An average citizen's guide to fight Fake News

Thevie Mortiniera thevie.mortiniera@epfl.ch
Timothe Bornet dit Vorgeat timothee.bornetditvorgeat@epfl.ch
Patrik Wagner patrik.wagner@epfl.ch

Abstract

With the exponential spread of data accessible to everyone, the proliferation of misleading information has made it a challenge to rightfully identify reliable news sources as seen in (Rada, 2017). This has led to an increase in the demand for automated tools to detect false news. In this paper, we proposed a guide for the average citizens to make them more critical about information, and help them make the distinction between the *safe* and *unreliable* topics and sources of information. We verified our assumptions by running machines learning techniques. We compared several of them along with different features extraction techniques. The best performance on the test set was achieved with Logistic Regression using Term Frequency-Inverted Document Frequency (TF-IDF) to extract our features. We obtained a **F1-score** of **71%**

1 Introduction

Since the beginning of the 21st century, the Web has become the primary source of information and local news. However, with the exponential spread of data accessible to everyone, the challenge of getting rightly informed is getting harder and harder. In particular, fake news and misleading articles phenomena are playing a significant role in political and social matters. Fake news are made to create doubt and discredit some people or organizations. We propose in this paper a guide for the average citizen to make him more critical about information. In a first part, in order to tell when we should be more suspicious, we will run an exploratory analysis to find recurrent patterns of fake news regarding their origin (geographic, context) and the most relevant topics in which they appear in. In a second part, we will propose and compare different machine learning classification models to assess and improve the guidelines given in the previous part.

2 Data Collection

The dataset was created in (William Yang Wang, 2017). Its scope only spans the United States of America, thus limiting our guide to the US citizens. The set is made up of 12800 manually labeled short statements in various contexts from POLITIFACT.COM. The labels are categorized according to their truthfulness ratings: pants-fire, false, barely-true, half-true, mostly-true, and true. Their distribution correspond to the one from the whole database of the site. In order to separate the quality of truth in the different contexts, we choose to keep only the data in the pants-fire, False and True categories. The pants-fire and false are joined together as False news in our work. The data is split in train, validation and test. But, given the small sample available, we decide to merge validation and test set together. Hence we are left with 4508 samples in training data, and 1907 for testing.

3 Exploratory Analysis

3.1 Can we discriminate some news regarding their geographical origin ?

One way to guide the people's distrusts of statements is to localize their origin and determine if there are regions where false information is more common than in others. To do this a states' score depending on the proportion of statements which have been labeled as fake.

As we can see the geographical origin of news is not that relevant in determining its validity. This can be explained by the virtualization of the human interactions by the word wide web as well as the explosion of the travelling power of the indi-

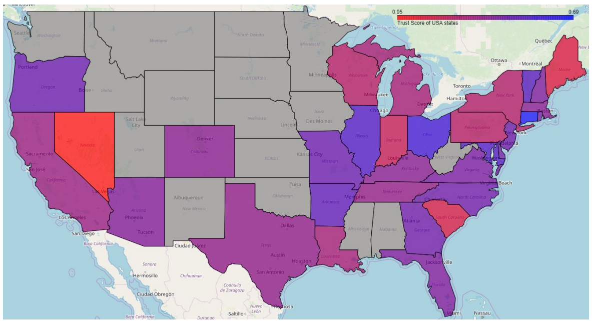


Figure 1: Caption

viduals. Indeed the creation of fake news can be done anywhere in the digital age.

3.2 Can we classify some contexts as less reliable than others ?

Statements coming from the void don't exist, indeed they always come from a person or group of people, as well as a way of communicating the information. Therefore the question of the influence of the context must be answered as it is usually easily identifiable.

We visualize the different contexts by their proportion of false occurrences to infer some groups of varying confidence levels.

