Text Classification Model — MediaHound

Andrea Chan
Senhao Chen
Cody Lu
John Ndolo

Background & Purpose

Main Problem

❏ MediaHound: start-up media company with no internal data to train the model

Model Objectives

❏ Build PoC model for MediaHound that classify topics from text.

Data

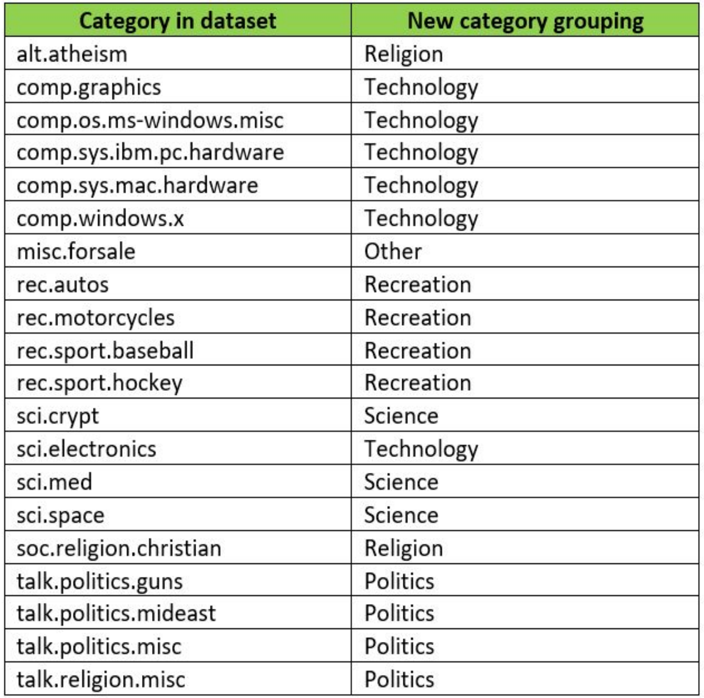
❏ Scikit Learn’s 20 Newsgroups dataset
❏ Comprises of 20,000 newsgroups posts
❏ 60/40 split between train and test

Model

❏ Mutinomial Navies Bayes
❏ Input: Text
❏ Output: Categories

Data Treatment

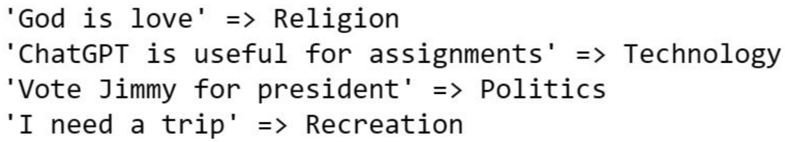
- Vectorize the data to turn the text into numerical values for statistical analysis
- Map Target_names from the original 20 categories to the new 6 categories



Category in dataset	New category grouping
alt.atheism	Religion
comp.graphics	Technology
comp.os.ms-windows.misc	Technology
comp.sys.ibm.pc.hardware	Technology
comp.sys.mac.hardware	Technology
comp.windows.x	Technology
misc.forsale	Other
rec.autos	Recreation
rec.motorcycles	Recreation
rec.sport.baseball	Recreation
rec.sport.hockey	Recreation
sci.crypt	Science
sci.electronics	Technology
sci.med	Science
sci.space	Science
soc.religion.christian	Religion
talk.politics.guns	Politics
talk.politics.mideast	Politics
talk.politics.misc	Politics
talk.religion.misc	Politics

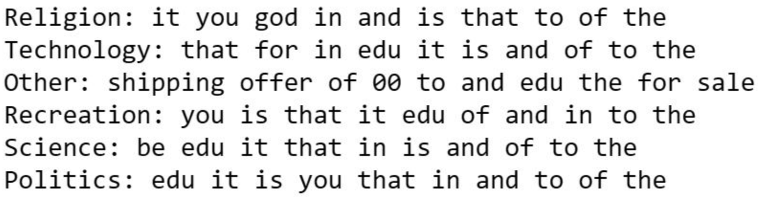
Model Result

- ❑ F1-score: a commonly used KPI for classification models, especially text data analysis
    - ❑ Weighted Average F1-Score: 0.905
- ❑ Sample Text Classifying



```
'God is love' => Religion
'ChatGPT is useful for assignments' => Technology
'Vote Jimmy for president' => Politics
'I need a trip' => Recreation
```

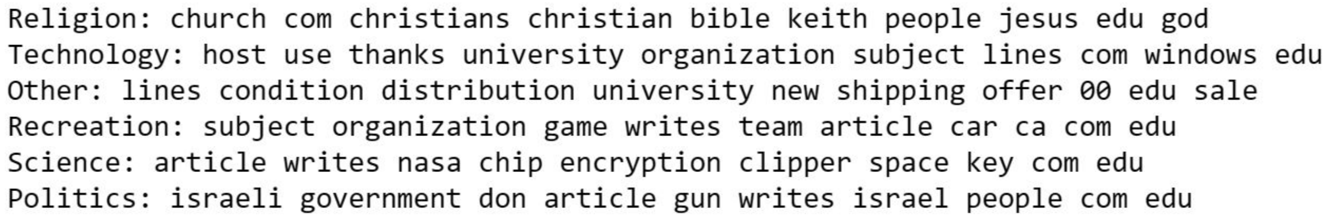
- ❑ Top 10 Informative Features



```
Religion: it you god in and is that to of the
Technology: that for in edu it is and of to the
Other: shipping offer of 00 to and edu the for sale
Recreation: you is that it edu of and in to the
Science: be edu it that in is and of to the
Politics: edu it is you that in and to of the
```

Model Validation and Data Treatment Cont'd

- ❑ Weighted Average F1-Score: 0.906
- ❑ Top 10 Features after remove stop_words



```
Religion: church com christians christian bible keith people jesus edu god
Technology: host use thanks university organization subject lines com windows edu
Other: lines condition distribution university new shipping offer 00 edu sale
Recreation: subject organization game writes team article car ca com edu
Science: article writes nasa chip encryption clipper space key com edu
Politics: israeli government don article gun writes israel people com edu
```

Strip Newsgroup Related Metadata by removing Header, Footer, and Quotes

- With such an abundance of clues that distinguish newsgroups, the classifiers barely have to identify topics from text at all, and they all perform at the same high level.
- The classifier lost over its F1-score and new weighted average F1-Score is 0.816, but the model is more realistic.

Future monitoring and internal data training

- ❑ High quality data collection and labelling
- ❑ Continuous monitoring and improvement of the model
- ❑ Human feedback: gather end user perspective for improvement
- ❑ Incorporation of domain expertise

Thank you for listening!

Q&A Session