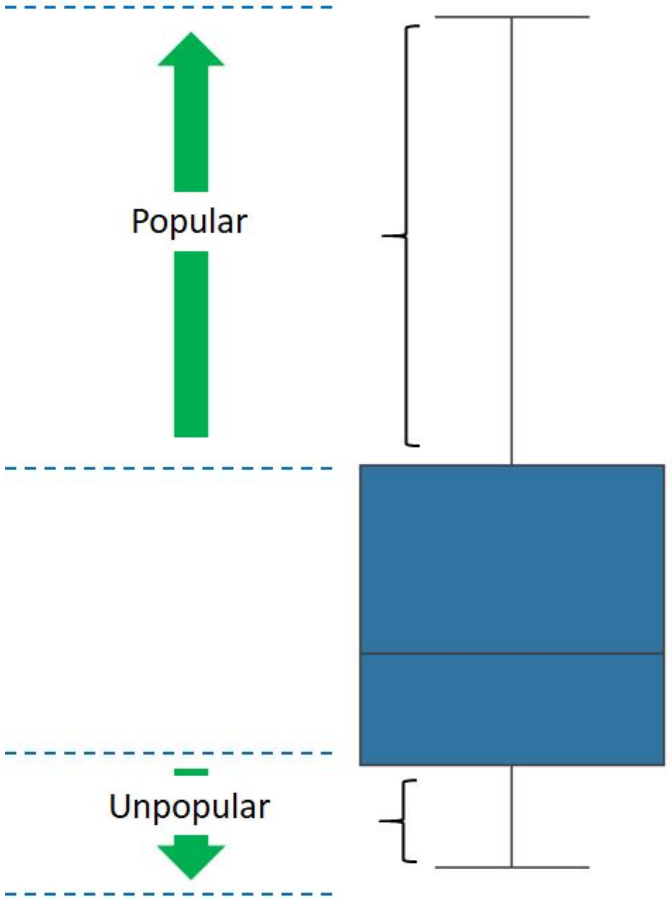


Problem Statement

TED Talks wants to find, organize, and share ‘ideas worth spreading’ for decades to come. For that reason, this analysis will attempt to model and predict the most popular talks from the past decade so that TED can focus on posting the most interesting talks.

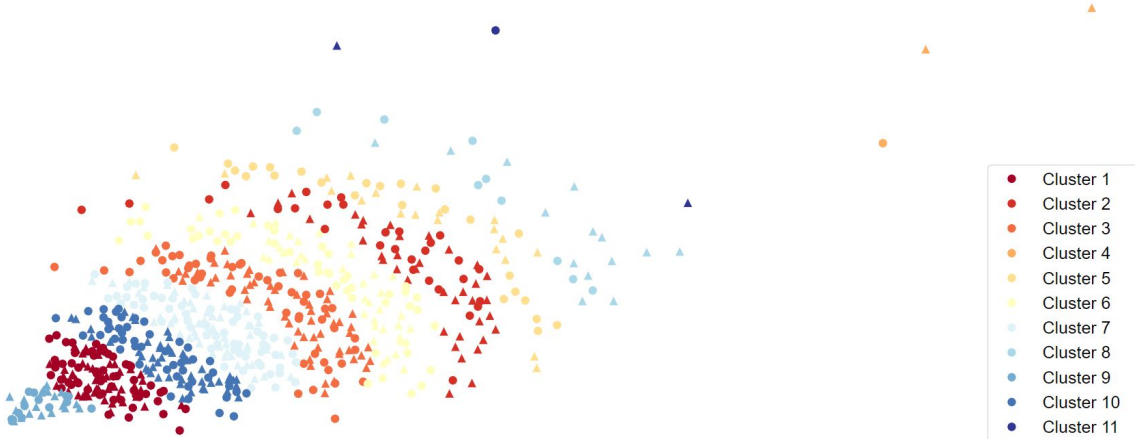
Data Definition

- Analyzed TED Talks published from 2010 to 2017.
- Popularity score compiled by total positive over negative ratings :
  - Greater than the 75th percentile is defined as Popular.
  - Less than the 25th percentile is defined as Unpopular.



