The Comparison of Local and Global Early Fake News Detection Methods

Maciej Pawlikowski, Hubert Ruczyński,
Bartosz Siński, Adrian Stańdo

01

Project goals

Main goals

1. Comparison of different **topic detection** models.
2. Comparison of **fake news detection** methods.
3. Introduction of **local** fake news detection methods.
4. Evaluation of the **local** approach, and comparison to corresponding **global** solutions.
5. Exploration of models **differences** between the two strategies with the usage of **XAI**.

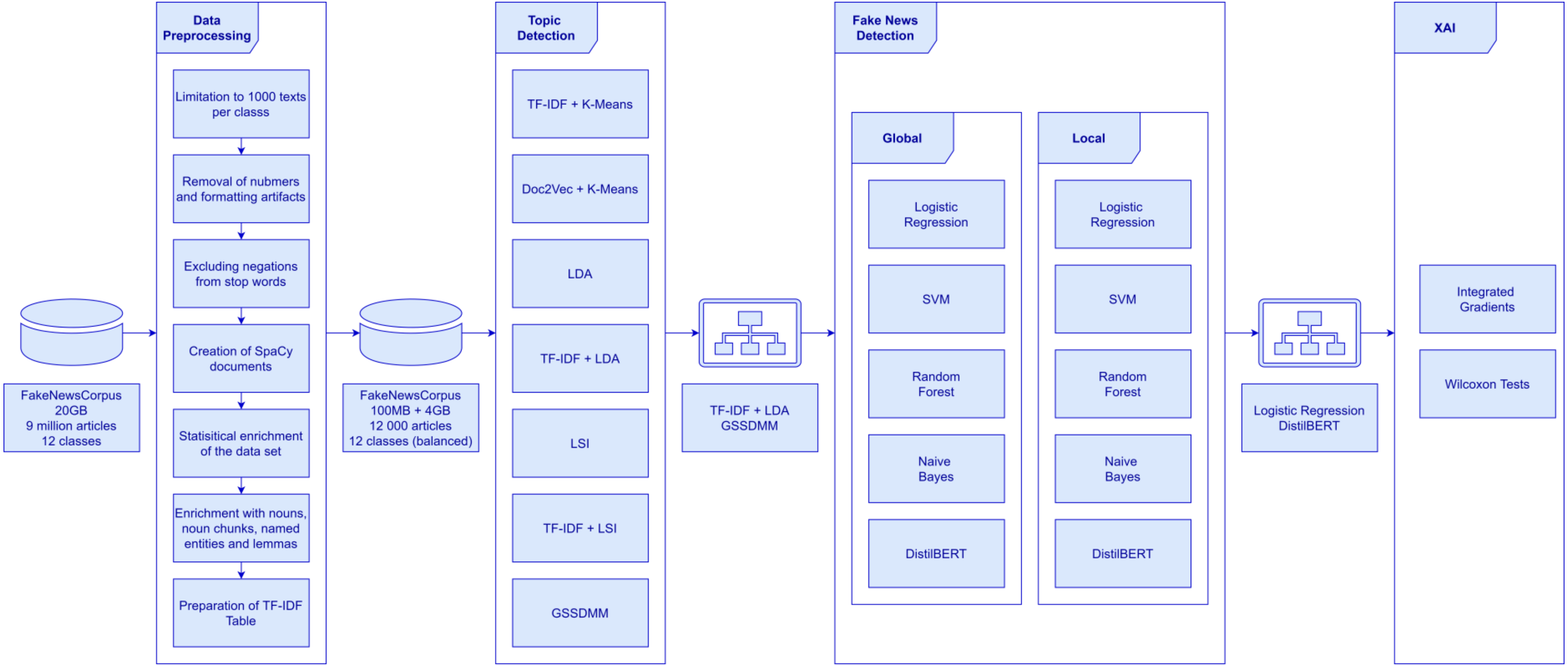
02

Proposed solution

Proposed solution

1. We used SOTA topic detection tools, namely: **K-Means**, **LDA**, **LSI**, and **GSSDMM**, based on **lemmas** and **noun chunks**.
2. We evaluated the clustering quality with SOTA approaches mentioned in the Literature Review, namely**: Coherence Score, Silhouette Score, and Calinski-Harabsz Index**.
3. We prepared a **train-test split** with regard to particular clusters (ex. 70% - 30%: training – testing).
4. We prepared a few (5) SOTA solutions for fake news detection methods, namely: **Logistic Regression**, **SVM**, **Random Forest**, **Naive Bayes**, and **DistilBERT**.
5. We trained and evaluated them on a **global** split.
6. We trained and evaluated them on **local** clusters.
7. We compared the results of global and local models in terms of **performance**.
8. We used **eXplainable AI (XAI)** methods to discover the most important indicators for global and local methods.

