Public Opinion Analysis on Politics

Prediction on Selection by NYT's Editors

Junqian Zhang

Data Science and Economics

May 31, 2021

Introduction

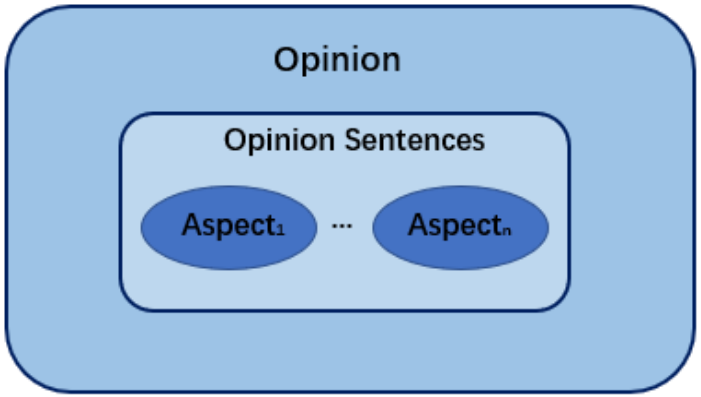
- Classify whether a comment can get the editors' selection under the topic of *Politics*
- Explore the relation between the Public's opinion and editors' opinion

Introduction

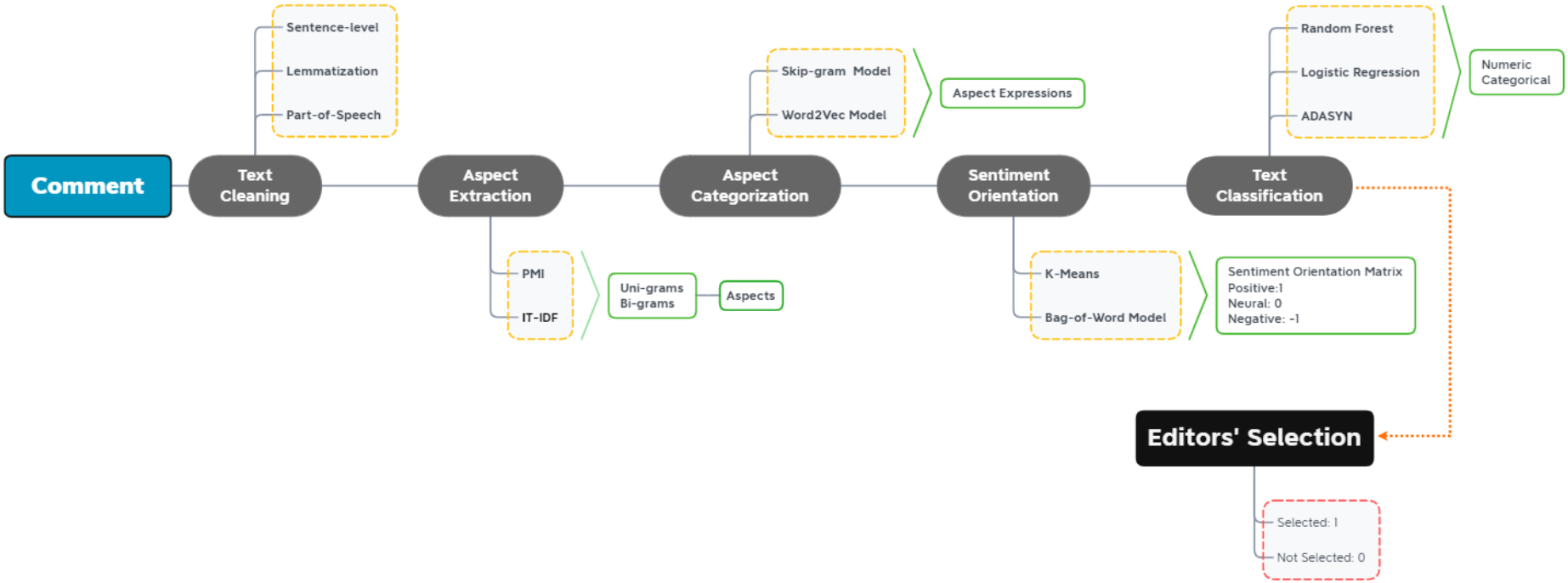
The *opinion* of the Public discussed in this project is a quadruple,

$$(e_i, a_{ij}, s_{ijk}, h_k)$$

where e_i is the topic of the comment, a_{ij} is an aspect of e_i, s_{ijk} is the sentiment orientation on aspect a_{ij} of topic e_i, h_k is the opinion holder. The topic e_i is fixed at *Politics*.



