Fake News Detection using Stacked Ensemble of Classifiers

James Thorne **Mingjie Chen** **Giorgos Myrianthous**
Jiashu Pu **Xiaoxuan Wang** **Andreas Vlachos**
Department of Computer Science
University of Sheffield, UK
{j.thorne,mchen33,gmyrianthous1,ppu1,xwang130,a.vlachos}
@sheffield.ac.uk

Abstract

Fake news has become a hotly debated topic in journalism. In this paper, we present our entry to the 2017 Fake News Challenge which models the detection of fake news as a stance classification task that finished in 11th place on the leader board. Our entry is an ensemble system of classifiers developed by students in the context of their coursework. We show how we used the stacking ensemble method for this purpose and obtained improvements in classification accuracy exceeding each of the individual models' performance on the development data. Finally, we discuss aspects of the experimental setup of the challenge.

1 Introduction

The distribution of news on social media is an influential factor in the public's political attitudes (Allcott and Gentzkow, 2017). Social networks offer platforms in which information and articles may be shared without fact-checking or moderation. Moderating user-generated content on social media presents a challenge due to both volume and variety of information posted. In particular, highly partisan fabricated materials on social media, *fake news*, is believed to be an influencing factor in recent elections (DiFranzo and Gloria-Garcia, 2017). Misinformation spread through fake news has attracted significant media attention recently and current approaches rely on manual annotation by third parties (Heath, 2016) to notify users that shared content may be untrue.

One of the challenges in detecting misinformation is that there does not yet exist a unified definition of fake news and the criteria required to label an article as true or false. As a consequence, there is no community-wide shared task in order to compare the various approaches proposed. Until recently, the evaluations related to fake news have had relatively little adoption. Even though there is valid criticism that shared tasks have the risk of focusing the community on a particular task definition and dataset, shared definition and evaluation platforms such as those developed for example by the CoNLL shared tasks[1] have largely stimulated progress.

The 2017 *Fake News Challenge*[2] (FNC) aims to provide a community-wide shared task and evaluation platform in order to stimulate progress in fake news detection. Acknowledging the complexity of the task even for human experts and following the task decomposition proposed by Silverman (2015), they propose to address a subtask in fake news detection, namely stance classification. Stance classification is the labeling of whether an article agrees with, disagrees with or simply discusses a 'fact'. It can be considered to be a form of textual entailment (Dagan et al., 2006), while it also bears similarity with stance classification in the context of sentiment analysis (e.g. Mohammad et al. (2016)) and . Stance classification serves as a first step in compiling lists of articles that corroborate or refute claims made on social media, allowing end-users to make a better informed judgment.

In this paper, we discuss our entry to the fake news challenge: an ensemble comprising five individual systems developed by students in the context of their natural language processing module at The University of Sheffield. We used stacking (Wolpert, 1992) as our ensembling technique as it has been applied successfully in other tasks (e.g. Riedel et al. (2011)) and show that it increases the ensemble score above the performance of any of

[1] http://www.conll.org/previous-tasks
[2] http://fakenewschallenge.org

Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism, pages 80–83
Copenhagen, Denmark, September 7, 2017.

the individual classifiers. Furthermore, we evaluate system accuracy against the upper performance bound of our ensemble, assuming a perfect oracle selecting the correct member of the ensemble to return the prediction.

2 The Fake News Challenge

The fake news challenge is a text classification task: given a headline and article body - the classifier must first predict whether the two are related and if so, must then further assign a stance label - whether the headline agrees with, disagrees with or is discussed by (observing) the article.

The evaluation for the FNC is as follows: for each stance, 0.25 points are available for correctly classifying whether the article and headline are related. A further 0.75 points are available for correctly labeling the relationship between a related headline-article pair. We report percentage scores as a proportion against the maximum possible score for correctly labeling a dataset.

The task dataset is derived from the Emergent project (Silverman, 2015) and is an extension of the stance classification task proposed by Ferreira and Vlachos (2016). It consists of 49972 labeled stances (headline and body pairs) constructed from 2582 articles and is publicly available on the organizers' website. In the FNC baseline, the organizers provide a dataset split between training data and hold-out development evaluation dataset (proportions: 0.8 training, 0.2 dev). The article bodies in this dataset split are disjoint, however, the headlines were not. An additional blind test set containing 25413 stances from 904 articles was used for evaluating the final solution. This was not made available until the competition closed and the winners were announced.

