Claim Extraction from Text using Transfer Learning

Acharya Ashish Prabhakar[1], Salar Mohtaj[1,2], and Sebastian Möller[1,2]

[1]Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
[2]DFKI Projektbüro Berlin, Berlin, Germany
a.prabhakar@campus.tu-berlin.de
{salar.mohtaj|sebastian.moeller} @ tu-berlin.de

Abstract

Building an end to end fake news detection system consists of detecting claims in text and later verifying them for their authenticity. Although most of the recent works have focused on political claims, fake news can also be propagated in the form of religious intolerance, conspiracy theories etc. Since there is a lack of training data specific to all these scenarios, we compiled a homogeneous and balanced dataset by combining some of the currently available data. Moreover, it is shown in the paper that how recent advancements in transfer learning can be leveraged to detect claims, in general. The obtained result shows that the recently developed transformers can transfer the tendency of research from claim detection to the problem of check worthiness of claims in domains of interest.

1 Introduction

The advent of social media and mobile-based messaging applications has led to a rapid spread of fake news and misinformation, which are having serious consequences in real-world scenarios. Manual fact-checkers are often overwhelmed by the number of sources that need to be verified. Facebook has released some statistics in its regular enforcement reports, as follow:

- In 2016, known fake news content was getting around 200 million engagements on Facebook each month
- In Q1 of 2018, Facebook removed 837 million pieces of spam
- In Q1 2018, Facebook also removed 583 million fake accounts

The automated fact-checking process can help to mitigate this problem. The very first step of a fact-checking system is to identify claims from the text, which can then be verified by querying a knowledge base or by converting it into a question and looking for suitable answers using a crawler (Vlachos and Riedel, 2014). Since identifying claims is the very first step of a fact-checking pipeline, the performance on this task has a major impact on the result of an end to end fact-checking system. As the performance of the state-of-the-art deep learning models are highly dependent to the availability of labelled data, in this paper, we introduce a novel dataset to assist in training models for the task of claim detection. Moreover, we investigate the performance of some of the recent transformers based language models (e.g., BERT) on the proposed data.

The paper is organized as follow. Some of the recently proposed models and datasets for the task of claim detection are presented in Section 2. Section 3 contains a brief introduction to transfer learning and the state-of-the-art models. The proposed approaches for compiling the dataset and detecting claims, and also the baseline model are described in details in Section 4.

2 Related Work

Since fact-checking and fake news detection is time-consuming and effort-full tasks, in many systems, claim detection and argument mining are considered as the preliminary modules that provide input for fact-checking module.

ClaimBuster, which is introduced in (Hassan et al., 2017), tries to rank political claims in debates based on their check-worthiness using supervised models. To this end, a labelled dataset of spoken sentences by presidential candidates is constructed. Each sentence in the dataset is given one of the three possible labels; it is not a factual claim; it is an unimportant factual claim; it is an important factual claim. Among various classification

Proceedings of the 17th International Conference on Natural Language Processing, pages 297–302
Patna, India, December 18 - 21, 2020.

Method	Source of training data	Model used	Precision	Recall	F1
ClaimBuster	Hand annotated US presidential debates	SVM	79%	74%	76%
ClaimRank	Popular fact checking organizations	FNN	93%	65%	77%
Full Fact	Crowd sourced annotations	LogReg	88%	80%	83%
Logically	Annotated news articles	Google USE	90%	89%	89%

Table 1: Comparison of the scores of previous works

methods which have been trained on the proposed dataset, support vector machine outperforms the other methods in the accuracy of finding important claims. Its *ClaimSpotter* component performs the task of claim detection with a precision of 79% and a recall of 74%.

Konstantinovskiy et al. (2018) focused on monitoring news sources and identifying "if a particular sentence constitutes a claim that could be fact-checked?". Since the definition of claim and also its importance are subjective tasks and rely on many factors (e.g. personal background), one of the most important aspects of their work is proposing an annotation guideline in which 'claim' and 'important claim' are defined to annotators. Their final results show that logistic regression classifier gives the highest overall F1 score, comparing to the other supervised models in a different setting.

Another recent approach to claim detection is *ClaimRank* (Jaradat et al., 2018). The authors claim that although the system originally trained on political debates, it works for any text, e.g. interviews and regular news articles. They compiled a dataset by taking the outputs of fact-checking of a political debate, published by nine reputable organizations simultaneously. Models were created to predict if the claim would be highlighted by at least one or by a specific organization. The modelling is done with a large variety of features from both the individual sentence and the wider context around it. To classify claims and rank them on their check worthiness, a two hidden layered neural network is trained. Adler and Boscaini-Gilroy (2019) used Google's universal sentence encoder to obtain the sentence embedding and passed it through a logistic regression layer to get the final classification. Their model has trained on a dataset based on news articles created by them.

