Fake News

Erdos Bootcamp
Spring 2023

Team:
Guillermo Lezama, Man Cheung Tsui, Pedro Fontanarrosa,
Shayne Plourde

Fake news

- **Fake news has emerged as a significant societal issue**
  - impacting public opinion, democratic processes, and social cohesion.
- **Studying fake news**
  - helps us understand its origins, effects, and mechanisms, empowering us to develop effective strategies to combat misinformation.
  - Insights gained from studying fake news can assist news organizations in improving their fact-checking processes and identifying potential vulnerabilities.
- **Policymakers**
  - can benefit from research findings to shape regulations and policies that address the challenges posed by fake news.

Examples and its consequences

- COVID-19 misinformation cost at least 2,800 lives and $300M between March and November of 2021 in Canada
  - January 26, 2023 - Council of Canadian Academies (CCA)
- Fake image of Pentagon explosion briefly sends jitters through stock market
  - May 23, 2023 - AP News: Fact Focus

https://cca-reports.ca/wp-content/uploads/2023/01/Report-Fault-Lines-digital.pdf
https://www.france24.com/en/tv-shows/truth-or-fake/20230523-fake-ai-generated-image-of-explosion-at-the-pentagon-goes-viral
https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4

Project Objective:

Develop a machine-learning model capable of analyzing news articles and classifying them as "real" or "fake."

Secondary Project Objective:
Understand if synthetic features can increase the accuracy of these models when we ignore the source of the news

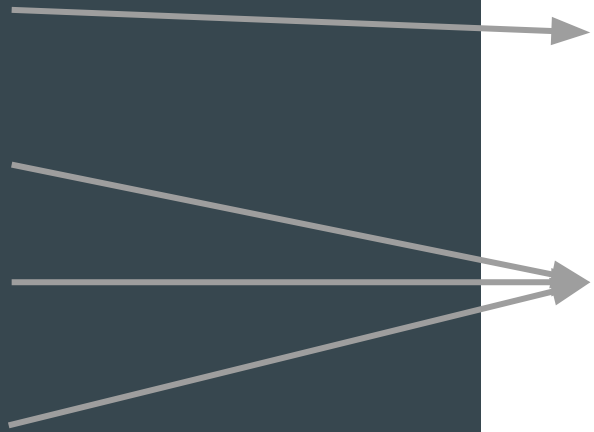
Stakeholders

News Organizations

Governments

Firms

Citizens



Why?

- Model → Check sources quickly
- Scaffold → Train on their own truth/fake
- Features → Know what to pay attention to

- Model → Check for spread of Fake News

Classification Problem

X[‘dataset’] consisting of news (in text)

Title	Text	Fake or Real
		0 (Fake)
		1 (Real)

The Fake News Dataset

	True (1)		Fake (0)	
Training Set	9816	(48.6%)	10387	(51.4%)
Testing Set	2737	(54.0%)	2334	(46%)

