Propaganda based Fake News

Abhigyan Ghosh

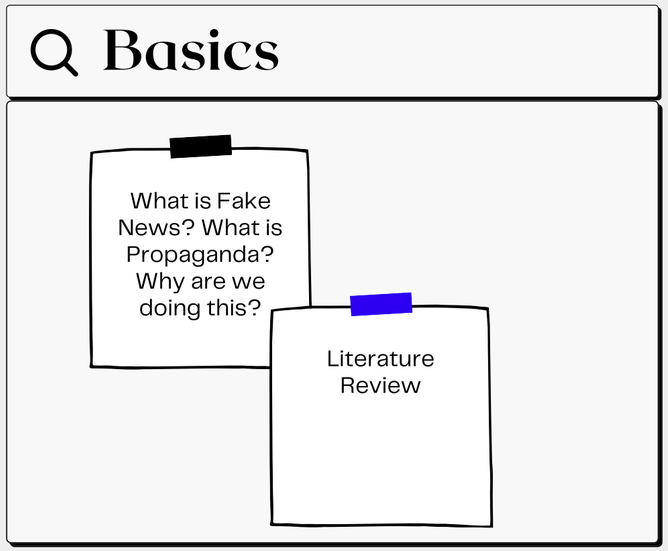
Shelly Jain

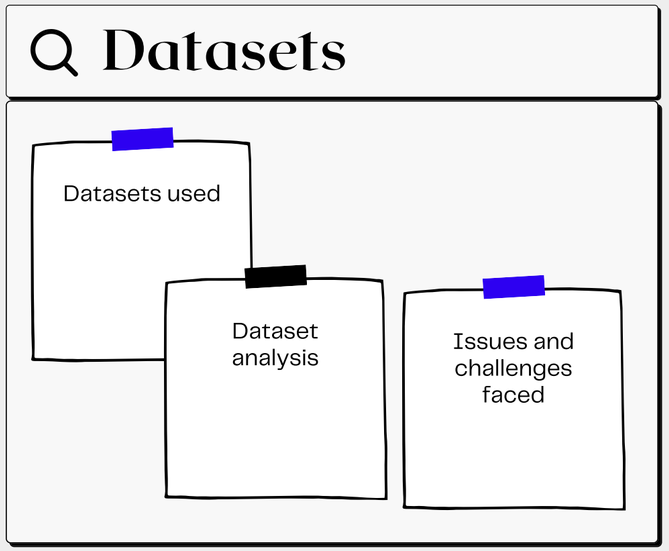
Sravani Boinepelli

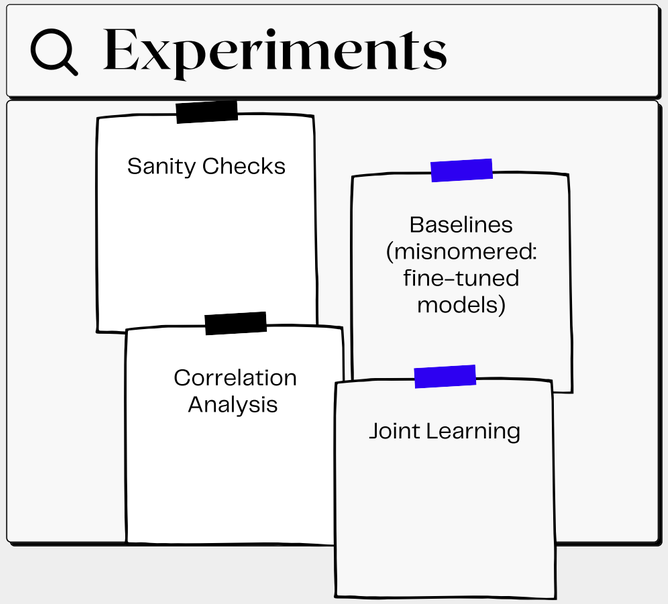
Zubair Abid

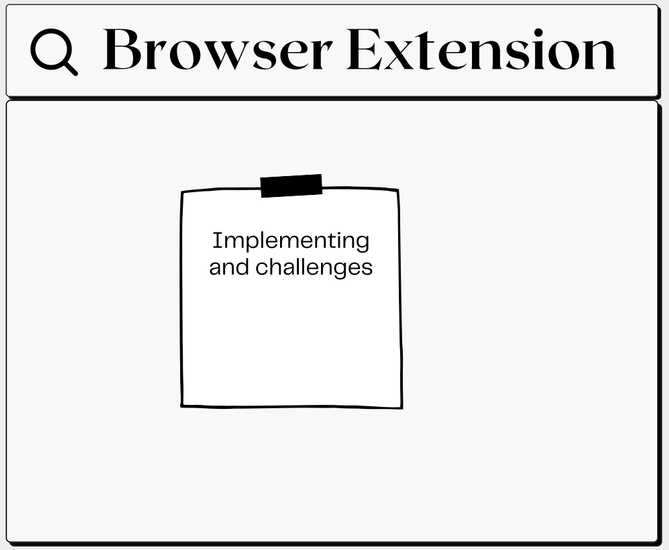
Structure

1. What is the problem space?
   - Fake news
   - Propaganda
2. What has been done?
   - Literature review
3. What did we use?
   - Datasets
4. What did we do?
   - Classifier models
   - Browser extension
5. What more can be done?
   - Future scope









Why, What

Fake news is casually observed to contain more propaganda than otherwise.

Running some experiments, we want to see if the addition of propaganda information benefits fake news detection.

We also made a browser extension.

Fake News

Fake news is the term given to any content which is false or misleading content presented as news and communicated in formats spanning spoken, written, printed, electronic, and digital communication.

Why Bother?

Fake news is harmful in several ways as:

- it undermines and consequently reduces the impact of legitimate news
- it has been observed to spread about six times as fast as real news.

It has been classified into many types, based on intent and description of the news:

- Clickbait
- Propaganda
- Misinformation
- Satire

Propaganda

Propaganda is different from fake news in that the information tends to benefit a particular party or agenda.

Propaganda need not be fake news. Likewise, fake news that does not benefit an agenda does not count as propaganda.

Notable Features

It is hard to spot logical falsities because they often sound like they might be true.

Some even deceive people based on the emotional language they use, so people overlook background research of the statement based on how the information is conveyed.

Other Work

Fake News Detection on Social Media: A Data Mining Perspective (Shu et al. 2017) – Comprehensive review of fake news detection in social media

Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking (Rashkin et al. 2017) – Characterises the language, truthfullness of political news

A Benchmark Study on Machine Learning Methods for Fake News Detection (Khan et al. 2019) – Compares performance of different ML models for detecting fake news

A Survey on Computational Propaganda Detection (Giovanni Da San Martino et al. 2020) – Propaganda detection from joint perspective of NLP and network analysis

Proppy: Organizing the news based on their propagandistic content (Barron-Cedeno et al. 2019) – Results show effectiveness of style to detect propaganda

Team QCRI-MIT at SemEval-2019 Task 4: Propaganda Analysis Meets Hyperpartisan News Detection (Saleh et al. 2019) – Variety of text features with logistic regression

SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles (G. Da San Martino et al. 2020) – Summary and analysis of all submissions

ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them (Jurkiewicz et al. 2020) – Ensemble of RoBERTa models

Datasets

1. The datasets all contain text related to American political news

2. For our models, we took a join of the Proppy and Liar datasets

NELA-GT-2019 (Gruppi, Horne, and Adalı 2020)

- 1.12M articles from 260 news sites
- Labels from 7 assessment sites
- Article meta data
- Aggregated labels based on MBFC – reliable, mixed, unreliable

Proppy 1.0 (Barron-Cedeno et al. 2019)

- 52K articles from 100+ news sites
- Postive (propagandistic) and negative (non-propagandistic)
- Labels obtained using distant supervision based on Media Bias/Fact Check
- Meta data regarding MBFC, location, URL

Liar, Liar Pants on Fire (Wang 2017)

- 12.8K short statements from PolitiFact
- Six labels (pants-fire, false, barely-true, half-true, mostly-true, true) changed to two labels (fake, true)
- Human labelled
- Meta data regarding speaker, political background, context of statement

