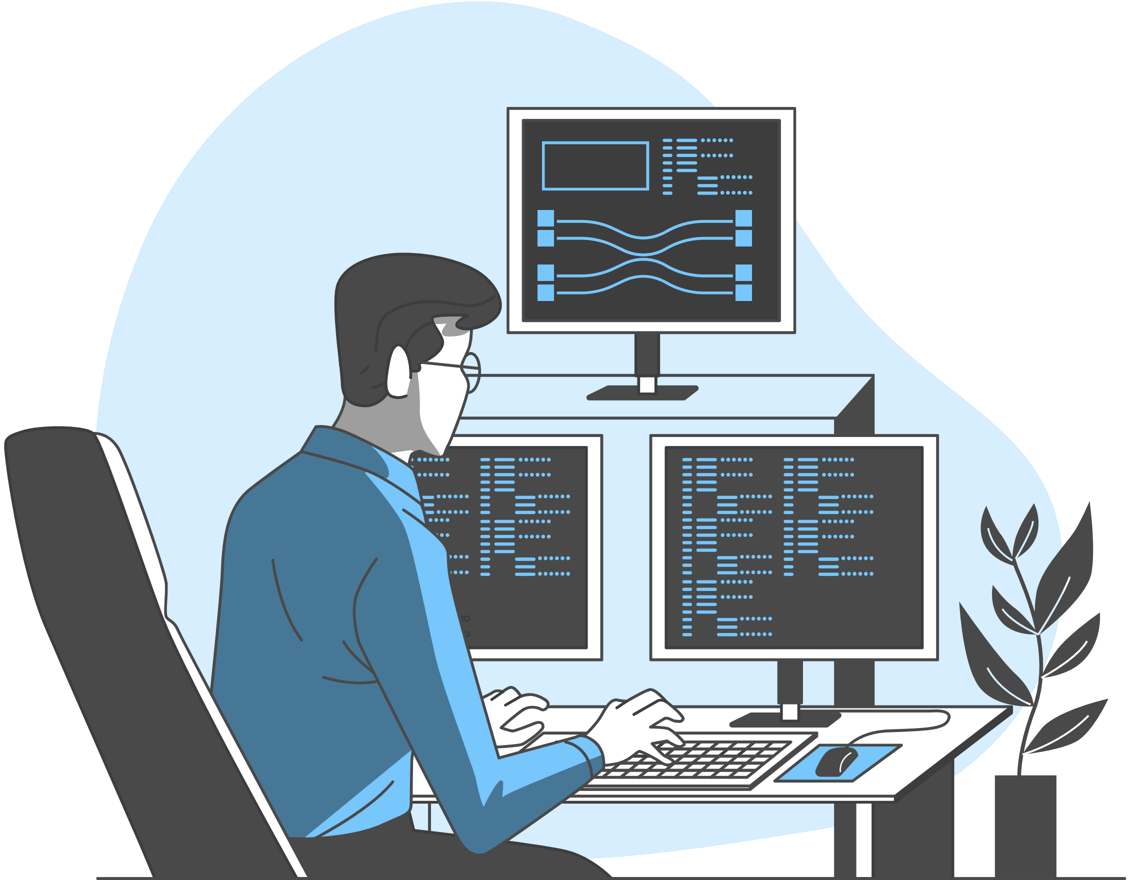


Research Popularity Prediction

Table of Contents

01 Introduction

02 Data Description

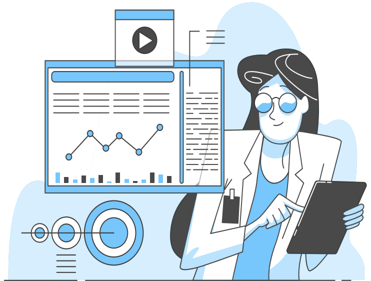
03 Data preprocessing

04 Topic Modeling

05 Regression

06 Classification

07 Conclusion



01

Introduction

Problem Description

We’re going to analyse what factors contribute to social science research articles popularity and predict the popularity of articles.

02

Data Description

- We scraped the data from researchgate.
- All articles are from social science departments.

Dataset

(1899,12)

column	Description	Type
title	article's title	string
author	author name	string
abstract	Abstract full-text	string
category	article, literature review, conference paper..etc	string
date_published	date the article was published	date
date_added	date the article was uploaded to researchgate	date
figures	1 if there're figures & 0 if not	int
full_text?	availability of full text (using download or Request full-text as keywords)	string
citation	number of times that paper was cited	int
reads	number of views (target variable)	int

Data Collection Limitations

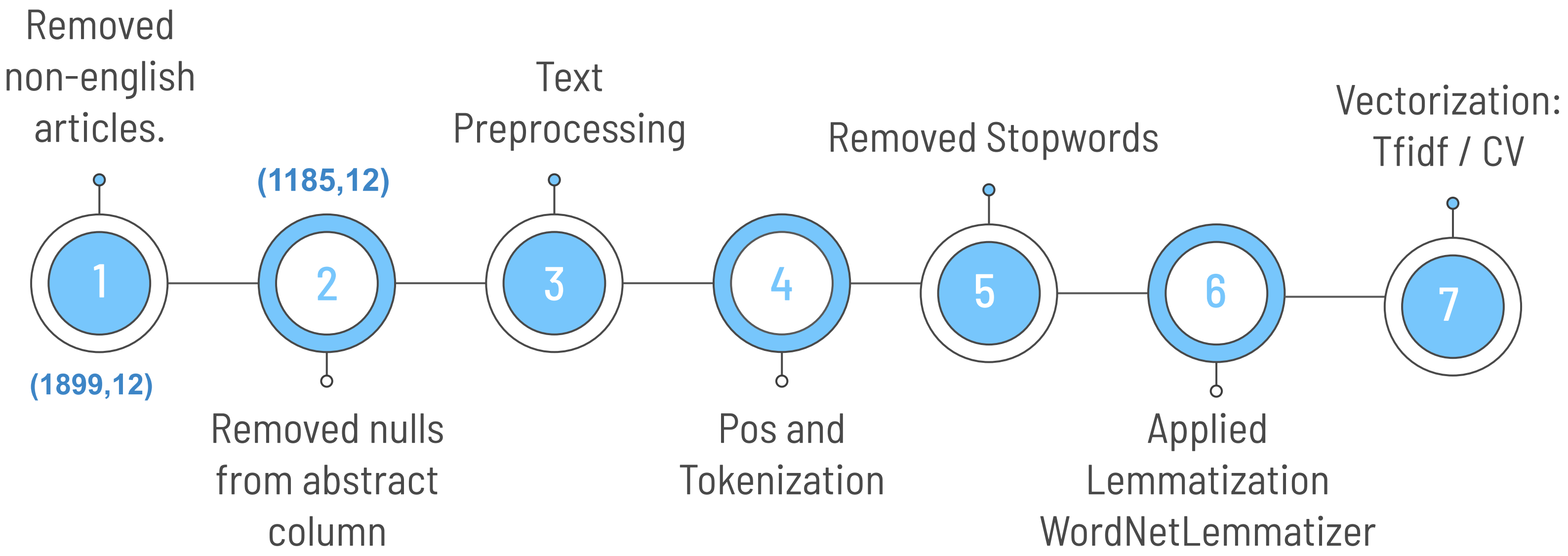
Have To use organization email Address

Limited Quota Per Day

03

Data preprocessing

Preprocessing flow



Removed non-english articles.

