

fake-real-news: A News Classifier in the Fight against Disinformation

Final Report

Huang, Sheng 3035534103 & Li, Yik Wai 3035566015

For better reading experience, we have more information documented in our GitHub Repo in Markdown.

This report is minimal. Please go to <https://github.com/vicw0ng-hk/fake-real-news> for **README.md**, **METHOD.md**, **FUNCTION.md** and **test.ipynb** and more.

Highlights

- We downloaded dataset from GitHub and conducted cleaning to make sure that the dataset we use for training contains and only contains rows that fits the criteria for training.
- We used a two-step approach to train our model. First, we trained a language model on the text using AWD-LSTM; then we added types to the training and produced a classifier model.
- We exported the model and built a web app on it. We also implemented a feedback mechanism and provided guidelines on continued model training.

Achievements

First, we conducted data engineering.

We first downloaded the dataset from GitHub, which is a huge file compressed in several zip multi-volume archive data. We first tried to use pandas to process the data, but failed. The data was so large, even the RAM allocated by GPU Farm wasn't enough. We then switched to use Dask because of its out-of-core characteristics. Since Dask is not compatible with some of the libraries we will use later, we saved the data into hundreds of smaller files, while deleting error lines.

We then used pandas to delete rows without any content and invalid types.

Then we started to train a language model.

After cleaning our data, we fed the content column in the dataset into fastai's `TextBlock`, which handles tokenization and numericalization automatically when passed to `DataBlock`. By doing so, we also set the splitter function, which help us split the dataset into training set and validation set. After passed on to `DataBlock`, we turn it into `dataloaders`.

And then we used fastai's `language_model_learner` to feed the data into the model and started training. After some time of training, we have a somewhat satisfactory result. We then saved the encoder.

We then started to train the classifier model.

The procedure is quite similar to the language model training, except we would also feed the corresponding categories into the learner. During training, we would also try to freeze certain number of layers to try to increase accuracy. After some time of training, we exported the model into a pickled file.

We then use the exported model to build a web app.

We used the minimal Flask web framework with some web programming technology such as jinja, jQuery, Bootstrap, CSRF Protection and database SQLite3. We also provided guidelines on continued model training.

Model Evaluation

We tested our model on several of the data files we saved in our data engineering step. The resulting accuracies vary by a lot from 30% to 98%. The model certainly could be further trained, but text model training requires a large amount of time and resources, and our dataset is extremely large. So, it's the kind of time and resources we simply did not have, even with quota provided by GPU Farm.

Another issue is that the dataset is a single-labeled one. See more in Limitations.

Limitations

The dataset we have is a single-label dataset. But this is not in accord with the reality. For example, many conspiracies are highly political, hence a lot of the articles with the conspiracy tag may also fit into the political tag. Hence, by this feature of the dataset, accuracy of training has not been very high for some of the test cases. And it is susceptible to overfitting if we train too much for higher accuracy, which is why we chose to present the predictions in the app by probabilities.

Also, the machine does not have the knowledge of common sense and general knowledge of society as humans, so some apparent signs may escape the model.

Of course, there is the time and resources problem we discussed in the previous segment. The RAM and GPU power provided cannot support our huge dataset in one go, so we had to separate into batches. More resources may improve the model's training performance.

News classifier in the fight against disinformation. Group Project for COMP3359 @ HKU.

AGPL-3.0 License

2 stars 0 forks

★ Unstar

👁 Unwatch ▾

- <> Code
- 🔔 Issues
- 🔗 Pull requests
- ▶ Actions
- 📁 Projects 1
- 🛡 Security
- 📈 Insights
- ⚙ Settings

🔗 main ▾

...

🐼 vicw0ng-hk Update README.md ...

7 minutes ago ⌚ 98

[View code](#)

☰ README.md

✎

fake-real-news

python v3.8.6 fastai v2.3.0 flask v1.1.2

A news classifier in the fight against disinformation by [HUANG, Sheng](#) & [LI, Yik Wai](#). 🤝

- Group project for [COMP3359 Artificial Intelligence Applications @ HKU](#) 🏢

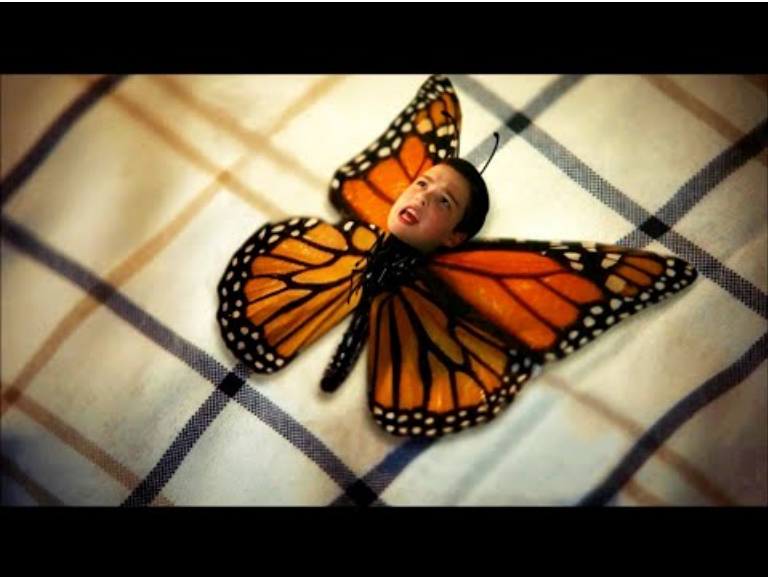
🗣 "If you tell a lie big enough and keep repeating it, people will eventually come to believe it. The lie can be maintained only for such time as the State can shield the people from the political, economic and/or military consequences of the lie. It thus becomes vitally important for the State to use all of its powers to repress dissent, for the truth is the mortal enemy of the lie, and thus by extension, the truth is the greatest enemy of the State."