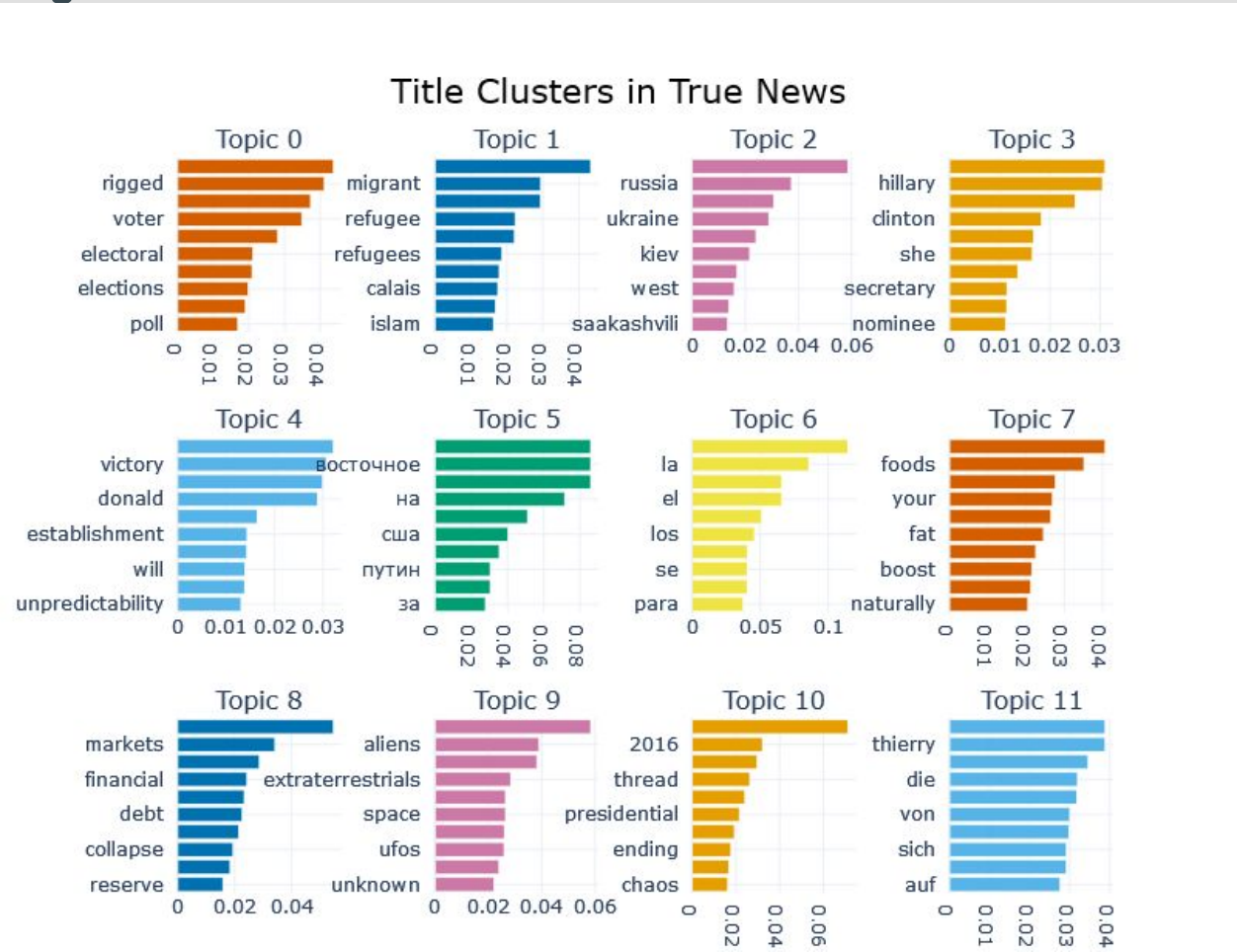
- Dataset obtained from Kaggle
- Preprocessing: Some NaN values in Title / Text present in training set.
- Features are: 'id', 'author', **'title'**, **'text'**, **'label'**

