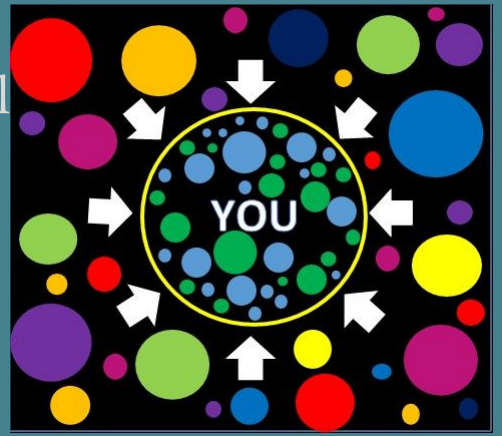
Data Science on Fake News

Keina Aoki

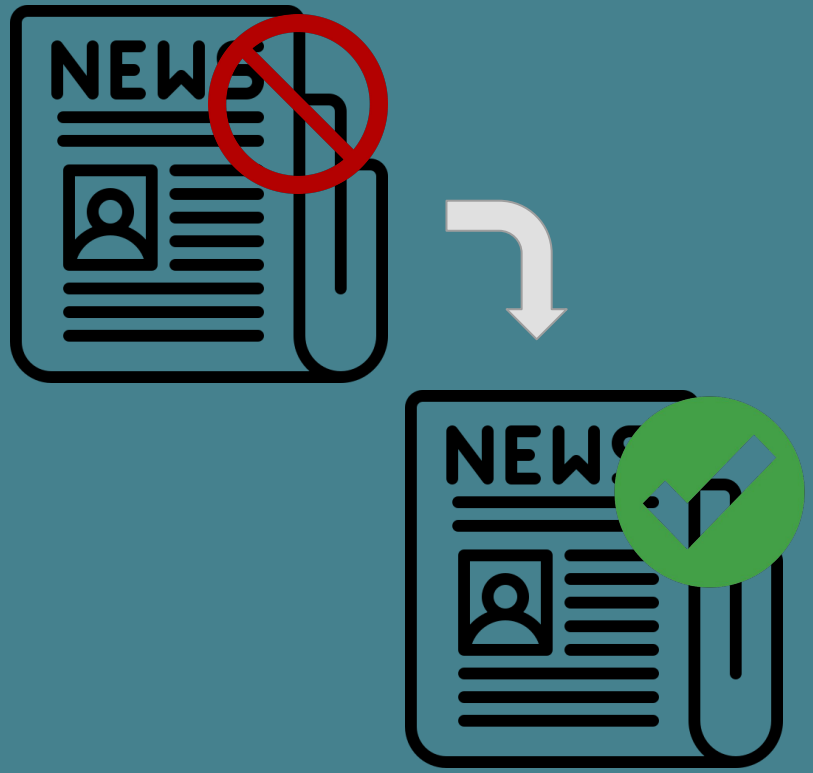
Overview

- There are vast amounts of information shared on a wide variety of digital platforms
    - Easily accessed, created and shared
- Misinformation and disinformation is poses a threat to democracy
    - Reinforces polarizing viewpoints
    - Generates mistrust in persons and institutions
    - Destabilizes society
- Develop tools to detect fake news to ensure we are well informed
- Use twitter data with labelled fake and real news



Data-Driven approach on Detecting Fake News

1. Data collection
2. Data preprocessing
3. Exploratory data analysis (EDA)
4. Feature selection
5. Model selection
6. Model training and testing
7. Classification



Impact of Data-Driven Approach

Speed → Faster the automation and detection, less harm fake news causes

Consistency → Decisions made based on patterns not on personal opinions

Cost-Effective → Compared to traditional fact-checking, expensive and time consuming