[Joseph Goebbels, Reich Minister of Propaganda, Nazi Germany](#)

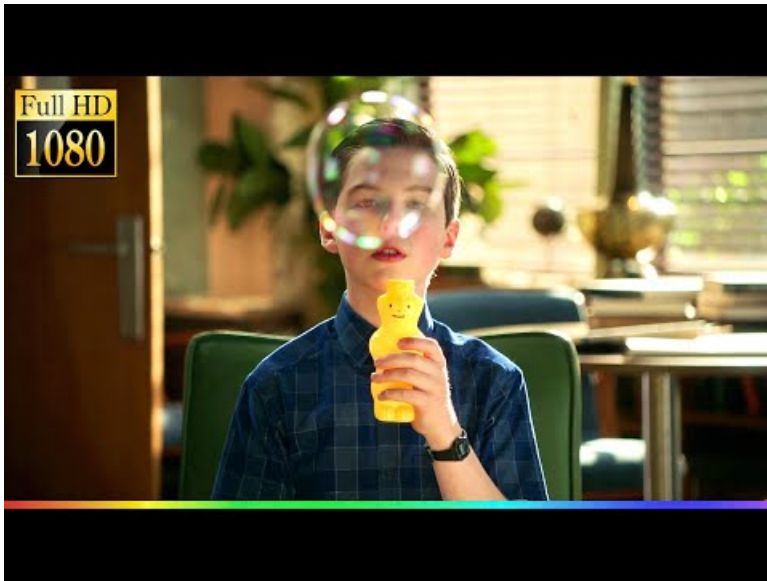
Mission 🚢

What is our relationship with the truth (or, the reality)? 😐 That is a philosophical question. 🤖

Great minds struggle with this question.



Young Sheldon Cooper struggled with this problem and wanted to switch major to philosophy. But in the end, he returned to science when he realized physics theories could explain more patterns in nature.

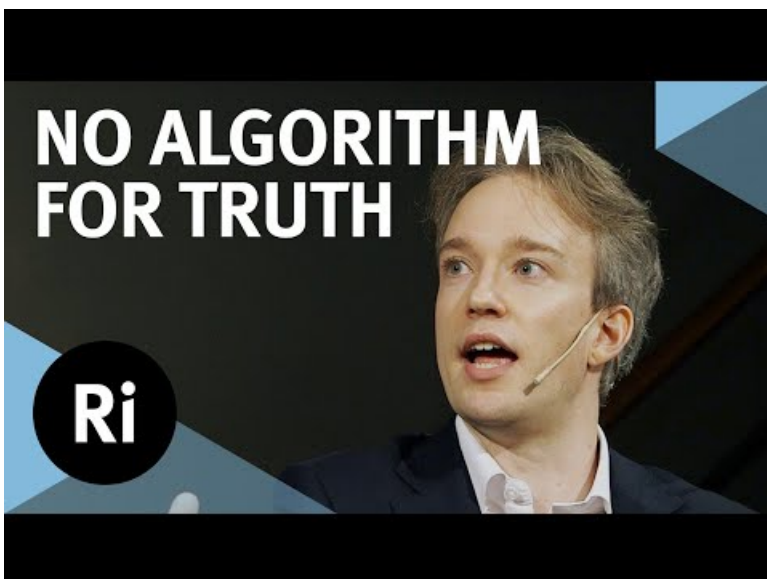


As fake news spreads wildly today, we have been in a similar crisis. 🤖 With so much information, how can you tell what is real and what is not? 🤖 Especially, with the cost of reading a news article so low and the cost of verifying the facts so high, how can anyone make judgments on the authenticity of the news content? 😞 Some may say nothing is real and lead a life without any thoughts on the world. 😞 Some may say they trust authoritative sources. But how do you define authoritative source? Is everything put out by these authoritative sources guaranteed to be true? 😞 Can we find patterns on these articles, also taking into account its sources, and then make a better judgment, or a more educated guess? 🤖

Indeed we CAN find linguistic patterns in news articles. 😊 These patterns may well have correlations to the realness or fakeness of these articles. But it's not that simple. 🤖 Remember that news media have biases. For example, in the United States, conservative media such as *Fox News* and liberal news media such as *MSNBC* and *CNN* have different styles when reporting news 🤖, but that doesn't automatically mean that one style is equal to fake news reporting. (Check out [Media Bias Ratings](#)) To avoid such biases when training our model, we recognize news articles that cannot be easily categorized as real or fake, such as pieces that are strong in opinions. (Check out the types we have [here](#) 🖱)

We concede that this approach is still flawed, which will be discussed in [Limitations](#). 😞 However, it can give us somewhat of a reference when we are judging the authenticity of an article, as we have given explicit descriptions on how we categorize the articles. 😊 Still, users should keep in mind that we are not the arbiter of truth and that the model cannot replace the work of a professional fact checker - it cannot visit the places where events happened; it cannot interview people involved in the stories; it cannot know the intention of the publishers when they put out the story. 🤖

THERE IS NO ALGORITHM FOR TRUTH.






Reports 📄

- Proposal 📄 pdf


- Interim Report  [pdf](#)

Running

It's highly  recommended to run the app on a Unix-like system (GNU/Linux, macOS, ...).  Using Windows may cause some issues when installing dependencies. 

0. Cloning the repository



```
git clone https://github.com/vicw0ng-hk/fake-real-news.git
```

Or, clone through SSH for better security. 

```
git clone git@github.com:vicw0ng-hk/fake-real-news.git
```

Or, clone with [GitHub CLI](#) 

```
gh repo clone vicw0ng-hk/fake-real-news
```

Due to the large size of our model, it is stored with [Git LFS](#), and because of [GitHub's bandwidth limit](#) , please use this [link](#)  to download [app/model/model.pkl](#) and replace the file in the cloned directory.