(1899,12)

Removed nulls from abstract column

(1185,12)

Text Preprocessing

Pos and Tokenization

Removed Stopwords

Applied Lemmatization WordNetLemmatizer

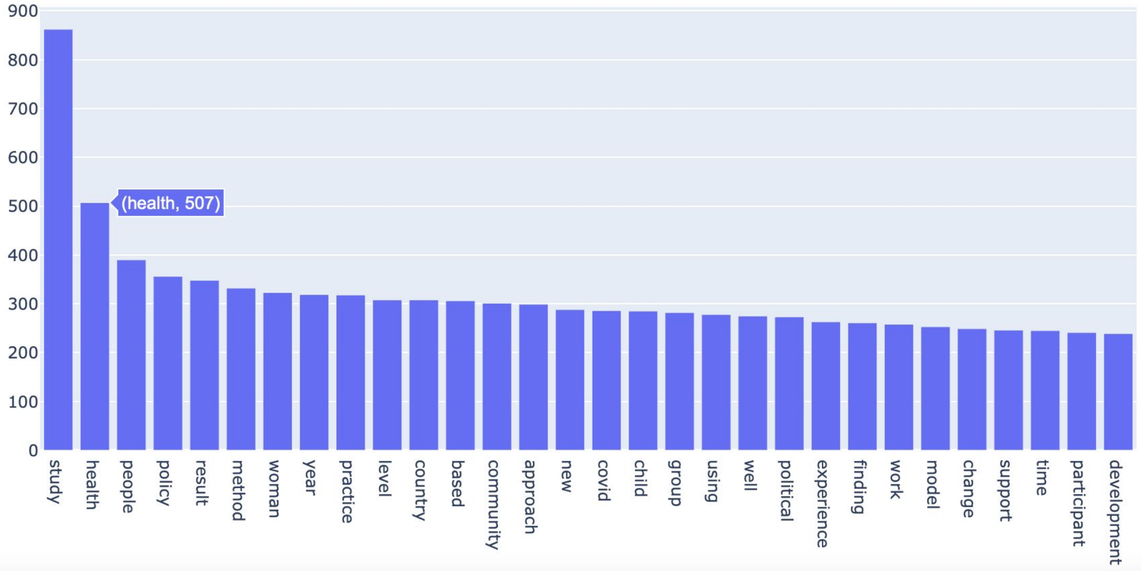
Vectorization: Tfidf / CV

04

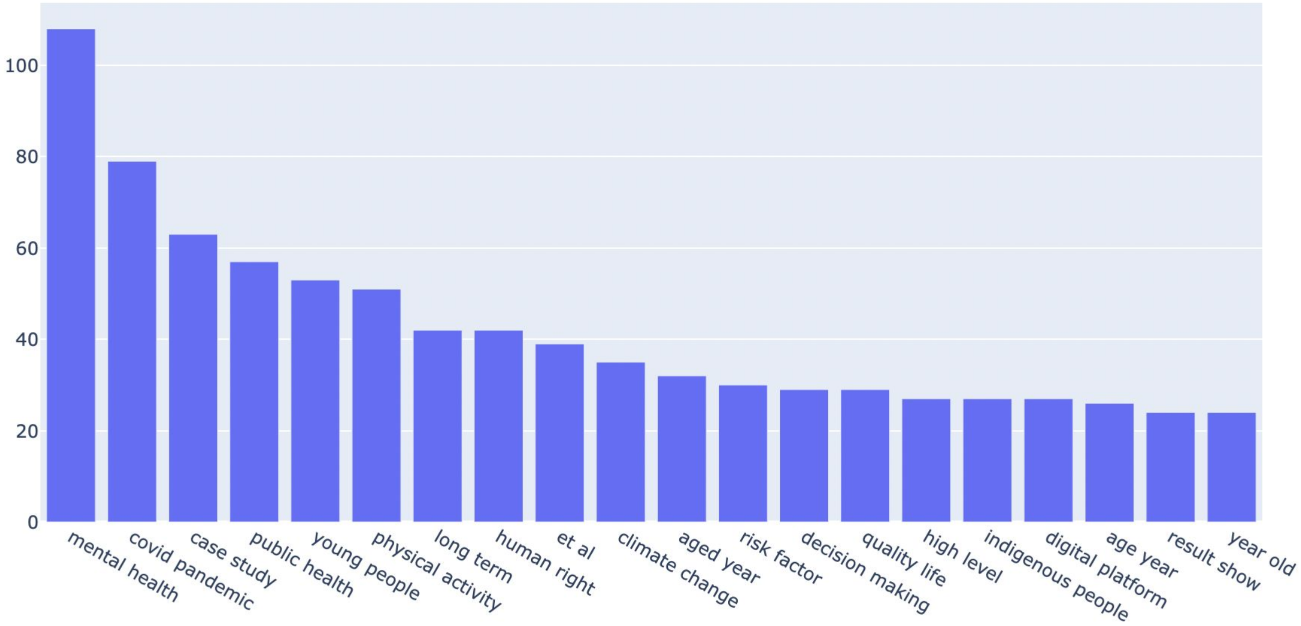
Topic Modeling

NMF, LDA, CorEx

Top 30 unigrams in the abstract after removing stop words and lemmatization



Top 20 bigrams in the abstract after removing stop words and lemmatization



NMF Topic Modeling

TFIDF Vectorization

16>12>10>9

Number of topics

NMF Topic Modeling

Topic 0
political, policy, state, new, economic, power, right

Topic 1
sexual, drug, violence, gender, minority, health, sex, sexuality, identity, lgbtq

Topic 2
covid, health, pandemic, mental, lockdown, country, financial, psychologica

Topic 3
woman, party, gender, labour, policy, migrant, feminist, men

Topic 4
police, crime, justice, knife, policing, criminal, violence, armed,

Topic 5
student, school, education, university, learning, digital, online, teaching

Topic 6
patient, age, study, factor, disease, ci, risk, cancer

Topic 7
child, parent, family, food, childrens, abuse, parental

Topic 8
care, people, service, community, disability, health

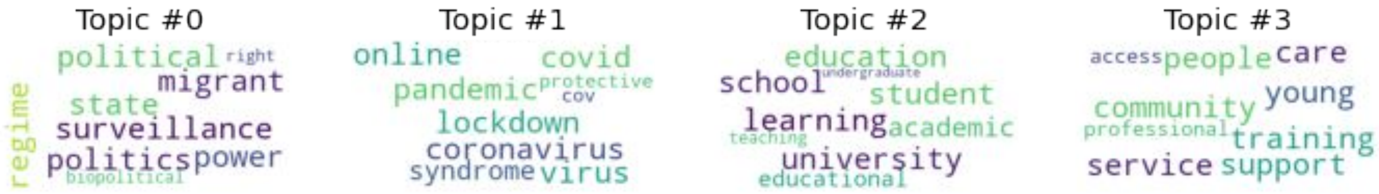
CorEX Topic Modeling

Count Vectorization

12
Number of topics

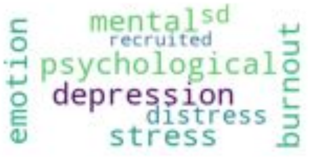
Used words from NMF as anchor words, with correlation strength of **7**

