A simple but tough-to-beat baseline for the Fake News Challenge stance detection task

Benjamin Riedel[1], Isabelle Augenstein[12], Georgios P. Spithourakis[1], Sebastian Riedel[1]
[1]Department of Computer Science, University College London, United Kingdom
[2]Department of Computer Science, University of Copenhagen, Denmark
benjamin.riedel.09@ucl.ac.uk,
{i.augenstein|g.spithourakis|s.riedel}@cs.ucl.ac.uk

Abstract

Identifying public misinformation is a complicated and challenging task. Stance detection, i.e. determining the relative perspective a news source takes towards a specific claim, is an important part of evaluating the veracity of the assertion. Automating the process of stance detection would arguably benefit human fact checkers. In this paper, we present our stance detection model which claimed third place in the first stage of the Fake News Challenge. Despite our straightforward approach, our model performs at a competitive level with the complex ensembles of the top two winning teams. We therefore propose our model as the 'simple but tough-to-beat baseline' for the Fake News Challenge stance detection task.

1 Introduction

Automating stance detection has been suggested as a valuable first step towards assisting human fact checkers to detect inaccurate claims. The Fake News Challenge initiative thus recently organised the first stage of a competition (FNC-1) to foster the development of systems for evaluating what a news source is saying about a particular issue [13].

More specifically, FNC-1 involved developing a model that, given a news article headline and a news article body, estimates the stance of the body towards the headline. The stance label to be assigned could be one of the set: 'agree', 'disagree', 'discuss', or 'unrelated' (see example of Figure 1). More information on the FNC-1 task, rules, data, and evaluation metrics can be found on the official website: `fakenewschallenge.org`.

The goal of this short paper is to present a description of UCL Machine Reading's (UCLMR) model employed during FNC-1, a summary of the model's performance, a brief overview of the competition, and our work going forward.

2 Model description

Our stance detection model is a single, end-to-end system consisting of lexical and similarity features fed through a multi-layer perceptron (MLP) with one hidden layer. Although relatively simple in nature, the model performs on par with more elaborate, ensemble-based systems of other teams [3, 7] (see Section 4).

The code for our model and instructions on how to reproduce our submission are available at UCLMR's public GitHub repository: `github.com/uclmr/fakenewschallenge`.

2.1 Features

We use two simple bag-of-words (BoW) representations for the text inputs: term frequency (TF) and term frequency-inverse document frequency (TF-IDF) [10, 8]. The features extracted from the headline and body pairs consist of only the following:

- The TF vector of the headline;
- The TF vector of the body;
- The cosine similarity between the TF-IDF vectors of the headline and body.

We tokenise the headline and body texts as well as derive the relevant vectors using `scikit-learn` [12].

Different vocabularies are used for calculating the TF and TF-IDF vectors. For the TF vectors, we extract a vocabulary of the 5,000 most frequent words in the training set and exclude stop words (the `scikit-learn` stop words for the English language with negation terms removed). For the TF-IDF vectors, a vocabulary of the 5,000 most frequent words is defined on both the training and test sets and the same set of stop words is excluded.

The TF vectors and the TF-IDF cosine similarity are concatenated in a feature vector of total size 10,001 and fed into the classifier.