1. Installing environment

This may be different depending on the virtualization technology you are using , but generally do

```
cd app/  
pip3 install -r requirements.txt
```

2. Run the app!

```
python3 app.py
```




Methodology

Check out the [Methodology](#) document.

Functionalities

Check out the [Functionalities](#) document.

Limitations

- One major limitation is from the categorization of our dataset.  The dataset we have is a single-label dataset. But this is not in accord with the reality. For example, many conspiracies are highly political, hence a lot of the articles with the `conspiracy` tag may also fit into the `political` tag. Hence, by this feature of the dataset, accuracy of training has not been very high for some of the test cases. And it is susceptible to overfitting if we train too much for higher accuracy, which is why we chose to present the predictions in the app by probabilities. (Check out [Functionalities](#))
- Another limitation is our development time and resources.  We have a very large dataset (Check out [Methodology](#)). However, we cannot make full use of it because we have limited time and resources allocated by GPU Farm is relatively restrictive compared to the size of our dataset. Hence, we used only a portion of the total data to train our model.
- There is also the limitation of the capabilities of machines.  We can only use the content (plus its URL, Title and Authors) to decide the categorization of new articles. For some articles, humans could easily tell their nature and authenticity based on common sense and general knowledge. However, the model cannot think that way, so some of the easy-to-recognize evidence to a human is difficult to find for the model.

Terms and Conditions

In addition to the restrictions of [GNU Affero General Public License v3.0](#) of this repo, you also agree to the following terms and conditions:

YOUR USE OF THIS WEB APP CONSTITUTES YOUR AGREEMENT TO BE BOUND BY THESE TERMS AND CONDITIONS OF USE.

1. The classification of the text you submit to this web app is in no way legal recognition. The web app and/or its authors bear no legal responsibilities for its result. If you choose to publish the result, the web app and/or its authors shall not bear any legal consequences relating to this action.
2. You shall be liable for the legal responsibilities of the copyright of the text you submit to this web app. You shall gain the right to copy the text before you submit it to the web app.
3. This web app shall not be used by any political organization and/or any entity, partially or entirely, directly or indirectly, funded and/or controlled by a political organization in any jurisdiction.
4. In case of any discrepancy with any other licenses, terms or conditions associated with this web app and/or its repository, this agreement shall prevail.

Languages



● Jupyter Notebook 70.5% ● HTML 24.2% ● Python 5.3%

Methodology 🤖

This project comprises of two 🙌 major parts: [Model](#) and [Web App](#).

We built both parts on GNU/Linux, with model trained on the [HKU CS GPU Farm](#) and web app developed on [Pop!_OS 20.10](#).



Model

Data

Undoubtedly, we need data to train our model. Thanks to [@serveral27](#) 🙏, we have a [FakeNewCorpus](#). 😊

Size of Original Corpus is 29 GB 🤖, but we don't need all the data.

In the first phase of the project, we deleted certain not-so-useful columns and trimmed the data down to 25 GB. 🙄

```
import dask.dataframe as dd
df = dd.read_csv('news_cleaned_2018_02_13.csv', usecols=['url', 'title', 'authors', 'content', 'type'],
                engine='python', error_bad_lines=False)
df.to_csv('fake.csv', index=False)
```

Due to the large size of the corpus, [Dask](#), instead of [Pandas](#), is used to process the initial data for its [out-of-core](#) characteristics.

👍 We saved the data into hundreds of smaller files.

In the second phase of the project, we concatenated some of the columns together as they can all be handled by a language model. 😊 We also deleted rows without a valid type and content and the size came down to 22 GB.

```
df = df.dropna(subset=['type', 'content'])
label = ['fake', 'satire', 'bias', 'conspiracy', 'junksci', 'hate', 'clickbait', 'unreliable', 'political', 'reliable']


for i in range(459):
    df = pd.read_csv('fake.csv/{0:03}.part'.format(i), engine='python', error_bad_lines=False, keep_default_na=False)
    df = df[df['type'].isin(label)]
    df['content'] = df['url'].str.cat(df[['title', 'authors', 'content']], sep=' | ')
    df = df.drop(['url', 'title', 'authors'], axis=1)
    df.to_csv('fake1.csv/{0:03}.part'.format(i), index=False)
```


Types

Note the possible categories from the dataset:

Type	Tag	Description
Fake News	fake	Sources that entirely fabricate information, disseminate deceptive content, or grossly distort actual news reports
Satire	satire	Sources that use humor, irony, exaggeration, ridicule, and false information to comment on current events.
Extreme Bias	bias	Sources that come from a particular point of view and may rely on propaganda, decontextualized information, and opinions distorted as facts.
Conspiracy Theory	conspiracy	Sources that are well-known promoters of kooky conspiracy theories.
Junk Science	junksci	Sources that promote pseudoscience, metaphysics, naturalistic fallacies, and other scientifically dubious claims.
Hate News	hate	Sources that actively promote racism, misogyny, homophobia, and other forms of discrimination.
Clickbait	clickbait	Sources that provide generally credible content, but use exaggerated, misleading, or questionable headlines, social media descriptions, and/or images.
Proceed With Caution	unreliable	Sources that may be reliable but whose contents require further verification.
Political	political	Sources that provide generally verifiable information in support of certain points of view or political orientations.
Credible	reliable	Sources that circulate news and information in a manner consistent with traditional and ethical practices in journalism (Remember: even credible sources sometimes rely on clickbait-style headlines or occasionally make mistakes. No news organization is perfect, which is why a healthy news diet consists of multiple sources of information).

Language Model

To build a text classifier model we must first build a language model. 