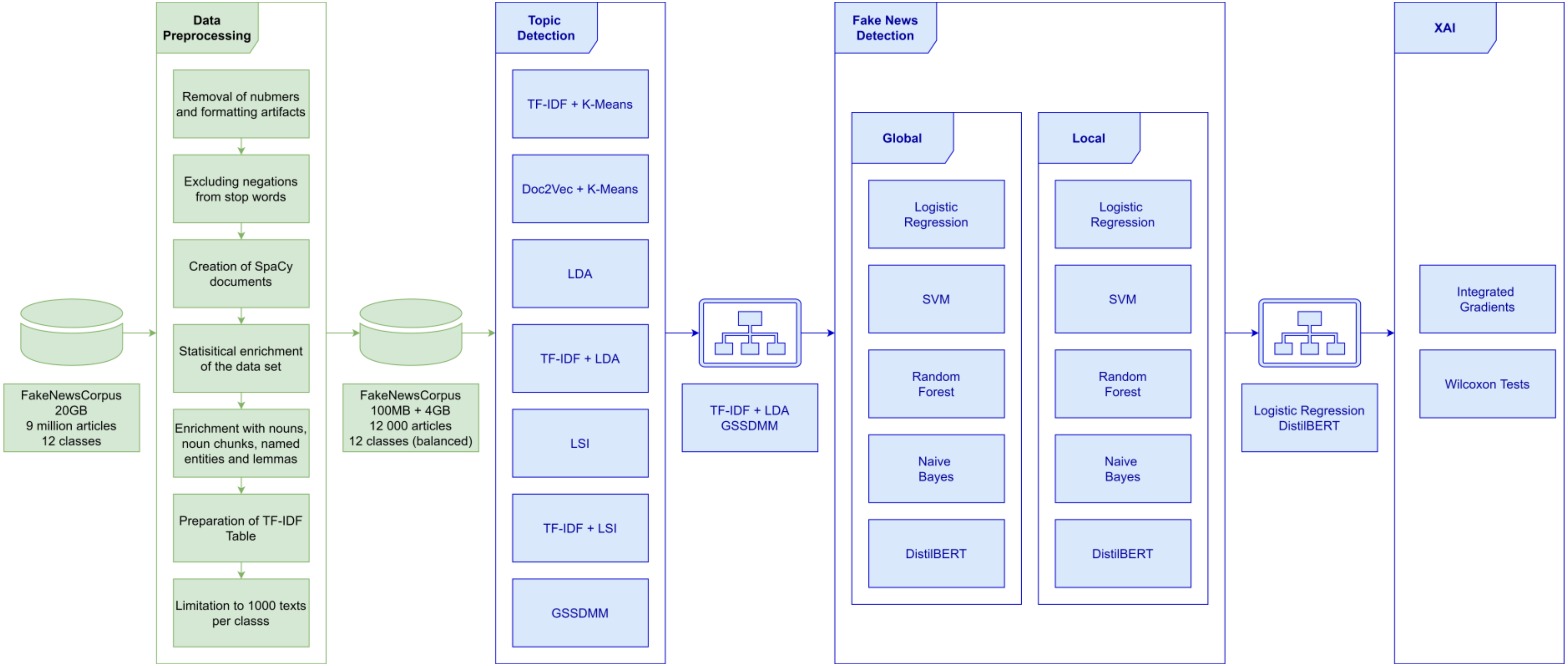
Proposed solution



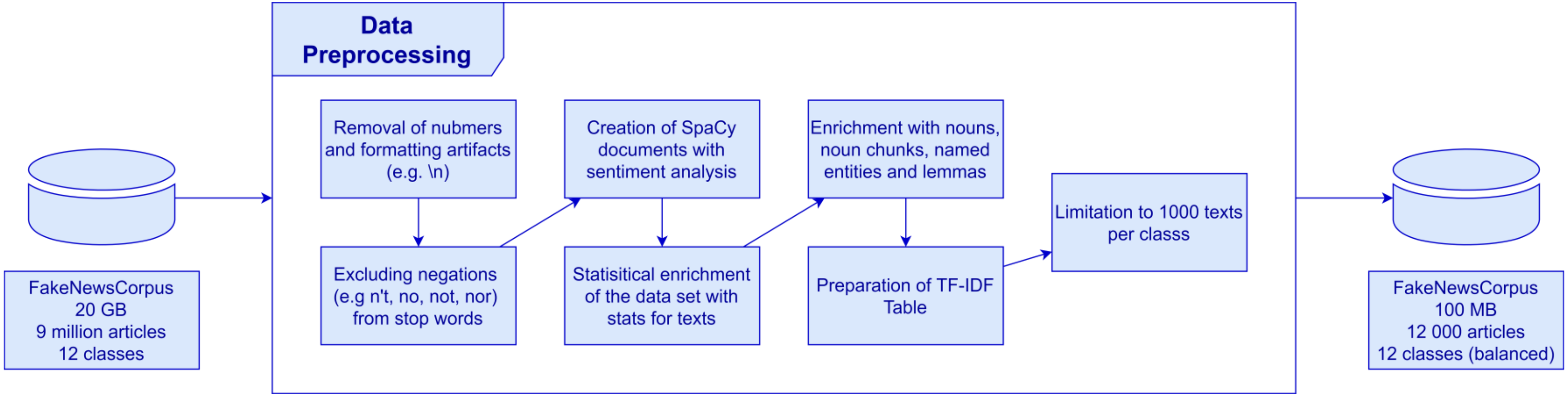
03

Data preprocessing + EDA

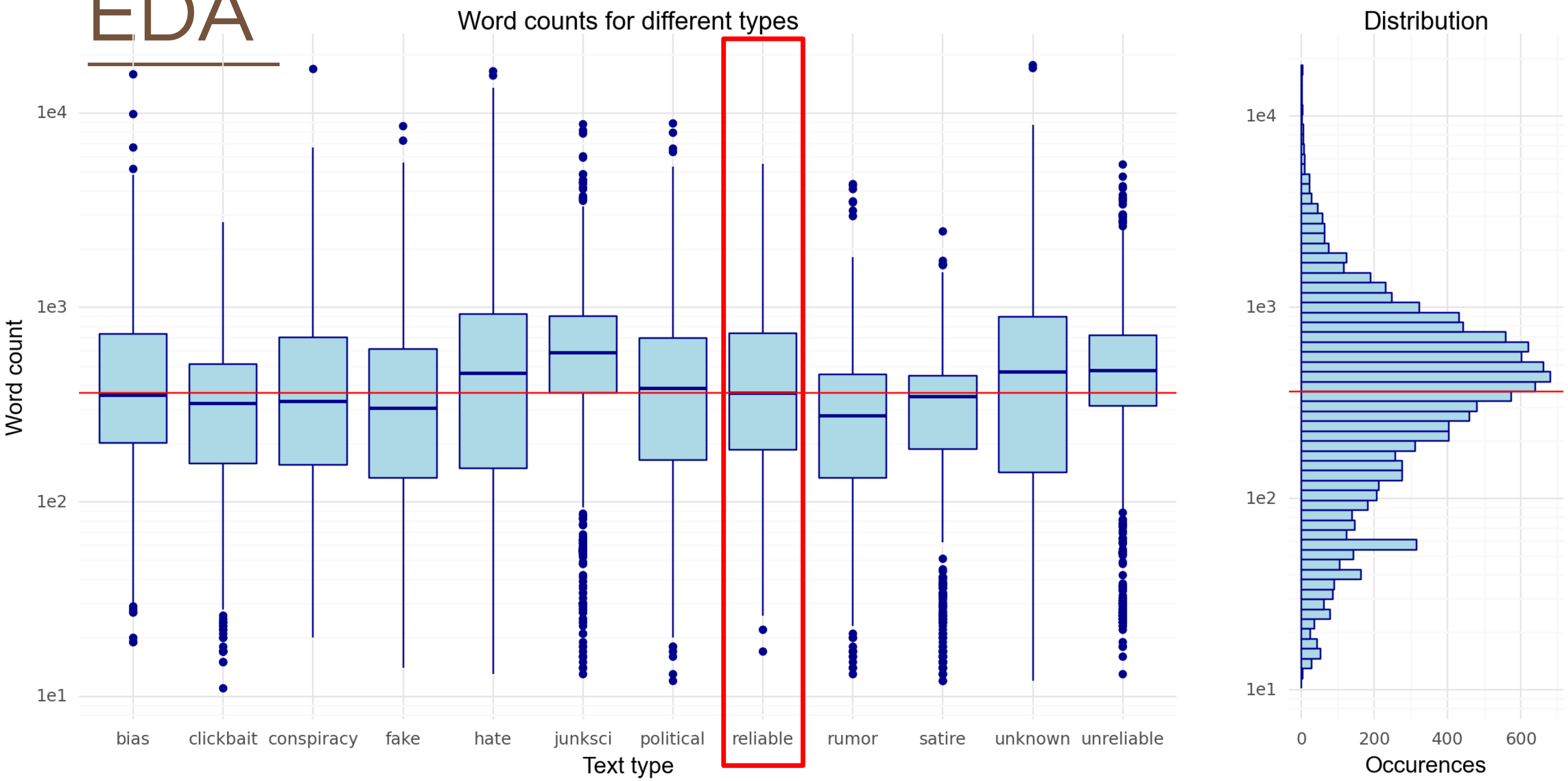
Proposed solution



Preprocessing design

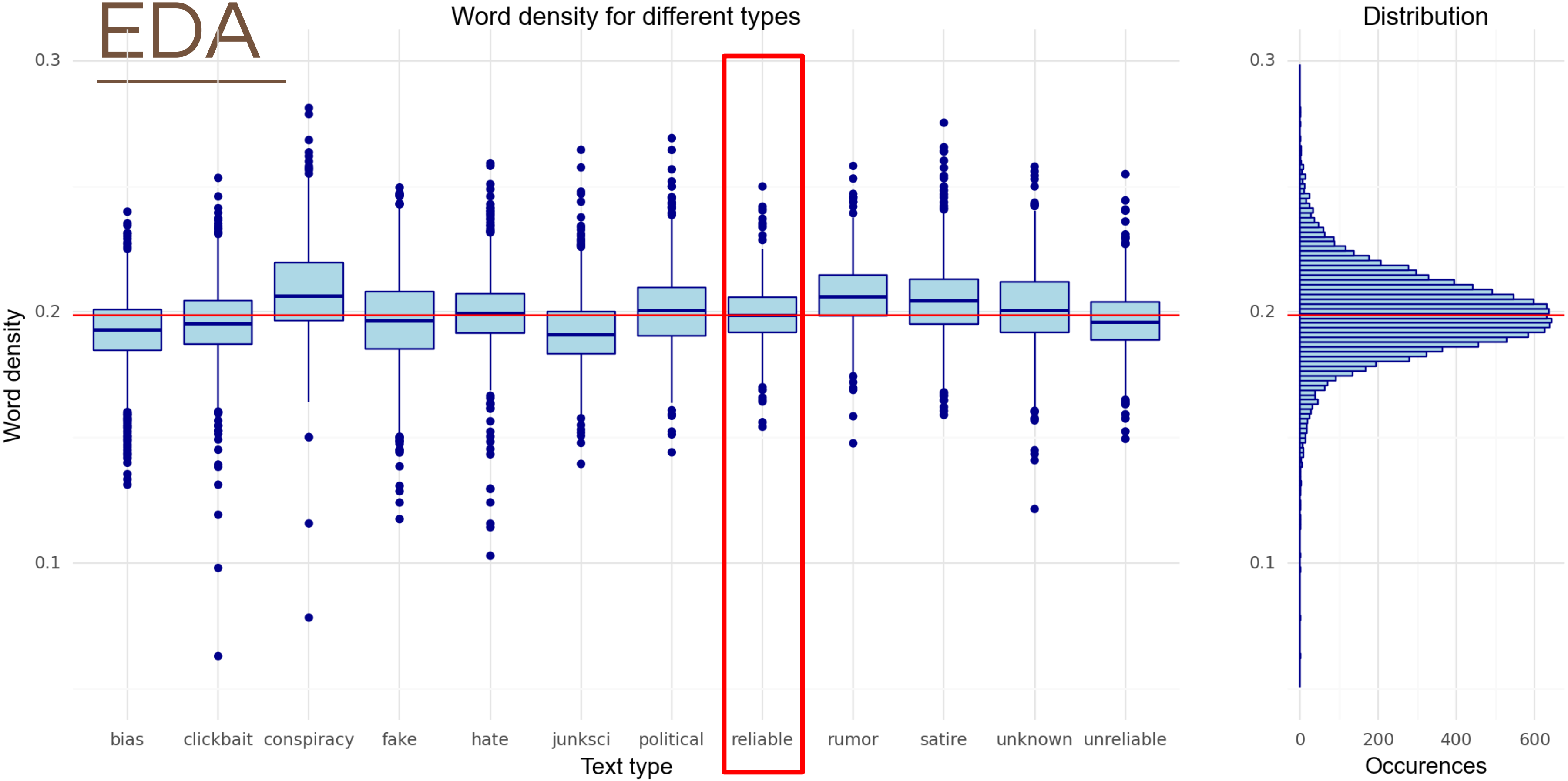


EDA



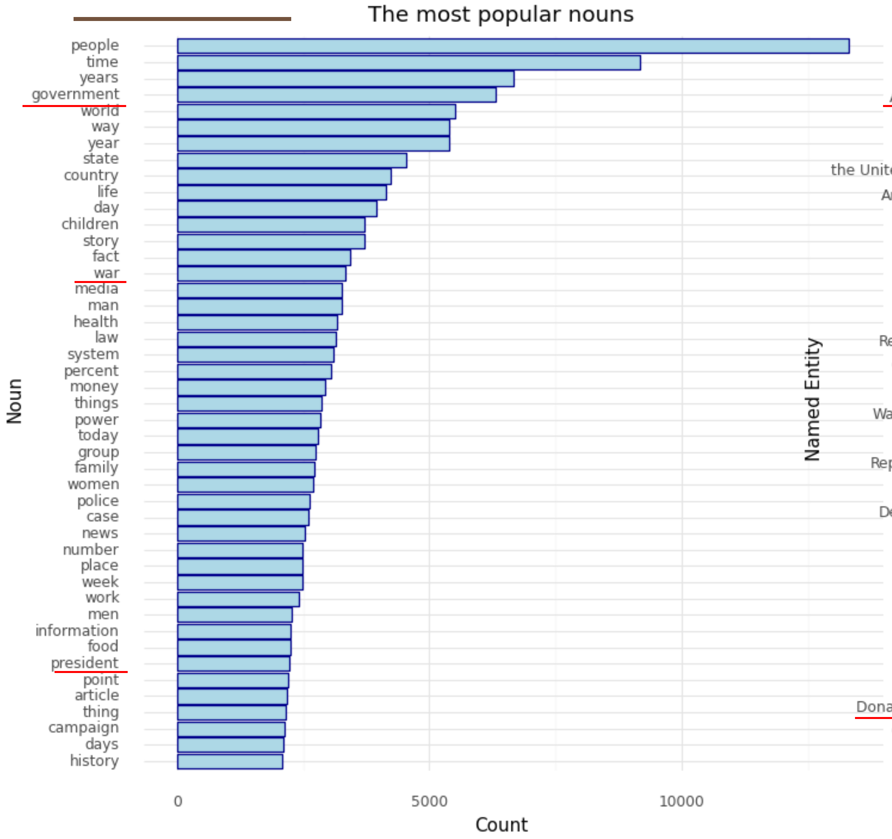
Word counts for different types
Word count
1e4
1e3
1e2
1e1
bias
clickbait
conspiracy
fake
hate
junksci
political
reliable
rumor
satire
unknown
unreliable
Text type
Distribution
1e4
1e3
1e2
1e1
0
200
400
600
Occurences

EDA

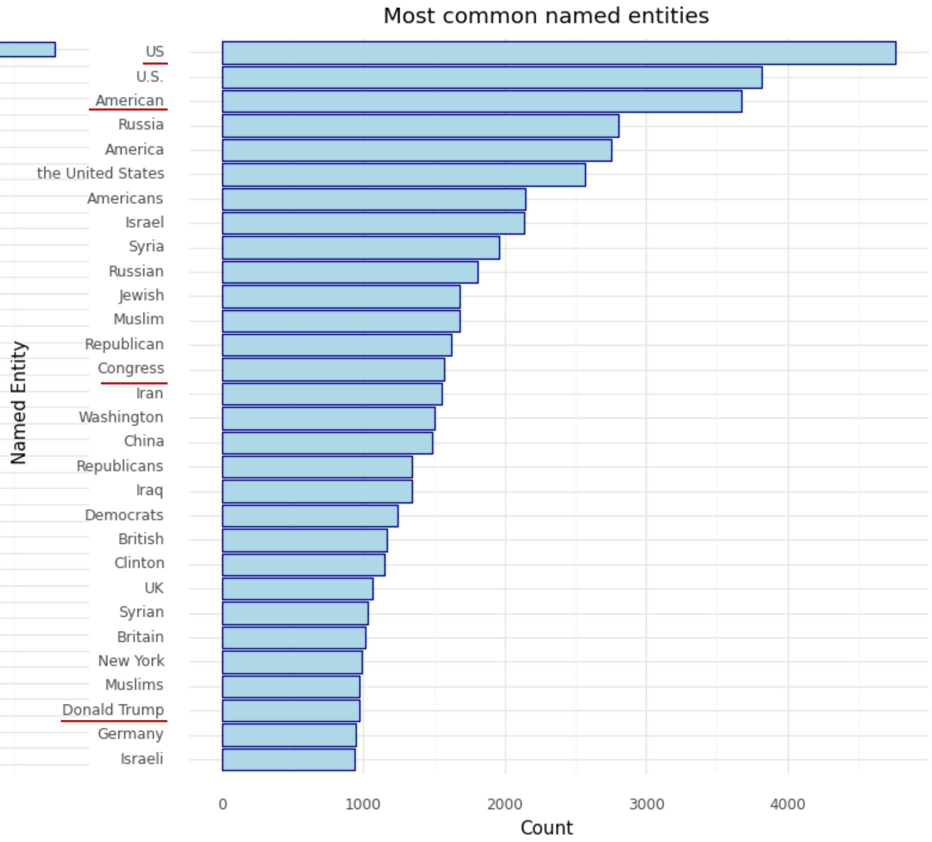