CorEX Topic Modeling

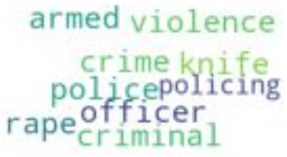


CorEX Topic Modeling

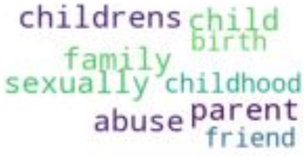
Topic #4



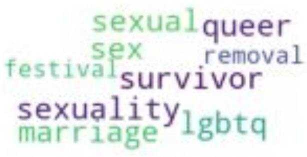
Topic #5



Topic #6

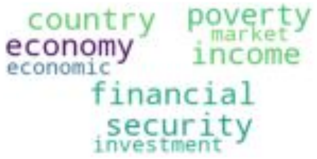


Topic #7

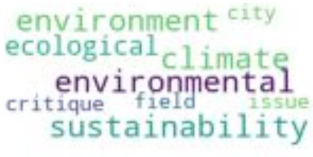


CorEX Topic Modeling

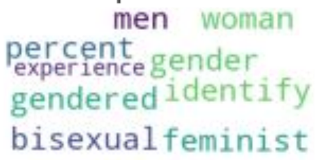
Topic #8



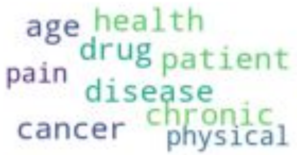
Topic #9



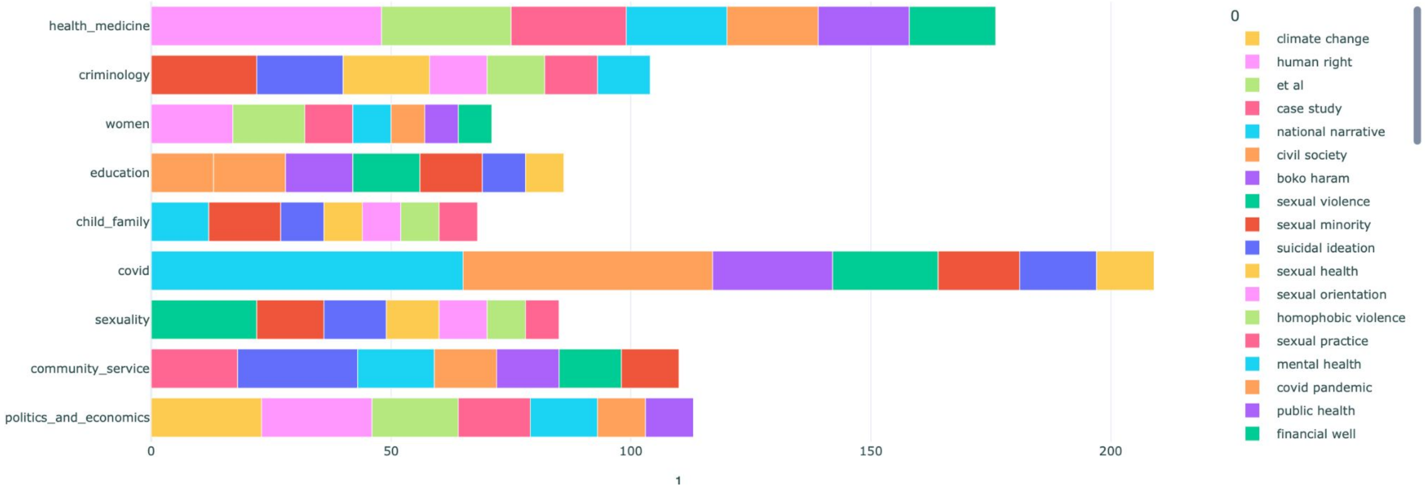
Topic #10



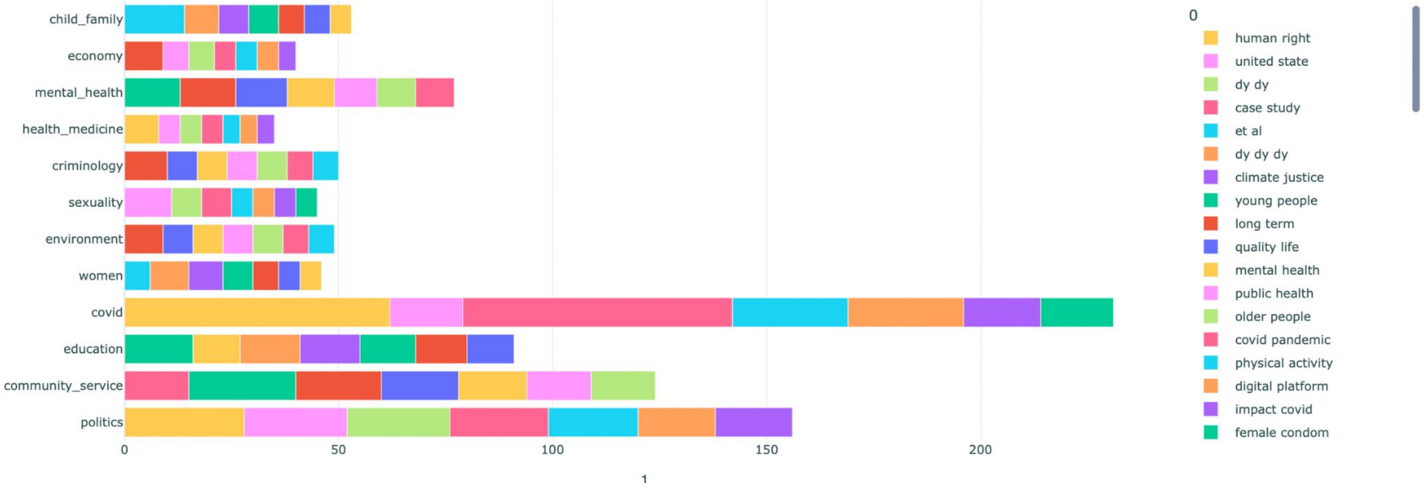
Topic #11



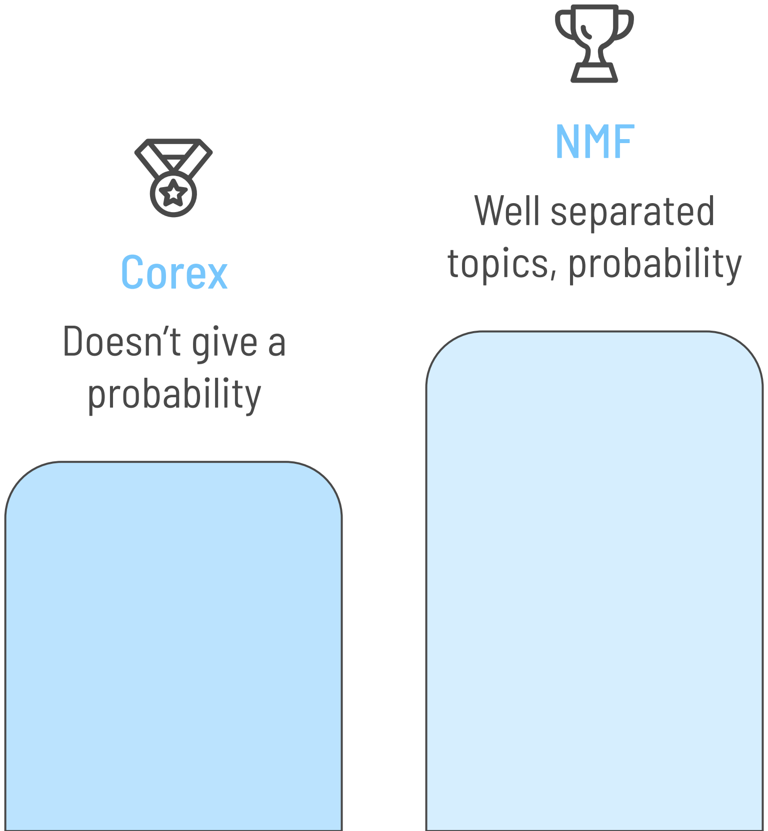
NMF Topics in Bigrams



CorEx Topics in Bigrams



Chosen Model



Corex

Doesn't give a probability

NMF

Well separated topics, probability

Topics

- Politics And Economics
- Sexology
- COVID
- Women
- Criminology
- Education
- Health & Medicine
- Family & Child
- Community Service