[fastai](#) handles tokenization and numericalization automatically when `TextBlock` is passed to `DataBlock`. All of the arguments that can be passed to `Tokenizer` and `Numericalize` can also be passed to `TextBlock`. 

```
from fastai.text.all import *


dls_lm = DataBlock(blocks=TextBlock.from_df('content', is_lm=True),
                    get_x=ColReader('text'), splitter=RandomSplitter(0.1)
                    ).dataloaders(df, bs=32, seq_len=80)
```

The reason that `TextBlock` is special is that setting up the numericalizer's vocab can take a long time (we have to read and tokenize every document to get the vocab).

fastai has some optimization on this: 

- It saves the tokenized documents in a temporary folder, so it doesn't have to tokenize them more than once.
- It runs multiple tokenization processes in parallel, to take advantage of your computer's CPUs.

Now that we have handled our data, let's fine-tune the pretrained language model.

To convert the integer word indices into activations that we can use for our neural network, we will use [embeddings](#). Then we'll feed those embeddings into a [recurrent neural network \(RNN\)](#), using an architecture called [AWD-LSTM](#). (Check out [Smerity et al.](#)) 

the embeddings in the pretrained model are merged with random embeddings added for words that weren't in the pretraining vocabulary. This is handled automatically inside `language_model_learner`:

```
learn = language_model_learner(
    dls_lm, AWD_LSTM, drop_mult=0.3,
    metrics=[accuracy, Perplexity()]).to_fp16()
```


The loss function used by default is cross-entropy loss, since we essentially have a classification problem. The perplexity metric used here is often used in NLP for language models: it is the exponential of the loss (`torch.exp(cross_entropy)`). We also include the accuracy metric to see how many times our model is right when trying to predict the next word, since cross entropy is both hard to interpret and tells us more about the model's confidence than its accuracy.

Then we can start to use `learn.fit_one_cycle()` to train our language model (we also used `learn.lr_find()` to find a good learning rate). After each time we trained, we would use `learn.save()` to save a copy of the state of our model as it takes quite a while to train an epoch. When come back to training after leaving for something else, we can use `learn = learn.load()` to load the state back, unfreeze the state by `learn.unfreeze()` and continue training. We would substitute new data using `learn.dls = dls_lm_new` as the RAM limits on GPU Farm does not allow us to use all data in one go. The accuracy of the language model came to upper 30s to 40s percent.

After training language model, we saved the model (excluding the final layer, of course) using

```
learn.save_encoder('finetuned') . 📁
```

Classifier Model

This is similar to what we have done above , with dataloaders as:

```
dls_clas = DataBlock(blocks=(TextBlock.from_df('content'),CategoryBlock),
    get_x=ColReader('text'), get_y=ColReader('type'),
    splitter=RandomSplitter(0.1)
).dataloaders(df, bs=32, seq_len=80)
```

The major difference is that `is_lm` is gone because this is no longer training for a language model, and there is an additional `CategoryBlock` .

We can now create a model using `text_classifier_learner` to classify our texts:

```
learn = text_classifier_learner(dls_clas, AWD_LSTM, drop_mult=0.5,
    metrics=accuracy).to_fp16()
```


And then load the encoder we saved earlier:

```
learn = learn.load_encoder('finetuned')
```

Then we start training using `learn.fit_one_cycle()` . We also used `learn.freeze_to()` to freeze some parameters and then train the model, after which we use `learn.unfreeze()` to unfreeze the parameters and then train for a few more epochs. The accuracy came to about 90% for the validation set of the data we used for training. However, since the RAM limits we could not use most of our data and the accuracy on some other data may vary from 30% - 98% (check out one [testing example](#)). 😊 This doesn't seem top-notch, but it's acceptable since news articles generally fit into multiple categories and yet we only have a single-label dataset. More on this in [our discussion of limitations](#). 🤔

After training, simply use `learn.export()` to export the model for the web app in the next step.

Web app

We used [Flask](#), a minimal Python web framework. 🐍 We love its simplicity. And you can see this by the only main part of our server side  `app/app.py` .

Some Highlights 📸

- Since Flask supports [Jinja](#), we used it for our [HTML templates](#), adding a dynamic component to static templates.
- We used [Bootstrap](#) to customize the look and feel of our web pages. We also included JavaScript and [jQuery](#) to implement certain styles.

- We implemented CSRF Protection ([CSRFProtect](#) from [Flask-WTF](#)) in POST form submissions.
- We used [Flask-SQLAlchemy](#) to interact with [SQLite3](#), which stores user feedback and we can use the feedback to better train our model. More on this in [our discussion of functionalities](#).
- We used `load_learner` from `fastai` to load the trained model and then use the model to get predictions on our input.