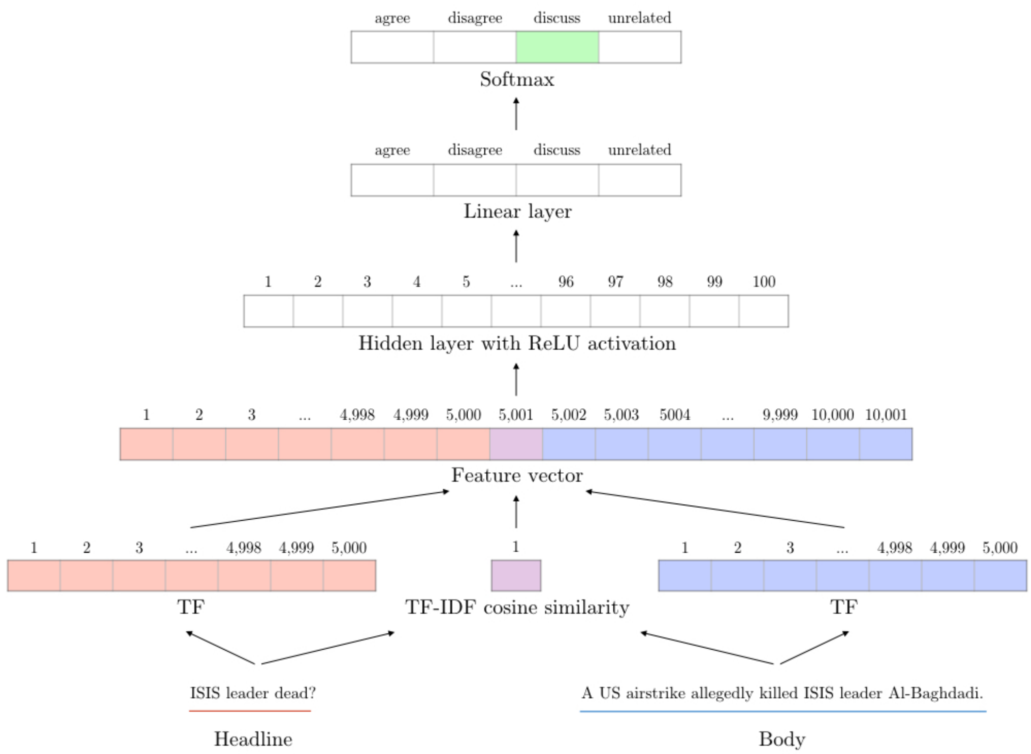

Figure 1: Schematic diagram of UCLMR's model.

2.2 Classifier

The classifier is a MLP [4] with one hidden layer of 100 units and a softmax on the output of the final linear layer. We used the rectified linear unit (ReLU) activation function [6] as non-linearity for the hidden layer. The model predicts with the highest scoring label ('agree', 'disagree', 'discuss', or 'unrelated'). The classifier as described is fully implemented in `TensorFlow` [1].

2.3 Training

Our training objective was to minimise the cross entropy between the model's softmax probabilities and the true labels. For regularisation, we included a L2 regularisation term for the MLP weights in the objective and applied dropout [14] during training on the output of both perceptron layers.

We trained in mini-batches over the entire training set with back-propagation using the Adam optimiser [9] and gradient clipping by a global norm clip ratio [11]. All of the aforementioned were implemented in `TensorFlow` [1].

Training was stopped early based on a qualitative criterion with respect to the plateau of the loss on the training set and the mean performance of the model on 50 random splits of the data into training and hold-out sets as defined in the official baseline setup [13].

2.4 Hyperparameters

The full set of hyperparameters of the model, their labels, their descriptions, the ranges of values considered, and corresponding optimised values are provided in Table 1. The hyperparameters were optimised during development using random search on a grid of combinations and (cross-validation on) various splits of the data.

Table 1: Details on hyperparameters of UCLMR's model.

Label	Description	Range	Optimised
`lim_unigram`	BoW vocabulary size	1,000 - 10,000	5,000
`hidden_size`	MLP hidden layer size	50 - 600	100
`train_keep_prob`	1 - dropout on layer outputs	0.5 - 1.0	0.6
`l2_alpha`	L2 regularisation multiplier	0.1 - 0.0000001	0.0001
`learn_rate`	Adam learning rate	0.1 - 0.001	0.01
`clip_ratio`	Global norm clip ratio	1 - 10	5
`batch_size`	Mini-batch size	250 - 1,000	500
`epochs`	Number of epochs	$\leq 1,000$	90

3 Results

Submissions to the competition were evaluated with respect to the FNC-1 score, as defined in the official evaluation metrics [13]. Our submission achieved a FNC-1 score of 81.72%.

The performance of our model is summarised by below confusion matrix for the label predictions submitted on the final test set (see Table 2). We conclude that although our model performs satisfactorily in general, this can mainly be attributed to the close to perfect classification of the instances into 'related' and 'unrelated' headline/body pairs (accuracy: 96.55%) and the more or less default 'discuss' classification of the 'related' instances.

Table 2: Confusion matrix of UCLMR's FNC-1 submission.

	Agree	Disagree	Discuss	Unrelated	% Accuracy
Agree	838	12	939	114	44.04
Disagree	179	46	356	116	6.60
Discuss	523	46	3,633	262	81.38
Unrelated	53	3	330	17,963	97.90
Overall					88.46

Our model's performance with respect to the 'agree' label is average at best, whereas the model's accuracy on the 'disagree' test examples is clearly quite poor. The disappointing performance is noteworthy since these two labels are arguably the most interesting in the FNC-1 task and the most relevant to the superordinate goal of automating the stance detection process.