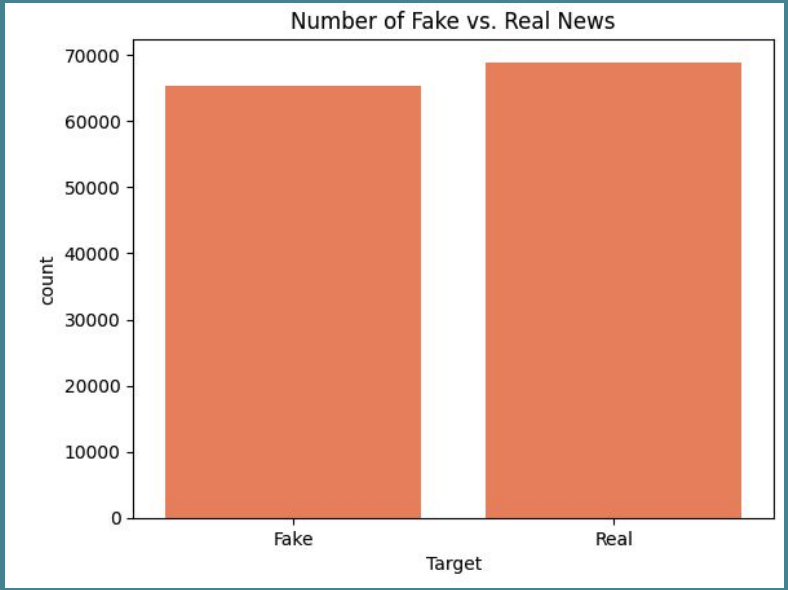
Overall, well-informed citizens

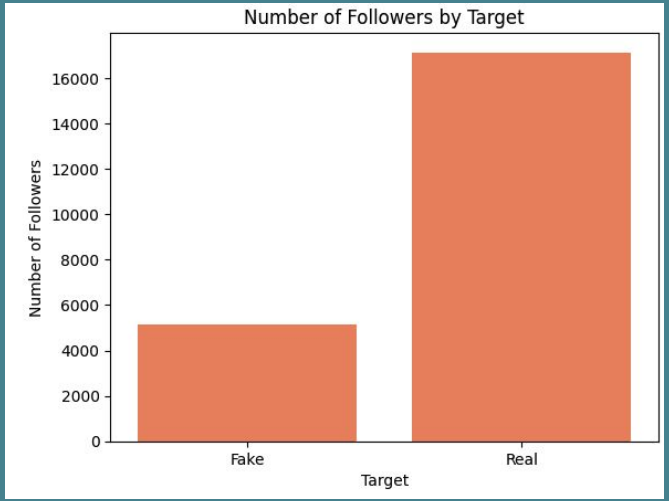
Dataset Overview

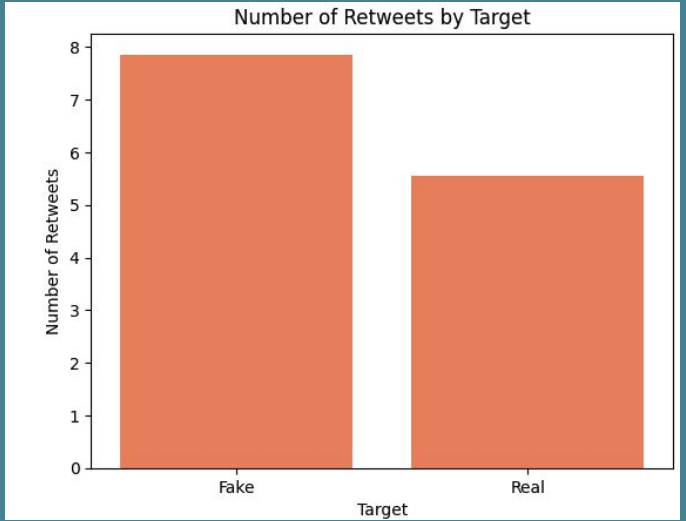
- Canadian Institute for Cybersecurity Truth Seeker Dataset 2023
  - https://www.unb.ca/cic/datasets/truthseeker-2023.html
- 134,198 records and 60 fields
- Features and target variable (fake = 0, real = 1)
- Data quality:
  - Data is clean (no missing values nor duplicated entries)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1		majority_target	statement	BinaryNumTarget	tweet	followers_count	friends_count	favourites_c	statuses_cou	listed_count	following	embeddings	BotScore	BotScoreBina	cred	normalize_ir	mentions	quotes	replies	retwe
2	0	TRUE	End of evicti	1	@POTUS	4262	3619	34945	16423	44	0	[[0 0 0 0 0 0	0.03	0	0.54079432	0.10460231	1	1	1	
3	1	TRUE	End of evicti	1	@SOSickRic	1393	1621	31436	37184	64	0	[[0 0 0 0 0 0	0.03	0	0.46217651	0.09443633	3	0	0	
4	2	TRUE	End of evicti	1	THE SUPREM	9	84	219	1184	0	0	[[0 0 0 0 0 0	0.03	0	0.09677419	0.03984638	0	2	5	
5	3	TRUE	End of evicti	1	@POTUS	4262	3619	34945	16423	44	0	[[0 0 0 0 0 0	0.03	0	0.54079432	0.10460231	1	0	0	
6	4	TRUE	End of evicti	1	@OhComfy I	70	166	15282	2194	0	0	[[0 0 0 0 0 0	0.03	0	0.29661017	0.06113481	1	0	0	
7	5	TRUE	End of evicti	1	I've said this	29010	5081	120924	89474	405	0	[[0 0 0 0 0 0	0.03	0	0.85095773	0.12532361	0	9	14	
8	6	TRUE	End of evicti	1	As many fac	33	21	432	62	0	0	[[0 0 0 0 0 0	0.03	0	0.61111111	0.04593827	0	0	0	
9	7	TRUE	End of evicti	1	@Thomas1	919	620	131682	216319	71	0	[[0 0 0 0 0 0	0.03	0	0.597141	0.09202534	3	0	0	
10	8	TRUE	End of evicti	1	@SocialismI	18002	15754	969539	497128	48	0	[[0 0 0 0 0 0	0.03	0	0.53329778	0.12214566	2	0	1	