vicw0ng-hk Update FUNCTION.md

🕒 History

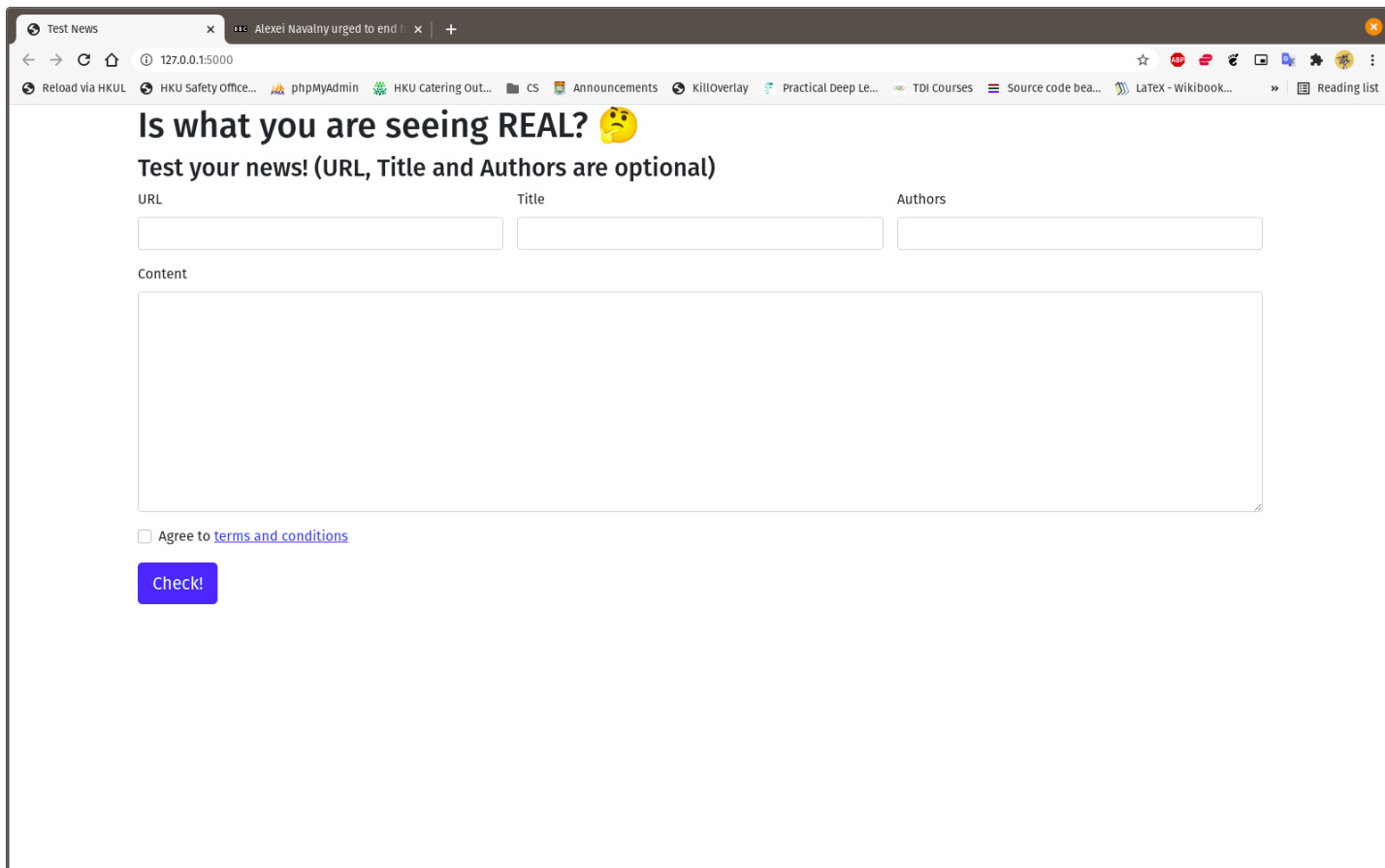
👤 1 contributor

67 lines (39 sloc) | 2.12 KB

Functionalities

After running app (Check [how to run](#)), go to the link prompted in the terminal, which is usually `localhost:5000`, and you will see...

0



Test News

127.0.0.1:5000

Reload via HKUL HKU Safety Office... phpMyAdmin HKU Catering Out... CS Announcements KillOverlay Practical Deep Le... TDI Courses Source code bea... LaTeX - Wikibook... Reading list

Is what you are seeing REAL? 🤔

Test your news! (URL, Title and Authors are optional)

URL Title Authors

Content

☐ Agree to [terms and conditions](#)

Check!

And then you will find out that `URL`, `Title` and `Authors` are optional fields...

1

Test News

127.0.0.1:5000

Reload via HKUL HKU Safety Office... phpMyAdmin HKU Catering Out... CS Announcements KillOverlay Practical Deep Le... TDI Courses Source code bea... LaTeX - WikiBook... Reading list

Is what you are seeing REAL? 🤔

Test your news! (URL, Title and Authors are optional)

URL Title Authors

Looks good! Looks good! Looks good!

Content

Please provide the content.

☐ Agree to [terms and conditions](#)
You must agree before submitting.

Check!

You can find a news article and fill in the information (I am going to use [this](#))...

2

Test News

127.0.0.1:5000

Reload via HKUL HKU Safety Office... phpMyAdmin HKU Catering Out... CS Announcements KillOverlay Practical Deep Le... TDI Courses Source code bea... LaTeX - WikiBook... Reading list

Is what you are seeing REAL? 🤔

Test your news! (URL, Title and Authors are optional)

URL Title Authors

<https://www.bbc.com/news/world-europe-5685426> Looks good! Alexei Navalny urged to end hunger strike immediately Looks good! BBC Looks good!

Content

Doctors of jailed Russian opposition leader Alexei Navalny have urged him to immediately end his hunger strike, warning that otherwise he could die.

In a statement, the five say they have been shown the results of his medical tests conducted on 20 April.

"If the hunger strike continues even for a minimal amount of time, unfortunately, we will simply have no-one to treat soon," they say.

Thousands of Russians rallied on Wednesday demanding Navalny be freed.

More than 1,000 people were arrested in a number of cities including the capital Moscow, according to Russia's monitoring group OVD-Info.

☒ Agree to [terms and conditions](#)

Check!

Agree to our [terms and conditions](#), and you can click on **Check!**...

3

Test Result

127.0.0.1:5000/result/

Is what you are seeing REAL?

Test result: 🤖

Alexei Navalny urged to end hunger strike immediately by BBC

Type	Probability
political	0.6912940144538879
conspiracy	0.11011889576911926
junksci	0.08063260465860367
bias	0.041956327855587006
clickbait	0.03865810111165047
fake	0.013845155946910381
unreliable	0.00985281728208065
satire	0.006140087265521288
reliable	0.003953966312110424
hate	0.0035480563528835773

Learn more about the types [here](#).

What do you think this article is most likely?

bias

Submit Feedback!

Back!

It is indeed a political coverage! But if you disagree, choose another type...

4

Test Result

127.0.0.1:5000/result/

Is what you are seeing REAL?

Test result: 🤖

Alexei Navalny urged to end hunger strike immediately by BBC

Type	Probability
political	0.6912940144538879
conspiracy	0.11011889576911926
junksci	0.08063260465860367
bias	0.041956327855587006
clickbait	0.03865810111165047
fake	0.013845155946910381
unreliable	0.00985281728208065
satire	0.006140087265521288
reliable	0.003953966312110424
hate	0.0035480563528835773

Learn more about the types [here](#).

