Master thesis

Simon Lorent

June 21, 2019

Contents

Master thesis

Fake news detection using machine learning

Simon Lorent

Acknowledgement

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Master thesis

Fake news detection using machine learning

Simon Lorent

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Chapter 1

Introduction

1.1 What are fake news?

1.1.1 Definition

Fake news have quickly become a society problem, being used to propagate false or rumorus informations in order to change behaviors of peoples. Before stating to work on detecting fake news, it is needed to first understand what they are. It have been show that propagation of fake news have had a non negligable influence of 2016 US presidential elections[1]. A few facts on fake news in the United States:

- 62% of US citizen get there news for social medias[2]
- Fake news had more share on facebook than mainstream news[3].

Fake news have also been used in order to influence the referendum in the United Kingdom for the "Brexit".

There are two aspects of fake news detection that need to be taken into account according to Shu et al[4]. The first is characterization or what are fake news and the second is detection. In order to build detection models, it is need to start by charaterization, indeed, it is need to understand what are fake news before trying to detect them.

Fake news definition is made of two part: authenticity and intent. Authenticity means that fake news content fale information that can be verified as such, which means that concpiracy theory is not included in fake news as there are difficult to be proven true or false in most cases. The second part, intent, means that the false information have been written with the goal of missleading the reader.

1.1.2 Fake News Characterization

Definition 1 *Fake news is a news article that is intentionally and verifiably false*

The part of the definition introducing the intent of missleading the reader automaticaly discard satire news medias, that is why this works will focus on the first part, the fact that the piece of information is verifiably false or true. Indded, even if satire news medias does not have the intent to misslead the readers, not all of them have the ability of making

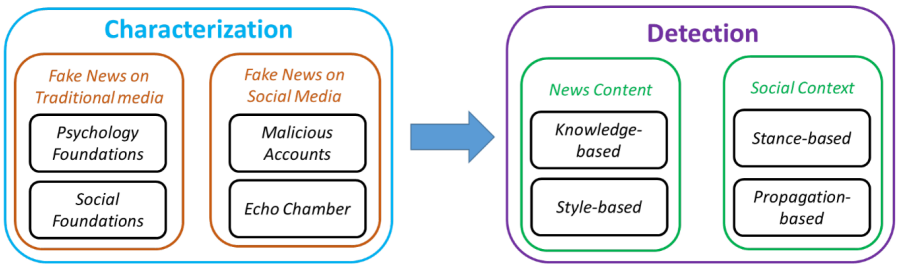

Figure 1.1: Fake news on social media: from characterization to detection.[4]

criticisme and not taking it to the first degree. On the other hand, in the case of political media, even if it clearly try to influence the consumer, verifying the authenticity of there claims is usualy harder as, in most of the cases, openly lies.

1.1.3 Feature Extraction

Now that fake news have been defined and the target have been set, it is needed to analize what features can be used in order to classify fake news. Starting by looking at news content, it can be seen that it is made of four principal raw components:

- **Source**: Where does the news come from, who wrote it, is this source reliable or not.
- **Headline**: Short summary of the news content that try to attract the reader.
- **Body Text**: The actual text content of the news.
- **Image/Video**: Usualy, textual information is agremented with visual information such as images, videos or audio.

Features will be extracted from these four basic components, with the mains features being linguistic-based and visual-based. As explained before, fake news are used to influence the consumer, and in order to do that, they often use a specific language in order to attract the readers. On the other hand, non fake news will mostly stick to a different language register, being more formal. This are linguistic-based features, to which can be added lexical features such as total number of words, frequency of large words or unique words.

The second features that need to be taken into account are visual features. Indeed, modified images are often used to add more weight to the textual information. For example, the **Figure 1.2** is supposed to show the progress of deforestation, but the two images are acutaly from the same original one, and in addition the WWF logo make it looks like to be from a trusted source.

1.2 Datasets

1.2.1 Fake News Corpus

This works uses multiples corpus in order to train and test different models. The main corpus used for training is called Fake News Corpus[5]. This corpus have been

Figure 1.2: The two images provided to show deforestation between two date are from the same image taken at the same time.

automatically crawled using `opensources.co` labels. In other words, domains have been labeled with one or more labels in

- Fake News
- Satire
- Extreme Bias
- Conspiracy Theory
- Junk Science
- Hate News
- Clickbait
- Proceed With Caution
- Political
- Credible

These annotations have been provided by crowdsourcing, which means that they might not be exactly accurate, but are expected to be close to the reality. Because this works focus on fake news detection against reliable news, only the news labels as fake and credible have been used.