Looking into Proppy

Bias labels against factuality labels

Label	Number	High	Mixed	Low/Very Low	NIL	unknown
NIL	14	9	2	3	0	0
unknown	65	0	0	0	0	65
extreme right	6	0	1	5	0	0
right	8	3	3	2	0	0
right center	6	5	1	0	0	0
least biased	7	7	0	0	0	0
left center	60	56	3	0	1	0
left	3	1	2	0	0	0

The labels were determined by MBFC, which provides only source level information

Amount of usable data (has MBFC label) was only 19.6K of 36K

Looking into Proppy

List of all sources labelled as propagandistic (only 10/169)

Source	Factuality Label	Bias Label	Link
Breaking 911	Very Low	-	https://mediabiasfactcheck.com/breaking911/
SHTFPlan	Mixed	Extreme Right	https://mediabiasfactcheck.com/shtfplan-com/
Clash Daily	Low	Extreme Right	https://mediabiasfactcheck.com/clash-daily/
Personal Liberty	Low	Extreme Right	https://mediabiasfactcheck.com/personal-liberty/
Frontpage Magazine	Low	Extreme Right	https://mediabiasfactcheck.com/frontpage-magazine/
The Washington Standard	-	-	-
VDare	Low	Extreme Right	https://mediabiasfactcheck.com/vdare/
Freedom Outpost	-	-	https://mediabiasfactcheck.com/fake-news/
Remnant	Low	Extreme Right	https://mediabiasfactcheck.com/the-remnant-magazine/
Lew Rockwell	Mixed	Extreme Right	https://mediabiasfactcheck.com/lew-rockwell/

Leveraging the datasets

1. Noticed source field was common in NELA and Proppy
2. This would help us as NELA has PolitiFact which would give us truth labels for articles and we could potentially have a dataset with both Propaganda and Fake News labels since the Proppy labels were source level anyways.
3. So we did a simple SQL Inner Join using the source field.
4. This proved to be a futile effort. The Proppy dataset was not properly formatted and there were issues with taking Inner Join. At the end we only found ~4000 usable article which is too little to run deep learning models on.

Stages of the project

1 Sanity Checks →

We ran standard classifiers for the detection of fake news and the detection of propaganda.

Provided input was classified as 'propaganda based fake news' based on the predictions of both the classifiers.

Sentence embeddings were used as features for model input.

2 Fine-Tuning →

As the standard classifiers were not very effective, we fine-tuned pretrained Huggingface models separately for both tasks, and this is what we use for our baseline.

3 Correlation →

The baseline classifiers were tested on each others data. The count of co-occurring 'fake' and 'propaganda' labels from each classifier were used to determine the correlation between fake news and propaganda in American political news.

4 Joint Learning →

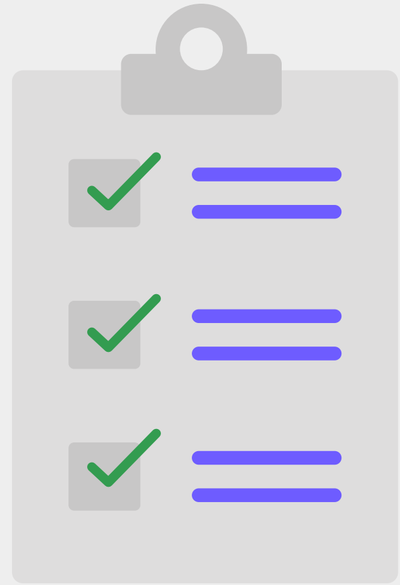
To see if we can get any improvements on the fine-tuned "baseline", we then conducted experiments by pretraining a combined propaganda based fake news model. This involved creating our own pretrained propaganda model to supplement to the fake news model.

5 Extension ✓

A browser extension was made for Mozilla Firefox using python FastAPI

The browser extension automatically obtains the text from the news articles in the web page and sends to the trained classifier. The prediction is returned as a pop-up.