EDA

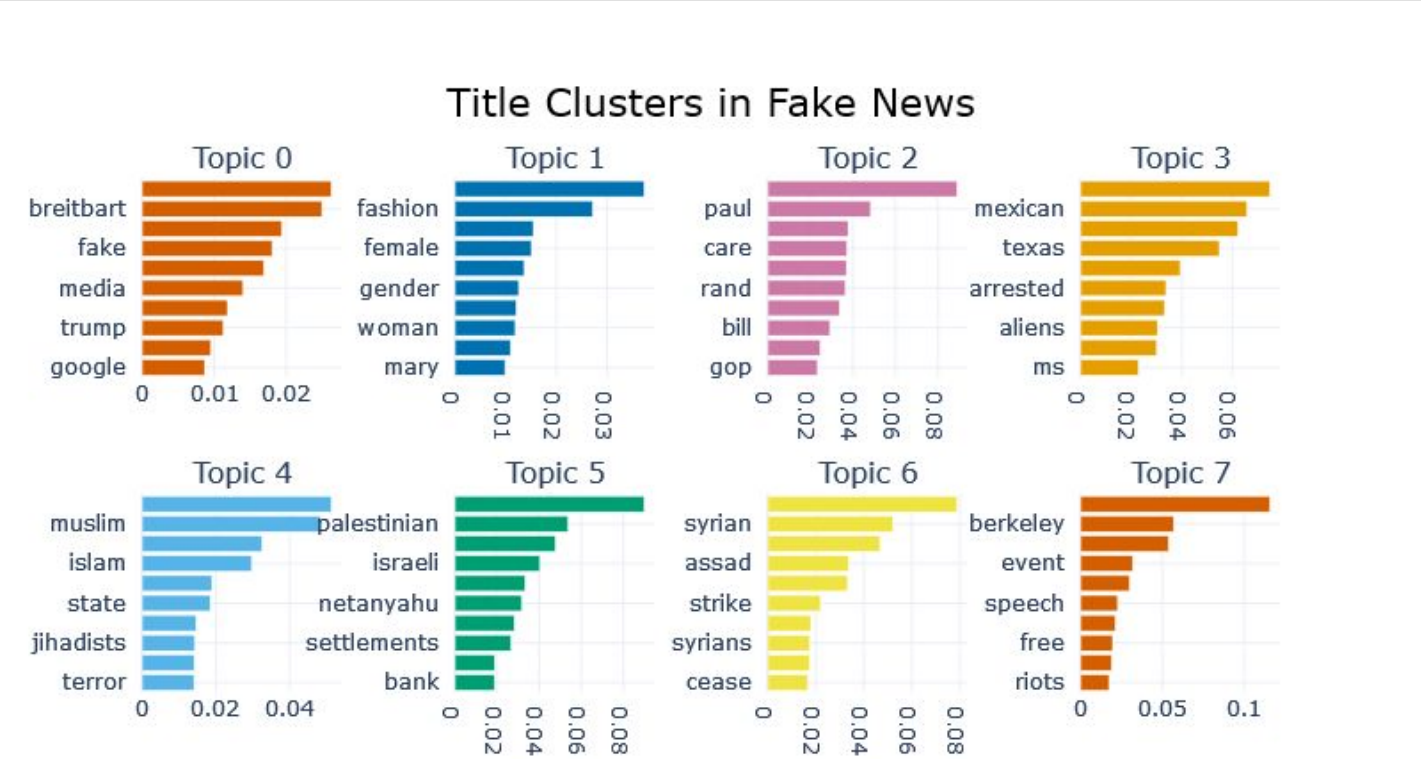
We defined features from reading about the subject:

- The number of grammatical and spelling errors
- **The number of times the article uses stop words**
- **The use of emotional language**
- The number of words used with capital letters
- The use of numbers
- The use of long words
- Topic modeling of the article using BERTopics

