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 Add a final date

Abstract

Add an 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 [7]. 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 [5].

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 [13]. 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 [19]. 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"* [5]. 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 [19]. 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. Furthermore, such an approach can help human fact-checking by targetting statements that are most likely to be false.

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 [15][6]. Although fake news classification has been attempted before [21][9], 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 Which way of pooling vectors to a fixed length works best for classifying fake news?

RQ2 At what padding sequence length do neural networks hold the highest accuracy when classifying fake news?

RQ3 How well do neural network classification architectures classify fake news compared to non-neural classification algorithms?

Add an overview of thesis section

2 Related Work

2.1 Automatic fake news detection

There have been several attempts in the past to create classifiers for automatic detection of lies and fake news. Wang used both linear and neural classifiers to classify statements from the Liar dataset into 6 possible gradations of truthfulness. Furthermore, he added speaker metadata to improve the result of his classifications. Both with and without introducing speaker metadata, the best performing architecture was found to be a convolutional neural network. With an accuracy of 27% without, and 27,4% with metadata on the test set, Wang was able to perform 6,2% and 6,6% better than the majority baseline of 20,8%[21].

From the same dataset, Khurana extracted linguistic features such as n-grams, sentiment, number of capital letters and POS tags to classify the data into 3 labels instead of the original 6 labels. For classification, she used a set of non-neural classifiers. Her best performing classifier, using gradient boosting, obtained an accuracy of 49,03%, which performed around 5% better than the majority baseline of 44,28% [9].

The British factchecking organization Full Fact has developed an architecture that is able to monitor and factcheck statements from the British Parliament and major media outlets in the United Kingdom. It can automatically factcheck the accuracy of statistical claims, for example [3]. For detecting factual claims from texts, the organization uses InferSent, which is a way of transfer learning that has been proved to perform well for the use case of Full Fact [11].

Various tools with regards to fake news detection are also available. Faker Fact is a tool which can classify texts into a set of categories ranging from satire to agenda-driven, the former identifying humorous intent, the latter identifying manipulation [1]. Hoaxy, on the other hand, allows for the visualization of unverified claims through Twitter networks [17].

2.2 Pooling

Linear classifiers need data in a two-dimensional shape to be able to perform calculations. In the case of raw text data, sentences in the dataset have variable word lengths, resulting in a different vector length when turning the text into a vector representation. To turn the vector representations into a uniform length, we can either cut off the vectors at a fixed length (*padding*), or we can perform calculations to reduce the length of the vectors (*pooling*).

In computer vision, feature pooling is used to reduce noise in data. The goal of this step is to transform joint feature representations into a new, more usable one that preserves important information while discarding irrelevant details. Pooling techniques such as max pooling and average pooling perform mathematical operations to reduce several numbers into one [4]. In the case of transforming the shape of the data, we can reduce vectors to the smallest vector in the dataset to create a uniform shape.

Scherer et al. compared performance of two pooling operations on a convolutional neural network architecture. The first pooling method extracted maximums and the second one was primarily based on working with averages. They have shown that a max pooling operation is vastly superior for capturing invariances in image-like data [16].

Shen et al. noted that in text classification, only a small number of key words contribute to the final prediction. As a result, simple pooling operations are surprisingly effective for representing documents [18]. Lai et al., Hu et al. and Zhang et al. use a max pooling layer in a (recurrent) convolutional neural network for identifying key features in text classification [12][8][23]. In the case of text classification, max pooling strategies seem to be the most popular.

2.3 Padding

When padding a sequence, a list of sequences is transformed to a specific length. Sequences longer than the desired length will be truncated to fit the requirement, while sequences shorter than the desired length will be padded by a specified value [2]. To fill the sequences, a value of zero is often used. Hu et al. also use zero values for padding their sequences [8].

Apart from controlling the size of the feature dimension, padding has other uses as well. Simard et al. make use of sequence padding for convolutional neural networks to center feature units, and concluded it did not impact the performance of the classifier significantly [20]. Wen et al. apply padding to convolutional network models to prevent dimension loss [22].

2.4 Neural text classifiers

Wang has shown that neural networks perform slightly better on classifying fake news than linear classifiers. In his research, he compared accuracies on support vector machines, logistic regressions, bidirectional LSTMs and convolutional neural networks with each other. With his 6 label classification, his support vector machine implementation was the best performing linear classifier, but the performance was slightly worse than the best performing neural network (25,8% for the former, and 26% for the latter) [21].