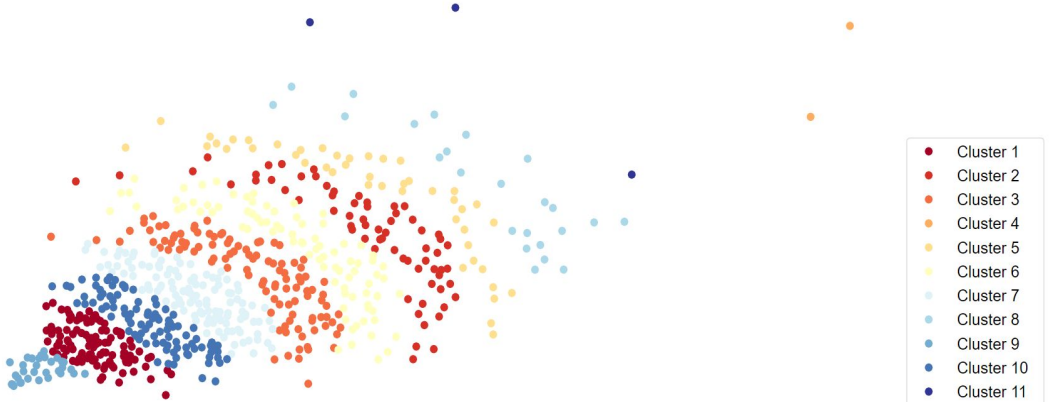
Clustering Analysis

K Means Clustering performed using Word Frequencies and Euclidean Distance method



