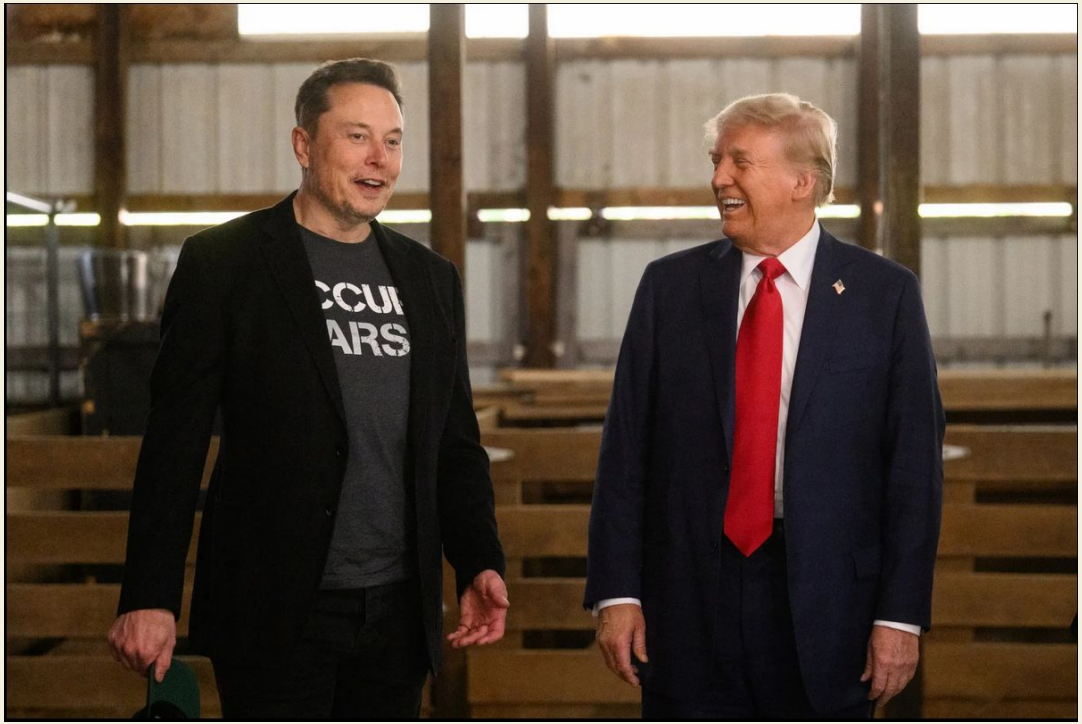
Group 2:
Fake News Detector

Ambrose
Guy
Aurele

1. Overview of the problem
2. Data pre-processing
3. Models exploration
4. Evaluation & critics
5. Deployment

Overview of the problem



Overview of the problem

- **Objective:** Build a Fake News Detector using NLP models
- **Required to:**
  - Test & assess multiple combinations of:
    - Data Cleaning
    - Natural Language Processing
    - Vectorization techniques
    - Modeling
  - Analyze results on test set to try to come up with optimizations

Data pre-processing

- **Data cleaning:**
  - Best model removes: URLs, numbers, double spaces, special characters...
  - Noticed a negative impact on removing stopwords for this exercise (Δ = 1.15% avg. acc.)