A detailed comparison of the previously proposed methods is presented in Table 1. In addition to develop a claim detection based on transfer learning approach, in this paper we compile a new balanced dataset for the claim detection task, containing sentences from different domain and contexts.

3 Transfer Learning

Neural networks need large scale datasets to be trained efficiently. They are difficult to apply where the available data is sparse. The Imagenet moment (Deng et al., 2009) where fine-tuning, a model trained on a large dataset could be applied in various applications with limited availability of data, created a breakthrough in computer vision.

Language modelling was seen as the most appropriate task to model this achievement by Imagenet in Natural Language Processing (NLP). When trained for language modelling, neural networks capture the basic structure and understanding of language and can thus be fine-tuned on downstream tasks with limited data. Many attempts have been made in creating a sophisticated model trained on a large dataset that can be fine-tuned easily on a downstream task. *ULMFit* (Howard and Ruder, 2018) was one of the earliest models to achieve this through an *AWD LSTM* (Merity et al., 2018). *ELMO* (Peters et al., 2018) used a bi-directional *LSTM* architecture with character level encoding of features to avoid out of vocabulary errors. *BERT* (Devlin et al., 2019) used the recently developed transformer architecture to perform language modelling.

Transformers have the advantage of being able to train faster due to possibility of parallelization. *DistilBert* model (SANH et al.) was created by a method called model distillation, and it is 40% smaller, and 60% faster than *BERT* and it retains 97% of its performance, making it more deployment friendly. In our experiments, *BERT* and *DistilBert* have been used to extract claims from text.

4 Proposed Approach

In this section we present the applied approaches for compiling the dataset and detecting claims. Moreover, the baseline model is also described in this section.

Description		No.
Document statistics	Total number of samples	395,057
	Total number of claim samples	197,528
	Total number of non-claim samples	197,529
Word/Char statistics	Average number of words per sample	83.61
	Average character per sample	408.94
	Average number of stop words per sample	39.18

Table 2: Statistics of our dataset.

4.1 Dataset

To train a neural network for detecting claims, we need a dataset with claim and non-claims classes. There are several datasets available for the task of claim detection. *FEVER* (Thorne et al., 2018) is the largest available dataset for this task. Their annotators performed several types of mutations of *Wikipedia* articles summary section to create claims. Wang (2017) collected claims from well-known fact-checking organizations. However, there are no significant datasets available for the negative class (i.e. non-claims). Since several machine learning methods expect a balanced training dataset to get the desired result. With this motivation, we have come up with the following methodology to complement the abundance of publicly available claim samples with non-claim samples.

We have created a large scale dataset of non-claim sentences and made it publicly available at the following Github repository [1]. Although such a large dataset is not needed for the transfer learning based models, we aim to assist future researchers in training simpler models which would expect a balanced ratio of classes.

These non-claim sentences have been obtained from *Wikipedia*. Wikipedia's citation policy states that;

> "Wikipedia's verifiability policy requires inline citations for any material challenged or likely to be challenged, and for all quotations, anywhere in article space."

The definition of a claim also happens to be;

> "A statement about the world that can be verified".

Since *Wikipedia* expects citations for anything verifiable, it inturn requisites that claims in their articles be cited. *Wikipedia* also expects quotations to be cited which may or may not be claims. Thus sentences without citations would be non-claims and sentences with citations can be both. By filtering out citations from Wikipedia articles we would be left with only non-claims. Since any user can edit a *Wikipedia* page, and it can happen that beginners are not entirely aware of its policy and ignore these rules. We only work with pages having pending changes protection, extended confirmed protection, semi-protection and full protection. Only verified users can edit these pages with these protection levels.

We created a web crawler to access Wikipedia's page contents and filter out the sentences containing citations leaving us with only non-claim sentences. This crawler is programmed using Python and uses libraries such as Spacy, Regex and Wikipedia's API features. For our final balanced dataset, we use the samples returned by the above-mentioned crawler as non-claims and take data samples of class claim from *FEVER* and Wang (2017). We summarize some of the statistics of our final combined dataset in Table 2. Moreover, some samples from the combined dataset are presented in Table 3.

4.2 Baseline model

We trained a two-layered Long short-term memory *(LSTM)* network (Hochreiter and Schmidhuber, 1997) as a baseline model. Its input words are converted into word embeddings using google's pre-trained *word2vec* model (Mikolov et al., 2013), which creates word embeddings of dimension 300. The network's configurations consist of a hidden layer size of 300, a learning rate of 0.001, Adam Optimizer (Kingma and Ba, 2015), negative log-likelihood as loss function and a mini-batch size of 64 for training. We train for the dataset mentioned in section 4.1, in a five-fold cross-validation set manner, on a Tesla GPU based computing cluster.