Wang used two neural network architectures both well known for their robustness and performance when it comes to text classification. The first model, the bidirectional Long Short Term Memory (LSTM) networks, are specifically tailored at keeping track of information for a long period of time. This makes those models able to keep track of the context in a more intelligent way when compared to a standard non-neural classification algorithm [14].

The second architecture, the convolutional neural network, apply a set of filters in its layers to local features. These models are shown to be effective in numerous natural language processing applications, such as semantic parsing, search query retrieval, sentence modeling and other traditional NLP tasks [10].

3 Methodology

3.1 Description of the data

For classifying fake news, Wang's Liar dataset will be used [21]. The Liar dataset contains 12.791 short statements from Politifact.com, which are labeled manually by a Politifact.com editor on truthfulness. The statements are an average of 18 tokens long, and the topics vary from different political subjects, as can be seen in figure 1. Truthfulness is evaluated by assigning one of 6 labels,

Check whether this number is still correct ranging from *pants-on-fire* to *true*. The distribution of statements across labels can be seen in figure 3.

Check whether this number is still correct

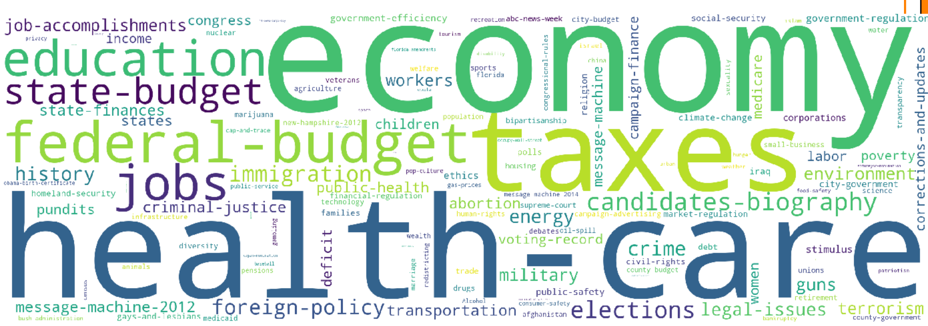

Figure 1: An overview of all statement topics in the Liar dataset.

For each statement, the dataset contains an id, a label, a subject, a speaker, the function title of the speaker, the affiliated state and political affiliation, the context of the statement and a vector with a truthfulness history. Wang introduced this truthfulness history to boost the prediction scores, as speakers with a track record of lying are expected to have a lower chance of speaking the truth when classifying new statements. However, for our application we are only interested in the statement itself and its corresponding label. Due to cheapness and spreadability, a large amount of fake news is spread over social media [19]. This means author information and metadata will not readily be available in real world circumstances.

The original dataset has been split beforehand into a test, train and validation set. The train set contains 80% of the total amount of statements, while the test and validation set both contain approximately 10% of the statements.

3.2 Data preprocessing and cleaning

3.2.1 Filtering statements

The original dataset contained statements ranging from 1 sentence to 19. On closer inspection of statements with the high amounts of sentences, it was found that not all statements were processed from source files into dataframes correctly. As a result, records of some different statements were joined together, forming a single string. To combat this, the following regular expression was used to filter those statements out:

```
\\.json\\t(mostly-true|true|half-true|false|barely-true|pants-fire)\
```

After applying this regular expression, the total amount of sentences in the statements were reduced from a maximum of 19 to a maximum of 11.

3.2.2 Reducing labels

Wang's main objective was to classify fake news into a fine-grained category of fakeness. His best model obtained a 27% accuracy on the validation set

