Fake News Classification

Carles Zato

Harbour.Space University

March 2022

Dataset

- **True news dataset**: data scrapping from Reuters.com website
- **Fake news dataset**: news from unreliable flagged by a 'fact-chacking' organization and Wikipedia.

Personal opinion: questionable methodology to create the datasets. All news in true dataset come from a single source.

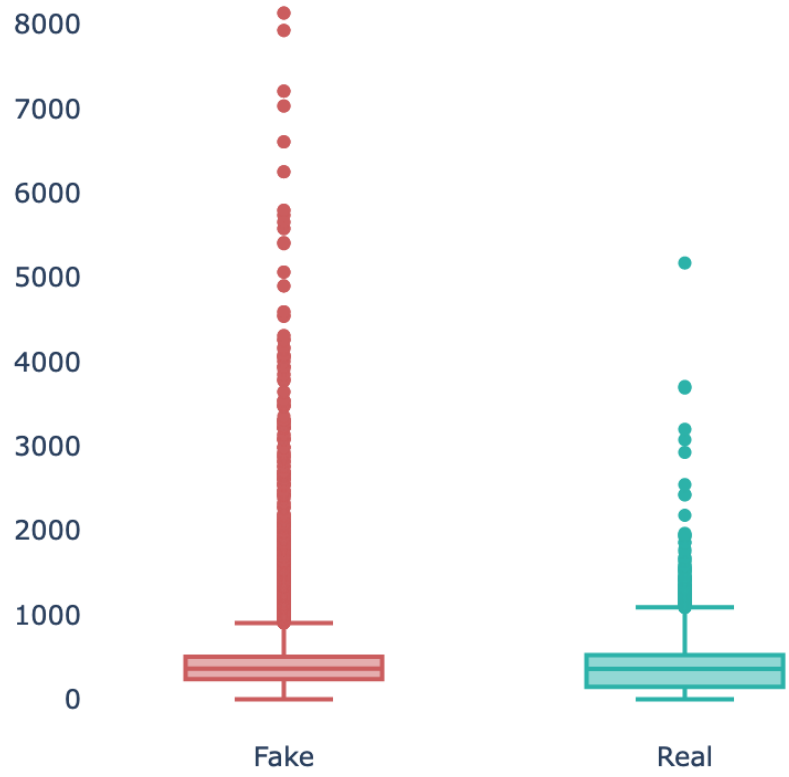

Figure: Token Distribution

Data Leakage

Right outside the box, if we run a SGD Classifier we get

- Accuracy on train data: 98.28%
- Accuracy on test data: 99.09%

```
1 pprint(true['text'][19])
```

```
('(Reuters) - A gift-wrapped package addressed to U.S. Treasury Secretary '
 'Steven Mnuchin's home in a posh Los Angeles neighborhood that was suspected '
 'of being a bomb was instead filled with horse manure, police told local '
 'media. The package was found Saturday evening in a next-door neighbor's '
 'driveway in Bel Air, the Los Angeles Police Department told the Los Angeles '
 'Times and KNBC television, the NBC affiliate in Los Angeles. The package '
 'also included a Christmas card with negative comments about President Donald '
 'Trump and the new U.S. tax law signed by Trump last week. Reuters could not '
 'reach LAPD officials for comment on Sunday. An LAPD bomb squad X-rayed the '
 'package before opening it and found the horse manure inside, police told '
 'local media. Aerial footage from KNBC showed officers investigating a large '
 'box in wrapping paper, then dumping a large amount of what they later '
 'identified as the manure and opening the card that was included inside.  '
 'Mnuchin, who KNBC said was not home when the package was discovered, is a '
 'former Goldman Sachs Group Inc executive and Hollywood film financier. A '
 'road in Bel Air was closed for about two hours, KNBC reported. The U.S. '
 'Secret Service is also investigating the incident, according to the TV '
 'station. ')
```

Figure: True news example

```
1 pprint(fake['text'][5])
```

```
('The number of cases of cops brutalizing and killing people of color seems to '
 'see no end. Now, we have another case that needs to be shared far and wide. '
 'An Alabama woman by the name of Angela Williams shared a graphic photo of '
 'her son, lying in a hospital bed with a beaten and fractured face, on '
 'Facebook. It needs to be shared far and wide, because this is '
 'unacceptable.It is unclear why Williams  son was in police custody or what '
 'sort of altercation resulted in his arrest, but when you see the photo you '
 'will realize that these details matter not. Cops are not supposed to beat '
 'and brutalize those in their custody. In the post you are about to see, Ms. '
 'Williams expresses her hope that the cops had their body cameras on while '
 'they were beating her son, but I think we all know that there will be some '
 'kind of convenient  malfunction  to explain away the lack of existence of '
 'dash or body camera footage of what was clearly a brutal beating. Hell, it '
 'could even be described as attempted murder. Something tells me that this '
 'young man will never be the same. Without further ado, here is what Troy, '
 'Alabama s finest decided was appropriate treatment of Angela Williams  '
 'son:No matter what the perceived crime of this young man might be, this is '
 'completely unacceptable. The cops who did this need to rot in jail for a '
 'long, long time  but what you wanna bet they get a paid vacation while the '
 'force  investigates  itself, only to have the officers returned to duty '
 'posthaste?This, folks, is why we say BLACK LIVES MATTER. No way in hell '
 'would this have happened if Angela Williams  son had been white. Please '
 'share far and wide, and stay tuned to Addicting Info for further '
 'updates.Featured image via David McNew/Stringer/Getty Images')
```