Word density for different types
Distribution
Word density
0.3
0.2
0.1
bias
clickbait
conspiracy
fake
hate
junksci
political
reliable
rumor
satire
unknown
unreliable
Text type
0
200
400
600
Occurences

EDA

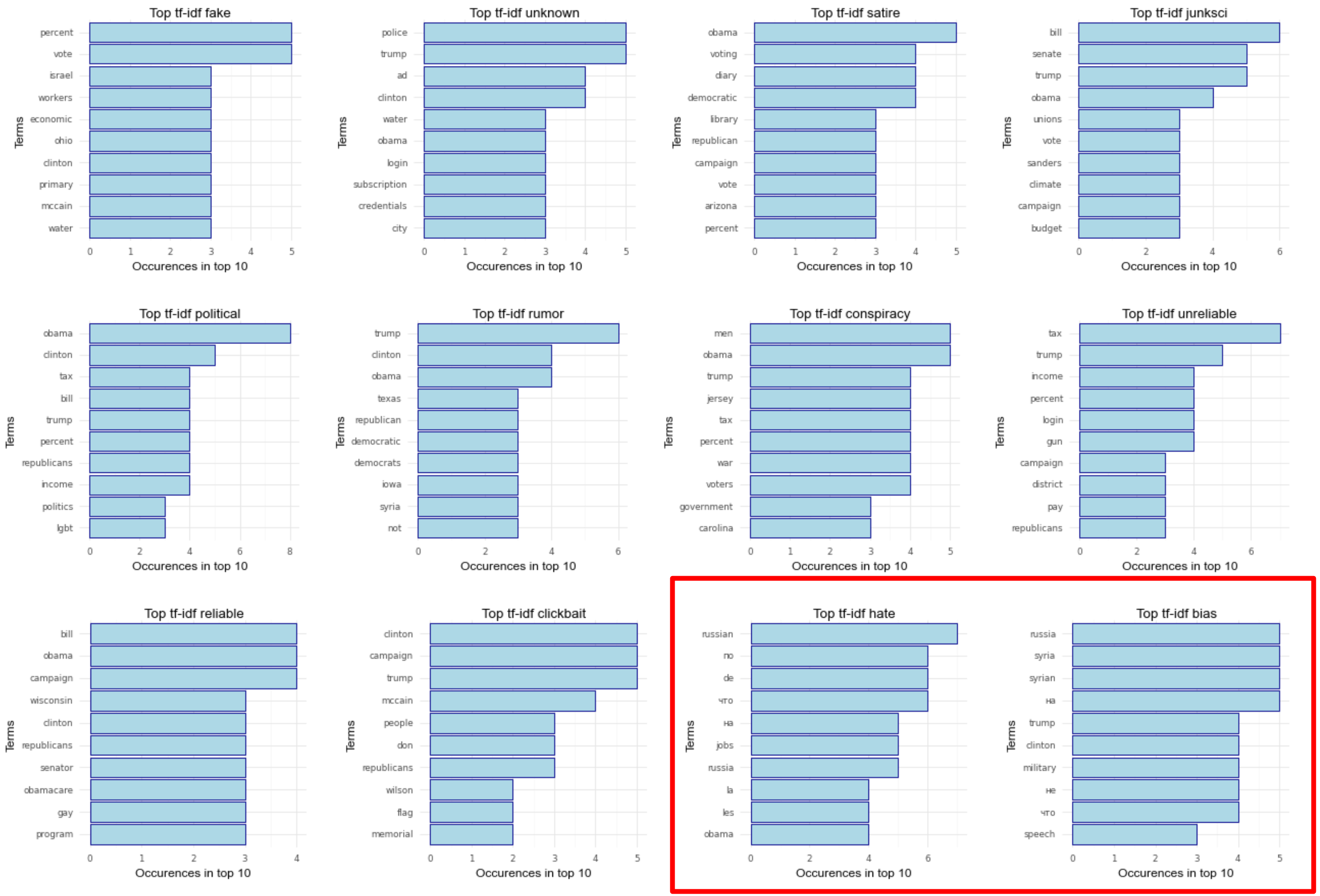


The most popular nouns
people
time
years
government
world
way
year
state
country
life
day
children
story
fact
war
media
man
health
law
system
percent
money
things
power
today
group
family
women
police
case
news
number
place
week
work
men
information
food
president
point
article
thing
campaign
days
history
Noun
0
5000
10000
Count