TODO: Expliquer comment le dataset a t nettoy et mis dans une base de donnes afin d'augmenter les performances.

1.2.2 Fake News Net

The second dataset used is fake news net[6, 7, 4]. This corpus is made of news from two different sources, PolitiFact and GossipCop. An older version also provide news from BuzzFeed. News are categorized in two classes: fake and non fake. Being quite smaller than fake news corpus, this dataset will be used as a test dataset.

1.2.3 Liar, Liar Pants on Fire

The third and last dataset is **Liar, Liar Pants on Fire** dataset[8], which is a collection of twelve thousand small sentences collected from various sources and hand labeled. They are devided in six classes:

- pants-fire
- false
- barely-true
- half-true
- mostly-true
- true

This set will be used a second test set. Because in this case there are six classes againt two in the other cases, a threshould should be used in order to fix which one will be concidered as true or false.

It should be noted that this one differ from the two other dataset is it is composed only on short sentences, and thus it should not be expected to have very good results on this dataset for models trained on Fake News Corpus which is made of full texts.

1.3 State of the Art

Chapter 2

Data Exploration

2.1 Introduction

A good starting point for the analysis is to make some data exploration of the data set. The first thing to be done is statistical analysis such as counting the number of text per class or counting the number of words per sentence. The second step consit of doing Latent Dirichlet Allocation[9] in order to make unsupervised clustering of the text and see if there is some kind of correlation between the clusters to which a text belongs and its labels.

2.2 Dataset statistics

2.2.1 Fake News Corpus

Because **Fake News Corpus** is the main dataset, the data exploration will start with this dataset. And the first thing is to count the number of items per class.

Because the dataset have been cleaned, numbers provided by the dataset creators and number computed after cleaning will be provided. We found the values given at **Table 2.1**. It shows that the number of fake news is smaller by a small factors with respect to the number of reliable news, but given the total number of items it should not cause any problems. But it will still be taken into account later on.

Type	Provided	Computed
Fake News	928,083	
Satire	146,080	
Extreme Bias	1,300,444	
Conspiracy Theory	905,981	
Junk Science	144,939	
Hate News	117,374	
Clickbait	292,201	
Proceed With Caution	319,830	
Political	2,435,471	
Credible	1,920,139	

Table 2.1: Number of texts per categories

To have a better view of the distribution of categories, an histogram is provided at **Figure 2.1**.

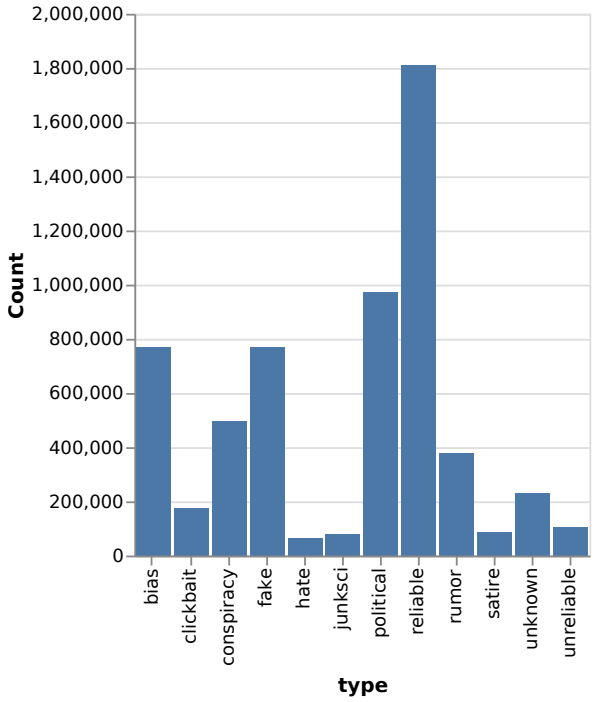

Figure 2.1: Histogram of text distribution along their categories on the computed numbers.

Bibliography

[1] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. In *Journal of Economic Perspective*, volume 31, 2017.

[2] Jeffrey Gottfried and Elisa Shearer. *News Use Across Social Medial Platforms 2016*. Pew Research Center, 2016.

[3] Craig Silverman and Lawrence Alexander. How teens in the balkans are duping trump supporters with fake news. *Buzzfeed News*, 3, 2016.

[4] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[5] Maciej Szpakowski. Fake news corpus. `https://github.com/several27/FakeNewsCorpus`. Accessed: 2018-10.

[6] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[7] Kai Shu, Suhang Wang, and Huan Liu. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 2017.

[8] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.