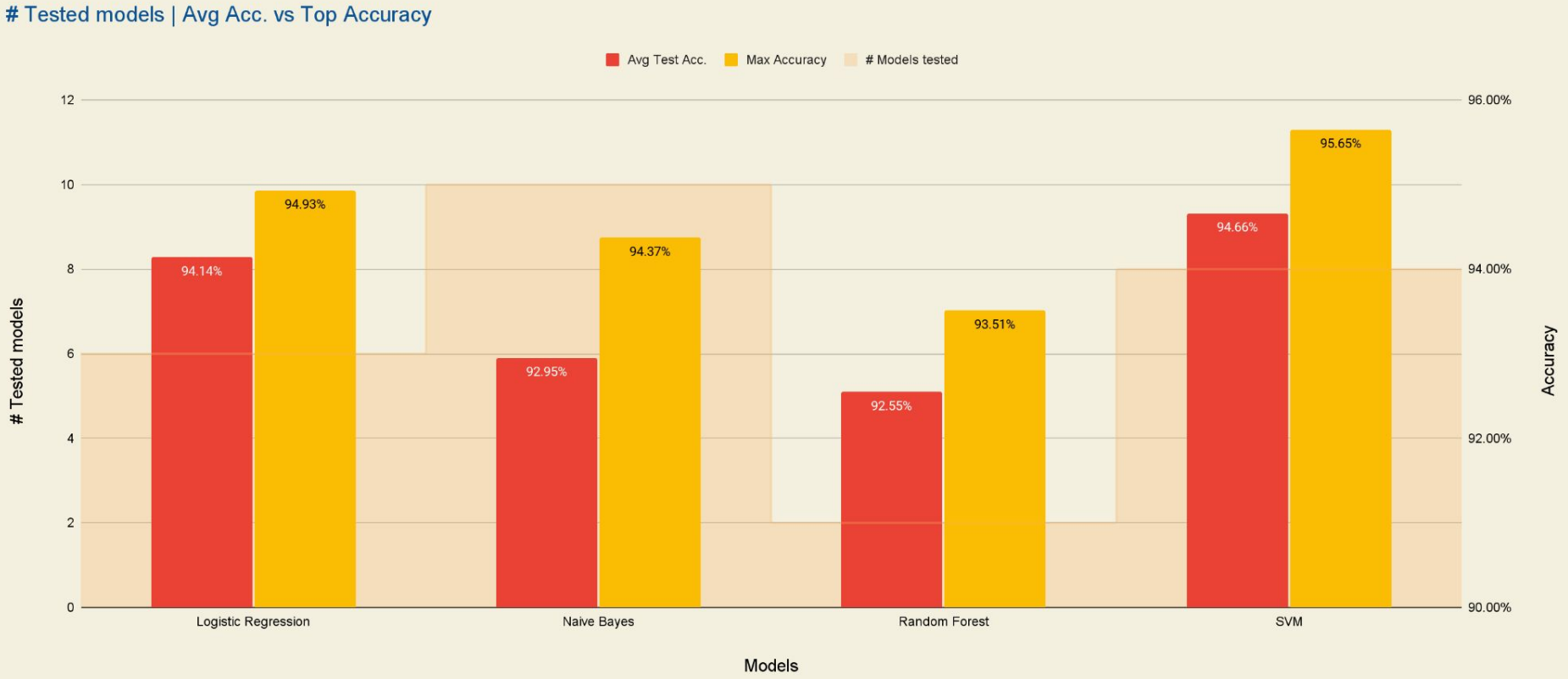
- **Word processing:**
  - Lemmatization avg. acc. = 93.65% ⇒ 20 models tested
  - Stemming avg. acc. = 93.97% ⇒ 6 models tested
  - Low average difference but overall lemmatization achieved top acc. in most cases.

- **Language processing:**
  - Bigram,
  - BoW, BoW Bigram/Trigram
  - TF-IDF, TF-IDF Bigram
  - TF-IDF + Trigram reached the best accuracy

Over 30 models tested

	Owner	Stopwords	Stem or Lemma	Bow / N-gram / TF-IDF	Model	%	Top accuracy
1	Aurele	Yes	Lemmanization	BoW	Naive Bayes		93.44%
2	Aurele	Yes	Lemmanization	Bigram	Naive Bayes		89.06%
3	Aurele	Yes	Lemmanization	TF-IDF	Naive Bayes		92.73%
4	Aurele	Yes	Lemmanization	TF-IDF	Random Forest		91.59%
5	Aurele	Yes	Lemmanization	TF-IDF	Logistic Regression		93.93%
6	Aurele	Yes	Stemming	BoW	Naive Bayes		92.79%
7	Aurele	Yes	Stemming	TF-IDF	SVM		93.73%
8	Aurele	Yes	Lemmanization	BoW	SVM		93.44%
9	Aurele	No	Lemmanization	BoW	Naive Bayes		94.37%
10	Aurele	No	Lemmanization	Bigram	Naive Bayes		92.13%
11	Aurele	No	Lemmanization	TF-IDF	Naive Bayes		93.91%
12	Aurele	No	Lemmanization	TF-IDF	Random Forest		93.51%
13	Aurele	No	Lemmanization	TF-IDF	Logistic Regression		94.48%
14	Aurele	No	Stemming	BoW	Naive Bayes		93.98%
15	Aurele	No	Stemming	TF-IDF	SVM		95.27%
16	Aurele	No	Lemmanization	BoW	SVM		95.05%
17	Aurele	No	Lemmanization	TF-IDF Bigram	SVM		95.63%
18	Aurele	No	Lemmanization	TF-IDF Trigram	SVM		95.65%

Result analysis over 26 selected models



Top test accuracy - 95.65%

- Stopwords included
- WordnNet Lemmatizer
- TF-IDF + Trigram
- SVM model
  (kernel='rbf', gamma='scale', C=100.0)