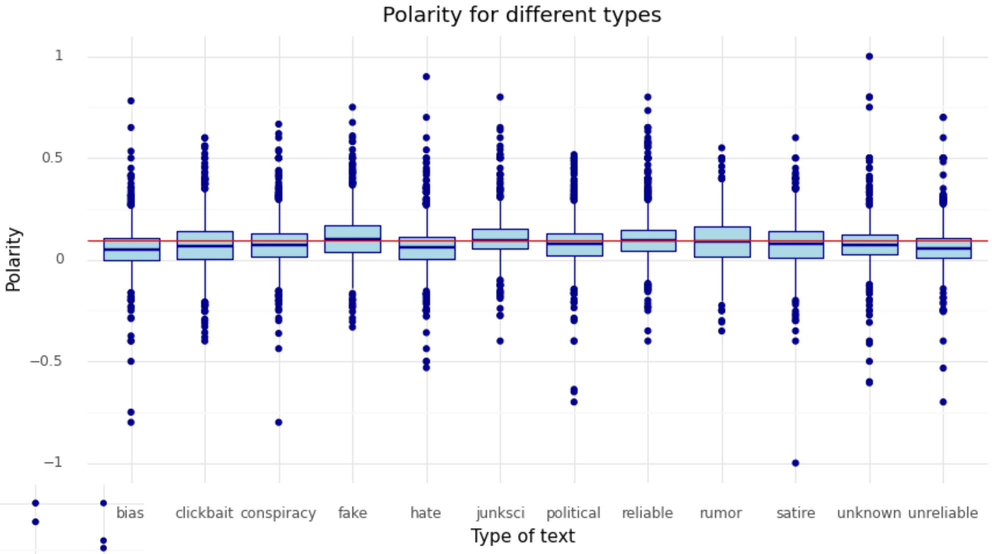


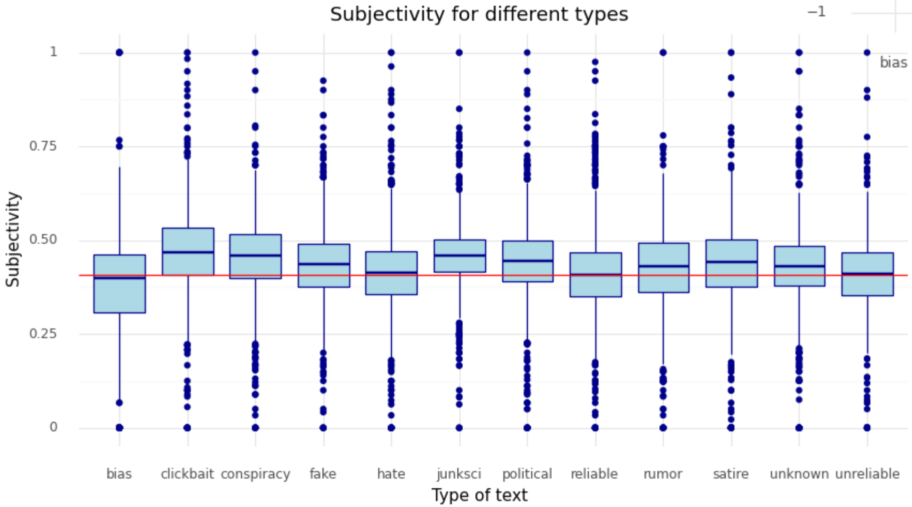
Most common named entities
US
U.S.
American
Russia
America
the United States
Americans
Israel
Syria
Russian
Jewish
Muslim
Republican
Congress
Iran
Washington
China
Republicans
Iraq
Democrats
British
Clinton
UK
Syrian
Britain
New York
Muslims
Donald Trump
Germany
Israeli
Named Entity
0
1000
2000
3000
4000
Count

EDA


Top tf-idf fake
percent
vote
israel
workers
economic
ohio
clinton
primary
mccain
water
Top tf-idf unknown
police
trump
ad
clinton
water
obama
login
subscription
credentials
city
Top tf-idf satire
obama
voting
diary
democratic
library
republican
campaign
vote
arizona
percent
Top tf-idf junksci
bill
senate
trump
obama
unions
vote
sanders
climate
campaign
budget
Top tf-idf political
obama
clinton
tax
bill
trump
percent
republicans
income
politics
lgbt
Top tf-idf rumor
trump
clinton
obama
texas
republican
democratic
democrats
iowa
syria
not
Top tf-idf conspiracy
men
obama
trump
jersey
tax
percent
war
voters
government
carolina
Top tf-idf unreliable
tax
trump
income
percent
login
gun
campaign
district
pay
republicans
Top tf-idf reliable
bill
obama
campaign
wisconsin
clinton
republicans
senator
obamacare
gay
program
Top tf-idf clickbait
clinton
campaign
trump
mccain
people
don
republicans
wilson
flag
memorial
Top tf-idf hate
russian
по
de
что
на
jobs
russia
la
les
obama
Top tf-idf bias
russia
syria
syrian
на
trump
clinton
military
не
что
speech
Terms
Occurences in top 10

EDA

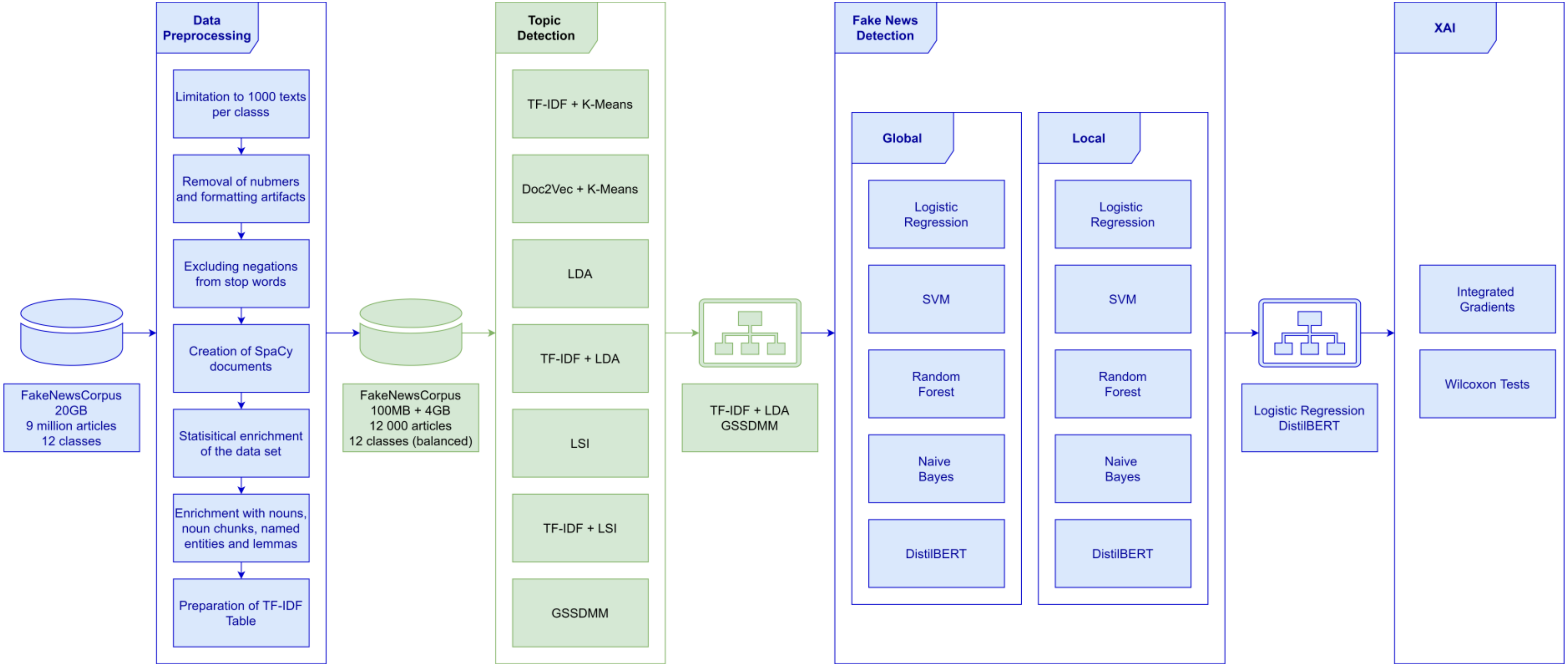

Polarity for different types
Polarity
1
0.5
0
−0.5
−1
bias
clickbait
conspiracy
fake
hate
junksci
political
reliable
rumor
satire
unknown
unreliable
Type of text