Generating Subtopics

Subtopics example: COVID

```
Topic  0 ( Covid & Mental Health )
mental, health, covid, psychological, distress, gi, symptom, impact, disorder, participant

Topic  1 ( Covid & economy)
financial, well, inclusion, study, economy, country, understand, effect, system, pandemic

Topic  2 ( lockdown )
behavior, lockdown, change, time, diary, risk, activity, location, uk, population

Topic  3 ( Policies )
health, policy, covid, response, pandemic, medium, crisis, state, news, public
```

Subtopics example: Sexology

Topic 0 (Sexual Health)
sexual, health, people, adult, young, minority, study, lockdown, older, identity

Topic 1 (?)
coumarins, coumarin, compound, various, widely, performed, perfume, crystalline, explicates, scent

Topic 2 (Sex & Drugs)
drug, lgbtq, gender, consumption, practice, sex, particular, effect, consumer, policy

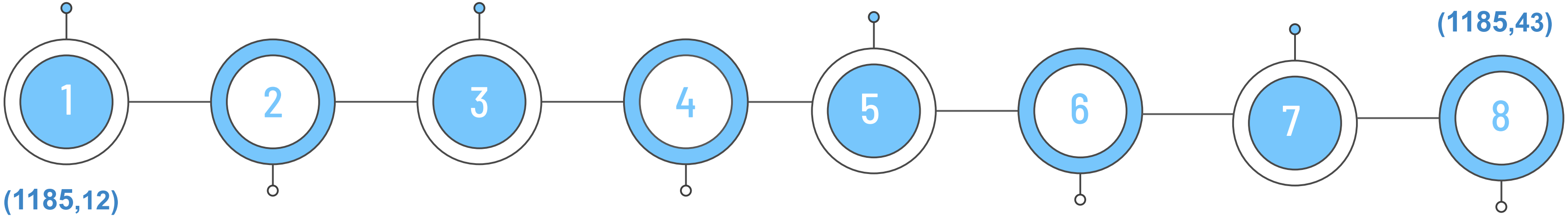
Topic 3 (Sexual Violence)
violence, festival, sexual, music, argue, feminist, harm, framework, experience, victim

05

Regression

Predicting the number of Reads

Feature Engineering



1. Adding dominant topic column

(1185,12)

2. Adding probability per topic columns

3. Title & abstract sentiment polarity

4. Title & Abstract length in words / sentences

5. Reformatting addition date to days

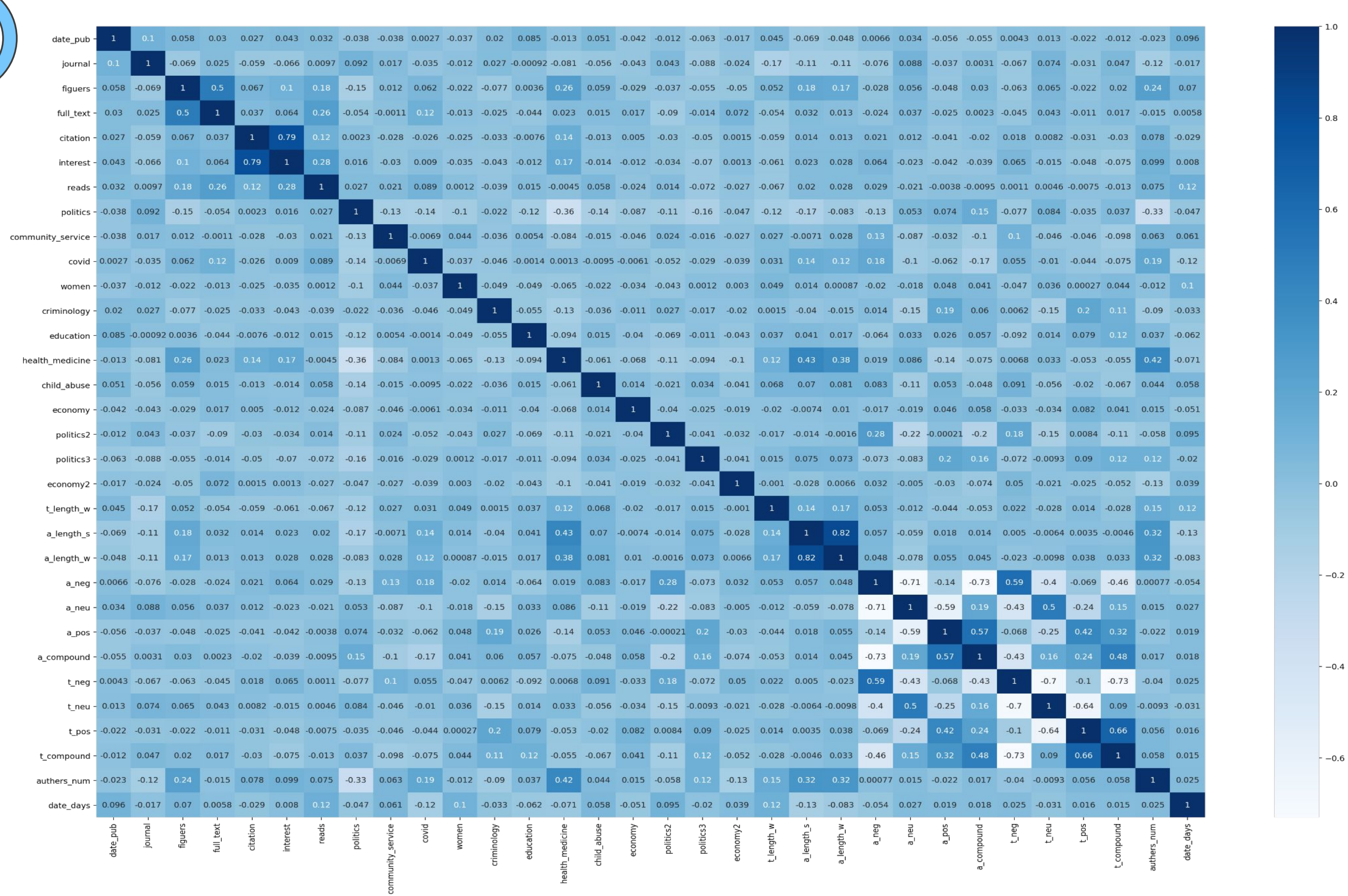
6. Number of authors

7. Encoding Journal column

8. Mapping Full_text column to 0 and 1

(1185,43)

