

Master thesis

Simon Lorent

June 10, 2019

Contents

1	Introduction	3
1.1	What are fake news?	3
1.2	Datasets	3
1.2.1	Fake News Corpus	3
1.2.2	Fake News Net	4
1.2.3	Liar, Liar Pants on Fire	4

Master thesis

Fake news detection using machine learning

Simon Lorent

Abstract

Lorem ipsum dolor sit amet, consectetur 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, consectetur 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?

Fake news have quickly become a society problem, being used to propagate false or rumorous 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 negligible 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".

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[4]. This corpus have been 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
- State News
- 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 work focuses on fake news detection against reliable news, only the news labels as fake and credible have been used.

TODO: Expliquer comment le dataset a été nettoyé et mis dans une base de données afin d'augmenter les performances.

1.2.2 Fake News Net

The second dataset used is fake news net[5, 6, 7]. This corpus is made of news from two different sources, PolitiFact and GossipCop. An older version also provides 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 divided in six classes:

- pants-fire
- false
- barely-true

- half-true
- mostly-true
- true

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

It should be noted that this one differs from the two other datasets as it is composed only of 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.

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] Maciej Szpakowski. Fake news corpus. <https://github.com/several27/FakeNewsCorpus>. Accessed: 2018-10.
- [5] 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.
- [6] Kai Shu, Suhang Wang, and Huan Liu. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 2017.
- [7] 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.
- [8] William Yang Wang. ”liar, liar pants on fire”: A new benchmark dataset for fake news detection.