Subjectivity for different types
Subjectivity
1
0.75
0.50
0.25
0
bias
clickbait
conspiracy
fake
hate
junksci
political
reliable
rumor
satire
unknown
unreliable
Type of text

04

Topic Detection / Clustering

Experimental design



Used Methods

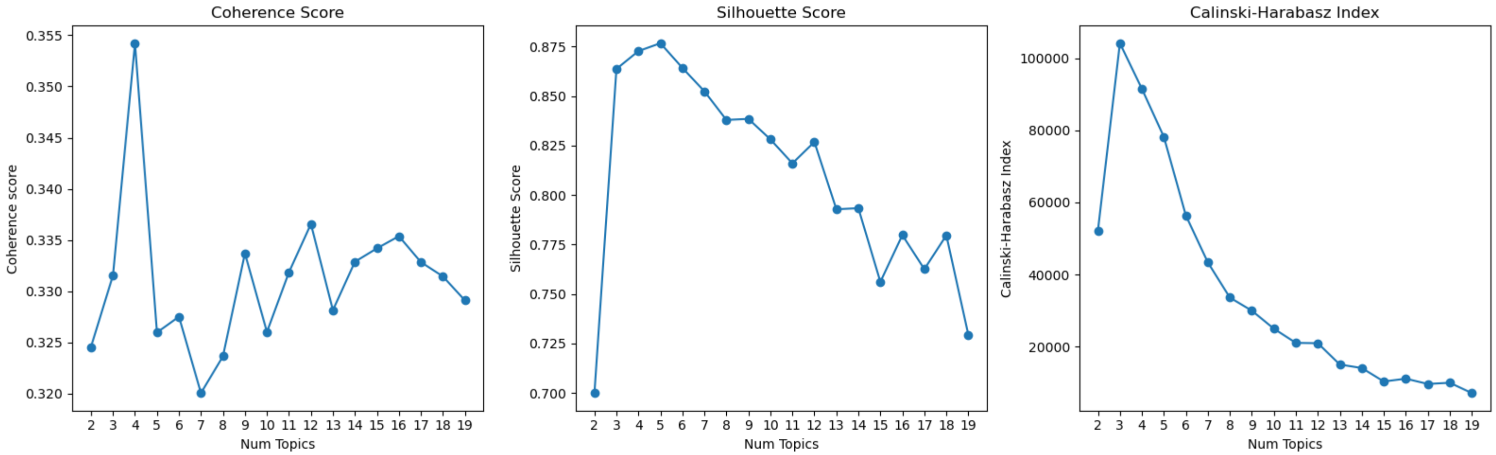
1. Classical clustering method with two document representations:
   a. Doc2Vec + k-means
   b. TF-IDF + k-means

2. Topic modeling method:
   c. Latent Dirichlet Allocation (LDA)
   d. Latent Semantic Indexing (LSI)
   e. Gibbs Sampling Dirichlet Multinomial Mixture (GSSDMM)

3. Clustering performed on extracted noun chunks and lemmas.

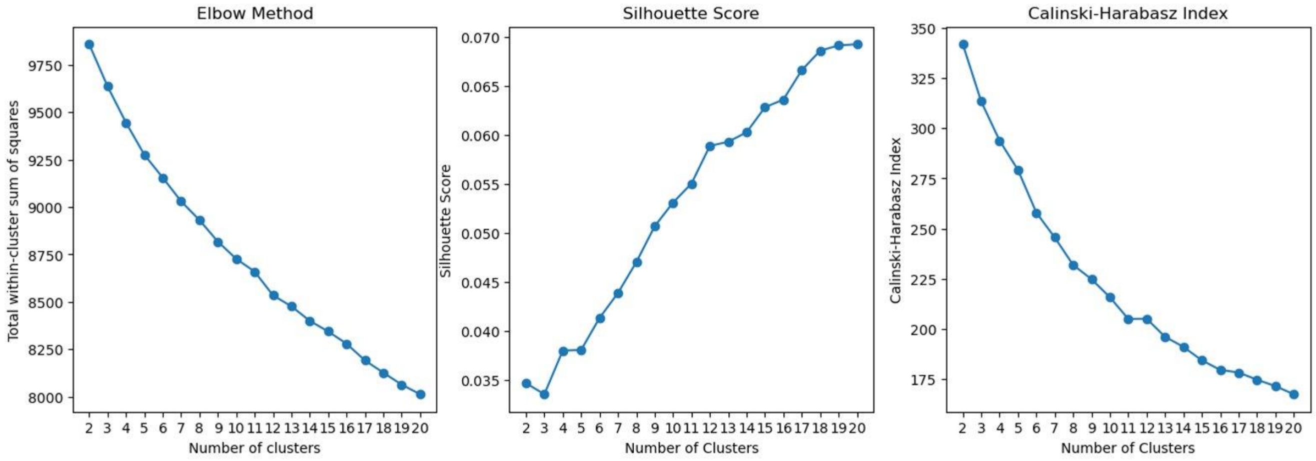
Choosing number of clusters

Metrics for TF-IDF + LDA clustering



Choosing number of clusters

Metrics for TF-IDF + k-means clustering



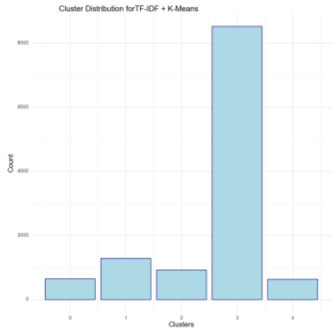
Results - lemmas

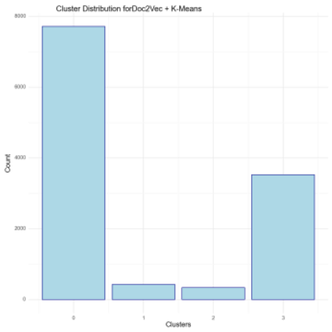
Clustering Algorithm	Silhouette Score	Calinski-Harabasz Score
TF-IDF + K-Means	0.038	293.9
Doc2Vec + K-Means	0.134	449.6
LDA	0.607	18668.8
TF-IDF + LDA	0.873	91490.7
LSI	-0.320	49.5
TF-IDF + LSI	0.469	1655.9
GSSDMM	0.714	529.4

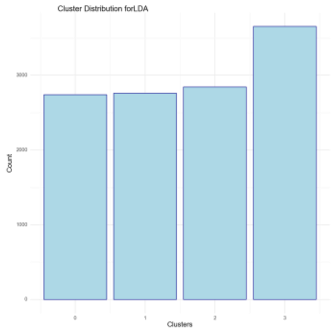
Results – noun chunks

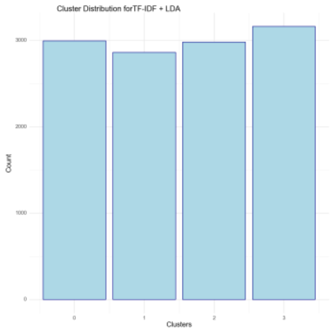
Clustering Algorithm	Silhouette Score	Calinski-Harabasz Score
TF-IDF + K-Means	0.067	314.1
Doc2Vec + K-Means	0.386	3473.24
LDA	0.883	110432.1
TF-IDF + LDA	0.929	323993.32
LSI	-0.512	132.0
TF-IDF + LSI	-0.290	393.5
GSSDMM	0.867	15681.4

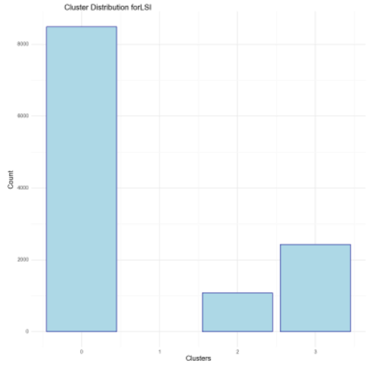
Results

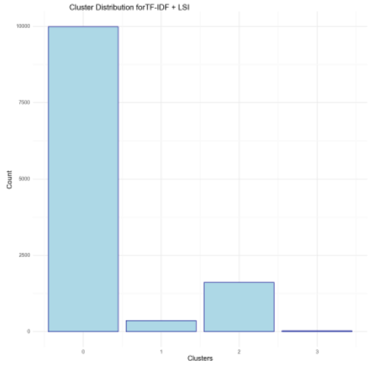

Cluster Distribution forTF-IDF + K-Means
Count
Clusters

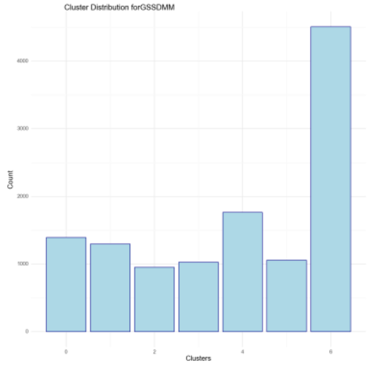

Cluster Distribution forDoc2Vec + K-Means
Count
Clusters


Cluster Distribution forLDA
Count
Clusters


Cluster Distribution forTF-IDF + LDA
Count
Clusters


Cluster Distribution forLSI
Count
Clusters


Cluster Distribution forTF-IDF + LSI
Count
Clusters


Cluster Distribution forGSSDMM
Count
Clusters

Results

Diving into GSDMM clusters.