Procedure


Comment
Text Cleaning
Sentence-level
Lemmatization
Part-of-Speech
Aspect Extraction
PMI
IT-IDF
Uni-grams
Bi-grams
Aspects
Aspect Categorization
Skip-gram Model
Word2Vec Model
Aspect Expressions
Sentiment Orientation
K-Means
Bag-of-Word Model
Sentiment Orientation Matrix
Positive:1
Neural: 0
Negative: -1
Text Classification
Random Forest
Logistic Regression
ADASYN
Numeric
Categorical
Editors' Selection
Selected: 1
Not Selected: 0

Data

- Comments made on the articles published in New York Times in April 2017
- `sectionName` describes the topic of the articles commented, only *Politics* are selectd in this analysis
- `commentBody` stores the contents of the comment
- `editorsSelection` describes whether a comment is picked by editors, and is the target variable, and **0** for not picked, **1** for picked

Data

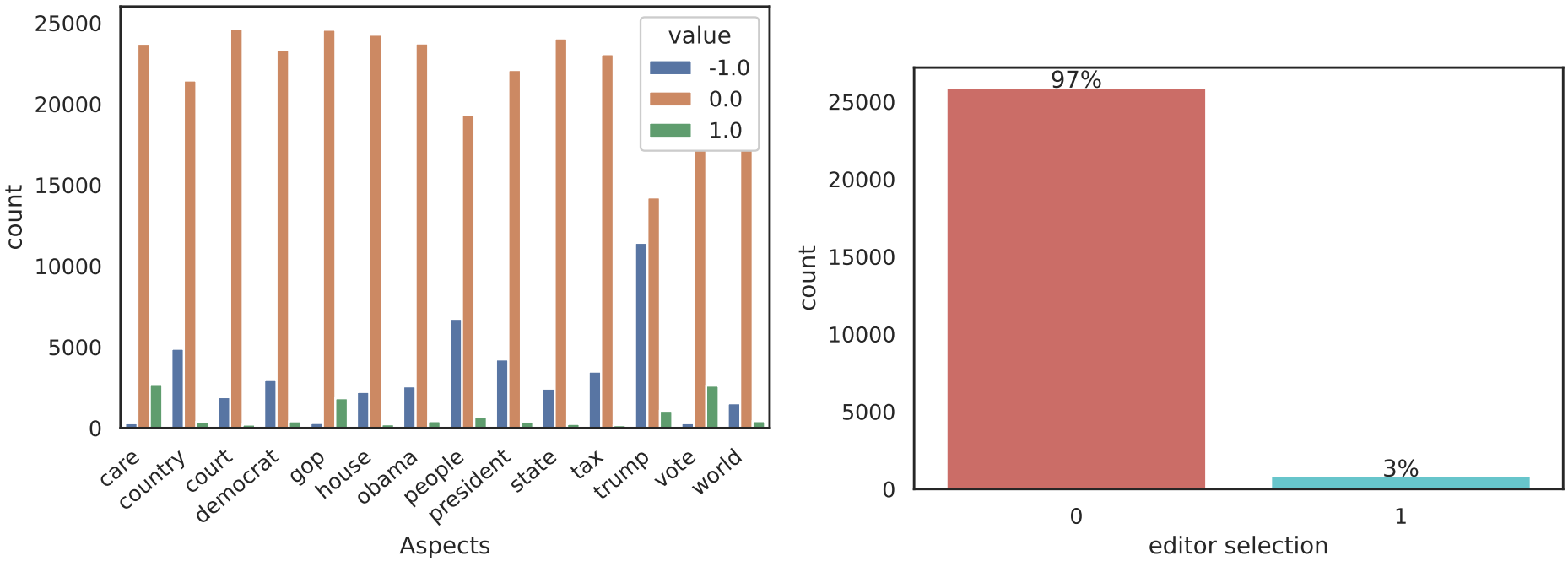


Figure: Distribution of Input variables and Target variable

- 14 aspects
- 0: 25,907; 1: 801

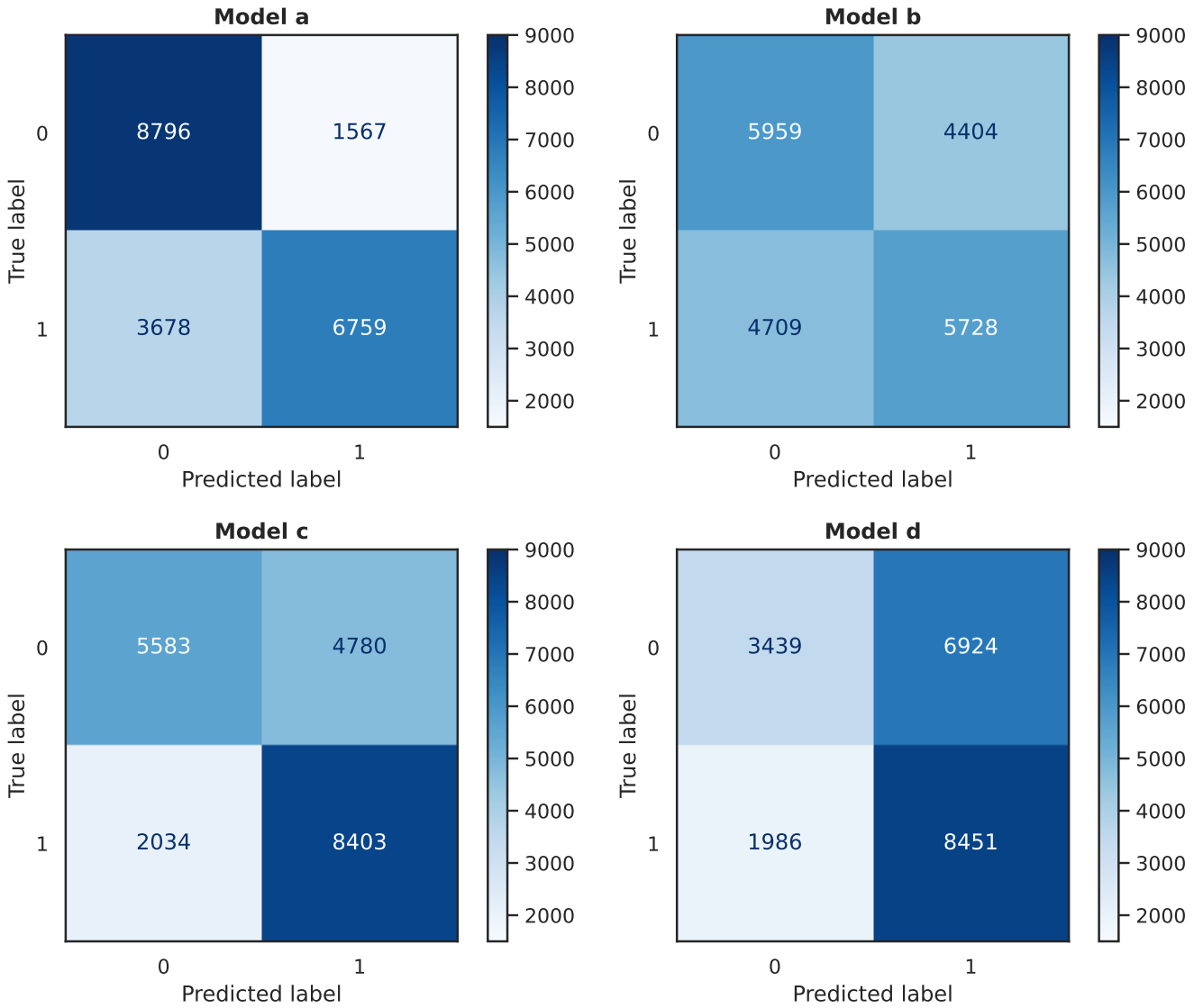
Results

- Evaluation
  - 0: 25,907; 1: 26,092
  - 10-fold Cross Validation
  - Training set: 60%

Table: Comparison of the results

No.	Description	CV Score	Test Score
Model a	Random Forest (Numeric)	0.750	0.747
Model b	Logistic Regression (Numeric)	0.562	0.561
Model c	Random Forest (Categorical)	0.676	0.672
Model d	Logistic Regression (Categorical)	0.592	0.600