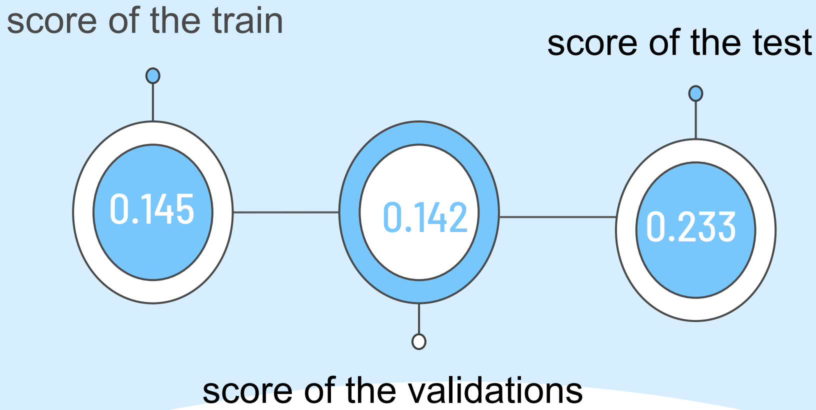
Correlation Between Reads and other Features



	date_pub	journal	figuers	full_text	citation	interest	reads	politics	community_service	covid	women	criminology	education	health_medicine	child_abuse	economy	politics2	politics3	economy2	t_length_w	a_length_s	a_length_w	a_neg	a_neu	a_pos	a_compound	t_neg	t_neu	t_pos	t_compound	authers_num	date_days
date_pub	1	0.1	0.058	0.03	0.027	0.043	0.032	-0.038	-0.038	0.0027	-0.037	0.02	0.085	-0.013	0.051	-0.042	-0.012	-0.063	-0.017	0.045	-0.069	-0.048	0.0066	0.034	-0.056	-0.055	0.0043	0.013	-0.022	-0.012	-0.023	0.096
journal	0.1	1	-0.069	0.025	-0.059	-0.066	0.0097	0.092	0.017	-0.035	-0.012	0.027	-0.00092	-0.081	-0.056	-0.043	0.043	-0.088	-0.024	-0.17	-0.11	-0.11	-0.076	0.088	-0.037	0.0031	-0.067	0.074	-0.031	0.047	-0.12	-0.017
figuers	0.058	-0.069	1	0.5	0.067	0.1	0.18	-0.15	0.012	0.062	-0.022	-0.077	0.0036	0.26	0.059	-0.029	-0.037	-0.055	-0.05	0.052	0.18	0.17	-0.028	0.056	-0.048	0.03	-0.063	0.065	-0.022	0.02	0.24	0.07
full_text	0.03	0.025	0.5	1	0.037	0.064	0.26	-0.054	-0.0011	0.12	-0.013	-0.025	-0.044	0.023	0.015	0.017	-0.09	-0.014	0.072	-0.054	0.032	0.013	-0.024	0.037	-0.025	0.0023	-0.045	0.043	-0.011	0.017	-0.015	0.0058
citation	0.027	-0.059	0.067	0.037	1	0.79	0.12	0.0023	-0.028	-0.026	-0.025	-0.033	-0.0076	0.14	-0.013	0.005	-0.03	-0.05	0.0015	-0.059	0.014	0.013	0.021	0.012	-0.041	-0.02	0.018	0.0082	-0.031	-0.03	0.078	-0.029
interest	0.043	-0.066	0.1	0.064	0.79	1	0.28	0.016	-0.03	0.009	-0.035	-0.043	-0.012	0.17	-0.014	-0.012	-0.034	-0.07	0.0013	-0.061	0.023	0.028	0.064	-0.023	-0.042	-0.039	0.065	-0.015	-0.048	-0.075	0.099	0.008
reads	0.032	0.0097	0.18	0.26	0.12	0.28	1	0.027	0.021	0.089	0.0012	-0.039	0.015	-0.0045	0.058	-0.024	0.014	-0.072	-0.027	-0.067	0.02	0.028	0.029	-0.021	-0.0038	-0.0095	0.0011	0.0046	-0.0075	-0.013	0.075	0.12
politics	-0.038	0.092	-0.15	-0.054	0.0023	0.016	0.027	1	-0.13	-0.14	-0.1	-0.022	-0.12	-0.36	-0.14	-0.087	-0.11	-0.16	-0.047	-0.12	-0.17	-0.083	-0.13	0.053	0.074	0.15	-0.077	0.084	-0.035	0.037	-0.33	-0.047
community_service	-0.038	0.017	0.012	-0.0011	-0.028	-0.03	0.021	-0.13	1	-0.0069	0.044	-0.036	0.0054	-0.084	-0.015	-0.046	0.024	-0.016	-0.027	0.027	-0.0071	0.028	0.13	-0.087	-0.032	-0.1	0.1	-0.046	-0.046	-0.098	0.063	0.061
covid	0.0027	-0.035	0.062	0.12	-0.026	0.009	0.089	-0.14	-0.0069	1	-0.037	-0.046	-0.0014	0.0013	-0.0095	-0.0061	-0.052	-0.029	-0.039	0.031	0.14	0.12	0.18	-0.1	-0.062	-0.17	0.055	-0.01	-0.044	-0.075	0.19	-0.12
women	-0.037	-0.012	-0.022	-0.013	-0.025	-0.035	0.0012	-0.1	0.044	-0.037	1	-0.049	-0.049	-0.065	-0.022	-0.034	-0.043	0.0012	0.003	0.049	0.014	0.00087	-0.02	-0.018	0.048	0.041	-0.047	0.036	0.00027	0.044	-0.012	0.1
criminology	0.02	0.027	-0.077	-0.025	-0.033	-0.043	-0.039	-0.022	-0.036	-0.046	-0.049	1	-0.055	-0.13	-0.036	-0.011	0.027	-0.017	-0.02	0.0015	-0.04	-0.015	0.014	-0.15	0.19	0.06	0.0062	-0.15	0.2	0.11	-0.09	-0.033
education	0.085	-0.00092	0.0036	-0.044	-0.0076	-0.012	0.015	-0.12	0.0054	-0.0014	-0.049	-0.055	1	-0.094	0.015	-0.04	-0.069	-0.011	-0.043	0.037	0.041	0.017	-0.064	0.033	0.026	0.057	-0.092	0.014	0.079	0.12	0.037	-0.062
health_medicine	-0.013	-0.081	0.26	0.023	0.14	0.17	-0.0045	-0.36	-0.084	0.0013	-0.065	-0.13	-0.094	1	-0.061	-0.068	-0.11	-0.094	-0.1	0.12	0.43	0.38	0.019	0.086	-0.14	-0.075	0.0068	0.033	-0.053	-0.055	0.42	-0.071
child_abuse	0.051	-0.056	0.059	0.015	-0.013	-0.014	0.058	-0.14	-0.015	-0.0095	-0.022	-0.036	0.015	-0.061	1	0.014	-0.021	0.034	-0.041	0.068	0.07	0.081	0.083	-0.11	0.053	-0.048	0.091	-0.056	-0.02	-0.067	0.044	0.058