Most frequent nouns in cluster 4:

1. Trump
2. Obama
3. people
4. Congress
5. Republicans
6. America
7. Donald Trump
8. Democrats
9. Clinton
10. Hillary Clinton

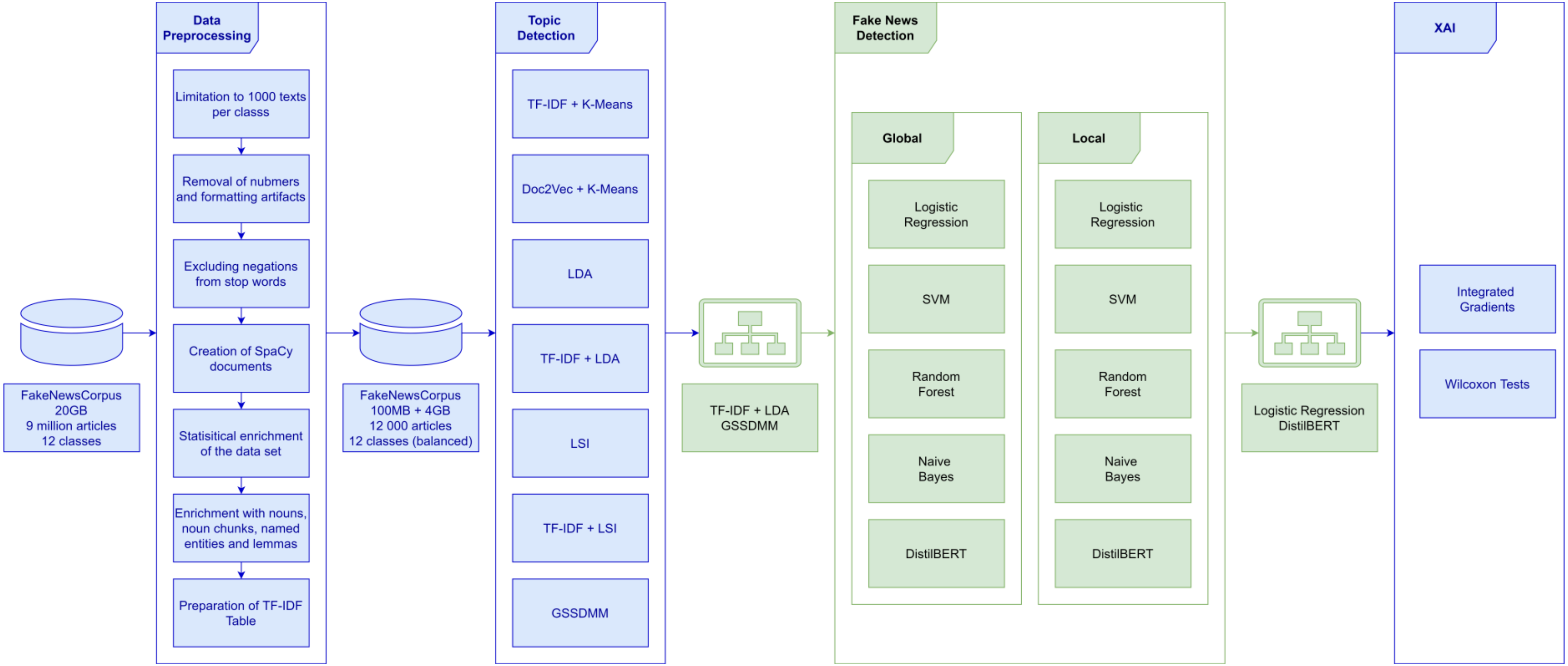
Most frequent nouns in cluster 0:

1. Russia
2. Syria
3. Israel
4. people
5. Iran
6. Iraq
7. Europe
8. China
9. Britain
10. ISIS

05

Fake News Detection

Experimental design



Tokenization

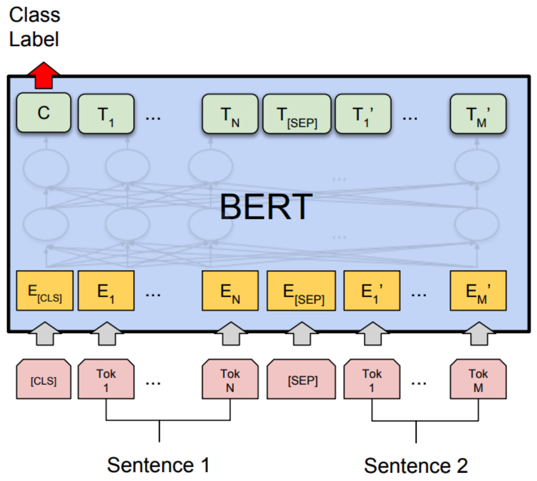
- Lowering the case of the entire string
- Adding special tokens to mark the start and end of the string
- Splits uncommon words into several tokens

  "tokenizer" -> "token" + "##izer"

- Punktuation removal
- Separating input text into individual sentences
- Removal of unnecessary whitespaces
- Padding and truncating

Model choice

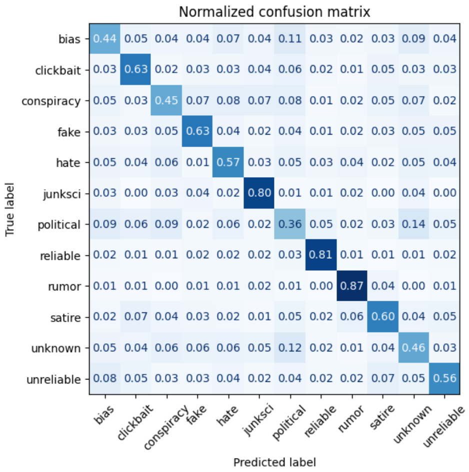
1. Statistical models
    - Logistic regression
    - Support Vector Machines
    - Random Forest
    - Naive Bayess
2. Transformers
    - DistilBERT