Confusion Matrix

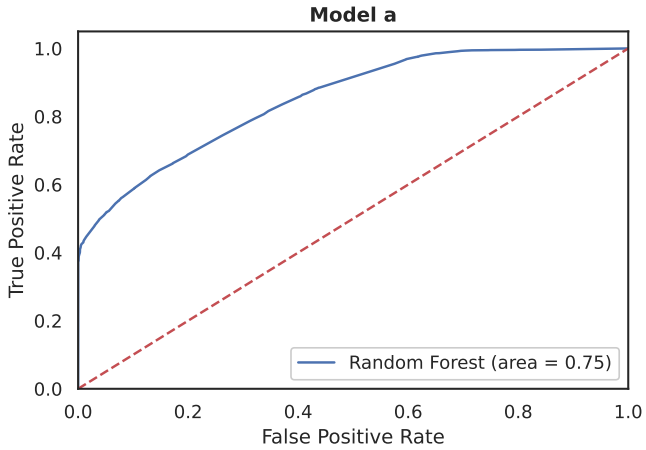


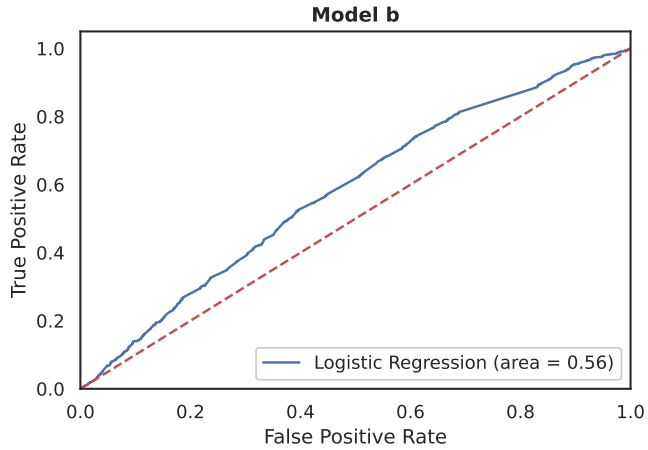
Performance

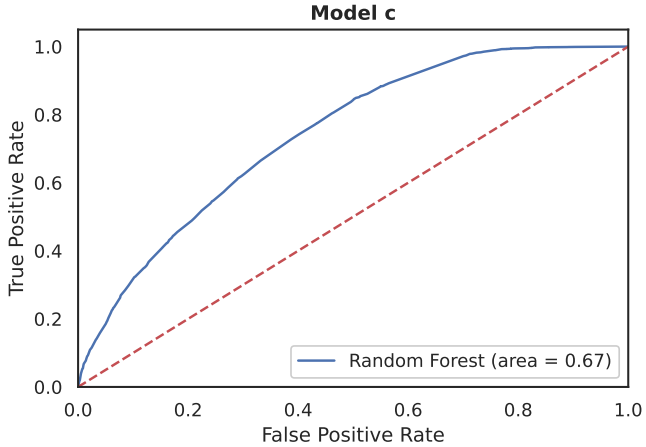
Model	Label	Precision	Recall	F1-score
Model a	0	0.70	0.85	0.77
	1	0.81	0.64	0.72
Model b	0	0.56	0.56	0.56
	1	0.56	0.56	0.56
Model c	0	0.75	0.51	0.61
	1	0.63	0.83	0.72
Model d	0	0.63	0.47	0.54
	1	0.58	0.73	0.65

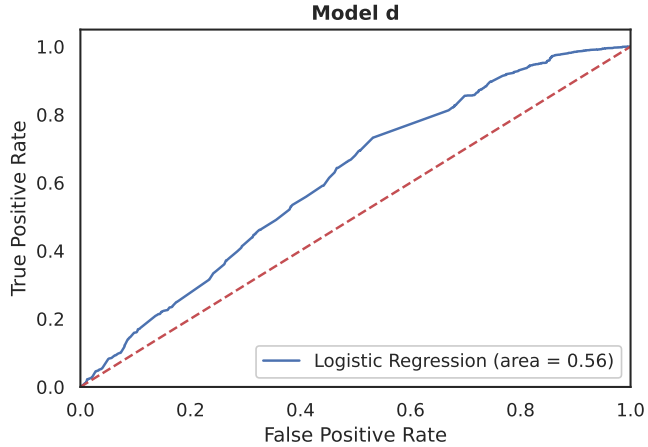
- 0: 10363; 1: 10437

ROC & AUC


Model a
True Positive Rate
False Positive Rate
Random Forest (area = 0.75)
0.0
0.2
0.4
0.6
0.8
1.0


Model b
True Positive Rate
False Positive Rate
Logistic Regression (area = 0.56)
0.0
0.2
0.4
0.6
0.8
1.0


Model c
True Positive Rate
False Positive Rate
Random Forest (area = 0.67)
0.0
0.2
0.4
0.6
0.8
1.0


Model d
True Positive Rate
False Positive Rate
Logistic Regression (area = 0.56)
0.0
0.2
0.4
0.6
0.8
1.0

Future Work

- Sentiment orientation judgement:
  - multi-label
  - other information retrieval tools
  - sarcasm and comparative words
- Implicit aspects
- Other Classifiers and ensemble methods

Thank you!