[1] https://github.com/ashish6630/Claim_extraction.git

Sentence	Claim
In Georgia, women earn 82 cents for every dollar earned by men.	Yes
Bermuda Triangle is in the western part of the Himalayas.	Yes
In Azerbaijan 53% of the population, according to polls, state that religion has little to no importance in their lives.	Yes
Tupac Shakur is my favourite Rapper.	No
Some praised Rogan for hosting a pragmatic discussion while others seemed rather stunned by Sanders decision to appear on the show at all.	No

Table 3: Some samples from the proposed dataset

The results are summarized in Table 4.

Further experiments were tried with a varying range of dropout and learning rates, but there was no increase in the scores. Increasing the size of the training data is the only option to improve model accuracy. The transfer learning based model is described in the next section.

4.3 Transfer learning models

We perform transfer learning by fine-tuning a pre-trained *BERT* base and *DistilBert* base model. These models have the advantage of having been pre-trained on a large corpus for the unsupervised task of language modelling. They also include multiple self-attention heads which encode how a word in a sentence relates to the other words, and this information can prove useful during the final classification.

Although *BERT* does not use character level embedding like *ELMO*, it is still able to avoid out of vocabulary errors by breaking words into subwords wherever possible. Unlike the LSTM, where the next time step of the computation requires information from the previous time step, making it challenging to parallelize, *BERT* being a transformer-based model processes the entire sentence at once. Training time is now significantly faster. The sentence embedding generated at the end is used for further steps, and the rest of the output is discarded. This embedding is passed through a linear transformation layer to map the embedding of size 768 to the class size of 2, i.e. claim and non-claim. These models were fine-tuned on our combined dataset shown in Table 3 for three epochs with a learning rate of 2e-5 with *Adam* optimizer on a Tesla GPU cluster. Table 4 contains the results after training for three epochs with 2000 samples (1000 claims and 1000 non-claims) and testing with the remaining samples from the dataset mentioned in section 4.1.

The ratio of the distribution of claims can vary depending on the scenario, e.g. A presidential debate transcript will mostly consist of claims. In contrast, some scenarios might only have a lesser percentage of claims. Taking that into account, we experimented with varying ratios of claims and non-claims, and the results are shown in Table 4. The results of the transfer learning model are robust regardless of the ratio of claims in the dataset. As highlighted in the table, both *BERT* and *DistilBert* obtain promising results with only a fraction of the dataset.

5 Conclusion and Future Work

In this work, we have presented our publicly available dataset and quantified the performance of *BERT* and *DistilBert* in detecting claims in general. These are one of the most advanced transfer learning methods available and can provide highly accurate results with fewer data. The transfer learning based pre-training of these models helped it to achieve high evaluation scores despite having been trained with a fraction of the available dataset.

Until now, Fake news detection has been thought of as a two-step process consisting of detecting claims and verifying them. The first part can be further subdivided into detecting claims and sorting them according to the check worthiness of a claim. Other research domains such as argument mining would benefit from this since they would want to sort claims according to argumentative claims rather than check worthiness. Researchers can thus decide themselves the type of claim to filter out. We will address the problem of check worthiness in our future work.

Model	Data distribution		Accuracy	Precision	Recall	F1
	% of claims	% of non-claims				
LSTM(Baseline)	50%	50%	70.32%	74.37%	77.82%	76.06%
BERT	10%	90%	95.32%	95.12%	96.24%	**95.68**%
BERT	25%	75%	96.13%	96.43%	96.89%	**96.66**%
BERT	50%	50%	98.42%	97.38%	98.61%	**97.99**%
BERT	75%	25%	97.06%	97.10%	97.79%	**97.44**%
BERT	90%	10%	95.54%	95.21%	95.88%	**95.54**%
DistilBert	10%	90%	95.36%	94.95%	95.39%	95.17%
DistilBert	25%	75%	95.85%	95.72%	95.91%	95.81%
DistilBert	50%	50%	97.78%	96.61%	98.37%	97.66%
DistilBert	75%	25%	96.12%	96.24%	96.31%	96.27%
DistilBert	90%	10%	94.47%	94.59%	94.63%	94.61%

Table 4: Final Results on the proposed dataset

References

Ben Adler and Giacomo Boscaini-Gilroy. 2019. Real-time claim detection from news articles and retrieval of semantically-similar factchecks. In *Proceedings of the Third International Workshop on Recent Trends in News Information Retrieval, co-located with 42nd International ACM Conference on Research and Development in Information Retrieval (SIGIR 2019), Paris, France, July 25, 2019*, volume 2411 of *CEUR Workshop Proceedings*, pages 36–41. CEUR-WS.org.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 1803–1812. ACM.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, Louisiana. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *CoRR*, abs/1809.08193.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and optimizing LSTM language models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF, and Hugging Face. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the Workshop on Language Technologies and Computational Social Science@ACL 2014, Baltimore, MD, USA, June 26, 2014*, pages 18–22. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 422–426. Association for Computational Linguistics.