Evaluation

We performed clustering of the titles according to repetitive keywords in the title.(The upper graph)

We got the following results:

- **Cluster 0:** north, korea, say, syria, russia - **217 Titles**
- **Cluster 1:** medium, social, ep, room, boiler - **189 Titles**
- **Cluster 2:** police', 'new', 'hillary', 'say', 'video - **7961 Titles**
- **Cluster 3:** donald', 'supporter', 'president', 'video', 'trump - **1205 Titles**
- **Cluster 4:** refugee', 'bangladesh', 'rohingya', 'myanmar', 'muslim- **412 Titles**

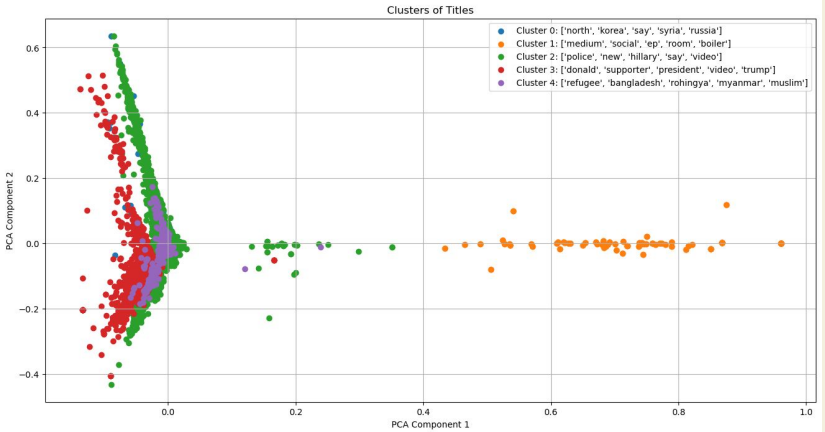
According to the keywords selected by the model, you can see that cluster number 1 has the least titles that are mostly far from being news titles

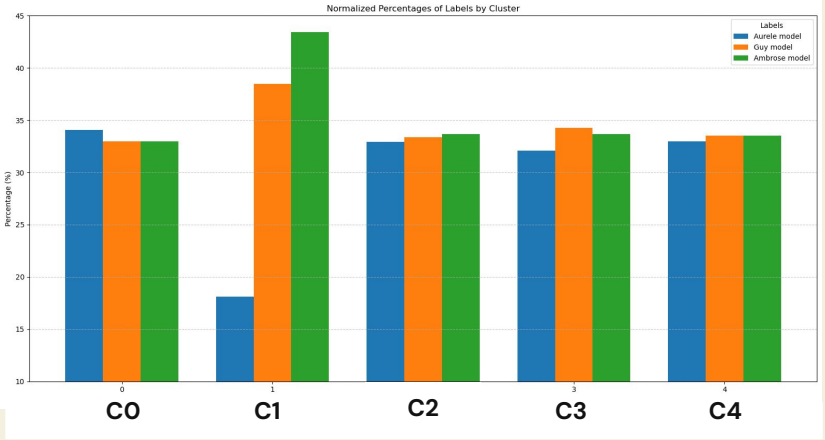
Following, we combined our best models accuracies results under the created clustering.(The bottom graph).

You can as well see that in the 1 cluster is a largest deviation between the models that could be caused but the title names that are not really related to the news titles

Conclusion:

1. The 1 cluster could be filtered out by preprocessing the data or by performing the clustering before data preprocess
2. If ignoring the 1 cluster we could see that the deviation between the models lies between 30 to 35 percent





Fake News Detector

Enter a News headline to check its veracity

Enter News Headline

Trump offers invites to foreign leaders through calls and back channels

Advice :

Likely News.

Clear | Submit | Flag

https://182581ea7dc6767f0e.gradio.live

CNN US World Politics Business Health Entertainment

Susan Walsh/AP

Even without Xi coming, Trump's invitation sheds light on the president-elect's confidence and ambition

- Trump offers invites to foreign leaders through calls and back channels
- 'The statements change every day': Capitol rioters try to parse Trump's pardon pledges
- 'Are you worried?': Acosta asks former January 6 committee chair about Trump's threat 2:51

https://182581ea7dc6767f0e.gradio.live

Project overview

- Data cleaning is critical to the training and success of the model
- Testing all the possible combinations takes a long time, especially when exploring models like SVM and Random Forest
- Improving further the model would require depper analysis of the training data set and of the test results