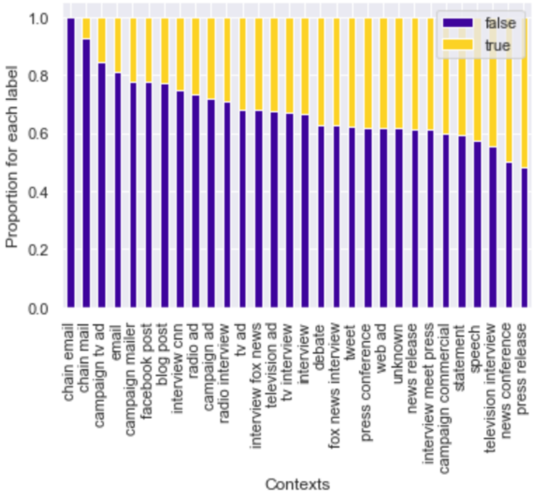


Figure 2: Distribution of fake and real news in 30 Most common Contexts

A subtle separation in three parts can be done. Indeed a baseline of truth, the situation where a normal level of distrust is needed, emerges as the central line where the proportion of true statements vary between forty and thirty percent. Both extreme cases of proportion (high and low) are also apparent. The safest contexts, where fifty to forty percent of true news are given, is a group needing the less amount of distrust of all. It contains information coming from official channels like press or news releases, television interviews (mostly from pundits). This is the complete opposite of the last case where the context is mostly online in e-mails like ads, chain mails, posts on Facebook or blogs (especially during critical times like elections, they can come from trolls which can be politically oriented or paid by foreign countries (Allcott, 2017)). This falseness haven is made up of contexts where less than thirty percent of statements are true.

One remark that must be made is the level of false information in the context of political campaign ads. Its value is too high for the well-being of a democracy. Elections shouldn't be won on lies and falsehood.

3.3 Are there suspicious topics where extra caution is advised ?

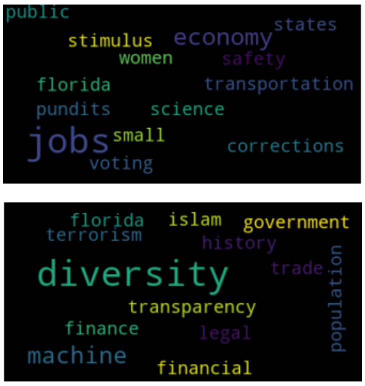


Figure 3: word cloud of 10 most common topics In Fake then Real News

We plotted the 10 most meaningful Topics using TF-IDF. It is a statistical metric used to measure how important a term is to a document in a dataset. One of the main characteristics of IDF is that it weights down the term frequencies while scaling up the rare ones. Hence, it allows us not to target topics which are widely spread independently from being in fake or real news, but to extract meaningful topics which are relevant depending on the subset of study. When we look at the WordCloud produced above, we notice the following :

- Most important topics in Real news refer to topics delivered from "serious" press release. Science or economy newspapers, statements and figures on Jobs, transportation and Safety concerns, in which the experts (”pundits” appearing in the Wordcloud) establishing the facts mentioned are quoted.

- On the other hand, it seems that when observing the WordCloud from Fake news, most of the information relates to conspiracy theory, with buzzwords such as ”terrorism”, ”transparency”, ”government” . Furthermore, ”Diversity” appears to be the most relevant word. As we know, immigration and asylum are becoming increasingly controversial issues. It seems, that hype and fear are used to promote differences and intolerance. Hence, it’s not surprising at all to find ”Diversity” as the most important topic in Fake news.

4 Classification Process

4.1 Choice of metric

Here the **positive case** is to predict a news to be **fake**.

In Fake news detection, news that are true (actual negative) but predicted as fake (positive case) are called false positives. And conversely the false negatives are the fake news predicted to be true. The precision is the proportion of true positive cases of all predicted positives so the the percentage of real fake news in the predicted fake set. The recall is the proportion of true positive cases of all real positives so the percentage of real fake news that are predicted in the real fake set.

Said more clearly news that are labelled as fake when they are true decrease the precision metric while news that are labelled as true while being false decrease the recall metric. Tee recall is the more sensible metric in our case since we really don’t want to mislabel fake news as being true. When obtaining machine learning models optimizing only one of the previously cited metrics the result is the prediction of all values to either fake or true which is equivalent to a dummy classifier, obviously not a good result. To avoid such a poor result the solution is to optimize the harmonic mean of the two metrics called the f1-score.

We still compare the different models using the recall but also the whole confusion matrix to avoid dummy classifiers.

4.2 Model Selection

A common quote is that with an increase in model complexity, model interpretability goes down at least as fast.

We will use here **Feature importance** to interpret our models.

We compare 4 different classes of machine learning models :

- In generalized linear models (logistic regression for example), we know that the weights that are obtained after training are a direct proxy of feature importance and they provide very concrete interpretation of the model internals. If the most important words/topics do not correspond to our intuition (e.g. names or stopwords, base on the analysis we performed in Part 2), it probably means that the model is fitting to noise in the dataset and it wont perform well on new data.

- Tree based models (Random Forest for example) also allow to obtain information on the feature importance. It even allows us to plot the tree as a path to follow, which would be great for us, as we want to construct a guide for the average citizen.

- Although kernel methods (SVC with linear kernel for example) are able to capture nonlinear relations between variables by projecting the features into kernel space, just looking at the weights as feature importance does not tell us much about their interaction.