economy	-0.042	-0.043	-0.029	0.017	0.005	-0.012	-0.024	-0.087	-0.046	-0.0061	-0.034	-0.011	-0.04	-0.068	0.014	1	-0.04	-0.025	-0.019	-0.02	-0.0074	0.01	-0.017	-0.019	0.046	0.058	-0.033	-0.034	0.082	0.041	0.015	-0.051
politics2	-0.012	0.043	-0.037	-0.09	-0.03	-0.034	0.014	-0.11	0.024	-0.052	-0.043	0.027	-0.069	-0.11	-0.021	-0.04	1	-0.041	-0.032	-0.017	-0.014	-0.0016	0.28	-0.22	-0.00021	-0.2	0.18	-0.15	0.0084	-0.11	-0.058	0.095
politics3	-0.063	-0.088	-0.055	-0.014	-0.05	-0.07	-0.072	-0.16	-0.016	-0.029	0.0012	-0.017	-0.011	-0.094	0.034	-0.025	-0.041	1	-0.041	0.015	0.075	0.073	-0.073	-0.083	0.2	0.16	-0.072	-0.0093	0.09	0.12	0.12	-0.02
economy2	-0.017	-0.024	-0.05	0.072	0.0015	0.0013	-0.027	-0.047	-0.027	-0.039	0.003	-0.02	-0.043	-0.1	-0.041	-0.019	-0.032	-0.041	1	-0.001	-0.028	0.0066	0.032	-0.005	-0.03	-0.074	0.05	-0.021	-0.025	-0.052	-0.13	0.039
t_length_w	0.045	-0.17	0.052	-0.054	-0.059	-0.061	-0.067	-0.12	0.027	0.031	0.049	0.0015	0.037	0.12	0.068	-0.02	-0.017	0.015	-0.001	1	0.14	0.17	0.053	-0.012	-0.044	-0.053	0.022	-0.028	0.014	-0.028	0.15	0.12
a_length_s	-0.069	-0.11	0.18	0.032	0.014	0.023	0.02	-0.17	-0.0071	0.14	0.014	-0.04	0.041	0.43	0.07	-0.0074	-0.014	0.075	-0.028	0.14	1	0.82	0.057	-0.059	0.018	0.014	0.005	-0.0064	0.0035	-0.0046	0.32	-0.13
a_length_w	-0.048	-0.11	0.17	0.013	0.013	0.028	0.028	-0.083	0.028	0.12	0.00087	-0.015	0.017	0.38	0.081	0.01	-0.0016	0.073	0.0066	0.17	0.82	1	0.048	-0.078	0.055	0.045	-0.023	-0.0098	0.038	0.033	0.32	-0.083
a_neg	0.0066	-0.076	-0.028	-0.024	0.021	0.064	0.029	-0.13	0.13	0.18	-0.02	0.014	-0.064	0.019	0.083	-0.017	0.28	-0.073	0.032	0.053	0.057	0.048	1	-0.71	-0.14	-0.73	0.59	-0.4	-0.069	-0.46	0.00077	-0.054
a_neu	0.034	0.088	0.056	0.037	0.012	-0.023	-0.021	0.053	-0.087	-0.1	-0.018	-0.15	0.033	0.086	-0.11	-0.019	-0.22	-0.083	-0.005	-0.012	-0.059	-0.078	-0.71	1	-0.59	0.19	-0.43	0.5	-0.24	0.15	0.015	0.027
a_pos	-0.056	-0.037	-0.048	-0.025	-0.041	-0.042	-0.0038	0.074	-0.032	-0.062	0.048	0.19	0.026	-0.14	0.053	0.046	-0.00021	0.2	-0.03	-0.044	0.018	0.055	-0.14	-0.59	1	0.57	-0.068	-0.25	0.42	0.32	-0.022	0.019
a_compound	-0.055	0.0031	0.03	0.0023	-0.02	-0.039	-0.0095	0.15	-0.1	-0.17	0.041	0.06	0.057	-0.075	-0.048	0.058	-0.2	0.16	-0.074	-0.053	0.014	0.045	-0.73	0.19	0.57	1	-0.43	0.16	0.24	0.48	0.017	0.018
t_neg	0.0043	-0.067	-0.063	-0.045	0.018	0.065	0.0011	-0.077	0.1	0.055	-0.047	0.0062	-0.092	0.0068	0.091	-0.033	0.18	-0.072	0.05	0.022	0.005	-0.023	0.59	-0.43	-0.068	-0.43	1	-0.7	-0.1	-0.73	-0.04	0.025
t_neu	0.013	0.074	0.065	0.043	0.0082	-0.015	0.0046	0.084	-0.046	-0.01	0.036	-0.15	0.014	0.033	-0.056	-0.034	-0.15	-0.0093	-0.021	-0.028	-0.0064	-0.0098	-0.4	0.5	-0.25	0.16	-0.7	1	-0.64	0.09	-0.0093	-0.031
t_pos	-0.022	-0.031	-0.022	-0.011	-0.031	-0.048	-0.0075	-0.035	-0.046	-0.044	0.00027	0.2	0.079	-0.053	-0.02	0.082	0.0084	0.09	-0.025	0.014	0.0035	0.038	-0.069	-0.24	0.42	0.24	-0.1	-0.64	1	0.66	0.056	0.016
t_compound	-0.012	0.047	0.02	0.017	-0.03	-0.075	-0.013	0.037	-0.098	-0.075	0.044	0.11	0.12	-0.055	-0.067	0.041	-0.11	0.12	-0.052	-0.028	-0.0046	0.033	-0.46	0.15	0.32	0.48	-0.73	0.09	0.66	1	0.058	0.015
authers_num	-0.023	-0.12	0.24	-0.015	0.078	0.099	0.075	-0.33	0.063	0.19	-0.012	-0.09	0.037	0.42	0.044	0.015	-0.058	0.12	-0.13	0.15	0.32	0.32	0.00077	0.015	-0.022	0.017	-0.04	-0.0093	0.056	0.058	1	0.025
date_days	0.096	-0.017	0.07	0.0058	-0.029	0.008	0.12	-0.047	0.061	-0.12	0.1	-0.033	-0.062	-0.071	0.058	-0.051	0.095	-0.02	0.039	0.12	-0.13	-0.083	-0.054	0.027	0.019	0.018	0.025	-0.031	0.016	0.015	0.025	1