Topic modeling via BERT

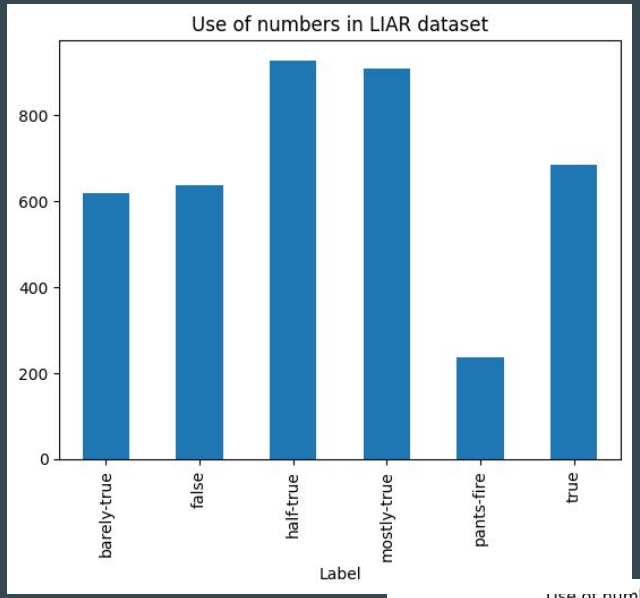


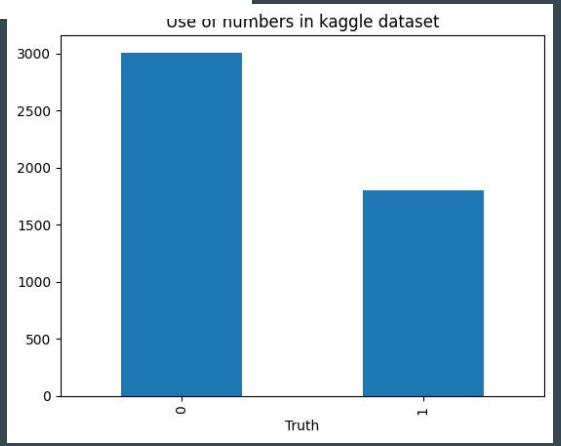
Topic modeling via BERT



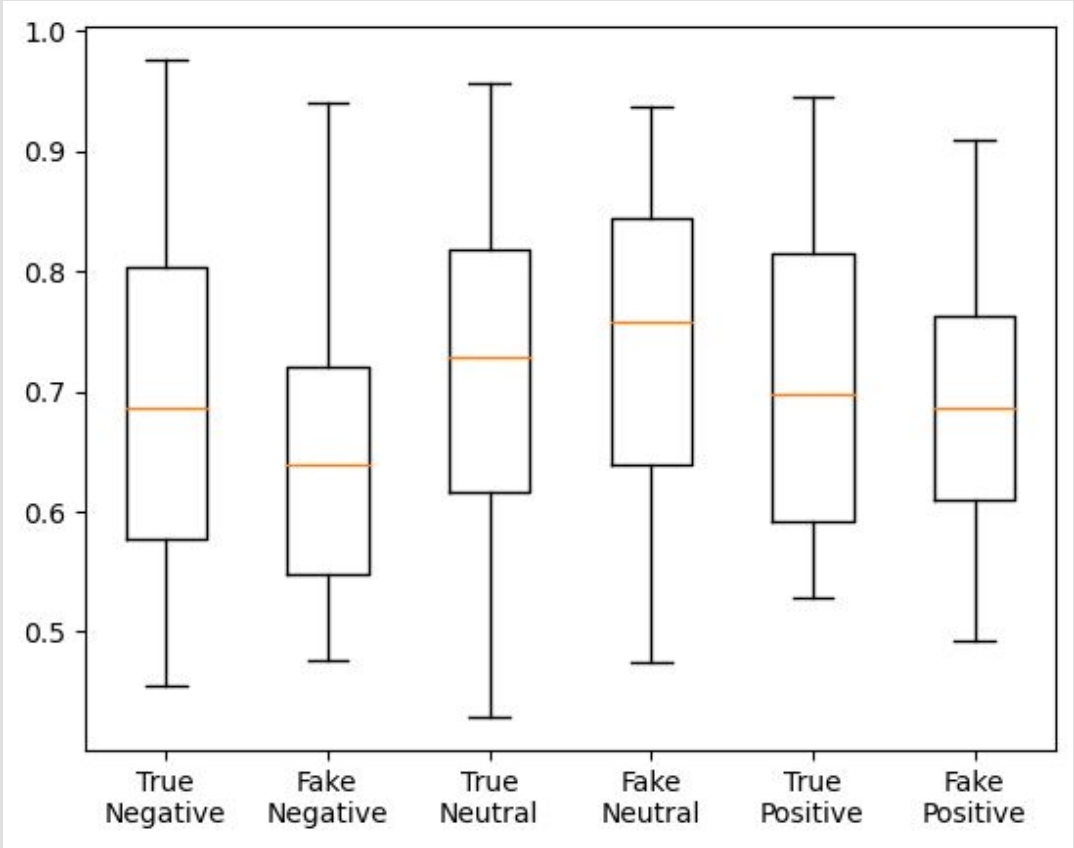
Use of numbers

- News with #s that are true for kaggle: 37.4%
  - Percentage that uses #s: 10.7%
- News with #s that are true for liar: 55.1%
  - Percentage of Liar that uses numbers: 39.2%
- The difference in number use for the Kaggle database was significantly different from true to false, but the overall usage of numbers was too low to make meaningful predictions from.





Sentiment analysis on titles via roBERTa



EDA

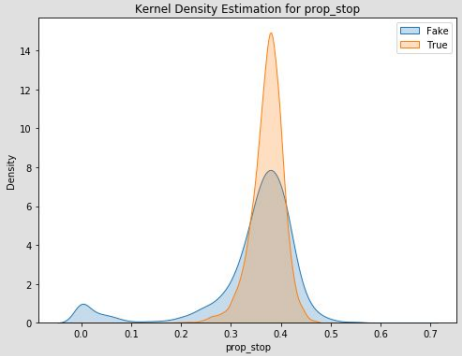
X['dataset']