Popular Talks

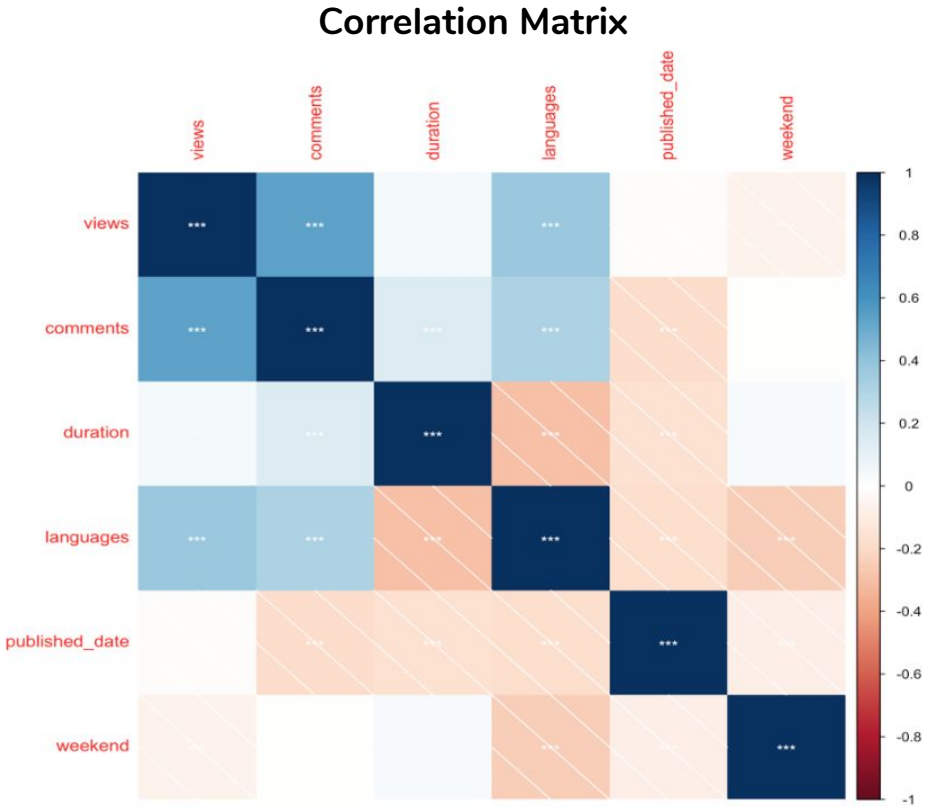
Unpopular Talks



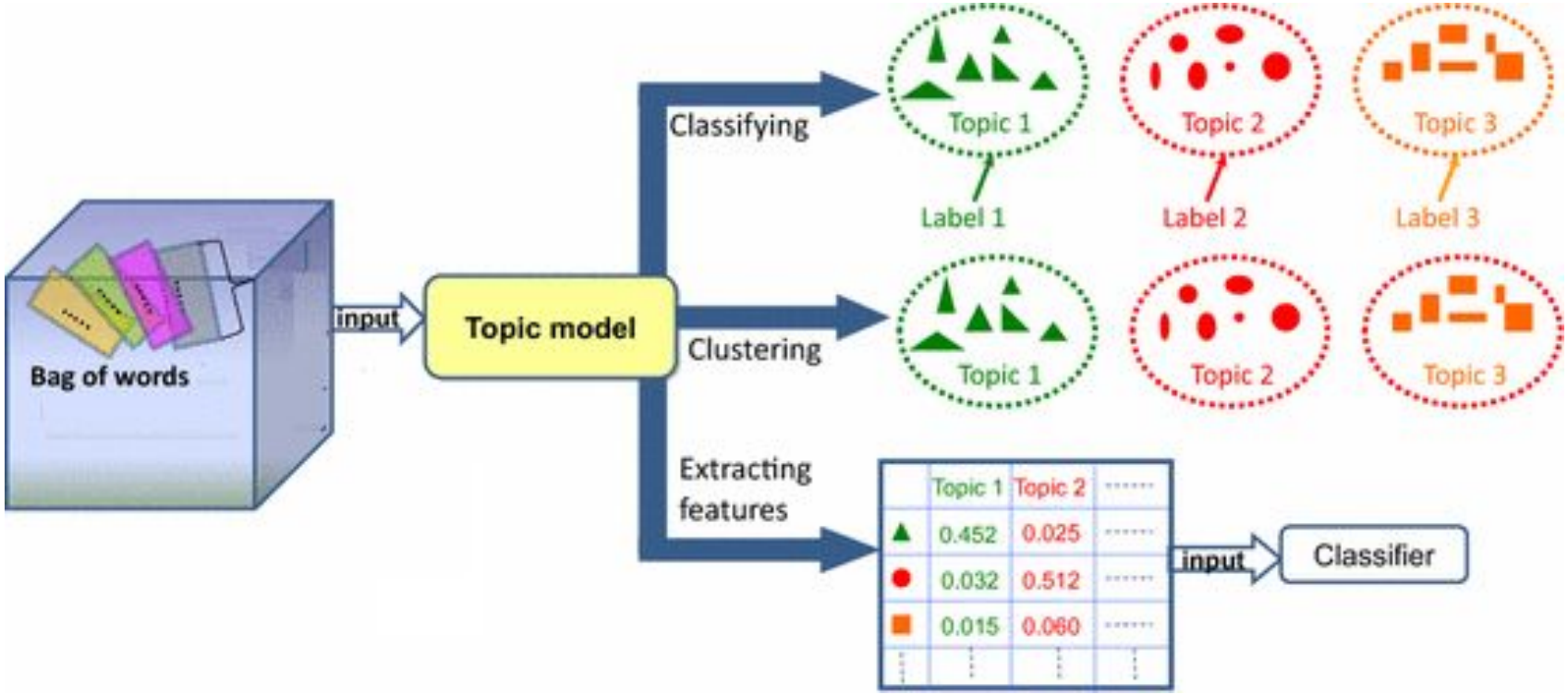
Cluster	Topics
Cluster 1	AI, Machine Learning and Design
Cluster 2	Science and Technology
Cluster 5	Science and Technology
Cluster 3	Culture and Creativity
Cluster 4	Global Issues and Politics
Cluster 11	Global Issues and Politics
Cluster 6	Research and Technology in Health care and Energy
Cluster 7	Climate Change, Disease and War (Threats)
Cluster 8	Arts, Nature and Architecture
Cluster 9	Entertainment and Music
Cluster 10	Astronomy, Math, Engineering and computer science

Features of a popular talk

- High number of comments.
- Translations in many languages.
- It shouldn’t be too short .
- Higher number of tags, ideally between 3–8!
- It would be uploaded on a weekday, preferably a Friday!