Results

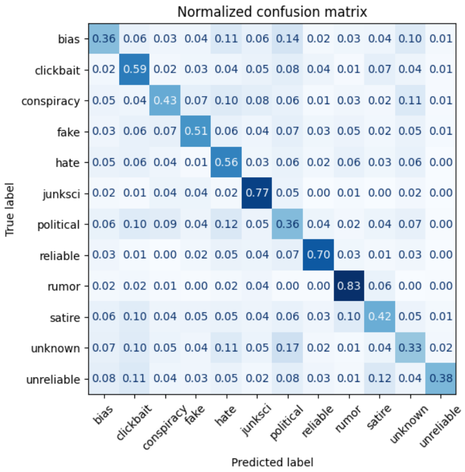
Before topic detection

Logistic regression



Accuracy = 0.598

SVM

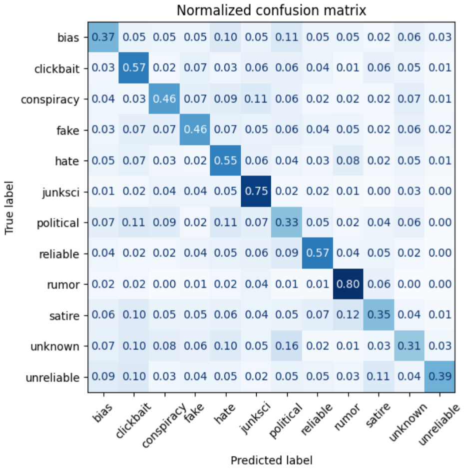


Accuracy = 0.518

Results

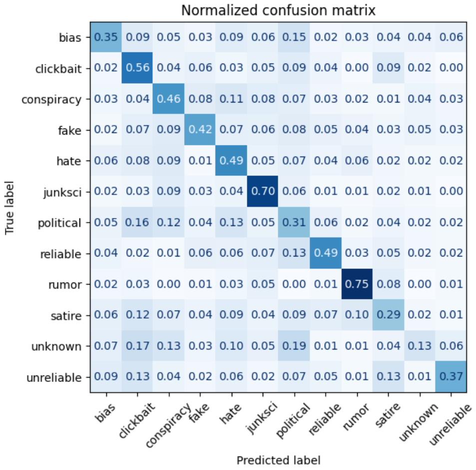
Before topic detection

Random forest



Accuracy = 0.491

Naive Bayes

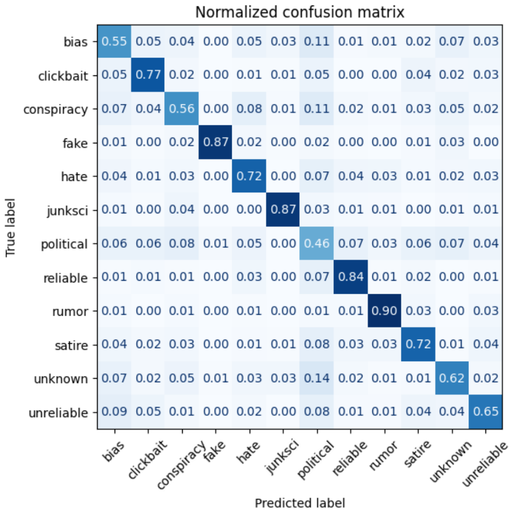


Accuracy = 0.444

Results

Before topic detection

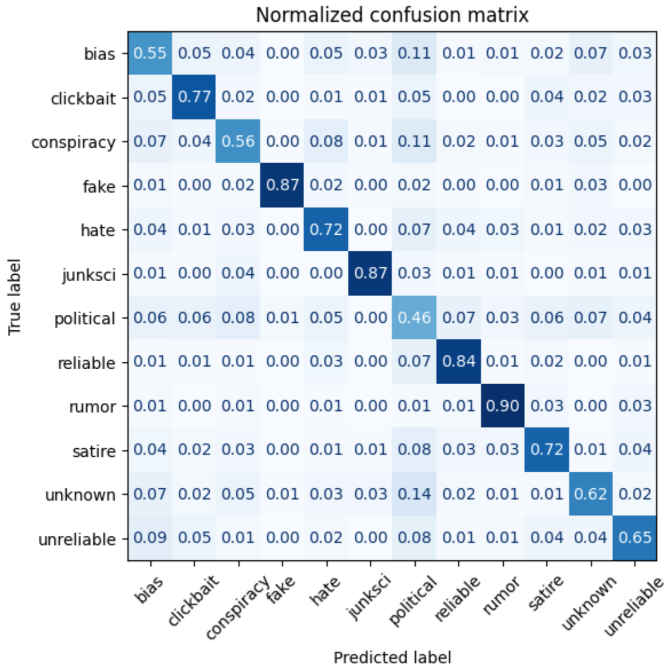
DistilBert



Accuracy = 0.710

Results

Clustering results



Weighted accuracy = accuracy * size / sum(size)

	bias	clickbait	conspiracy	fake	hate	junksci	political	reliable	rumor	satire	unknown	unreliable	accuracy	accuracy w/o political	size
0	0.44	0.61	0.41	0.49	0.61	0.61	0.38	0.74	0.77	0.51	0.41	0.50	0.54	0.58	2995
1	0.29	0.60	0.46	0.42	0.36	0.74	0.28	0.68	0.78	0.49	0.40	0.46	0.50	0.55	2862
2	0.47	0.58	0.45	0.53	0.52	0.63	0.33	0.68	0.79	0.56	0.42	0.51	0.53	0.59	2981
3	0.29	0.62	0.38	0.57	0.68	0.75	0.28	0.76	0.81	0.38	0.27	0.46	0.52	0.59	3162

Results

LDA clustering

Logistic regression

Weighted accuracy = 0.523

	bias	clickbait	conspiracy	fake	hate	junksci	political	reliable	rumor	satire	unknown	unreliable	accuracy	accuracy without political	size
0	0.44	0.61	0.41	0.49	0.61	0.61	0.38	0.74	0.77	0.51	0.41	0.50	0.54	0.58	2995
1	0.29	0.60	0.46	0.42	0.36	0.74	0.28	0.68	0.78	0.49	0.40	0.46	0.50	0.55	2862
2	0.47	0.58	0.45	0.53	0.52	0.63	0.33	0.68	0.79	0.56	0.42	0.51	0.53	0.59	2981
3	0.29	0.62	0.38	0.57	0.68	0.75	0.28	0.76	0.81	0.38	0.27	0.46	0.52	0.59	3162

DistilBERT

Weighted accuracy = 0.619

	bias	clickbait	conspiracy	fake	hate	junksci	political	reliable	rumor	satire	unknown	unreliable	accuracy	accuracy without political	size
0	0.53	0.70	0.53	0.92	0.71	0.78	0.52	0.60	0.87	0.54	0.49	0.63	0.65	0.70	2995
1	0.37	0.74	0.61	0.89	0.55	0.80	0.17	0.72	0.82	0.71	0.57	0.47	0.63	0.69	2862
2	0.49	0.64	0.56	0.82	0.57	0.80	0.34	0.65	0.76	0.70	0.43	0.46	0.59	0.67	2981
3	0.30	0.79	0.33	0.88	0.74	0.91	0.54	0.78	0.78	0.37	0.48	0.43	0.61	0.71	3162

Results

GSDMM clustering

Logistic regression

	bias	clickbait	conspiracy	fake	hate	junksci	political	reliable	rumor	satire	unknown	unreliable	accuracy	accuracy without political	size
0	0.69	0.33	0.36	0.50	0.60	0.00	0.32	0.45	0.73	0.45	0.32	0.41	0.52	0.54	1393
1	0.21	0.25	0.69	0.56	0.52	0.55	0.33	0.61	0.58	0.41	0.25	0.25	0.49	0.54	1296
2	0.80	0.41	0.56	0.88	0.25	0.33	0.10	0.79	0.96	0.57	0.00	0.63	0.70	0.75	951
3	0.10	0.53	0.32	0.63	0.30	0.73	0.28	0.72	0.83	0.49	0.14	0.15	0.54	0.58	1028
4	0.14	0.74	0.26	0.28	0.43	0.20	0.61	0.42	0.00	0.57	0.44	0.10	0.46	0.52	1766
5	0.15	0.70	0.68	0.40	0.50	0.45	0.05	0.80	0.84	0.61	0.23	0.41	0.58	0.63	1057
6	0.40	0.53	0.24	0.35	0.72	0.86	0.27	0.72	0.74	0.30	0.38	0.76	0.55	0.61	4509

Weighted accuracy = 0.540

Results

GSDMM clustering

DistilBERT

	bias	clickbait	conspiracy	fake	hate	junksci	political	reliable	rumor	satire	unknown	unreliable	accuracy	accuracy without political	size
0	0.71	0.00	0.47	0.81	0.29	0.00	0.00	0.05	0.92	0.83	0.46	0.21	0.50	0.54	1393
1	0.00	0.42	0.55	0.97	0.67	0.97	0.60	0.63	0.75	0.27	0.11	0.00	0.57	0.64	1296
2	0.76	0.14	0.00	0.89	0.00	0.00	0.00	0.78	0.95	0.69	0.00	0.00	0.58	0.59	951
3	0.00	0.86	0.27	0.76	0.17	0.89	0.06	0.61	0.70	0.73	0.00	0.00	0.55	0.60	1028
4	0.22	0.76	0.10	0.81	0.50	1.00	0.81	0.05	0.00	0.52	0.55	0.17	0.53	0.67	1766
5	0.15	0.77	0.65	0.67	0.50	0.20	0.00	0.90	0.77	0.69	0.00	0.00	0.57	0.61	1057
6	0.45	0.68	0.47	0.94	0.76	0.85	0.38	0.73	0.64	0.58	0.60	0.79	0.67	0.75	4509

