Andreas Hanselowski [Follow]
Aug 1 · 7 min read

Team Athene on the Fake News Challenge

In this blog post, we would like to report about our experience in participating in the Fake News Challenge Stage One (FNC-1). We have secured the second place in the competition and would like to share our insights into the problem setting and to describe our approach to the task.

The Team

We would like to start by introducing ourselves as a group of junior researchers at the UKP Lab TU Darmstadt in Germany, which is led by Prof. Dr. Iryna Gurevych. Benjamin Schiller and Felix Caspelherr are master students who implemented the model and performed experiments. Andreas Hanselowski and Avinesh P.V.S. are Ph.D. students at UKP/AIPHES and have been in charge of organization and supervision of the project. Debanjan Chaudhuri is a Ph.D. student at the University of Bonn. He was advising the team in regular meetings.

Our team name is inspired by the emblem of TU Darmstadt, namely the goddess Athene. Athene is considered to be the goddess of wisdom, craft and war. She is believed never to fight without a purpose and only to fight for just causes. We were thinking, this is the kind of support we need in our battle against fake news. And indeed, Athene did not let us down. Our system reached high performance on the task, and out of 50 teams, we have only been beaten by the team SOLAT in the SWEN (not sure which Gods they worship).

Athene, antiquity's patron goddess of the Arts and Sciences, is the official emblem of Technische Universität Darmstadt. [www.reference.com]

The Fake News Challenge

In the past couple of years, there has been a significant increase of fake news articles on the web, which has come into focus of public debate. The spread of false information is not just considered as a threat to the reputation of news providers in general but even as a risk factor to democratic order. People are increasingly concerned about the influence of fake news on elections in the US and Europe.

In order to validate the false information circulating on the web, fact-checking became an essential tool, and today, there are numerous websites such as fullfact.org, politifact.com, and snopes.com devoted to the topic.
On these websites, journalists or professional fact-checkers manually resolve controversial claims, by collecting evidence that either support and refute them. Although it is an important instrument in the fight against fake news, manual fact-checking is not up to the challenge due to a vast number of fake news articles being generated at a high rate. Moreover, expeditious identification of fake news is also an important factor. Due to the high speed fake news spreads through social media, it is important to intervene in the proliferation process as soon as possible.
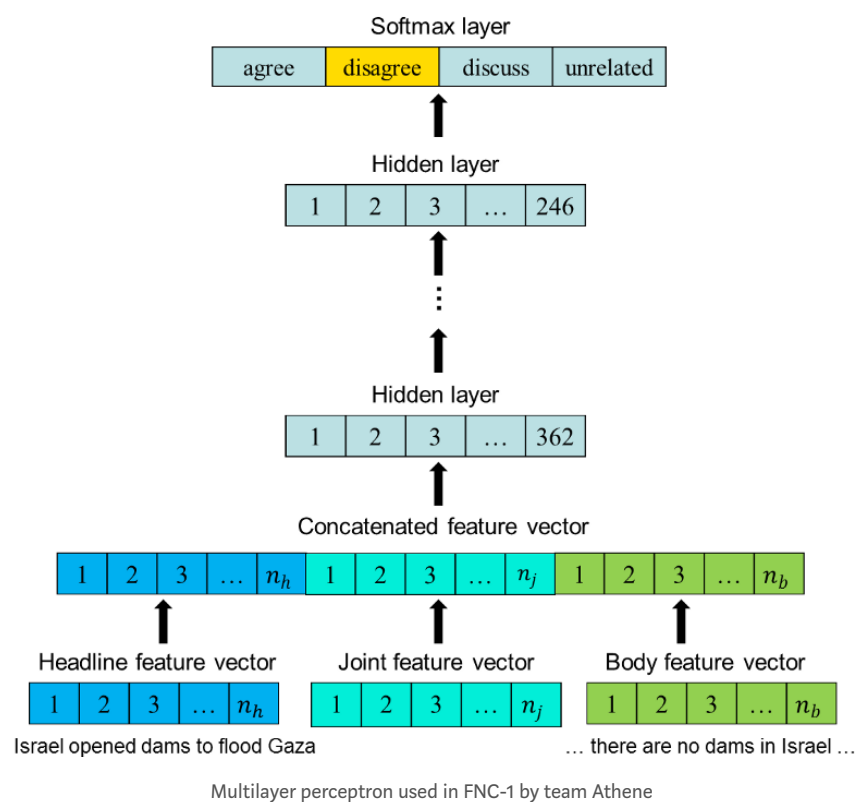
The recent surge of advancements in AI has opened up new opportunities in a wide range of different fields also suggesting that the technology can be leveraged to tackle the fake news problem. In December 2016, partners from academia and industry joined forces in order to foster the development of AI methods to help solve the fake news problem by launching the Fake News Challenge. Since the

task of fake news debunking is difficult, the problem is split into a number of sub tasks, each of which will be tackled in a separate competition. The first competition of the Fake News Challenge was concerned with the problem of identifying the stance of a given news article body with respect to a given headline. The systems should be able to classify whether the article body agrees, disagrees, discusses the topic of the headline or is unrelated to it.