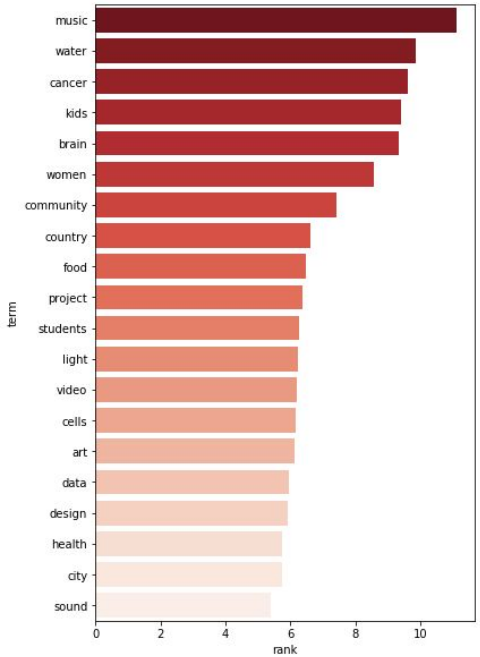


Topic Modelling

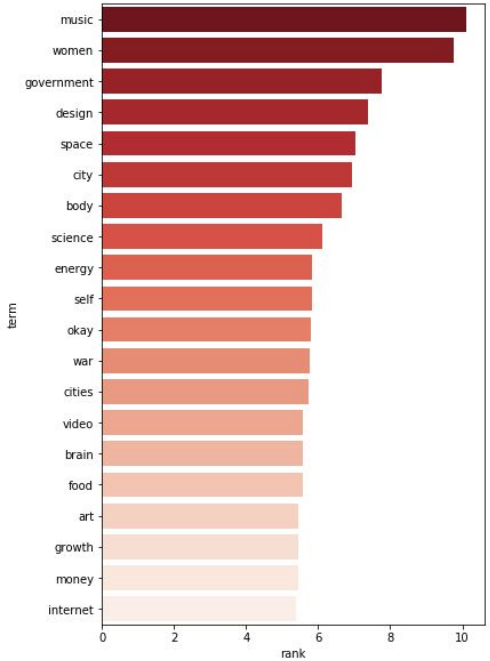


Topic Modeling - Top Ranked Words

Popular

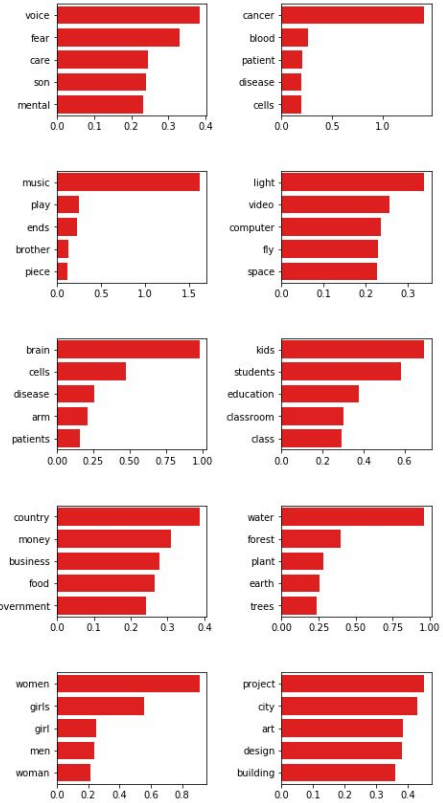


Unpopular

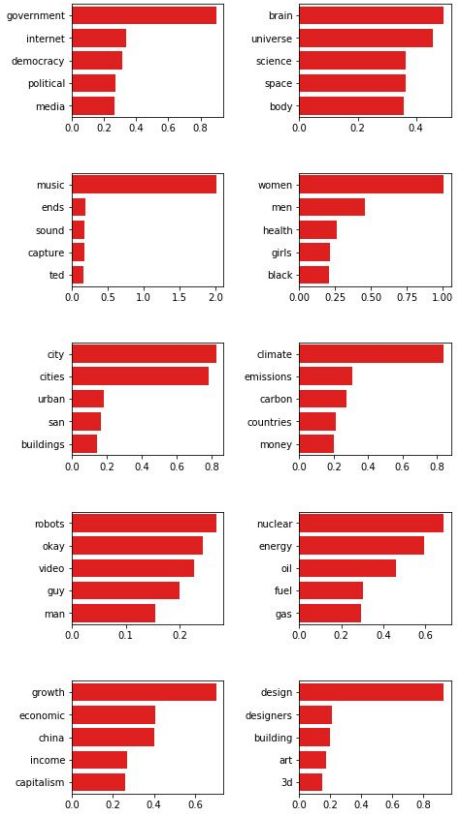


Topic Modeling - Top Ranked Topic Words