Sanity Checks



- Simple classification with traditional classifiers
- sentence-transformers embeddings used as features for classification
- Multiple experiments run with various models: MLP, RF, and SVM

Sanity Checks - Results

Classifier	Proppy			Liar		
	Propaganda	Not Propaganda	Overall	Fake	Not Fake	Overall
MLP	0.64	0.96	0.80	0.58	0.64	0.61
SVM	0.60	0.96	0.78	0.57	0.69	0.63
RF	0.15	0.94	0.54	0.50	0.66	0.58

```
parameters = {
    'activation': ['identity', 'tanh', 'relu'],
    'solver': ['lbfgs', 'adam'],
    'max_iter': [100, 200, 500, 1000],
    'hidden_layer_sizes': [(100,), (50, 100, 50), (100, 50, 50, 100)],
}
```

The parameters used for the grid search

- The data for both Proppy and Liar was run through sentence-transformer's distilbert-base-nli-stsb-mean-tokens
- A grid search was run over several parameters, without significant improvement over the defaults
- We noted that the Random Forest Classifier was both the slowest, and worst performing across the board

Fine Tuning

- In order to get substantial results for our baseline, we then fine-tuned the BERT-Small model on a section of the datasets.
- Separate models were trained.
- Improvement in results was significant, albeit not strictly "propaganda-based" as of yet.

Metric	**Proppy**			**Liar**		
	Propaganda	Not Propaganda	Overall	Fake	Not Fake	Overall
Macro-F1 Score	0.91	0.99	0.95	0.58	0.70	0.64

Table 3: Macro F1 scores for the fine-tuned classifiers

Correlation

- Aim is specifically to identify propaganda **based** fake news
- Intuitively, one expects that propaganda is fake news. This assumption is explored using the original baseline classifiers.
- Tested in two directions:
  - Lair dataset tested on propaganda classifier (**fake -> propaganda**)
  - Proppy dataset tested on fake news classifier (**propaganda -> fake**)
- The co-occurrences of fake news and propaganda labels are analysed

Limitations

- Questionable veracity of datasets
  - Ground truth for Proppy is publication level, which is unreliable
  - Liar has only few lines of text per sample, not entire news articles
- Classifiers used for correlation analysis are themselves trained on these unreliable datasets

Correlation - Results

Analysis on the results of cross prediction in both directions

Contingency Tables

	Propaganda	Not Propaganda
Fake	83	533
Not Fake	45	623

	Fake	Not Fake
Propaganda	99	476
Not Propaganda	540	4010

The row labels give the ground truth and the columns state the classifier prediction

In order to make more sense of the data, we look at the percentages of:

- Fake (and true) news that was classified as propaganda
  - This shows the degree of **fake -> propaganda** entailment
- Propaganda (and non-Propaganda) news that was classified as fake
  - This shows the degree of **propaganda -> fake** entailment

Correlation - Results

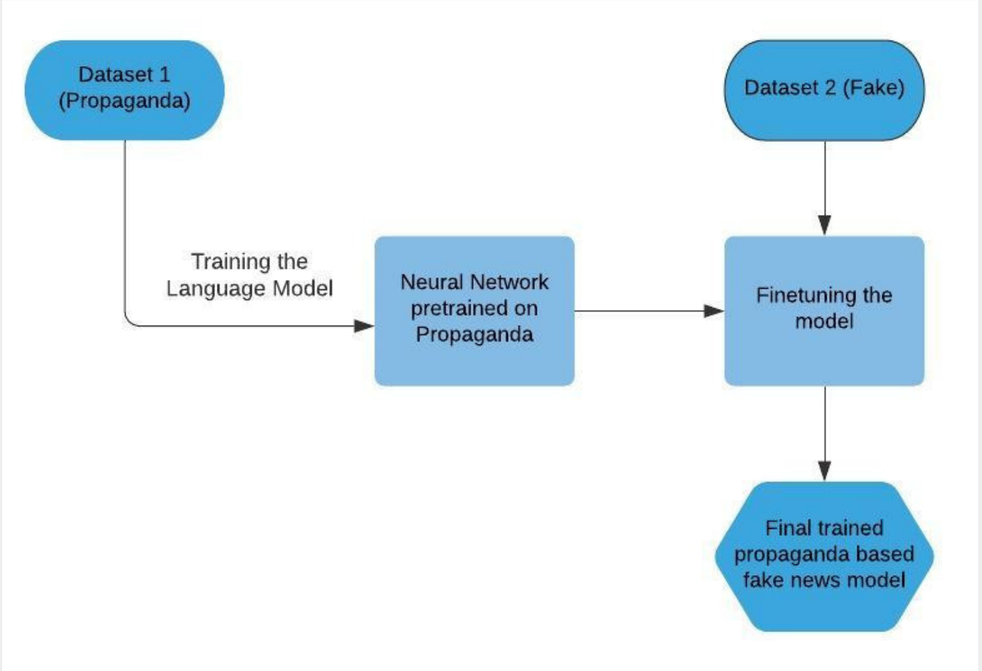
Analysis on the results of cross prediction in both directions