The official baseline (Galbraith et al., 2017) makes heavy use of task-specific feature engineering and applies a gradient boosted decision tree classifier to the fake news challenge dataset - achieving a score of 79.5% on the dev dataset. Features included in this approach include ngram overlap between the headline and article and the presence of refuting words (such as *fake* or *debunk*) in the headline or the article. While this baseline was good in distinguishing between the related/unrelated classes, the recall for the disagree label was poor.

The classification accuracy of the baseline is limited by the range of features used. While further feature engineering may be used to improve performance of the classifier, this requires human effort and judgment and biases the classifier to the domain in which the features were observed. Zeng et al. (2017) applied and compared three recurrent neural models which negate the need for feature engineering. While these have high FNC scores, they don't necessarily capture the aspects of the task that manually engineered features do. Bird et al. (2017) combine a deep convolutional network with feature engineering through an evenly weighted ensemble of two classifiers. Riedel et al. (2017) simply use term-frequency vectors and the tf-idf cosine similarity as features for a shallow multi-layer perceptron.

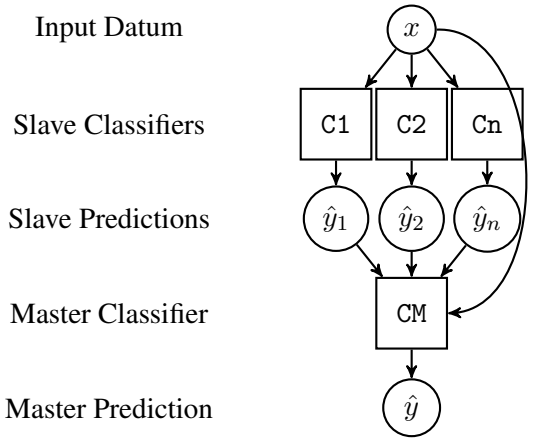

Figure 1: Stacked ensemble classification architecture where circle nodes represent data, rectangles represent classifiers and arrows indicate data flow

3 Our Solution

We present our solution to the Fake News Challenge, a stacked ensemble of five independent classifiers developed by students in the context of the natural language processing module assignments. The stacked ensemble is a two-layer classifier architecture that leverages predictions from weaker slave classifiers as features in a stronger master classifier. The architecture is illustrated in Figure 1. We provide an overview of the five slave classifiers (C1-C5) and master classifier (CM) used in the ensemble:

C1: Concatenate average word2vec vectors for headline and article body, cosine similarity between headline and article body tf-idf vectors and counts of refuting words. 4-way classification us-

ing a (300,8) multi-layer perceptron (MLP) with ReLU activation function.

C2: Average word2vec embeddings for headline words and article words excluding stop words, indicator features for punctuation, word overlap, counts of refuting words. 4-way classification using a (1010,6) MLP with ReLU activation function.

C3: 4-way classification using one-vs-all logistic regression with L2 regularization over word unigram and bigram tf-idf vectors.

C4: Concatenate word2vec embeddings for headline and article words. 4-way classification using (256,128,128) MLP with dropout probabilities of (0.5,0.3,0.1) between layers and ReLU activation function.

C5: Official FNC baseline classifier

CM: Gradient boosted decision tree classifier using as features the values predicted from C1-C5 and all the features from the FNC baseline classifier.

The master classifier is trained using 2 fold cross validation using the following regime: The dataset is randomly split into two equal sizes. Two instances of C1-C5 are instantiated and are trained independently on each data fold. The predictions are concatenated to the original input data to form one dataset - the master training data used to train CM. New instances of C1-C5 are trained on the entire original training dataset and used to provide input to CM at test time.

4 Results

We present the results for our stacked ensemble and slave classifiers trained and evaluated on the fake news challenge baseline data split (dev) and the final test set in Table 1. In the dev setup, the training set contains 40350 stances over 1345 unique articles and we evaluated on 9622 stances over 336 unique articles. The article bodies were disjoint between the training and development sets.

Because the test dataset was blinded, the risk of building a biased system was mitigated against.

[3](Galbraith et al., 2017)
[4](Bird et al., 2017)
[5](Hanselowski et al., 2017)
[6](Riedel et al., 2017)

System	Dev %	Test %
Official Baseline[3]	79.53	75.20
SOLAT in the SWEN[4]	-	82.02
Athene[5]	-	81.97
UCL Machine Reading[6]	-	81.72
C1	88.09	75.77
C2	86.68	75.08
C3	87.48	77.99
C4	87.36	58.69
C5	79.25	75.22
Our Ensemble (CM)	90.05	78.04
CM Upper Limit	*97.25*	*90.89*

Table 1: FNC score comparison on development evaluation dataset. The performance difference between C5 and the official baseline is caused by different k-fold training regimen.