Popular

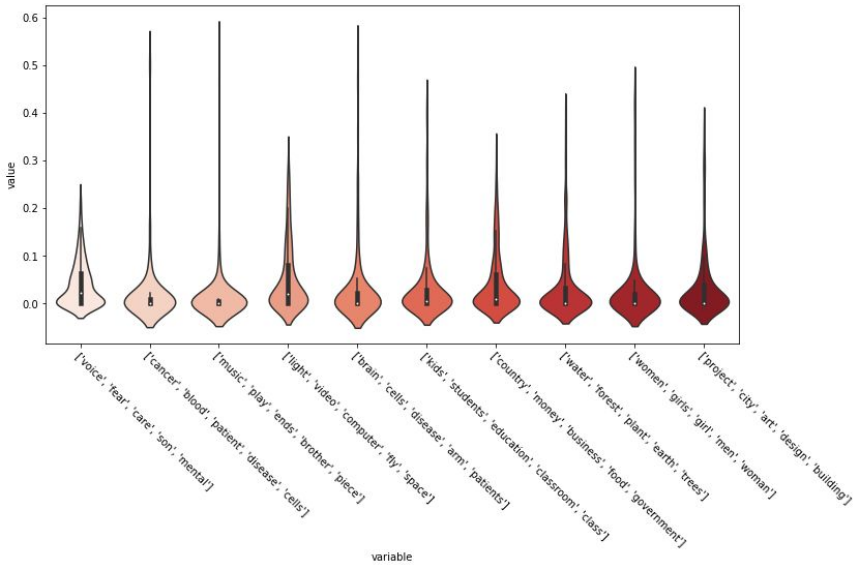


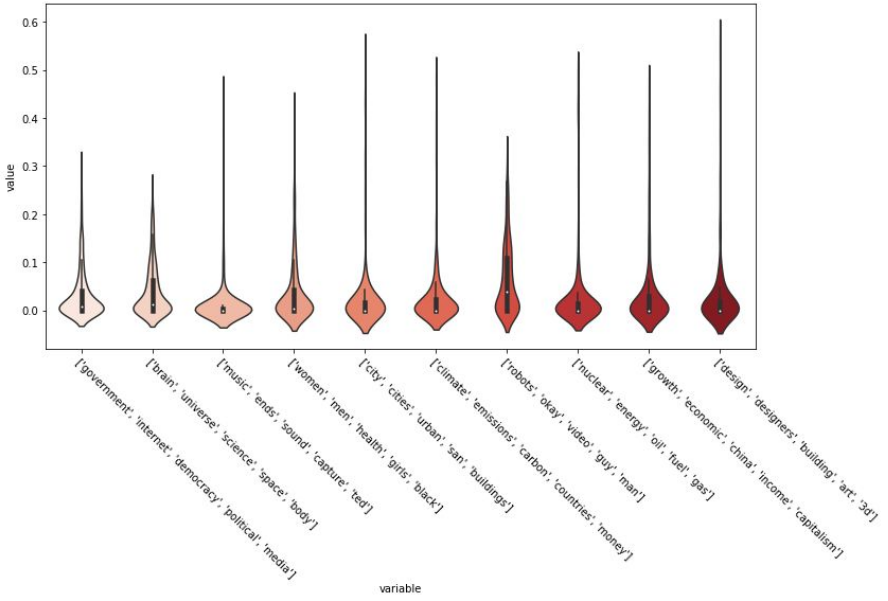
Unpopular



Topic Modeling - Topic Distribution

Popular

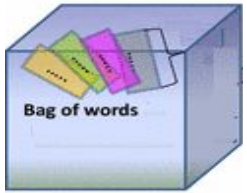




Unpopular

Popularity Predictions - Supervised Classifications

Count Vectors - Freq.