Weighted accuracy = 0.593

Results

Comparison of DistilBERTs

Accuracy w/o clustering = 0.710

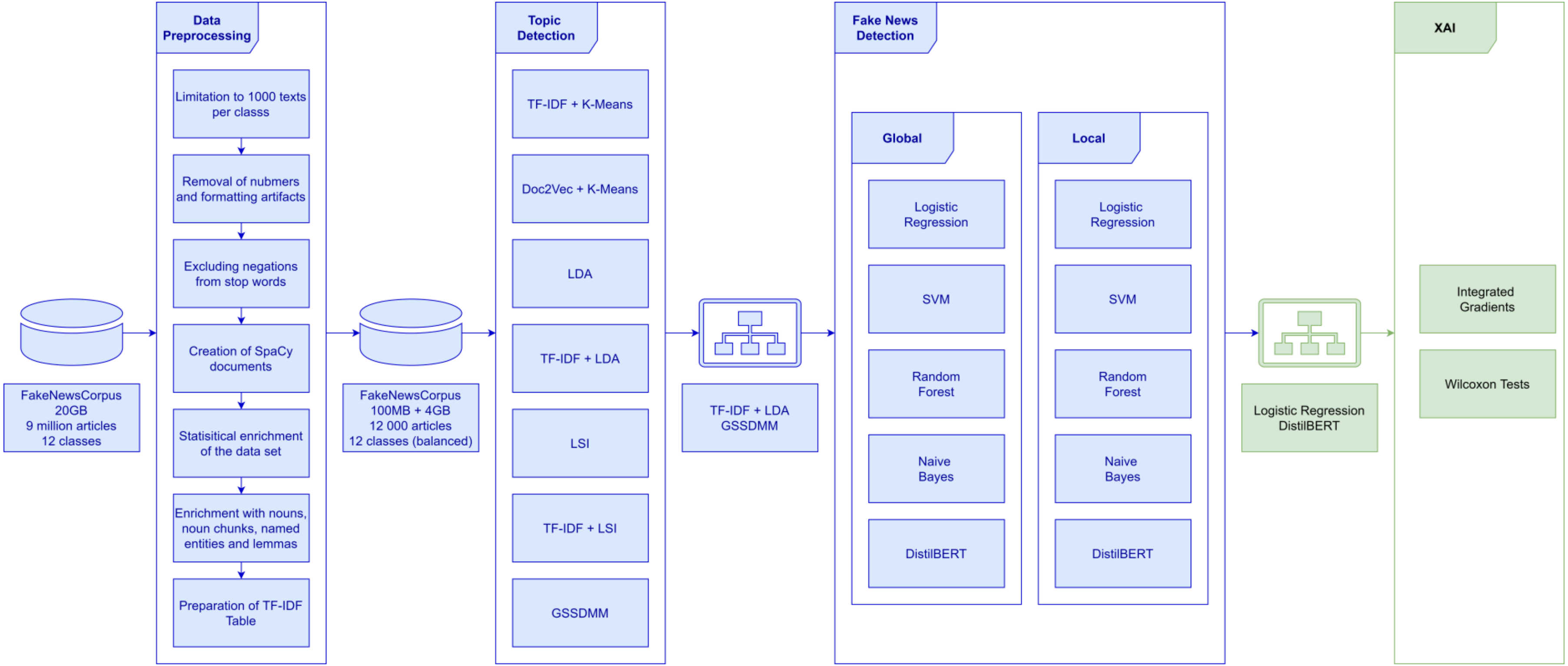
Weighted accuracy LDA = 0.619

Weighted accuracy GSDMM = 0.593

06

Models explanations

Experimental design

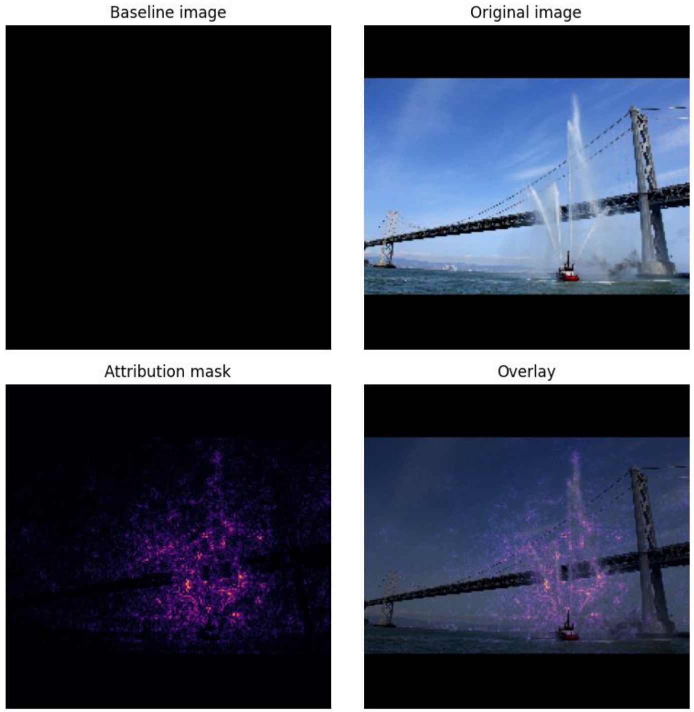


Explanations

1. The best model turned out to be the fine-tuned DistilBERT.
2. Consequently, only these models will be used to create explanations.
3. The explanation method is based on Integrated Gradients (IG). It is given by the following formula, where i – label, x – input, x' – baseline input.

$$IntegratedGrads_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

4. The equality of explanations will be determined by running the Wilcoxon statistical test with FDR correction.



Explanations

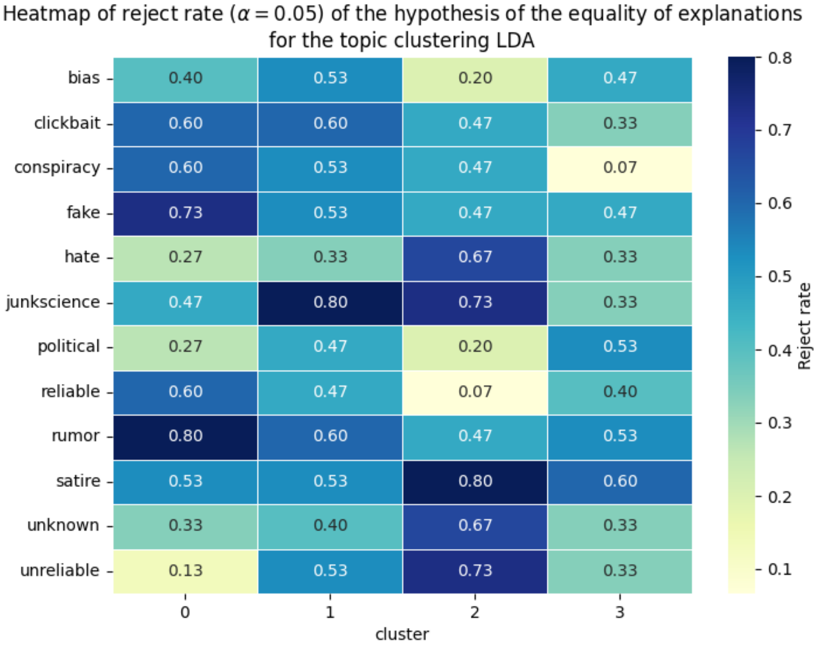
[CLS] donald trump welcomed the new england patriots to the white house wednesday afternoon to cong ##rat ##ulate them for their historic comeback victory and fifth super bowl championship for the franchise . the president addressed the patriots players , coaches and owner robert kraft on the south lawn in recognition of their thrilling comeback win in which they overcame a - deficit to push the game into overtime and take home another lombard ##i trophy . (video : rob gr ##on ##kowski crashes white house press briefing) “ with your backs against the wall – and

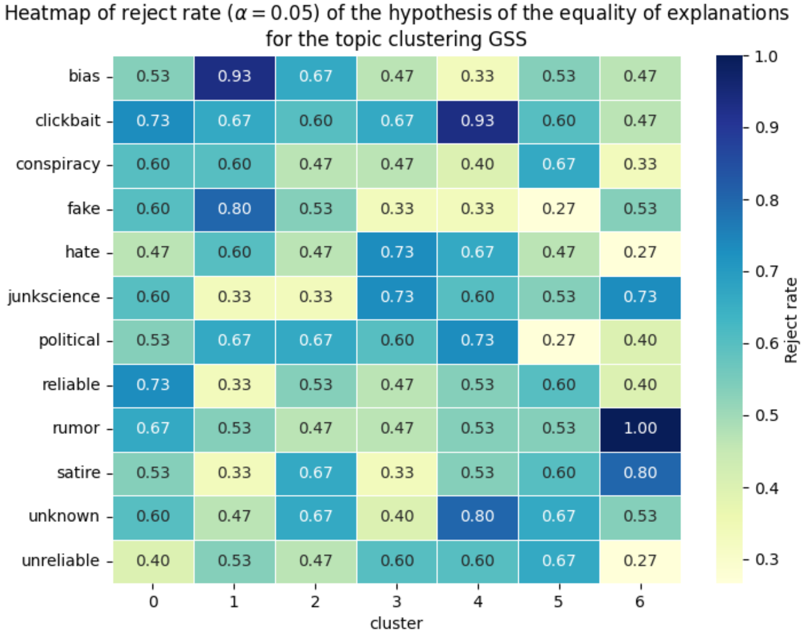
Global model

[CLS] donald trump welcomed the new england patriots to the white house wednesday afternoon to cong ##rat ##ulate them for their historic comeback victory and fifth super bowl championship for the franchise . the president addressed the patriots players , coaches and owner robert kraft on the south lawn in recognition of their thrilling comeback win in which they overcame a - deficit to push the game into overtime and take home another lombard ##i trophy . (video : rob gr ##on ##kowski crashes white house press briefing) “ with your backs against the wall – and

Local model

Explanations





07

Conclusions

Conclusions

1. We trained a variety of models. The best ones turned out to be DistilBert.
2. We split the input dataframe by using topic detection methods. Further, we trained models to detect classes in each of the clusters.
3. The performance of models trained on the specific clusters turned out to be, unfortunately, worse. This could be a result of similar concept (topics) inside clusters or smaller amount of data per model.
4. The models were compared also based on the explainability technique Integrated Gradients. This approach proved that topic-specific models put focus on different words than a global model. Moreover, the Wilcoxon test results showed that the differences are real.


08
Further Works

Further Works

1. The local approach proves to be ineffective for now. To fix this we plan to enhance the amount of data (5 times), and reduce numer of fake news types (end up with 2 and 4 classes).
2. To gain more time for training we will limit ourselves to the DistilBERT at first.
3. If the approach works, we plan to try a new, bigger model (maybe some LLM).
4. We want to introduce a new clustering method of BERTopic, which may contribute positively to the final results, as BERTopic is easily interpretable.


Thank you for your attention!