Figure: Fake news example

Data Leakage (cont'd)

Did you notice anything strange? True news have the source at the beginning of the text and fake news have some sort of tag like 'Featured image'. Fun fact: a guy in Kaggle discussing the issue did a classifier with one for-loop and one if-statement[1]. Notice 99.21% accuracy.

In [43]:
```
result = []
for text in concat2['text_pre']:
    if 'reuters' in text:
        result.append(0)
    else:
        result.append(1)
accuracy_score(concat2['is_fake'], result)
```

Out[43]:
```
0.9920931890061918
```

Figure: One-liner classifier

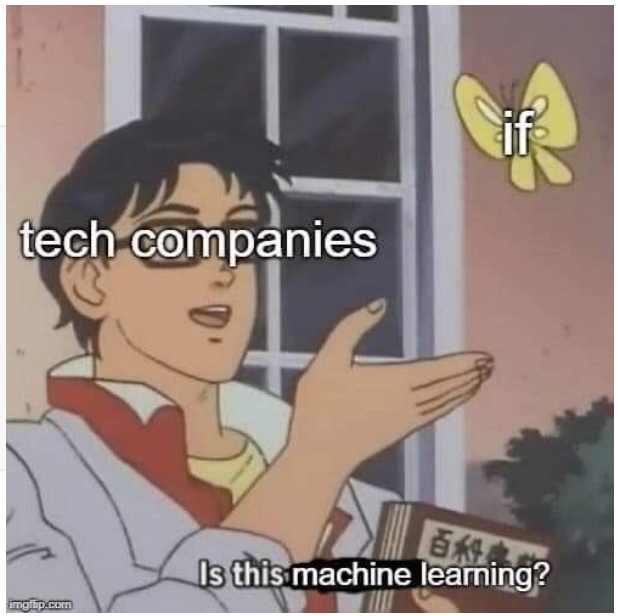

Figure: ML meme

[1] Code unashamedly taken from the aforementioned Kaggle guy

SGD Classifier

After removing sources, tags and twitter users, we get the following results

- Accuracy on train data: 95.62%
- Accuracy on test data: 96.81%

Why? The dataset quality is just to poor.

```
from hashlib import sha256
from tqdm import tqdm
list_ = [ ]
for text in tqdm(concat['text']):
    hash_ = sha256(text.encode('utf-8')).hexdigest()
    list_.append(hash_)
concat['hash'] = list_
t = concat.groupby(['hash']).size().reset_index(name='count')
duplicate = t[t['count']>1]
print('there are ',duplicate.shape[0], 'duplicate texts')
```

```
100%|██████████| 44898/44898 [00:00<00:00, 87222.12it/s]

there are  5140 duplicate texts
```

Figure: No comments[2]

[2] Unashamedly taken from the aforementioned Kaggle guy

Topic Modelling

Dataset contains a category called 'Subject' with 3 categories: news, politics and other. Instead of using this (which is a very vague categorization), I tried to find optimal number of topics (k) by looking at the coherence (based on PMI) and guessing categories myself my looking at top words.

Higher coherence is acheived with $k = 5$, despite clusters are not very differentiated. My guess, since we don't have ground truth:

- **Topic 1:** generic
- **Topic 2:** international relations
- **Topic 3:** economics/finance
- **Topic 4:** information
- **Topic 5:** elections/politics

Conclusions

Moral of the story:

- **Data source:** can't really express how much I've realized the importance of diving into the origin of your data. I learned why people say that 80% of the time invested in data science is preprocessing and cleaning.
- **Data quality:** it's something that you can overlook if you are just practicing and having fun in Kaggle. Definitely, cannot be overlooked in professional projects.
- **Transfer Learning:** naive data scientists might think that they have a cutting-edge ML model which gets 99% accuracy in fake news detection. That model wouldn't stand a chance if it had to classify any news outside this dataset as it's way too overfitted. Always be aware of the limitations of your model.