However, the classification difficulty of the test set was far greater than that of the development data split which impacted results. In the development data split, article bodies were disjoint but there was some overlap between article headlines. In the training set, both article bodies and headlines were entirely disjoint. The more successful entries for this competition, such as Riedel et al. (2017), built their own entirely disjoint development split and used this for cross-validation. We found that cross-validating against the development split yielded classifiers that were not able to generalize to the unseen articles in the test set, harming the classification accuracy.

On the development dataset, the ensemble classifier yielded an absolute improvement by at least 1.6% over any of the individual constituent slave classifiers. This performance gain, however, did not transfer to the blind test set.

The CM upper limit uses a scoring oracle that awards FNC score if at least one of the slave classifiers correctly labels the input stance. This acts as a measure that describes the maximum possible score that CM could give assuming that it always selected a correct label from one of the slaves. In this case, the upper limit was 90.89% - exceeding the top ranked system. While this result is encouraging, it highlights the need to build a stronger master classifier less prone to over-fitting and more resilient to the noisy predictions made by the slaves.

The performance of some of the slave classifiers (the student projects C1-4) was variable and

highly dependent on the network topology, feature selection and dataset/split. The most resilient classifier, C5, used entirely non-lexical features whereas C4, which used only averaged word vectors and a large network topology, suffered the greatest loss in performance on the unseen test data.

The best performing system (Bird et al., 2017) is an ensemble of a convolutional neural model and a decision tree classifier. This system simply averaged the two predictions with equal weighting. The master meta-classifier in our entry leverages additional information about which slave predictions to favor given a certain headline and article pair. While the two classifiers in (Bird et al., 2017) are strong, further improvements could be obtained by incorporating stacking.

5 Conclusions

In this paper, we presented a stacked ensemble of 5 classifiers developed by students. The performance gains observed in the development set did not materialize in the competition though due to a much more difficult blind test set. One factor limiting our assessment of the ability our model(s) to generalize is the overlap of headlines between the training and development evaluation dataset. Future evaluations could consider temporal splits, i.e. deriving training, development and test sets from articles from different periods, which would also mimic to an extent how these models might be used in practice.

Acknowledgements

We would like to thank the challenge organizers, Dean Pomerleau and Delip Rao, for their efforts in this community-wide event. We are looking forward to future versions of the challenge addressing more issues in fake news detection. Andreas Vlachos is supported by the EU H2020 SUMMA project (grant agreement number 688139).

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research.

Sean Bird, Doug Sibley, and Yuxi Pan. 2017. Talos targets disinformation with Fake News Challenge victory .

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, Springer, pages 177–190.

Dominic DiFranzo and Kristine Gloria-Garcia. 2017. Filter Bubbles and Fake News. *XRDS* 23(3):32–35. https://doi.org/10.1145/3055153.

William Ferreira and Andreas Vlachos. 2016. Emergent : a novel data-set for stance classification. *Naacl2016* (1):1163–1168. https://doi.org/10.18653/v1/N16-1138.

Byron Galbraith, Humza Iqbal, HJ van Veen, Delip Rao, James Thorne, and Yuxi Pan. 2017. Baseline FNC implementation. `http://github.com/fakenewschallenge/fnc-1-baseline`.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, and Felix Caspelherr. 2017. Description of the system developed by Team Athene in the FNC-1 .

Alex Heath. 2016. Facebook is going to use snopes and other fact-checkers to combat and bury 'fake news'. `http://businessinsider.com/facebook-will-fact-check-label-fake-news-in-news-feed-2016-12`. [Online; accessed 01-June-2017].

Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets pages 31–41.

Benjamin Riedel, Isabelle Augenstein, George Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *CoRR* abs/1707.03264. http://arxiv.org/abs/1707.03264.

Sebastian Riedel, David McClosky, Mihai Surdeanu, Christopher D. Manning, and Andrew McCallum. 2011. Model combination for event extraction in bionlp 2011. In *Proceedings of the Natural Language Processing in Biomedicine NAACL 2011 Workshop (BioNLP '11)*.

Craig Silverman. 2015. Lies, damn lies and viral content. `http://towcenter.org/research/lies-damn-lies-and-viral-content/`.

David H Wolpert. 1992. Stacked generalization. *Neural networks* 5(2):241–259.

Qi Zeng, Quan Zhou, and Shanshan Xu. 2017. Neural stance detectors for fake news challenge.