What do you think this article is most likely?

reliable

Submit Feedback!

Back!

And click on Submit Feedback...

5

Test Result

127.0.0.1:5000/result/

Is what you are seeing REAL?

Test result: 😲

Alexei Navalny urged to end hunger strike immediately by BBC

Type	Probability
political	0.6912940144538879
conspiracy	0.11011889576911926
junksci	0.08063260465860367
bias	0.041956327855587006
clickbait	0.03865810111165047
fake	0.013845155946910381
unreliable	0.00985281728208065
satire	0.006140087265521288
reliable	0.003953966312110424
hate	0.0035480563528835773

Learn more about the types [here](#).

What do you think this article is most likely?

reliable

Feedback recorded!

Back!

And you will see that it has been recorded, and you can go back now by clicking no **Back!**...

Sustainability 🌍

What happens to the recorded data? 😊 If you check `app/app.py`, you will see that it has been recorded in a database called `feedback.sqlite3`. You can use the following code to get recorded feedbacks:

```
import pandas as pd
import sqlite3

conn = sqlite3.connect('feedback.sqlite3')
df = pd.read_sql_query('SELECT * from Feedback', conn, index_col='id').reset_index(drop=True, inplace=True)
```

You can further clean the data as in [Methodology](#) and then use

```
from fastai.text.all import *

dls_clas = DataBlock(blocks=(TextBlock.from_df('content'), CategoryBlock),
                      get_x=ColReader('text'), get_y=ColReader('type'),
                      splitter=RandomSplitter(0.1)
                      ).dataloaders(df, bs=32, seq_len=80)

learn = load_learner('model.pkl')
learn.dls = dls_clas
```

to load the data into the model and start training as discussed in [Methodology](#). After that, simply export the model using `learn.export('model.pkl')` and replace the `app/model/model.pkl` with the newly trained one.

```
In [1]: import pandas as pd
        from fastai.text.all import *

In [2]: learn = load_learner('model.pkl')

In [3]: df = pd.read_csv('fake.csv/test.part')

        df.head()

Out[3]:
```

	type	content
0	fake	http://beforeitsnews.com/space/2016/11/pluto-has-a-subsurface-antifreeze-ocean-2503540.html Pluto Has a Subsurface 'Antifreeze' Ocean Universe Today Pluto Has a Subsurface 'Antifreeze' Ocean\n\nHeadline: Bitcoin & Blockchain Searches Exceed Trump! Blockchain Stocks Are Next!\n\nThe evidence keeps growing for a large subsurface ocean at Pluto, which also provides clues how the iconic 'heart' of Pluto was formed.\n\nWe reported in early October that thermal models of Pluto's interior and tectonic evidence suggest an ocean may exist beneath Pluto's heart-shaped Sputnik Planitia. Now, ne...
1	fake	http://beforeitsnews.com/space/2016/11/new-theory-of-gravity-does-away-with-need-for-dark-matter-2503546.html New Theory of Gravity Does Away With Need for Dark Matter Universe Today New Theory of Gravity Does Away With Need for Dark Matter\n\nHeadline: Bitcoin & Blockchain Searches Exceed Trump! Blockchain Stocks Are Next!\n\nErik Verlinde explains his new view of gravity\n\nLet's be honest. Dark matter's a pain in the butt. Astronomers have gone to great lengths to explain why it must exist and exist in huge quantities, yet it remains hidden. Unknown. Emitting no visible energy...
2	fake	http://beforeitsnews.com/space/2016/11/weekly-space-hangout-november-18-2016-dr-jason-wright-and-tabby-star-2503598.html Weekly Space Hangout - November 18, 2016: Dr. Jason Wright and Tabby's Star Universe Today Weekly Space Hangout - November 18, 2016: Dr. Jason Wright and Tabby's Star\n\n% of readers think this story is Fact. Add your two cents.\n\nHeadline: Bitcoin & Blockchain Searches Exceed Trump! Blockchain Stocks Are Next!\n\nHost: Fraser Cain (@fcain)\n\nSpecial Guest:\n\nDr. Jason Wright is Professor in Penn State University's Department of Astronomy and Astrophysics. Jaso...
3	political	http://www.washingtonexaminer.com/susan-collins-hints-trumps-backing-could-get-obamacare-bills-through-house/article/2642535 Susan Collins hints Trump's backing could get Obamacare bills through House Robert King Sen. Susan Collins hinted Monday that backing from President Trump would help get two Obamacare stabilization bills passed in the House.\n\nThe Maine Republican said she received an "ironclad commitment" from Senate Majority Leader Mitch McConnell, R-Ky., and President Trump that the two bills would become law by the end of the year. Missing from that statement is a commitme...
4	political	http://www.washingtonexaminer.com/trent-franks-resigns-immediately-from-congress-cites-wifes-hospitalization-after-surrogacy-revelations/article/2642995 Trent Franks resigns immediately from Congress, cites wife's hospitalization after surrogacy revelations Al Weaver Rep. Trent Franks, R-Ariz., announced Friday that he is resigning from Congress effective immediately instead of waiting until the end of January.\n\nFranks said Thursday he would leave next month after it emerged that he discussed surrogate parenthood with two female staffers - some reports said he asked them directly t...

```
In [4]: dls_clas = DataBlock(blocks=(TextBlock.from_df('content'),CategoryBlock),
        get_x=ColReader('text'), get_y=ColReader('type'),
        splitter=RandomSplitter(0.1)
        ).dataloaders(df, bs=32, seq_len=80)

        dls_clas.show_batch(max_n=3)
```