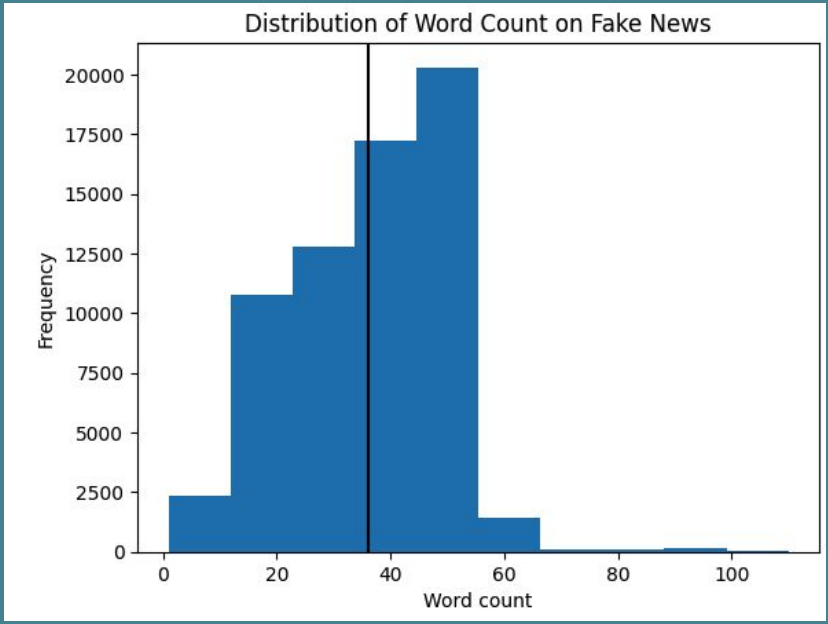
Preliminary EDA

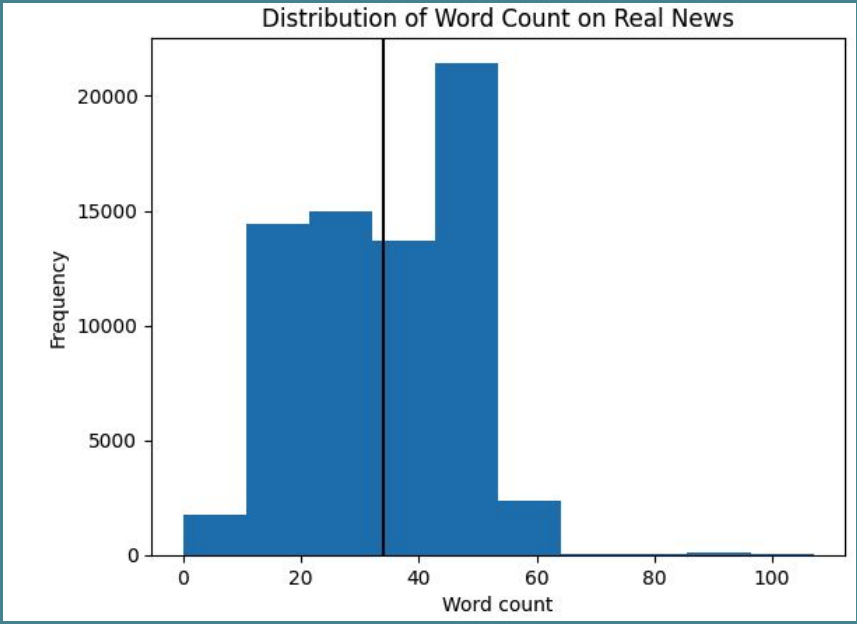






Preliminary EDA





Next Steps

- Approach 1: predict fake vs real news based on given features such as length of statement, influence score, credibility, number of followers
- Approach 2: use sentiment analysis
  - Determine the emotion attached to each statement to see if more negative or positive emotions are a good indicator for fake vs real news
- Compare accuracy scores