Predicting authenticity of news article using NLP and Machine Learning

Karthik Hubli

May 8th, 2018

Abstract

This paper describes the work done during as part of directed research under Prof. Haim Levkowitz in the Spring 2018 semester. The purpose of this study is to estimate the effectiveness of Natural Language Processing and Machine learning in predicting the authenticity of online news/blog articles. With free access to content creation, there is an exponential increase in fake and propaganda news articles. With large funding and affordable infrastructure to host one's content online, it is virtually impossible to vet all the content before it reaches out to the public. This project is an attempt to find patterns in fake news and articles which help in deciding its authenticity. Machines are much better than humans at identifying hidden patterns in text data. Hence, I have employed various machine learning techniques to help us in the project.

Dataset

There are two data sets used in the below experiment. Both are sourced from 'Kaggle.com'

- **Fake News Dataset**

  This contains the title, body and the label (fake/not fake) of the news/blog. The body is just the summary of the entire article. The dataset contains approximately 4000 samples with close 70% being fake articles. Below is a sample of the dataset.

URLs	Headline	Body	Label
http://www.bbc.com/news/world-us-canada-41419190	Four ways Bob Corker skewered Donald Trump	Image copyright Getty ImagesOn Sunday morning, Donald Trump went off on a Twitter tirade against a member of his own party.This, in itself, isn't exactly huge news. It's far from the first time the president has turned his rhetorical cannons on his own ranks.This time, however, his attacks were particularly biting and personal.	1

Fake Domain Dataset

This is a list of domains which are well established to be not a genuine source of information. There are 328 domains categorized as fake, partly fake, imposter etc.

Domain	Type	Registration
LadyLibertysNews.com	Fake news	Kirkland, Wash.***
LastDeplorables.com	Fake news	Scottsdale, Ariz. **
LearnProgress.org	Some fake stories	Scottsdale, Ariz. **
Liberty-Courier.com	Fake news	Orem, Utah~
LiberalPlug.com	Fake news	Scottsdale, Ariz. **

Experiment and Result

Analysis is divided into four stages, *classification* using multinomial naive Bayes, *sentiment analysis* of the title and body, *network analysis* and *prediction* using a Convolutional Neural Network (CNN). First three stages run in parallel and independent of each other. Classification and sentiment analysis are conducted on the title and the body independently. The network analysis is conducted on the body alone as the title is too concise to yield desired outpour. Finally, the results of the first three stages are fed to CNN for approximation.

Classification

I used Multinomial naive Bayes (MnB), a probabilistic learning method to classify the title and the body of the news article to into two groups, fake and not fake. The text data is vectorized before it is classified. The data is tokenized using a technique call term frequency–inverse document frequency (Tf-idf). Below is a sample text tokenization.

['He watches basketball and baseball', 'Julie likes to play basketball', 'Jane loves to play baseball']	{u'me': 8, u'basketball': 1, u'julie': 4, u'baseball': 0, u'likes': 5, u'loves': 7, u'jane': 3, u'linda': 6, u'more': 9, u'than': 10, u'he': 2} [[0.57735027 0.57735027 0.57735027 0. 0. 0. 0. 0. 0. 0. 0.] [0. 0.68091856 0. 0. 0.51785612 0.51785612 0. 0. 0. 0. 0.] [0.62276601 0. 0. 0.62276601 0. 0. 0. 0.4736296 0. 0. 0.]]

The tokenizer is implemented on the entire data set and is used to train the MnB model with 80-20 split in the data set. The model was able to achieve more than 94% efficiency in classifying the text into fake and not fake group. One interesting observation from the experiment was the discrepancy in the nature of the title and the body of the articles. That is, when the MnB models for title and body, the true false (not fake as not fake) predictions were consistent, but false positives (not fake as fake) and false negatives (fake as not fake) was not consistent and the efficiency dropped to 62%.