/userhome/cs/vicw0ng/anaconda3/lib/python3.8/site-packages/numpy/core/_asarray.py:83: VisibleDeprecationWarning: Creating a ndarray from ragged nested sequences (which is a list-or-tuple of lists-or-tuples-or ndarrays with different lengths or shapes) is deprecated. If you meant to do this, you must specify 'dtype=object' when creating the ndarray

```
return array(a, dtype, copy=False, order=order)
```

		text	category
0		xxbos http : / / beforeitsnews.com / religion / 2013 / 07 / xxunk xxup how xxup the xxup apostles xxup deceived xxup humanity xxup about xxup god xxup and xxup jesus xxup how xxup the xxup apostles xxup deceived xxup humanity xxup about xxup god xxup and xxup jesus \n\n % of readers think this story is xxmaj fact . xxmaj add your two cents . \n\n xxmaj headline : xxmaj bitcoin & xxmaj blockchain xxmaj searches xxmaj exceed xxmaj trump ! xxmaj blockchain xxmaj stocks xxmaj are xxmaj next ! \n\n xxup how xxup the xxup apostles xxup deceived xxup humanity xxup about xxup god xxup and xxup jesus \n\n xxmaj for the past two thousand years , xxmaj christianity has led mankind to believe that what the apostles wrote about xxmaj jesus in the xxmaj new xxmaj testament of the xxmaj bible , especially	fake
1		xxbos http : / / xxrep 3 w .unz.com / article / california - and - the - end - of - white - america / xxmaj california and the xxmaj end of xxmaj white xxmaj america xxmaj ron xxmaj unz , xxmaj sayed xxmaj hasan , xxup a . xxmaj graham , xxmaj the xxmaj rubin xxmaj report xxup summary \n\n xxmaj californians of xxmaj european xxunk a minority near the end of the 1980s , and this unprecedented ethnic transformation is probably responsible for the rise of a series of ethnically - charged political issues such as immigration , affirmative action , and bilingual education , as seen in xxmaj propositions 187 , 209 , and 227 . xxmaj since xxmaj america as a whole is undergoing the same ethnic transformation delayed by a few decades , the experience of these controversial campaigns tells us much	bias
2		xxbos http : / / beforeitsnews.com / alternative / 2016 / 03 / as - larger - and - more - numerous - fireballs - scream - across - the - sky - and - strong - winds - blow - all - over - the - world - here - is - what - you - need - to - know - to - make - it - through - post - nibiru - and - through - events - leading - up - to - xxunk xxmaj you xxmaj can xxmaj make xxmaj it xxmaj through post - nibiru xxmaj and xxmaj poleshift ! ! xxmaj here xxmaj is xxmaj what xxmaj you xxmaj need xxmaj to xxmaj know xxmaj as xxmaj fireballs xxmaj scream xxmaj across xxmaj the xxmaj sky xxmaj and xxmaj strong xxmaj winds xxmaj blow xxmaj all xxmaj over xxmaj the xxmaj world	fake

```
In [5]: learn.dls = dls_clas

In [6]: learn.show_results()
```