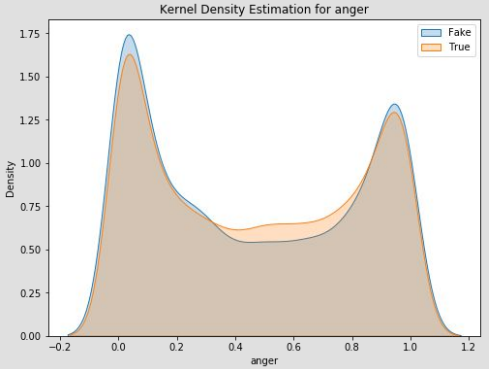
Features

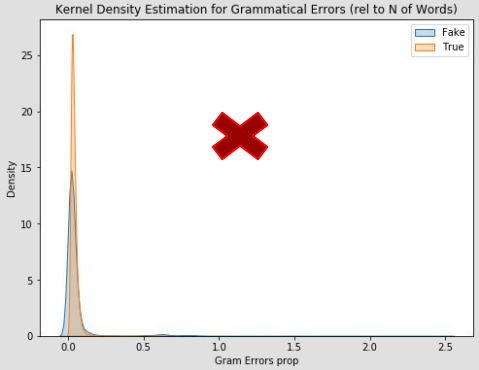
- N of stop words
- Emotions
- Grammatical Errors
- N of numbers
- N of Caps