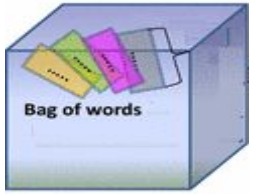


Supervised Learning

Count Vectors -TFIDF

Supervised Learning

Boolean Vectors



Supervised Learning

K-fold CV

MNB

SVM

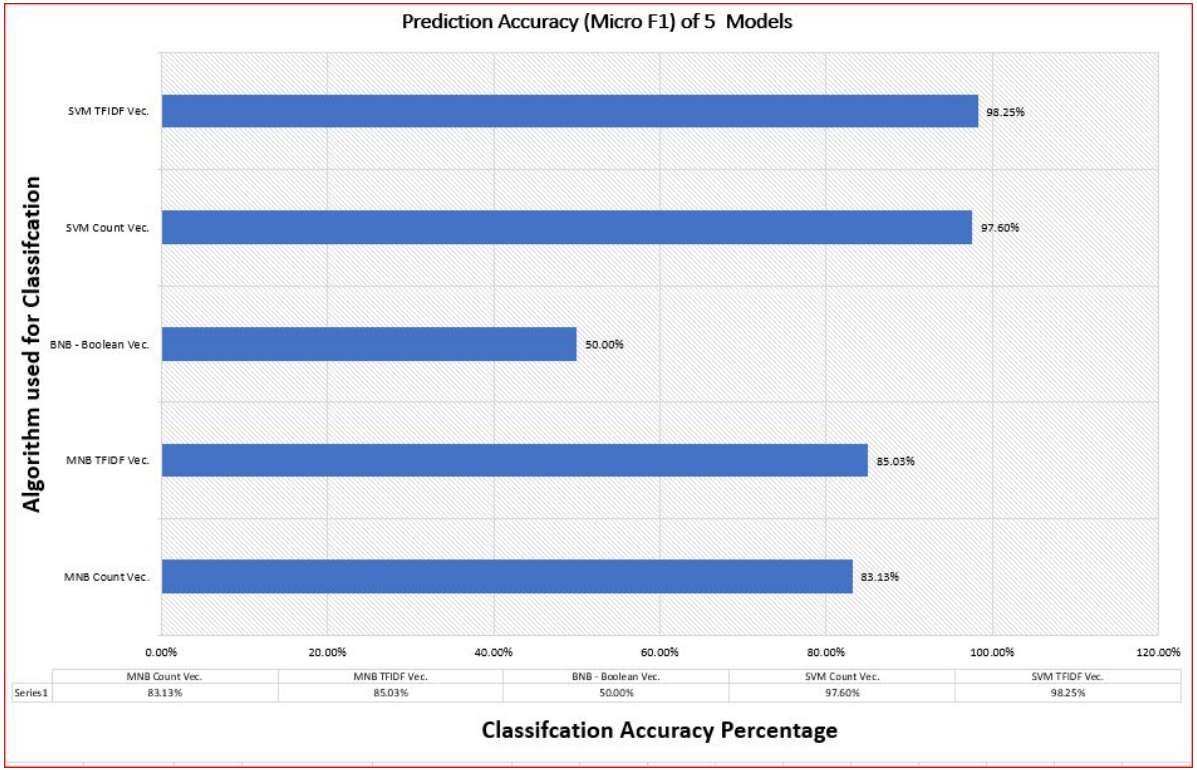
MNB

SVM

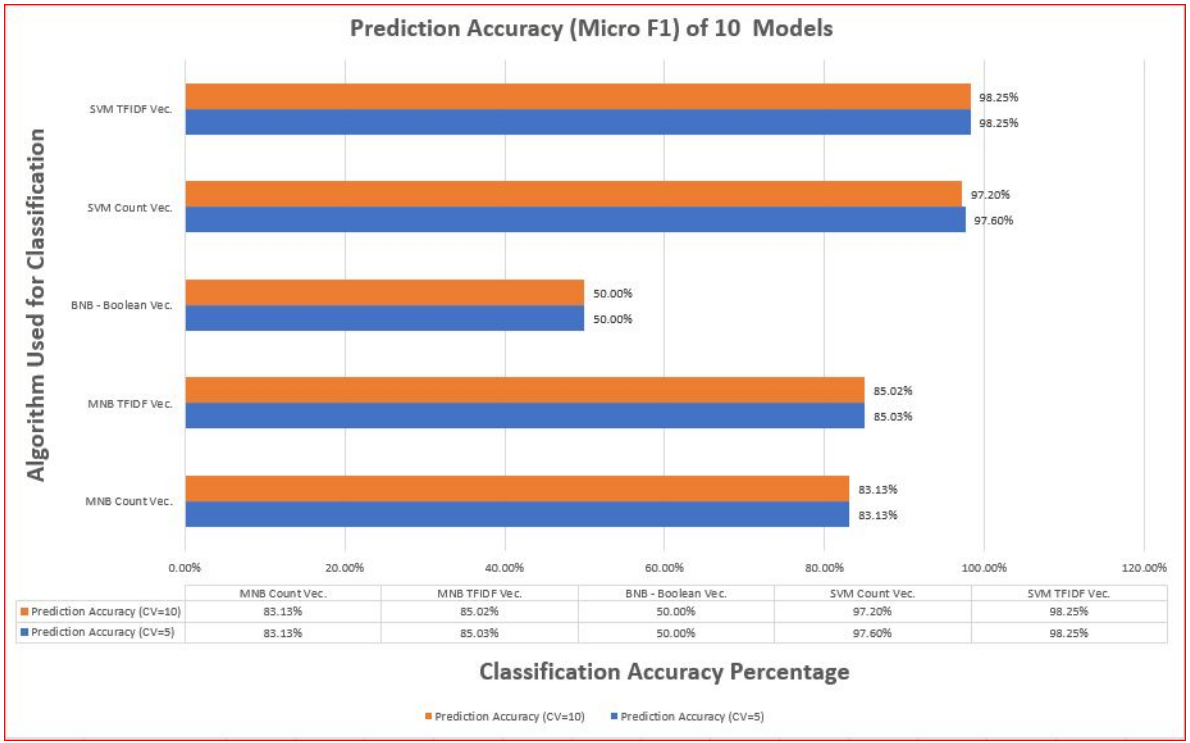
BNB

Predictions and subsequent Evaluation of Accuracy

Model Accuracy Comparison - I



Model Accuracy Comparison - II



- No significant difference found between accuracies obtained using 5 and 10-fold cross validation.

Conclusion

- Model will have to be updated yearly because trends over time
- Recommend to publish talk on Fridays
- Topics that were consistently popular over the past decade:
    - Music
    - Cancer Research
    - Children Education