- Fake news is twice as likely as true news to be predicted as propaganda
  - Most of fake news is non-propaganda, due to distribution of dataset (which is mostly labelled as non-propaganda)
- Similarly, propaganda is more likely than non-propaganda to be predicted as fake news
- $\chi 2$ test gives p-value < 0.05 in both directions, indicating significant correlation between fake news and propaganda
- Unfortunately, no specific direction of entailment
  - Might not improve performance,
  - This is corroborated by later experiments

	Percentage
Fake news classified as propaganda	13.47%
True news classified as propaganda	6.74%
Propaganda classified as Fake news	17.22%
Not Propaganda classified as Fake news	11.97%

Joint Learning

- Leverage both the propaganda and fake news datasets to create our final propaganda **based** fake news model.
- Pre-training from scratch:
  - Pre-trained weights from HuggingFace aren't used
  - Custom weights obtained by training from scratch on the text from the propaganda dataset
  - Fine-tuning the fake news model to load custom weights



Thus, this gives a model for classification of propaganda based fake news when there is lack of a primary dataset that is annotated for that specific task.

Pre-training

Results and Observations

1. We train a byte-level BPE (byte-pair encoding) tokeniser, with the same special tokens as RoBERTa, rather than using a WordPiece tokeniser like BERT.

   This will build its vocabulary from an alphabet of single bytes, so all words will be decomposable into tokens (no more <unk> tokens).

2. We then initialise our trainer with various parameters that were changed and run in order to test on several experiments.

   A few of these experiments involved varying the strength of weight decay, batch size, epochs etc.

Model	Params		Results (in %)			
	Epochs	Weight Decay	Acc	Macro F1	Precision	Recall
Prop-Liar	3	default	60.0	63.2	60.8	65.4
	3	0.01	62.0	59.2	61.0	57.4
Liar	3	default	63.0	56.2	66.0	48.0
	5	default	62.4	58.7	62.1	55.6

Table 4: Performance of pretrained Propaganda based Fake news model vs Simple Fake news model over various parameters

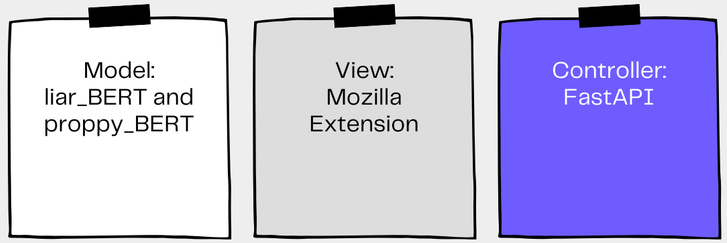
Stage 5

April 15th, 2021

Browser Extension

The browser extension has been created to alert the user in case the news they are reading has propaganda or fake content