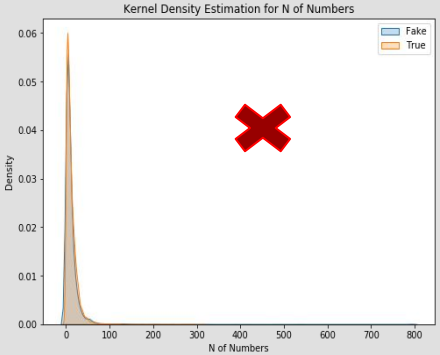
Roberta distillBERT titles

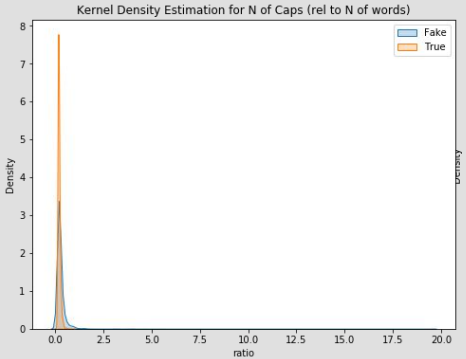
LanguageTool

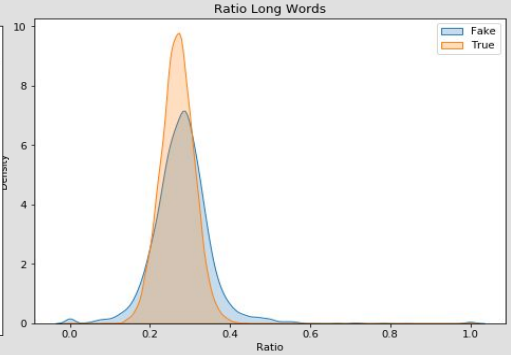








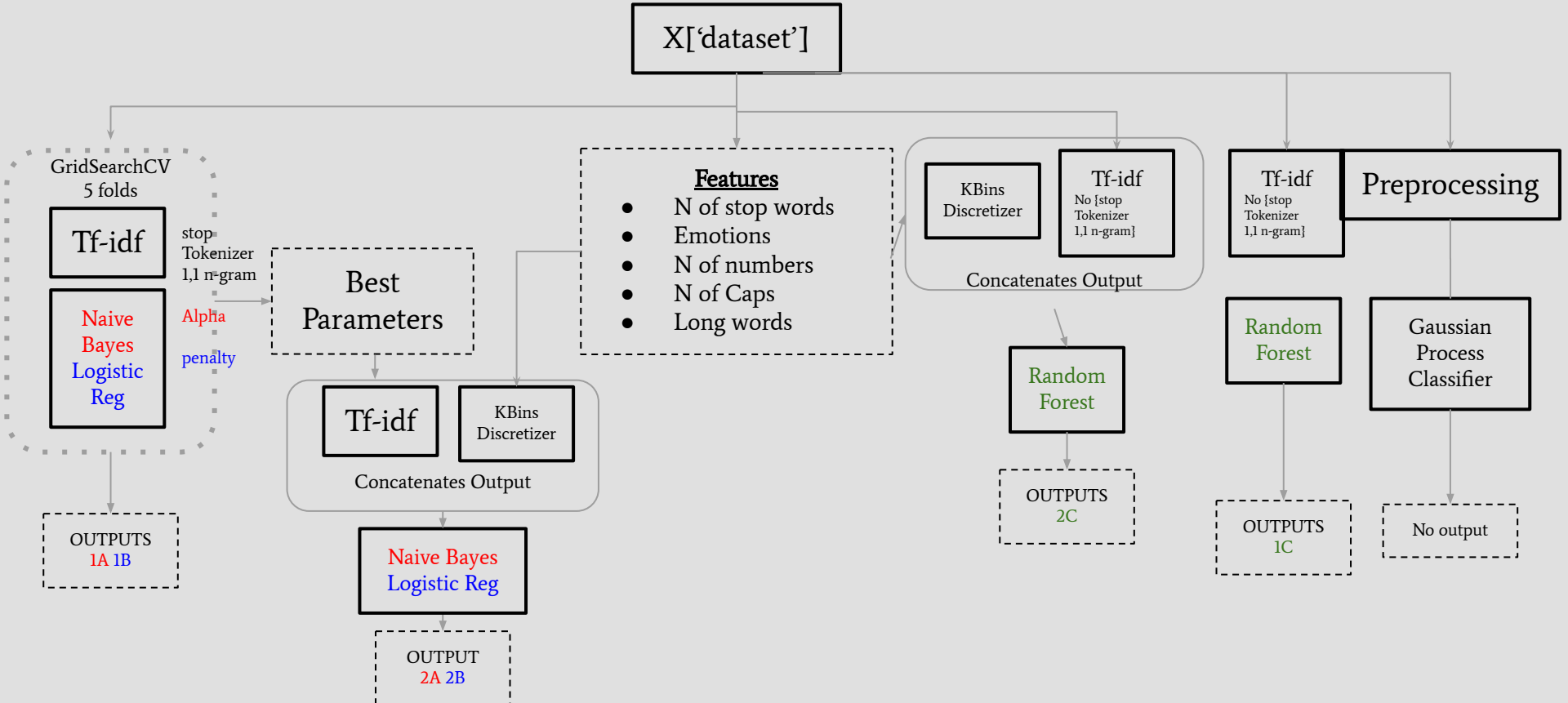




Mean fake: 0.05; Mean True: 0.04

Mean fake: 0.28; Mean True: 0.20

4x2 Modelling Approaches



How to Evaluate Them?

Accuracy

1) Simplicity: Stakeholders want a straightforward metric that is easy to understand.
2) Commonly used in previous similar tasks.
3) Clear performance indication: Accuracy provides a clear indication of how well a classifier is performing.