	text	category	category_
--	------	----------	-----------

	text	category	category_
0	<p>xxbos http : // beforeitsnews.com / new - world - order / 2015 / 11 / xxunk xxup nwo xxmaj mass xxmaj hypnosis xxmaj program xxup nwo xxmaj mass xxmaj hypnosis xxmaj program \n\n xxmaj headline : xxmaj bitcoin & xxmaj blockchain xxmaj searches xxmaj exceed xxmaj trump ! xxmaj blockchain xxmaj stocks xxmaj are xxmaj next ! \n\n 1 xxmaj new xxmaj world xxmaj order xxmaj underworld xxmaj mass xxmaj hypnosis xxmaj program xxmaj introduction 1 xxmaj mass xxunk and requests to reprint an earlier news article published earlier on xxup bin concerning real estate scams involving use of xxmaj new xxmaj mexico xxup llc companies to launder illicit real estate monies earned by a world meditation company and laundered by an xxmaj american businessman who created a multimillion dollar real estate empire from his relationship with the company generated a more detailed news -</p>	fake	junksci
1	<p>xxbos https : // xxrep 3 w .naturalnews.com / xxunk xxmaj the xxmaj mothers xxmaj act xxmaj disease xxmaj mongering xxmaj campaign - xxmaj part xxup iv xxmaj the a xxmaj team \n\n xxmaj old xxmaj chemical xxmaj imbalance in the xxmaj brain xxmaj scam \n\n xxmaj sad xxmaj xxunk \n\n xxmaj respected xxmaj researchers support the xxmaj mothers xxmaj act \n\n xxmaj and xxmaj speaking of xxmaj harvard \n\n xxmaj disease xxmaj mongering in the xxmaj media \n\n xxmaj time xxmaj magazine xxmaj blasted \n\n xxmaj amy xxmaj liked xxmaj it \n\n xxmaj plan of xxmaj attack \n\n xxmaj internet xxmaj battle xxmaj breaks xxmaj out \n\n xxmaj grohol 's xxmaj internet xxmaj one - stop \n\n (naturalnews) xxmaj this is part four of an article series by xxmaj evelyn xxmaj pringle . xxmaj find previous parts here : xxmaj part xxmaj one</p>	junksci	junksci
2	<p>xxbos http : // xxrep 3 w .vdare.com / radios / radio - derb - somalis - salvadorans - and - strategic - deportation - etc xxmaj radio xxmaj derb : xxmaj somalis , xxmaj salvadorans , xxmaj and xxmaj strategic xxmaj deportation , xxmaj etc . xxunk — xxmaj somali averages are terrible . (and we 've imported them .) \n\n xxunk — xxmaj somalis ' twofer privilege . (un - identifiable , un - xxunk .) \n\n xxunk — xxmaj it ai n't over till the alien wins . (also un - deportable .) \n\n xxunk — xxup ms-13 among the kulaks . (news from xxmaj xxunk .) \n\n xxunk — xxmaj immigrants are better than us , series # xxunk . (burn down the xxmaj ivy xxmaj league !) \n\n xxunk — xxmaj</p>	bias	political
3	<p>xxbos http : // beforeitsnews.com / power - elite / 2016 / 03 / if - you - want - to - know - why - americans - seem - so - cold - and - heartless - lately - then - read - what - the - ruling - khazars - have - done - through - their - drug - companies - and - our - xxunk xxmaj if xxmaj you xxmaj want xxmaj to xxmaj know xxmaj why xxmaj americans xxmaj seem xxmaj so xxmaj cold xxmaj and xxmaj heartless xxmaj lately , xxmaj then xxmaj read xxmaj what xxmaj the xxmaj ruling xxmaj khazars xxmaj have xxmaj done , xxmaj through xxmaj their xxmaj drug xxmaj companies xxmaj and xxmaj our xxmaj fda xxmaj the xxmaj vatic xxmaj project xxmaj headline : xxmaj bitcoin & xxmaj blockchain xxmaj searches xxmaj exceed xxmaj</p>	fake	fake
4	<p>xxbos http : // beforeitsnews.com / spirit / 2016 / 06 / xxunk xxmaj the " spirit of xxmaj man " — xxmaj do xxmaj you xxmaj know what that " spirit " xxmaj is ? xxmaj the " spirit of xxmaj man " — xxmaj do xxmaj you xxmaj know what that " spirit " xxmaj is ? \n\n xxmaj headline : xxmaj bitcoin & xxmaj blockchain xxmaj searches xxmaj exceed xxmaj trump ! xxmaj blockchain xxmaj stocks xxmaj are xxmaj next ! \n\n xxup the xxup spirit — " of " — xxup man \n\n xxmaj this xxmaj truth found in the xxmaj word of xxmaj god will be quickly rejected by " nominal " xxmaj christians — it invalidates all their doctrines . xxmaj the " spirit — xxup of — xxmaj man " is — not — an immortal ' soul '</p>	fake	fake
5	<p>xxbos http : // xxrep 3 w .rense.com / general96 / xxunk xxmaj appeal xxmaj of xxmaj ban xxmaj aspartame xxmaj petition xxmaj and xxmaj imminent xxmaj health xxmaj hazard xxmaj to xxup fda - xxmaj dr . xxmaj joseph xxmaj thomas with copy to xxmaj dr . xxmaj david xxmaj hattan \n\n\n\n xxmaj dear xxmaj dr . xxmaj thomas : \n\n\n\n xxmaj this is the second request appealing the original request for ban aspartame that took xxup fda 14 years to answer . xxmaj let me remind you of the law . xxmaj you have 180 days to answer . xxmaj the xxmaj imminent xxmaj health xxmaj hazard petition is suppose to be answered in a week or ten days and i filed in 2007 and its never been answered . \n\n\n\n xxmaj you have found a loop - hole where you can always answer</p>	conspiracy	conspiracy
6	<p>xxbos http : // beforeitsnews.com / christian - news / 2016 / 08 / xxunk xxup the xxup limits xxup of xxup the xxup church by xxmaj fr xxmaj georges xxmaj florovsky xxmaj monks xxmaj and xxmaj mermaids , a xxmaj benedictine xxmaj blog xxup the xxup limits xxup of xxup the xxup church by xxmaj fr xxmaj georges xxmaj florovsky \n\n xxmaj headline : xxmaj bitcoin & xxmaj blockchain xxmaj searches xxmaj exceed xxmaj trump ! xxmaj blockchain xxmaj stocks xxmaj are xxmaj next ! \n\n " the xxmaj limits of the xxmaj church " by xxmaj fr . xxmaj georges xxmaj florovsky \n\n xxmaj the following piece by xxmaj xxunk xxmaj georges xxup v. xxmaj florovsky was originally published in 1933 in xxmaj church xxmaj quarterly xxmaj review . xxmaj where xxmaj florovsky does not translate foreign phrases , we have supplied a translation</p>	fake	fake
7	<p>xxbos http : // beforeitsnews.com / survival / 2014 / 04 / xxunk xxmaj was xxmaj franklin xxmaj roosevelt a xxmaj communist ? b xxmaj mans xxmaj revolt xxmaj was xxmaj franklin xxmaj roosevelt a xxmaj communist ? \n\n xxmaj headline : xxmaj bitcoin & xxmaj blockchain xxmaj searches xxmaj exceed xxmaj trump ! xxmaj blockchain xxmaj stocks xxmaj are xxmaj next ! \n\n xxmaj was xxmaj franklin xxmaj roosevelt a xxmaj communist ? by xxup dc xxmaj dave \n\n xxmaj john xxmaj beaty in his 1951 book , xxmaj the xxmaj iron xxmaj curtain xxmaj over xxmaj america , asks the following rhetorical question (p. 187) : \n\n xxmaj in solemn truth , do not seven persons share most of the responsibility for establishing the xxmaj communist grip on the world ? xxmaj are not the seven : (1) xxmaj marx</p>	fake	fake
8	<p>xxbos http : // beforeitsnews.com / financial - markets / 2017 / 01 / new - year - naps - top - stocks - xxunk - how - to - play - the - xxunk xxmaj new xxmaj year xxup naps — xxmaj top xxmaj stocks for 2017 and how to play the xxmaj joker xxmaj new xxmaj year xxup naps — xxmaj top xxmaj stocks for 2017 and how to play the xxmaj joker \n\n xxmaj headline : xxmaj bitcoin & xxmaj blockchain xxmaj searches xxmaj exceed xxmaj trump ! xxmaj blockchain xxmaj stocks xxmaj are xxmaj next ! \n\n 2016 has been a remarkable year in the markets with xxmaj brexit and xxmaj trump creating so much market volatility . xxmaj in this climate , it 's been proven to be a year for the stock xxunk . xxmaj the wonderful xxmaj paul xxmaj</p>	fake	fake

In [7]: `learn.validate()[1]`

Out[7]: 0.9370078444480896