Fake News detection
Machine Learning for Natural Language Processing 2021

Bastien Billiot
ENSAE Paris
bastien.billiot@ensae.fr

Rémy Deshayes
ENSAE Paris
remy.deshayes@ensae.fr

Abstract

In this project we focus on fake news and their significant impact on various aspects of our society, let it be damaging someone's reputation, create controversy and even manipulate political outcomes. With the rise of false or misleading information presented as genuine news, all news publishers, news aggregators and social media platform tried to curtail the phenomena by introducing automatic text classification techniques. In this project, we first explore the ISOT dataset (H. Ahmed, 2018) and then train a LSTM and a CNN with PyTorch to address the above-mentionned binary fake news detection task. Lastly, leveraging the very high precision and recall yielded, we discuss the results of our model in the light of crucial topics such as censorship and information overflow.

1 Problem Framing

False or misleading information presented as genuine news have been used to serve various opinion spamming purpose over a wide range of crucial topics such as politics, safety or even public health and has been used by various actors stretching from government agencies (Wong, 2020) to private activists and individuals. With the recent surge of information sources and the general overflow of articles, disputes around news trustworthiness have been dramatically rising.
Indeed, the way we consume news has been radically modified over the last decade and a great share of headlines are now provided through news aggregators and social media platforms. This allowed malicious actors to leverage social-network theory concepts such as the *power law*, holding that messages can replicated quickly if targeted at a few well-chosen individuals (Andrews, 2019) but also mass spreading techniques such as bots hiding behind the information overflow.

In this context and in the light of a growing number of controversies such as the 2016 presidential election in the US and widespread false belief during the 2020-Covid pandemic, computer automated fake news detection became very popular among a variety of actors in the media industry and have yielded rather good results so far (N. Mihalcea, 2018).

Until now, two classical approaches have been widely adopted. Before being shared on a social media platform, the article is evaluated by an algorithm that can either output a quality score i.e how certain is the model about the truthfulness of the article or act as a filter and discard the article that is believed to be false.

For our project, we step in the shoes of a social media platform and will adopt the later approach i.e implementing a binary classification filter to discard fake news.

2 Experiments Protocol

The ISOT dataset (H. Ahmed, 2018) we are leveraging contains two types of articles labeled as fake and verified news. Verified articles have been collected over a two-year period from 2015 to 2017, on the website of the press agency Reuters whereas the fake news were collected on dubious websites flagged by the fact-checking agency Politifact. Each article comes with its title, date of publication and a label subject such as *Politics*. Most subjects revolves around hot topics which makes the dataset interesting for our project problematic. Samples of the articles can be found on the notebook[1] supporting this project. Fake and verified news articles are well balanced with around 20,000 articles in each category.

[1] https://github.com/remydeshayes/NLP_Pytorch.git

A brief exploratory analysis allowed us to identify a few sources of data leakage among which :

- The subject variable have instances that are specific to verified news and fake news only (e.g category *Politics subjects* only encompasses fake news)
- The presence of the regular expression "City name (Reuters) - " at the beginning of verified news articles only
- The presence of duplicate articles published at different dates which could lead to the same article being in the training and testing sets

After clearing those points, further pre-processing procedures are undertaken such as removal of URLs and Twitter verbose in the articles' body. This can be seen in more details with corresponding comments in the notebook supporting this project.

We then define a task-specific model implemented in PyTorch. After loading a GloVe embedding model pre-trained[2] on Wikipedia with an embedding dimension of 300, we proceed to retrieving our vocabulary word's corresponding vector, we define our model which is composed of an Embedding Layer of dimension (vocab_size, 300), two stacked LSTM layers with 256 hidden units and a Dense Layer. We use an Adam optimiser which parameters can be seen in the notebook an a Binary Cross Entropy as loss (with Logits which has a sigmoid layer included before BCE computation).

Further details on the CNN can be found in the notebook.

3 Results based on the LSTM model

With an overall accuracy score of 99.58% and no blatant overfitting, our model performances are great. Only 16 test samples were misclassified.

Given that data are balanced we use accuracy to evaluate our model as it gives a general idea of the performance. However, as introduced in section 1, depending on the user purpose, accuracy might not be the optimal metric to use.
In some cases precision could more interesting for some users. Namely, coming back to the 2016 US presidential election, one can imagine that institutions and governments would like to avoid at all costs fake news to be published and influence the political outcomes. In that case one would want the number of True Positive to be very high compared to samples classified as positive i.e. one would want high precision. High precision indicates that overall most published news are true.
However aiming for a 100% precision is subject to risks. Indeed, such a model could result in censorship of genuine news and yield a controversy.

Detailed results encompassing confusion matrix, classification report and discussion around the F_β scores can be found in the notebook.

4 Discussion/Conclusion

Our model yields excellent results whatever problematic - control over what the reader sees (high precision) or prevent censorship (high recall) - is chosen. Although we did our best to avoid data leakage, the excellent results of our rather simple model raise concerns on the soundness of both the dataset and the resulting model.
To ensure that our model is robust it could be interesting to test our model against a more homogeneous dataset by diversifying the sources of fake and verified news or test it against computer generated fake news that are especially designed to fool our model (adversarial model).

Finally, having a look at a few misclassified examples - here is an extract of a false negative :

"Topless Femen activist tries to snatch Jesus statue from Vatican crib"

thrill-seeker words could influence the model to lean towards classifying a headline as fake whilst being true.
A further development would then be to expand our results and model evaluation with explainability methods such as Lime or SHAP values for local (example) explainability. It would thus give us quantitative results to interpret an example's classification.

[2] As it wasn't specified whether to use an embedding trained exclusively on our data or a pretrained model, we assumed we had the liberty to choose the latter. *Remark : our model has an option to fine-tune the GloVe pretrained weights on our data*

References

S. Saad H. Ahmed, I. Traore. 2018. Detecting opinion spams and fake news using text classification. *Journal of Security and Privacy, Volume 1, Issue 1*.

B. Kleinberg A. Lefevre N. Mihalcea, V. Perez-Rosas. 2018. Fake news detector algorithm works better than a human. *27th International Conference on Computational Linguistics*.

E. Andrews. 2019. How fake news spreads like a real virus. *Stanford Engineering Magazine*.

J. Carrie Wong. 2020. Russian agency created fake leftwing news outlet with fictional editors. *The Guardian UK*.