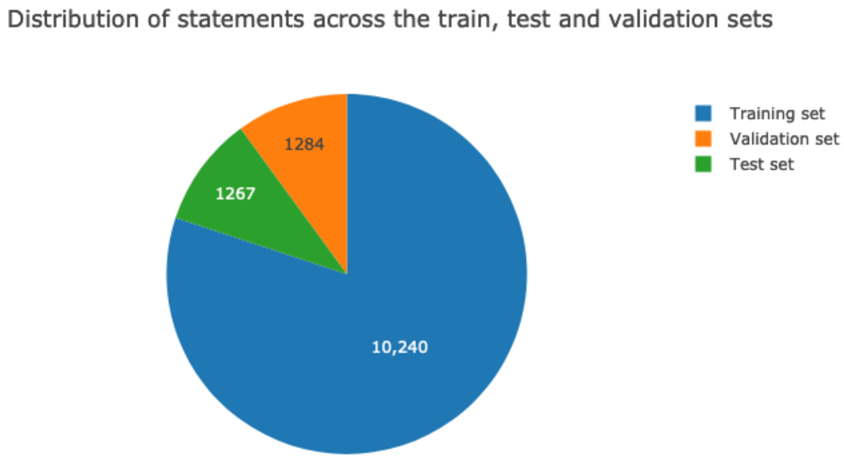

Figure 2: Distribution of the total dataset.

[21]. For our main research question, we only aim to predict whether the statements are fake news or not. Because of this, for the rest of this research, the prediction labels from the original dataset will be reduced from 6 labels to 3 labels, according to the recoding schema used by Khurana [9]. The 6 labels are reduced to the following 3 labels: `true`, `half-true` and `false`.

3.3 Methods

3.3.1 RQ1

3.3.2 RQ2

3.3.3 RQ3

4 Evaluation

4.1 RQ1

Which way of pooling vectors to a fixed length works best for classifying fake news?
The answer to this question will be given by comparing accuracy of a logistic regression with different pooling strategies (max, min, average).

4.2 RQ2

At what padding sequence length do neural networks hold the highest accuracy when classifying fake news?
The answer to this question will be given by comparing accuracy of a bidirectional LSTM architecture with variable padding lengths.

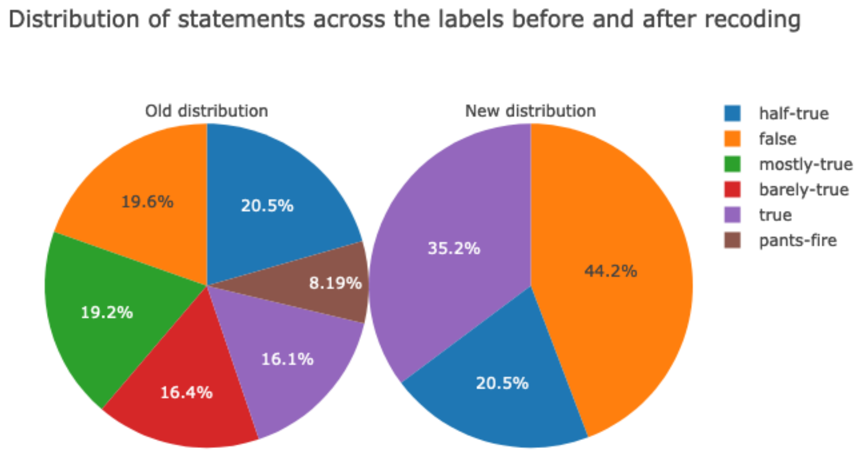

Figure 3: Distribution of the statements over the labels, before and after reducing the labels from 6 to 3.

By comparing all embedding techniques, I plan on getting a peak padding length with the best overall performance shared by all embedding methods. At this moment, the optimal length averages at around a length of 22.

4.3 RQ3

How well do neural network classification architectures classify fake news compared to non-neural classification algorithms?
The answer of this question will be given by comparing two linear classifiers with two neural classifiers.

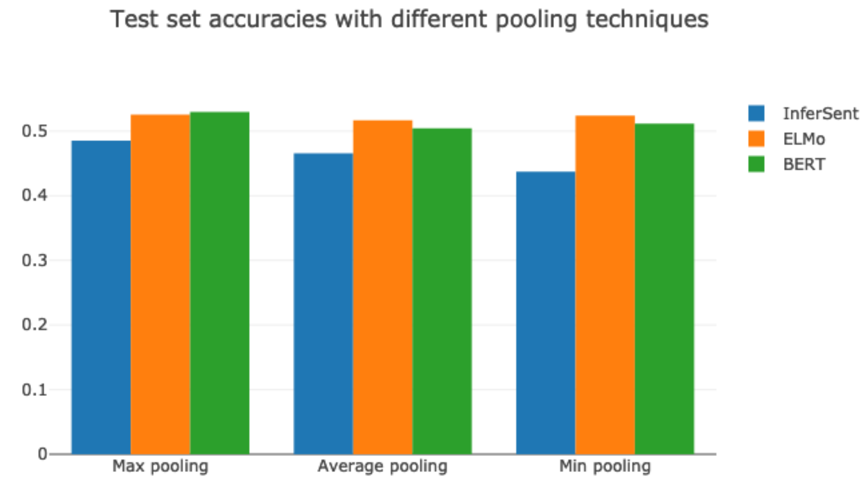

Figure 4: Comparing different pooling strategies.