Our Approach

In the course of the competition, we have tried out a number of different methods and our model has evolved over a number of steps. The changes in part reflect the ideas from the discussion of the methods on the fake news challenge slack. We started working with standard classifiers like SVM, XGBoost and implemented a number of features based on paraphrase detection, sentiment lexica and others. The performance of the methods did not significantly improve on the published baseline and we moved to more sophisticated techniques.

As often remarked by Christopher Manning, the standard approach to NLP problems these days are BiLSTMs. We were tempted also to take this route and performed the first experiments. However, the discussion on Slack did show us that recurrent models might not be the right choice for the task after all. The experiments performed by students from the Stanford class CS224 have confirmed this observation. They did a great job by trying out a large number different LSTM model structures and publishing their experiments.

However, special gratitude goes to **Richard Davis and Chris Proctor** who, to our knowledge, **first came up with the idea of applying a multiplayer perceptron and BoW features to the FNC-1**. This is surely the major advantage of a collaborative effort to solve a problem. We assume that many of the teams attending FNC-1 picked up on the idea as we did. Indeed, the MLP suggested by Richard Davis and Chris Proctor was the starting point for the development of our final system. The system structure has been further optimized for our features by a random search whereby the hyper-parameters have been adjusted. The resulting model structure is illustrated below. To shorten the illustration 5 of the 7 hidden layers are skipped.

Softmax layer

agree	disagree	discuss	unrelated

Hidden layer

1	2	3	…	246

Hidden layer

1	2	3	…	362

Concatenated feature vector

1	2	3	…	n_h	1	2	3	…	n_j	1	2	3	…	n_b

Headline feature vector   Joint feature vector   Body feature vector

1	2	3	…	n_h

1	2	3	…	n_j

1	2	3	…	n_b

Israel opened dams to flood Gaza   … there are no dams in Israel …

Multilayer perceptron used in FNC-1 by team Athene

Depending on the feature type, either individual feature vectors for the
article body and the headline are created and then concatenated, or a joint feature vector. In addition to the baseline features we have used the following features.

Features:
BOW: Bag of Words uni-grams

NNF: Non-Negative Matrix Factorization

LSI: Latent Semantic Indexing

LSA: Latent Semantic Analysis

PPDB: Paraphrase Detection based on Word Embeddings

The constructed multiplayer perception was used as a basis for a model ensemble consisting of 5 multilayer perceptrons, of which the weights have been randomly initialized. The final predictions have been determined by hard voting among the multilayer perceptrons.

The results of our system are listed below. The confusion matrix indicates that the system is able to predict the classes **Discuss** and **Related** with a high accuracy. However, the **Disagree** and **Agree** classes are predicted poorly. To improve on this classes would surely be a promising starting point for further research.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	851	69	826	157	44.72
Disagree	241	66	241	149	9.47
Discuss	466	37	3611	350	80.89
Unrelated	19	4	115	18211	99.25
Overall					89.5

Confusion matrix and accuracy

Additional information about our model as well as the code can be found on github.

Analysis and our take away from the challenge

As discussed on slack, there are a number of standard machine learning approaches, which did not to work for the challenge as they do for other NLP tasks. Among those were word embeddings, and

neural networks based on these. Instead, it has been observed that methods based on bag-of-words, Tf-Idf, n-grams, and topic models work well for FNC-1. This phenomenon is typically encountered for small corpora, and indeed, the FNC-1 corpus with about 1600 unique article bodies is rather small. We believe this was the reason why deep learning techniques like Bi-LSTMs or CNNs were outperformed by a feature based multilayer perceptrons. In fact, the team from UCL, which came in third, was also using a multilayer perception. *(The CNN by SOLAT in the SWEN was only used as a part of an ensemble and is not competitive on its own.)*

Rank	Team	Method	Score
1.	SOLAT in the SWEN	CNN-classifier ensemble	82.02
2.	Athene UKP Lab	MLP ensemble	81.97
3.	UCL Machine reading	MLP	81.72

It follows, that the developed methods are going to perform poorly on a new data-sets, which have different statistics compared to the FNC-1 corpus. Indeed, the drop of performance we experienced for our system from 92% on the training set to about 82% on testing set was enormous. This problem has also been noticed by the organizers on Slack. **Dean Domerleau:** *"What it all means is that the top teams did really well, but there is definitely still room for improvement!"*

In general, deep learning techniques like Bi-LSTM and CNNs with attention reach state-of-the-art results in many NLP tasks, and they are also expected to work well for document based stance classification if the "data bottleneck" is solved. It would be therefore advantageous if, in the coming challenges, the provided corpora would be significantly larger. The extension of the data sets would attract more researcher and the developed methods would be more robust and therefore more generally applicable.

Conclusion

The fake news challenge has taught us many important lessons in machine learning and NLP, and we had a great experience in participating in the competition. We would like to thank the organizers for a great setup, but also the community for valuable discussions on slack. In order to further progress research on FNC-1, we are currently running a number of experiments on the data set.

The results will be published in a paper in the coming months.