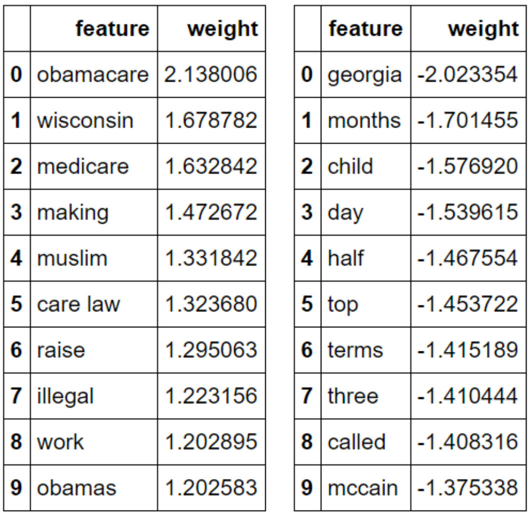
- Deep learning models are known to be uninterpretable due to the number of parameters to tune and the complex approach to extracting and combining features, even if they tend to achieve better performance than others on a lot of tasks.

Based on these observations, we choose to use Logistic regression and Random Forest.

4.3 Results

4.4 Logit model

We reached **67% of recall** and **71% of F1-score**. As for interpretation the following figure describe the most important features that helps the model to predict if a statement is false (positive predictions so the weights are positive) of true (negative prediction so the weights are negative). These features follow the intuition that controversial subjects are determinant in the prediction of fake news. These are everything surrounding president



	feature	weight
0	obamacare	2.138006
1	wisconsin	1.678782
2	medicare	1.632842
3	making	1.472672
4	muslim	1.331842
5	care law	1.323680
6	raise	1.295063
7	illegal	1.223156
8	work	1.202895
9	obamas	1.202583

	feature	weight
0	georgia	-2.023354
1	months	-1.701455
2	child	-1.576920
3	day	-1.539615
4	half	-1.467554
5	top	-1.453722
6	terms	-1.415189
7	three	-1.410444
8	called	-1.408316
9	mccain	-1.375338

Figure 4: Most important features of Logit model

Obama (remember the allegations concerning his birth certificate and the fake news about his belonging to the muslim religion) or his work like his health care reform. Another intuition is confirmed, indeed the words which are determinant in the prediction of the true news are mostly objective terms used to describe facts surely they are made by really objective statements that cannot be attacked of they truthfulness.

4.5 Random Forest

We reached a **recall of 69%** and a **F1-score of 71%** with our best random forest model. However as the model is not stable, since it is an ensemble model based on the decisions trees algorithm, we cannot obtain a sustainable tree as an interpretable guide through the features.

5 Discussion

One drawback of the work comes from the dataset used to create our guide. Indeed it is a subsample of the whole POLITIFACT.COM database, so many biases could have been added. The first ones come from the site. The others come from the sampling.

Let's be presumptuous and say that our guide is a success and many people use its content to better their judgment. Then, surely, malicious people like russian trolls will use it so as to improve their fake news generation. An endless fight between intelligent classifiers and fake news generators will begin similarly to the fight between spammers and anti-spam filters.

A feature engineering of the text by analyzing the grammatical value of each words and the extraction of their frequencies can be done to improve the models. A better choice of metric can be used, either find a good beta and its equivalent FB-score or use the area under the receiver operating characteristic curve (AUC ROC).

The lack of data is the biggest drawback to our guide. Indeed after extracting only the true, false and pants-fire false the amount of data is low for good and interpretable machine learning results.

6 Conclusion

The problem of fake news grows each days, indeed at each elections or other politically tense periods there are news of foreign interventions into the public speech domain. They appear through trolls or the spreading of rumors with political damage intentions. This is a direct threat to the democracies of the world as they are made on the basis of a well-informed population.

In this paper we proposed a guide to help people to be more critical regarding the truthfulness of some presented news. We answered multiples questions during our investigations. We concluded that geographical origin does not affect the truthfulness of statements, while the contexts does. Indeed we found more or less three different kinds of contexts; with varying levels of truthfulness. Finally, it seems that controversy brings tends to bring misinformation. We also verified our assumptions with machine learning techniques, such as Logit Model which gave us a **F1-Score of 71%** while its feature importance aligned with our assumptions.

Our hope is that the guide presented in this report helps people in placing their distrust at the right occasions in order to be fooled the minimum amount of time.

References

William Yang Wang. 2017. *”Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection*.

Ahmed, Hadeer, Traore, Issa, Saad, Sherif. 2017. *Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques*.

Allcott, Hunt, and Matthew Gentzkow. 2017. *Social Media and Fake News in the 2016 Election*. Prentice-Hall, Englewood Cliffs, NJ.

Perez-Rosas, Veronica, Kleinberg, Bennett, Lefevre, Alexandra, Mihalcea and Rada. 2017. *Automatic Detection of Fake News*.