5 Conclusions

5.1 Acknowledgements

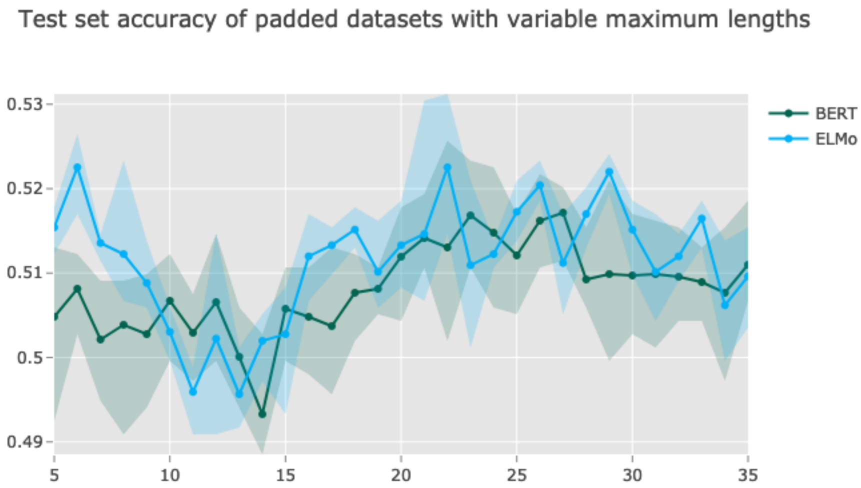

Figure 5: Comparing different maximum padding lengths.

References

[1] About fakerfact.

[2] Sequence preprocessing.

[3] Mevan Babakar and Will Moy. The state of automated factchecking. Technical report, Full Fact, 2016.

[4] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010.

[5] 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.

[6] 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.

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

[8] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050, 2014.

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

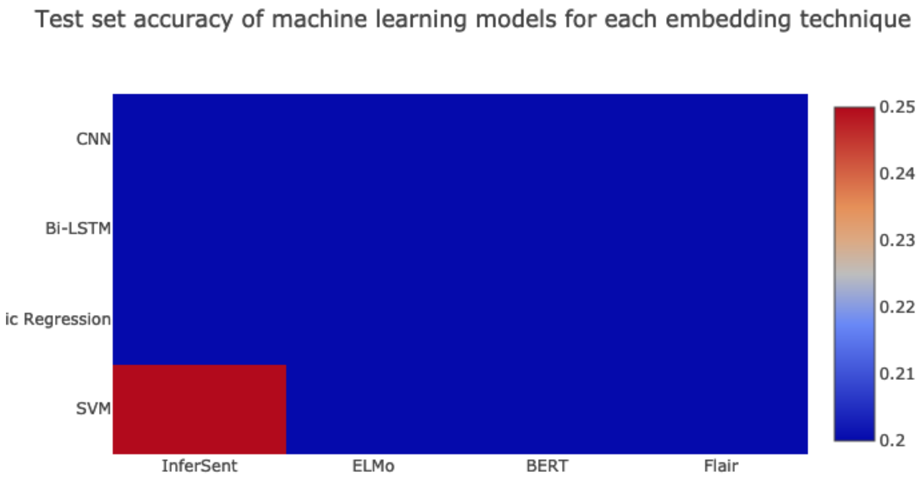

Figure 6: Comparing linear classifiers with neural classifiers.

[10] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[11] Lev Konstantinovskiy. Sentence embeddings for automated factchecking - lev konstantinovskiy.

[12] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

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

[14] Christopher Olah. https://colah.github.io/posts/2015-08-understanding-lstms/, August 2015.

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

[16] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer, 2010.

[17] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 745–750, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[18] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *CoRR*, abs/1805.09843, 2018.

[19] 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.

[20] Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3, 2003.

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

[22] L. Wen, X. Li, L. Gao, and Y. Zhang. A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Transactions on Industrial Electronics*, 65(7):5990–5998, July 2018.

[23] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.