Baseline Model

Fill null values with zero



score of the train

0.145

0.142

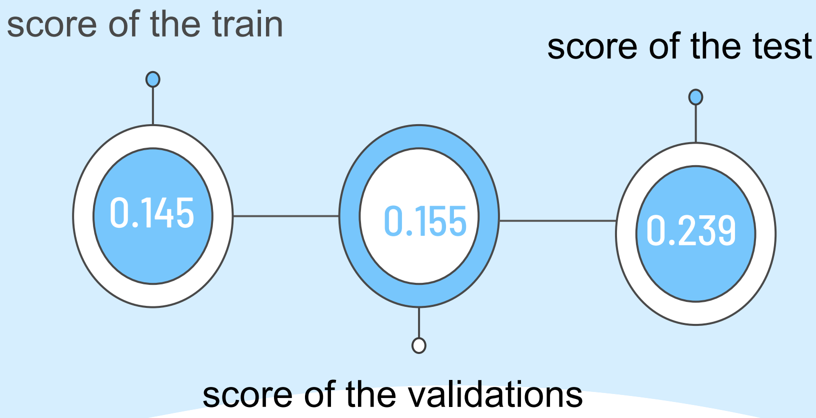
0.233

score of the test

score of the validations

Data prep & Experimentation

Fill null values with mode/mean



score of the train

0.145

0.155

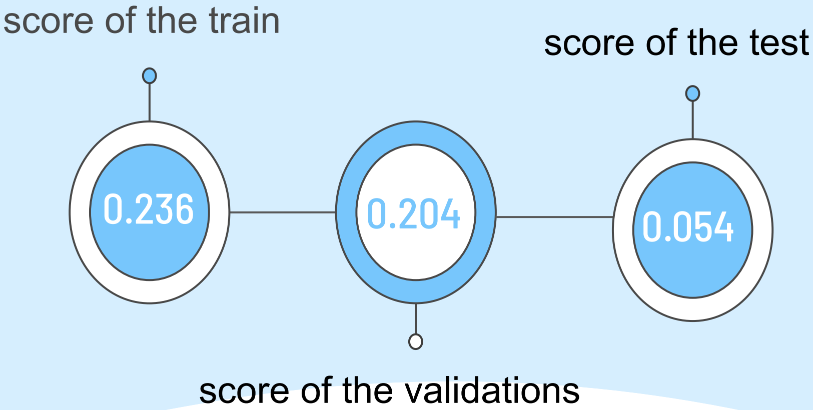
score of the validations

score of the test

0.239

Data prep & Experimentation

Feature Engineering



score of the train

0.236

0.204

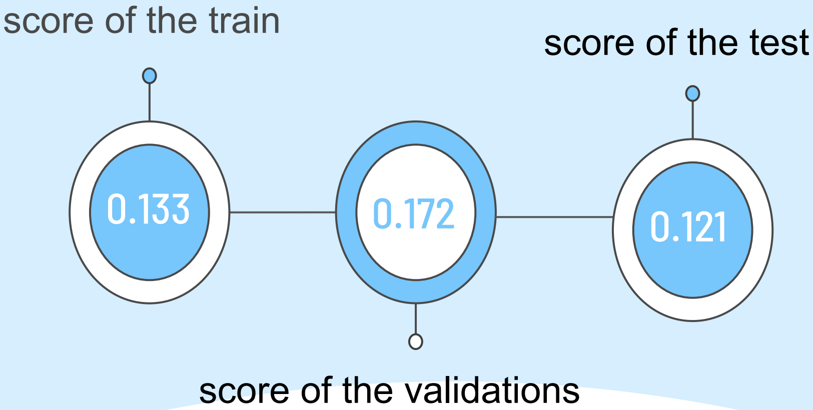
score of the validations

score of the test

0.054

Data prep & Experimentation

Ridge



score of the train

0.133

0.172

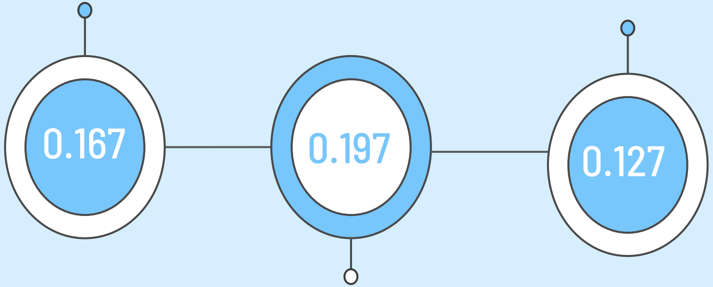
score of the test

0.121

score of the validations

Data prep & Experimentation

Lasso



score of the train

0.167

score of the validations

0.197

score of the test

0.127

Based on the unsatisfactory results in the linear regression, we decided to for the problem as a classification problem.

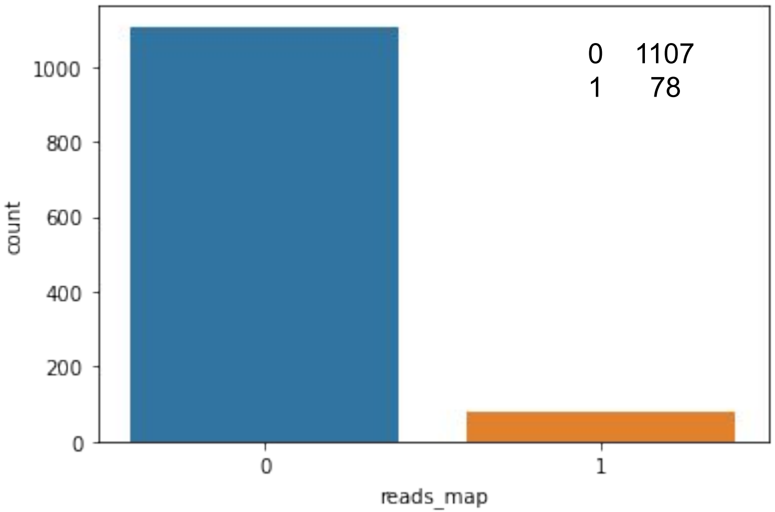
06

Classification

Data Distribution After Mapping

Average #reads = 57

Popular if Reads >= 200



```
0    1107
1      78
```

Discretization!

