Fake news: an algorithmic perspective on fact-checking

Martijn B.J. Schouten
11295562

Bachelor thesis
Credits: 12 EC

Bachelor's degree Information Science

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Dr. M. J. Marx

ILPS, IvI
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

2019-06

Abstract

Contents

1 Introduction

The ability to broadcast information on a large scale has been in the hands of large publishing organizations in the pre-Internet era, but nowadays everyone can share news via social media [5]. This introduces risks on validity and authenticity of news, as social media and digital platforms can speed up the spread of falsehoods without much effort from the author [1].

As a matter of fact, 63% of adults in the United States prefer to read their news on the Internet. Young adults take the lead: 76% of adults between the ages 18 and 49 get their primary news consumption via the web, compared to just 43% for adults of 50 years and older [8]. As time passes by, social media is slowly becoming the primary source of news for more and more people.

The main danger of this development is that human perception is often skewed with regards to objectivity of facts. Naïve realism let consumers of news belief that their perception is right, while other's perceptions are uninformed. Furthermore, confirmation bias results in consumers preferring information that confirms beliefs they already have [11]. This makes consumers vulnerable for the spread of misinformation or fake news.

According to the European Commission, *"disinformation - or fake news - consists of verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm"* [1]. The answer to the problem of fake news as of recently has been to manually fact-check statements on validity, but, as Shu et al. underlines, one of the downsides to this approach is that fake news typically relates to newly emerging, time-critical events. This means the real news may not be fully verified by proper knowledge bases due to a lack of contradicting claims [11]. An automated approach would both help in solving the problem of human subjectivity and the speed at which false information is spread in the current news spreading landscape.

Natural language processing has been in rapid development over the past years. With the releases of OpenAI's GPT-2 model in February of this year and Google's BERT in the autumn of 2018, state-of-the-art pre-trained textual embedding techniques have shown promising results on various classification tasks [10][4]. Although fake news classification has been attempted before [13][6], performance has been rather low. However, these new pre-trained textual embeddings have not yet been used in the fight against disinformation.

This thesis is focussed on the following research question: what is the performance of combinations of pre-trained embedding techniques with machine learning algorithms when classifying fake news? This main question will be answered through the results of the following subquestions:

RQ1

RQ2

RQ3

Overview of thesis

2 Related Work

2.1 RQ1

Fake news as a term only caught public attention starting from the end of 2016, during the Presidential Elections of the United States [12].

2.2 RQ2

In the last couple of years, using transfer learning for natural language processing has given promisable results. The following sentence embeddings will be used to detect fake news:

- Bag of Words as a baseline for performance of non-pretrained embeddings;
- Facebook's InferSent [2];
- ELMo from the Allen Institute for Artificial Intelligence [9];
- OpenAI's GPT-2 [10];
- Transformer-XL [3];
- Microsoft's MT-DNN [7];
- and Google's BERT [4].

2.3 RQ3

Aligned with the original research on this dataset by Wang [13], the following machine learning algorithms will be used to test the applicability of the abovementioned embedding techniques:

- SVMs;
- Logistic regression;
- Bi-LSTMs;
- CNNs.

3 Methodology

3.1 Description of the data

3.2 Data charts

3.3 Methods

3.3.1 RQ1

3.3.2 RQ2

4 Evaluation

Evaluation.

5 Conclusions

Conclusions.

5.1 Acknowledgements

Many thanks.

References

[1] European Commission. Fake news and online disinformation. `https://ec.europa.eu/digital-single-market/en/fake-news-disinformation`, 2018. Retrieved on 8th of April, 2019.

[2] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[3] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Lee Howell et al. Digital wildfires in a hyperconnected world. *WEF Report*, 3:15–94, 2013.

[6] Urja Khurana. The linguistic features of fake news headlines and statements, 2017.

[7] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.

[8] Amy Mitchell and Hannah Klein. Americans still prefer watching to reading the news – and mostly still through television. 2018.

[9] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018.

[10] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[11] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *CoRR*, abs/1708.01967, 2017.

[12] Google Trends. Explore. https://trends.google.nl/trends/explore?date=today%205-y&q=fake%20news, 2019. Retrieved on 16th of April, 2019.

[13] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *CoRR*, abs/1705.00648, 2017.