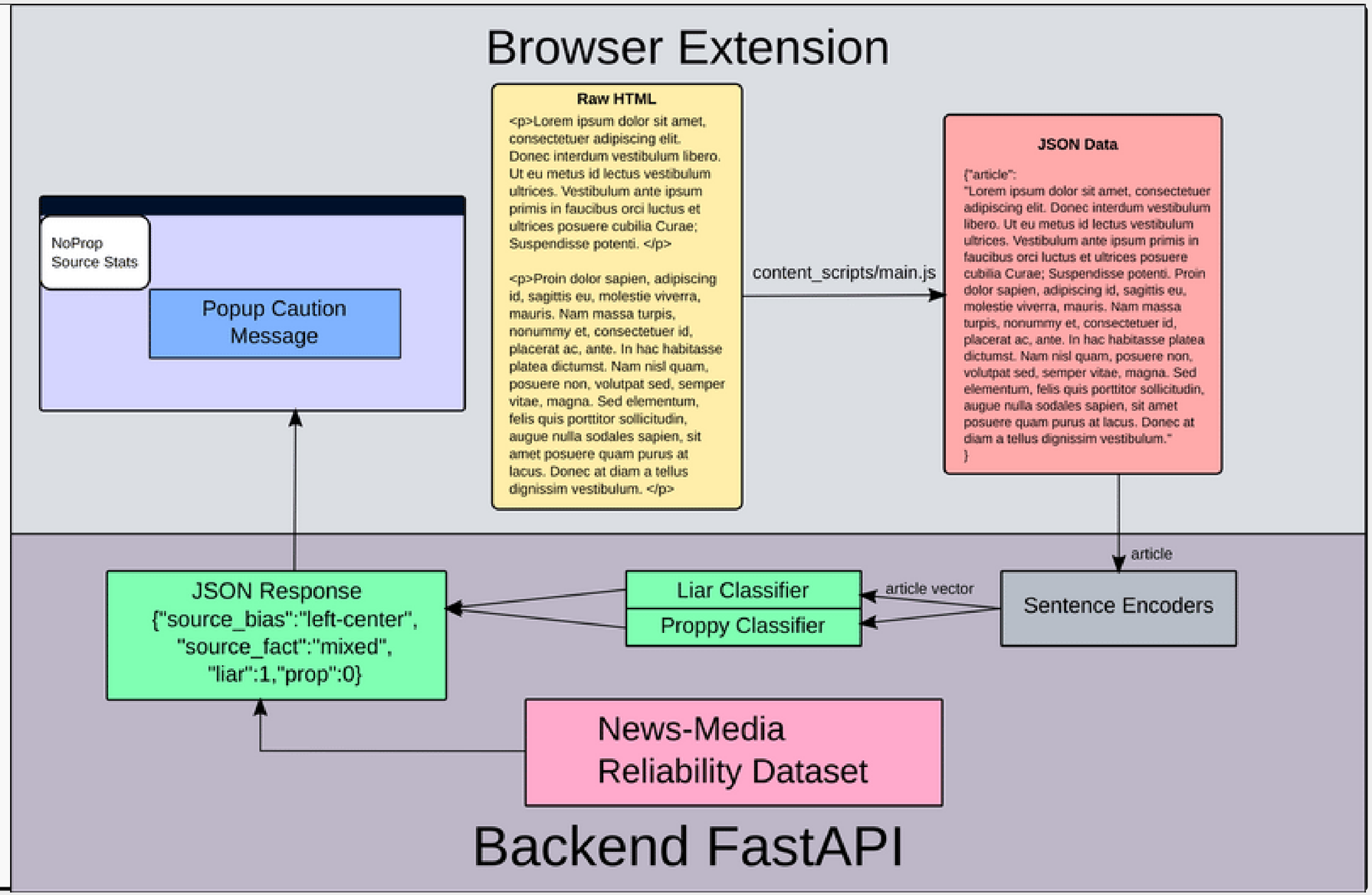
1. Uses MVC Architecture

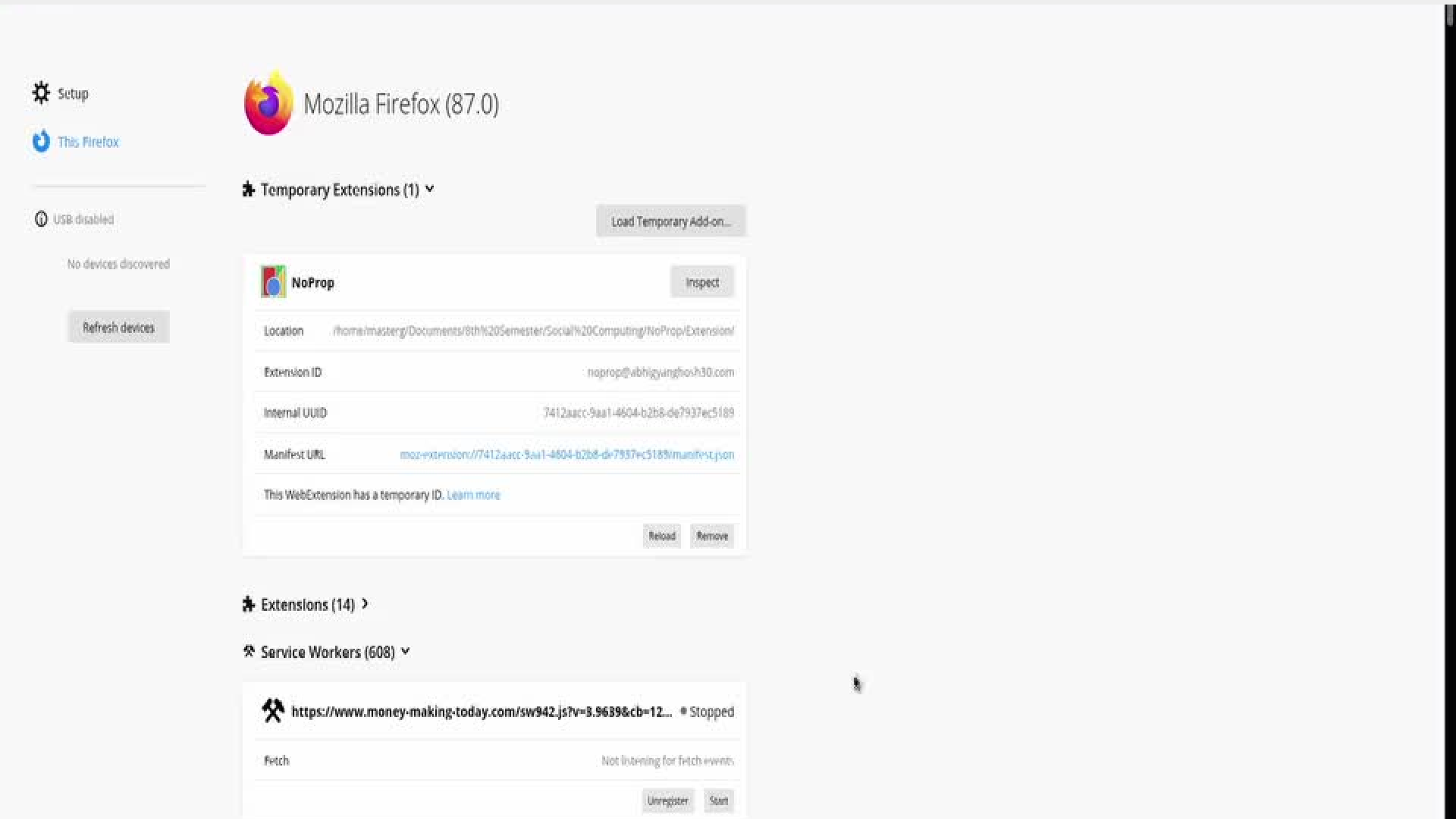


2. Gives an alert to the user in case of marked content

3. Option to turn off the feature as a control over the monitoring process

Architecture




Setup
This Firefox
USB disabled
No devices discovered
Refresh devices
Mozilla Firefox (87.0)
Temporary Extensions (1)
Load Temporary Add-on...
NoProp
Inspect
Location
/home/masterg/Documents/8th%20Semester/Social%20Computing/NoProp/Extension/
Extension ID
noprop@abhigyanghosh30.com
Internal UUID
7412aacc-9aa1-4604-b2b8-de7937ec5189
Manifest URL
moz-extension://7412aacc-9aa1-4604-b2b8-de7937ec5189/manifest.json
This WebExtension has a temporary ID. Learn more
Reload
Remove
Extensions (14)
Service Workers (608)
https://www.money-making-today.com/sw942.js?v=8.9689&cb=12...
Stopped
Fetch
Not listening for fetch events
Unregister
Start

Future Scope

1. Further finetuning of the pretrained model
2. Identification of span of propagandistic text
3. Identification of categories of propagandistic fake news (name calling/glittering generalities/card stacking/etc)
4. Verification of the factual content (using knowledge bases or other alternatives)

5. Percentage of veracity of text
6. Extend to domains beyond American political news

Thank you