4 Competition

A total of 50 teams actively participated in FNC-1. The final top 10 leaderboard (see Table 3) shows our submission (UCLMR) placed in third position.

Table 3: Top 10 FNC-1 leaderboard. UCLMR submission in **bold**.

Team	% FNC-1 score
SOLAT in the SWEN	82.02
Athene	81.97
UCL Machine Reading	**81.72**
Chips Ahoy!	80.21
CLUlings	79.73
unconscious bias	79.69
OSU	79.65
MITBusters	79.58
DFKI LT	79.56
GTRI - ICL	79.33

The competition was won by team 'SOLAT in the SWEN' from Talos Intelligence, a threat intelligence subsidiary of Cisco Systems, and the second place was taken by team 'Athene' consisting of members from the Ubiquitous Knowledge Processing Lab and the Adaptive Preparation of Information from Heterogeneous Sources Research Training Group at Technische Universität Darmstadt (TU Darmstadt). The respective FNC-1 scores of these teams were 82.02% and 81.98%.

The team from Talos Intelligence employed a 50/50 weighted average ensemble of (i) two one-dimensional convolutional neural networks on respectively word embeddings of the headline and body feeding into a MLP with three hidden layers and (ii) five overarching sets of features fed into gradient boosted decision trees [3].

The team from TU Darmstadt used an ensemble of five separate MLPs, each with seven hidden layers and fed with seven overarching sets of features. Predictions were based on the hard vote of the five separate, randomly initialised MLPs [7].

The submission of our team performed almost on par with the top two teams and with considerable distance to the remaining teams. In contrast to other submissions, we achieved competitive results with a simple, single, end-to-end system.

From discussions with other teams, including those from Talos Intelligence and TU Darmstadt, the test set performance of other models on the labels of key interest ('agree' and 'disagree') was not much better either, if at all.

5 Future work

Our goal going forward is to carry out in-depth analyses of our model. The added benefit of our straightforward setup, as opposed to more sophisticated neural network architectures, is that it provides an opportunity to try to understand how it works, what contributes to its performance, and what its limitations are.

A particular focus of these analyses will be to try and identify what the mediocre performance of the model with respect to the 'agree' and 'disagree' labels can potentially be traced back to, next to the limited size of the data set overall and the small number of instances of the labels of specific interest.

Notwithstanding this, we would like to propose our model as the 'simple but tough-to-beat baseline' [2] for FNC-1 stance detection task given the model's competitive performance and basic implementation. We accordingly welcome researchers and practitioners alike to employ, improve, and/or extend our work thus far.

Acknowledgements

We would like to thank Richard Davis and Chris Proctor at Stanford University for the description of their FNC-1 development efforts. The model presented here is based on their setup [5].

Furthermore, we are grateful for insightful discussions with the following individuals during model development and the official competition.

- Florian Mai at the Christian-Albrechts Universtität zu Kiel
- Anna Seg from the FNC-1 team 'annaseg'
- James Thorne at the University of Sheffield
- Sean Bird from Talos Intelligence
- Andreas Hanselowski at the Technische Universität Darmstadt
- Jingbo Shang at the University of Illinois at Urbana-Champaign

This work was supported by a Marie Curie Career Integration Award, an Allen Distinguished Investigator Award, and Elsevier.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `http://tensorflow.org/`.

[2] S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference for Learning Representations*, 2017.

[3] S. Baird, D. Sibley, and Y. Pan. Talos targets disinformation with Fake News Challenge victory, 2017. URL `https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html`.

[4] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 1989.

[5] R. Davis and C. Proctor. Fake news, real consequences: Recruiting neural networks for the fight against fake news, 2017. URL `https://web.stanford.edu/class/cs224n/reports/2761239.pdf`.

[6] R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 2000.

[7] A. Hanselowski, A. PVS, B. Schiller, and F. Caspelherr. Description of the system developed by team Athene in the FNC-1, 2017. URL `https://github.com/hanselowski/athene_system/blob/master/system_description_athene.pdf`.

[8] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 1972.

[9] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015.

[10] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1957.

[11] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, 2013.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[13] D. Pomerleau and D. Rao. Fake News Challenge, 2017. URL `http://www.fakenewschallenge.org/`.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014.