grouping continuous data into a
number of intervals or bins

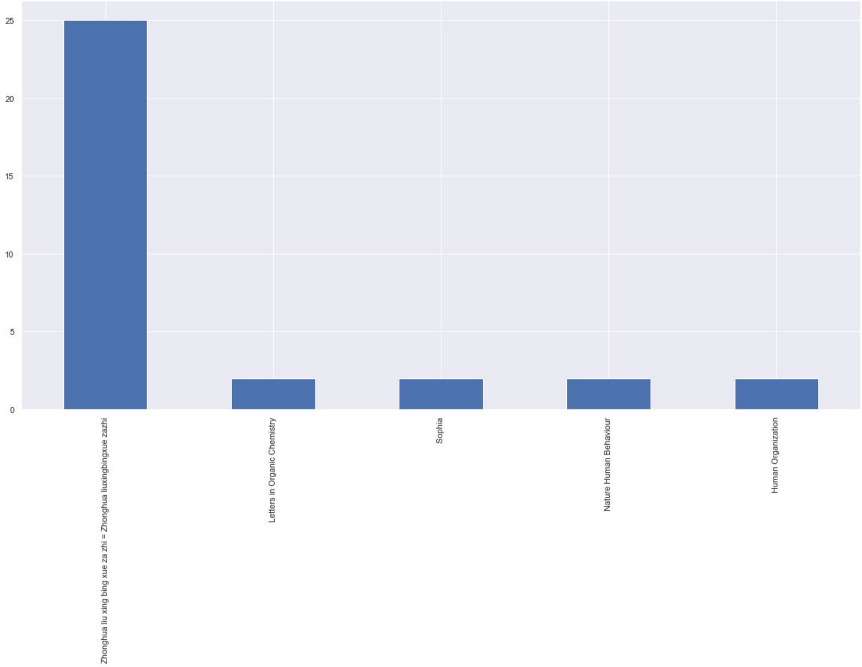
Popular Articles numerical description

Count : 78

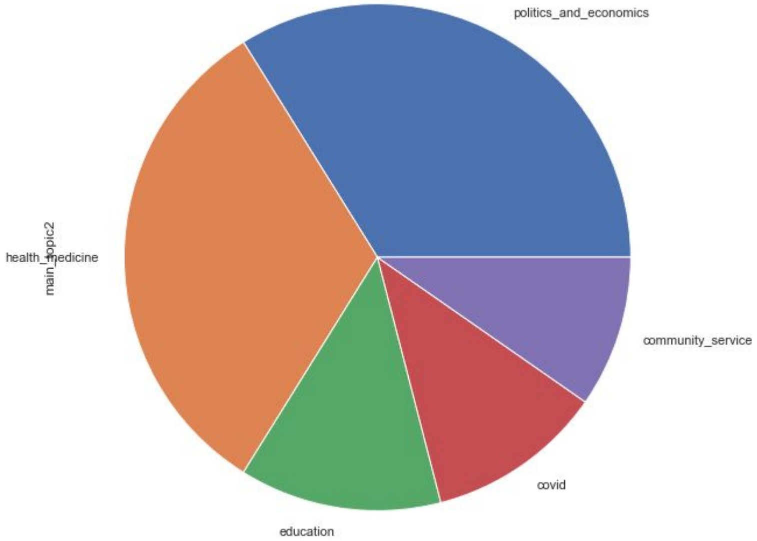
Mean Reads : 398

Max Reads : 936

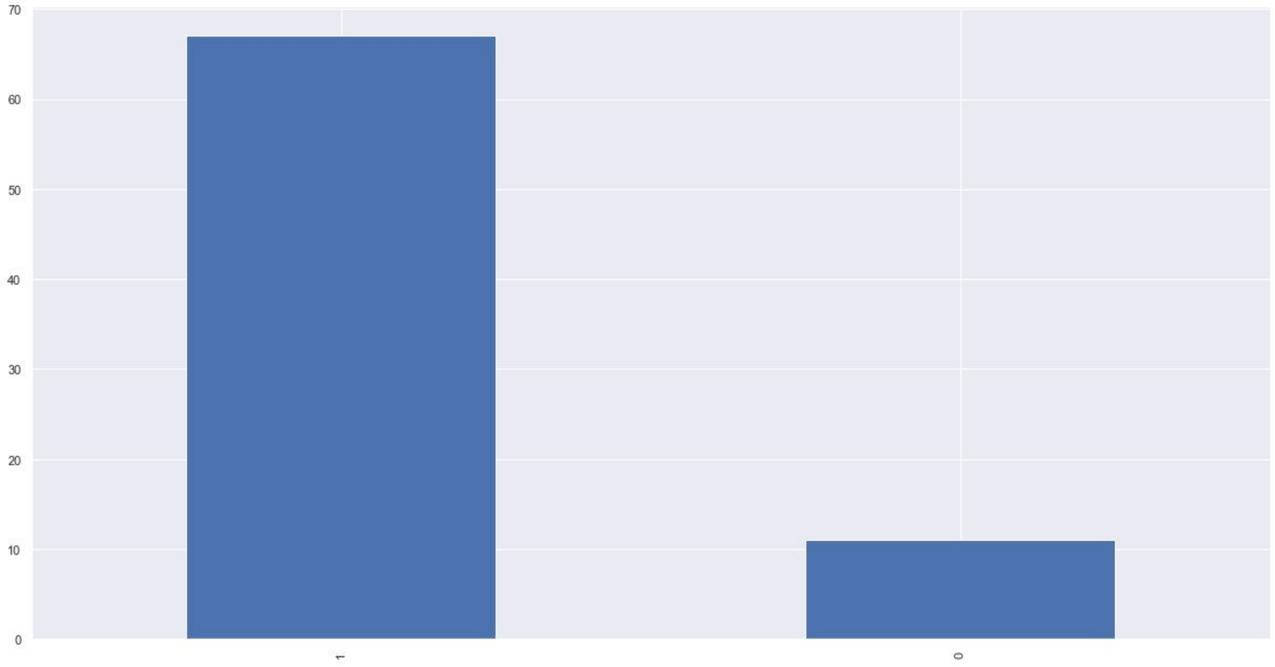
EDA : journal


25
20
15
10
5
0
Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi
Letters in Organic Chemistry
Sophia
Nature Human Behaviour
Human Organization

EDA : Popular Topics



EDA : Full Text availability



Experiments

Uninclude Features

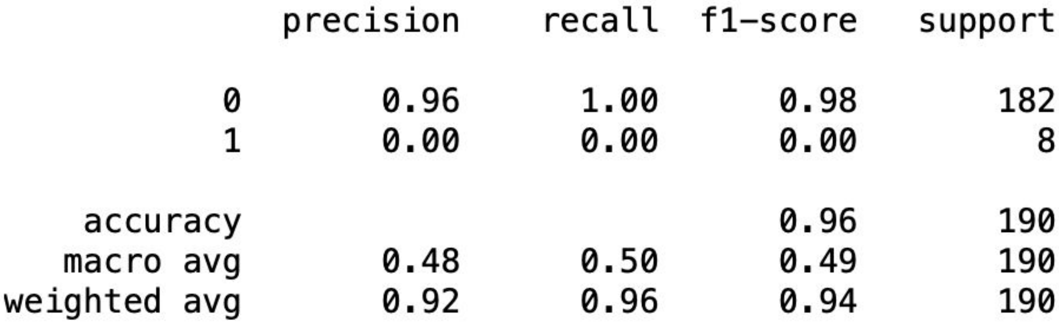
Research Interest

Calculated from other features

Citation

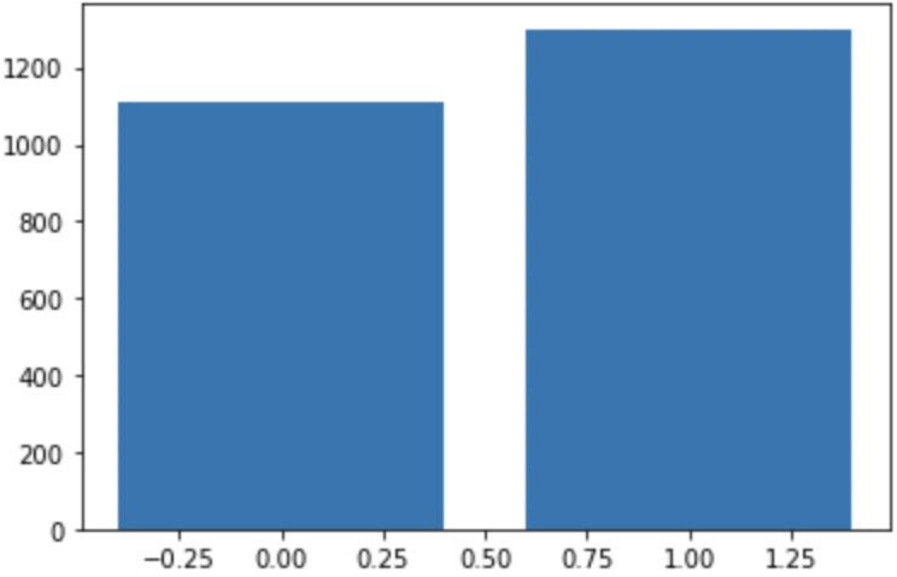
Popular articles are usually cited

Result Before Balancing Techniques



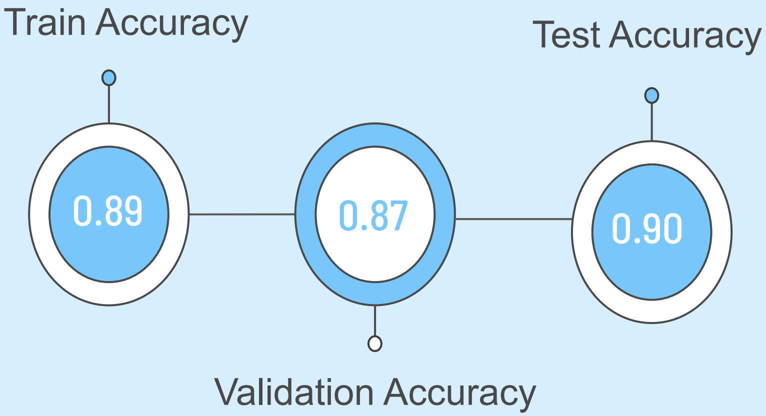
```
              precision    recall  f1-score   support

           0       0.96      1.00      0.98       182
           1       0.00      0.00      0.00         8

    accuracy                           0.96       190
   macro avg       0.48      0.50      0.49       190
weighted avg       0.92      0.96      0.94       190
```

balanced data : BorderLine SMOTE



Class=0, n=1107 (45.991%)
Class=1, n=1300 (54.009%)

Results



Train Accuracy

0.89

0.87

Validation Accuracy

Test Accuracy

0.90

Conclusion

We found 9 main topics in the articles, predicting the reads with the available data isn’t possible, more data needs to be collected, but it’s possible to classify articles according to popularity.

Thanks!

Do you have any questions?