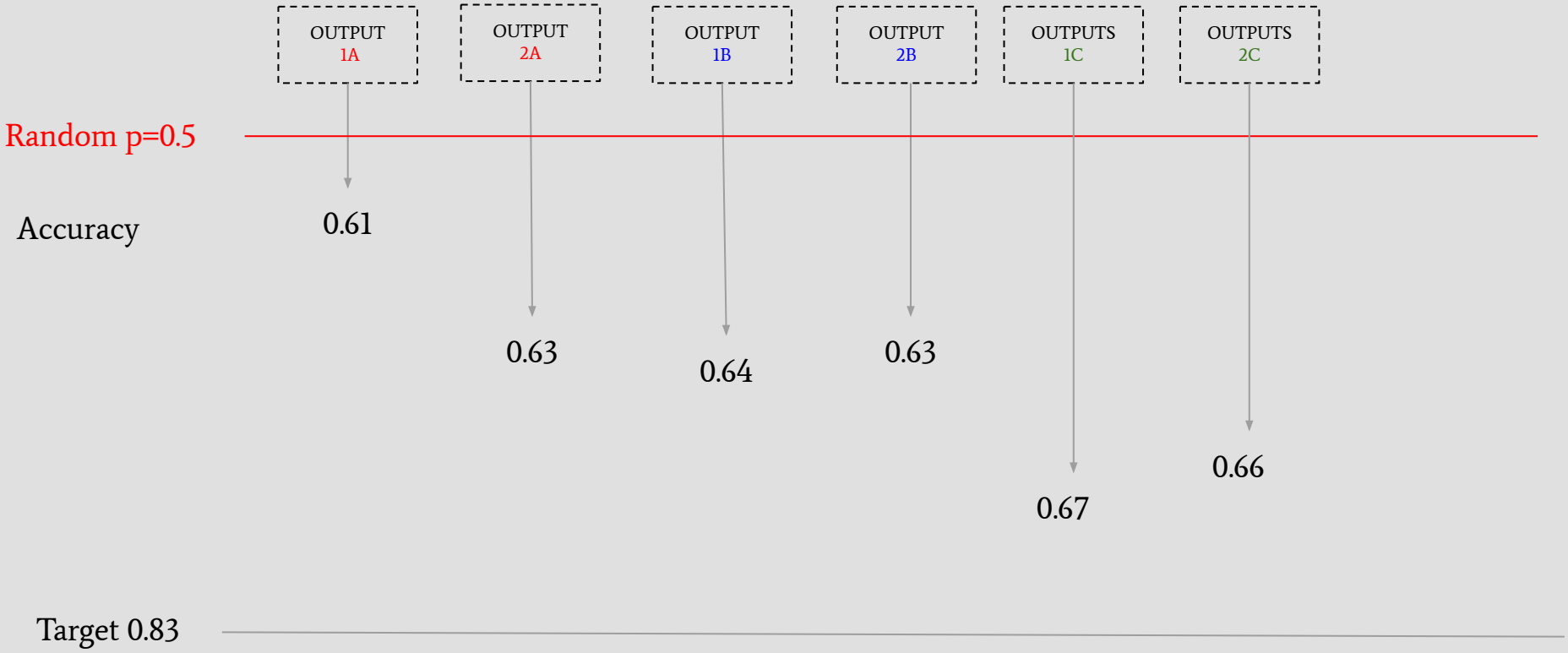
Target:

- Top 10 in Kaggle Competition performed with this dataset (accuracy 0.83681)
- Improving Base Classification (models A) with added features (models B)

Other targets (forthcoming)

- Accuracy higher better than random in out-of-sample testing
- Beat ChatGPT (better than random)

Results


OUTPUT 1A
OUTPUT 2A
OUTPUT 1B
OUTPUT 2B
OUTPUTS 1C
OUTPUTS 2C
Random p=0.5
Accuracy
0.61
0.63
0.64
0.63
0.67
0.66
Target 0.83

Comments

- Because we did not use text from the titles, or author information, we had an expected drop in accuracy. However, we intentionally dropped them because we wanted an “agnostic” news classificator, “agnostic” as to where the news came from.

- We did not use the text from the titles.
  - This dataset contains information of the news organization in the title.
  - Classification is biased.
  - The sample of news organizations is not representative for our clients.

- We did not use the author information.
  - Same reasons that above

Conclusions

- We developed a machine-learning model capable of analyzing news articles and classifying them as "real" or "fake." than is better than 50%.
- Accuracy did not meet the competitive standards set by the Kaggle competition.
- Exploration of synthetic features highlighted the importance of some attributes.
- Incorporating them was not as helpful as we expected, however other synthetic features should be explored.

Future Work

- Moving forward, further research and experimentation could potentially enhance our model's accuracy, bridging the gap to be competitive Kaggle competitions and other areas in which this issue is being discussed.

- Automatize feature extraction into the pipeline such that clients could feed the algorithm with their own data (Health News, Tweets about the Ukraine War, news about the stock market, etc.) and get a classifier that helps them to analyze news.

- Create a Word Cloud showing how when certain words (separate or in conjunction) are used the % fake probability increases

- Which sentence(s) or phrase(s) made a given news false?

- Clickbait Detection: Integrate a component for detecting clickbait titles. Fake news often comes with sensational headlines to attract more clicks.

- User Feedback Integration: Develop a mechanism to incorporate user feedback on model predictions. This real-time feedback can help the model learn and improve over time.

What we learned / takeaways

- Running machine-learning models on personal computers for large datasets can be **extremely time-consuming**. We've encountered delays of hours or even days. To overcome this, we'll utilize cloud-based machine learning platforms for faster and more efficient model training

Thanks!
Questions?