- **Sentiment Analysis**

In this section, the analysis was conducted in 2 stages. Once on the title of the article and once on the description. Sentiment analysis consists of two different sub-categories. Polarity and subjectivity. Polarity give us the emotional orientation of the text i.e. whether the given text is positive or negative in emotion. While subjectivity defines whether the given text is subjective or objective, i.e. if the article is ones' opinion on some other action or incident. Subjectivity was a lot more difficult and inconsistent to determine compared to polarity. Especially in case of the title as there isn't enough text to decide the subjectivity accurately. Though there was no established relation between the polarity/subjectivity and the authenticity of the article, there usually was a discrepancy in the polarity of title and body in case of most of the fake instances.

Text	Polarity	Subjectivity
Days before President Trump's decision on the Iran nuclear deal, Israel's leader called the pact "fatally flawed," while Iran's president warned of "historic regret" if it is ripped up.	-0.25	0.25
Google and Microsoft outline their roadmaps and advances for developers. Don't be surprised if two tech giants AI-infused visions rhyme.	0.1	0.9

- **Network Analysis**

In this section, I am trying to estimate whether the source of the information used by the article is genuine or a known source of unreliable information. In many articles, where there is a deliberate attempt to spread false news, the reference material originates to known sources of fake news. The system scans for all references and links in the body of the article. The list of these articles is compared with list of domains known to be source of false and biased information. The title is not analyzed in this case and titles usually do not contain any external references. The analysis is not done on the summary of the body, but on the entire body.

Though this is very useful and straight forward process to execute. But the outcome is not always indicative of the authenticity of the article. It is not clear as what purpose was the reference serving in the article. We cannot be sure if the article is using the reference to establish some fact, which in turn may not be authentic or if the article mentions the lack of credibility in the source and there-by leading to the article itself being more credible. Also, there is the issue of domains which are known to produce some-fake articles. We cannot determine with surety that the reference made here is a genuine source or a fake one.

- **Prediction**

For prediction, I have made use of support vector machine (SVM) model. The independent variables here are the results from the above three stages and it categorizes the inputs into fake or not fake. The model places a separation hyperplane in a 4-dimensional space. The model achieves an efficiency of nearly 92.8%

Conclusion and Future Work

The results of the experiment are promising yet is just the tip of the iceberg. Thought the accuracy in predicting the authenticity of the content is very high (>93%), in practice, it doesn't show the entire picture. The experiment can be treated as the base for deeper research into this field. The efficiency of prediction depends of several factors like the category of the news/blog, the nature of the article (report, satire etc), number of characters in the article. Etc. The model archives highest efficiency with political news published or relevant in last 2-3-year duration. This is partly due to the data set and part the nature or structure of such articles. There is not a lot of difference in the structure, tone, sentiment and other parameters of the fake news in other categories such as science, technology and sports. Additionally, the frequency of fake news in these categories is a lot lower than pollical news. Hence the data set is more biased to predict fake news in political category.

In the future, I would like to revisit this problem with a few improvements. I would like to find or create a data-set which has more sample for categories other than politics. This would help the model to be more generalized and improve the overall efficiency. I would also like to have different models for each categories of the news. This would again improve the efficiency as each category will have its own benchmark for stop-words, polarity, subjectivity and will have separate sets of sources of fake domains.

Another interesting experiment would be recognizing and categorizing informative news article from parody and satire articles. With the current model, these articles are most likely to be categorized as fake as these have a similar polarity and stop words.

Reference

[1] Pedregosa, Fabian and Varoquaux, Gael and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent and others. *Scikit-learn: Machine learning in Python* (2011)

[2] Bird, Steven and Loper, Edward. *NLTK: the natural language toolkit* (2004)

[3] Kibriya, Ashraf M and Frank, Eibe and Pfahringer, Bernhard and Holmes, Geoffrey. *Multinomial naive bayes for text categorization revisited* (2004)

[4] Tang, Bo and He, Haibo and Baggenstoss, Paul M and Kay, Steven. *A Bayesian classification approach using class-specific features for text categorization* (2016)

[5] Khan, Aamera ZH and Atique, Mohammad and Thakare, VM. *Combining lexicon-based and learning-based methods for Twitter sentiment analysis* (2015)

[6] Pang, Bo and Lee, Lillian and others. *Opinion mining and sentiment analysis* (2008)

[7] Yun-tao, Zhang and Ling, Gong and Yong-cheng, Wang. *An improved TF-IDF approach for text classification* (2005)