Combining Neural, Statistical and External Features for Fake News Stance Identification

Gaurav Bhatt
IIT-Roorkee, India
gauravbhatt.cs.iitr@gmail.com

Aman Sharma
IIT-Roorkee, India
amanvcks@gmail.com

Shivam Sharma
IIT-Roorkee, India
shivamlmniit@gmail.com

Ankush Nagpal
IIT-Roorkee, India
ankushnagpal.cs@gmail.com

Balasubramanian Raman
IIT-Roorkee, India
balarfma@iitr.ac.in

Ankush Mittal
Graphic Era University, India
dr.ankushmittal@gmail.com

ABSTRACT

Identifying the veracity of a news article is an interesting problem while automating this process can be a challenging task. Detection of a news article as fake is still an open question as it is contingent on many factors which the current state-of-the-art models fail to incorporate. In this paper, we explore a subtask to fake news identification, and that is stance detection. Given a news article, the task is to determine the relevance of the body and its claim. We present a novel idea that combines the neural, statistical and external features to provide an efficient solution to this problem. We compute the neural embedding from the deep recurrent model, statistical features from the weighted n-gram bag-of-words model and hand crafted external features with the help of feature engineering heuristics. Finally, using deep neural layer all the features are combined, thereby classifying the headline-body news pair as agree, disagree, discuss, or unrelated. Through extensive experiments, we find that the proposed model outperforms all the state-of-the-art techniques including the submissions to the fake news challenge.

KEYWORDS

External features; Statistical Features; Stance Detection; Fake news; Deep learning

ACM Reference Format:
Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining Neural, Statistical and External Features for Fake News Stance Identification. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.1145/3184558.3191577

1 INTRODUCTION

Fake news being a potential threat towards journalism and public discourse has created a buzz across the internet. With the recent advent of social media platforms such as Facebook and Twitter, it has become easier to propagate any information to the masses within minutes. While the propagation of information is proportional to growth of social media, there has been an aggravation in the authenticity of these news articles. For an instance, in the US presidential election of 2016, the fake news has been cited as the foremost contributing factor that affected the outcome [18]. A possible reason for the failure of current security systems is the open domain nature of the problem of fake news. The recently organized Fake News Challenge (FNC-1) [9] is an initiative in this direction. The aim of this challenge is to build an automatic system that has the capability to identify whether a news article is fake or not. More specifically, given a news article the task is to evaluate the relatedness of the news body towards its headline.

The idea behind building a countermeasure for fake news is to use machine learning and natural language processing (NLP) tools that can compute semantic and contextual similarity between the headline and the body, and classify the pairs into one of four categories. Deep learning models such as recurrent neural networks (RNN) and its variants [7, 13, 17] and convolution neural networks (CNN) [14] have been efficacious in solving many NLP problems that share similarities to fake news which includes but not limited to - computing semantic similarity between sentences [15, 22], community based question answering [24, 25], etc. A deep architecture encodes the given sequence of words into fixed length vector representation which can be used to score the relevance of two textual entities, in our case, relevance of each headline-body pair. Similarly, these days it is a common practice to use embeddings from a pre-trained model such as skip-thought [15] and compute other text-based features such as bag-of-words [20] and lexical and semantic features [27]. Key advantages of feature engineering heuristics is that they do not need large amount of data for training and are computationally quick to compute. In this paper, we combine external features introduced in the baseline with some more heuristics that have been shown to be successful in other NLP tasks, and demonstrate their effectiveness over state-of-the-art techniques.

Finally, the main contributions of the paper can be summarized as

(1) We combine statistical, neural and feature engineering heuristics which achieves state-of-the-art performance on the task of fake news stance identification.
(2) The performance of the proposed model is evaluated on fake news challenge dataset. We also analyze the applicability of several state-of-the-art deep models on FNC-1 dataset.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191577

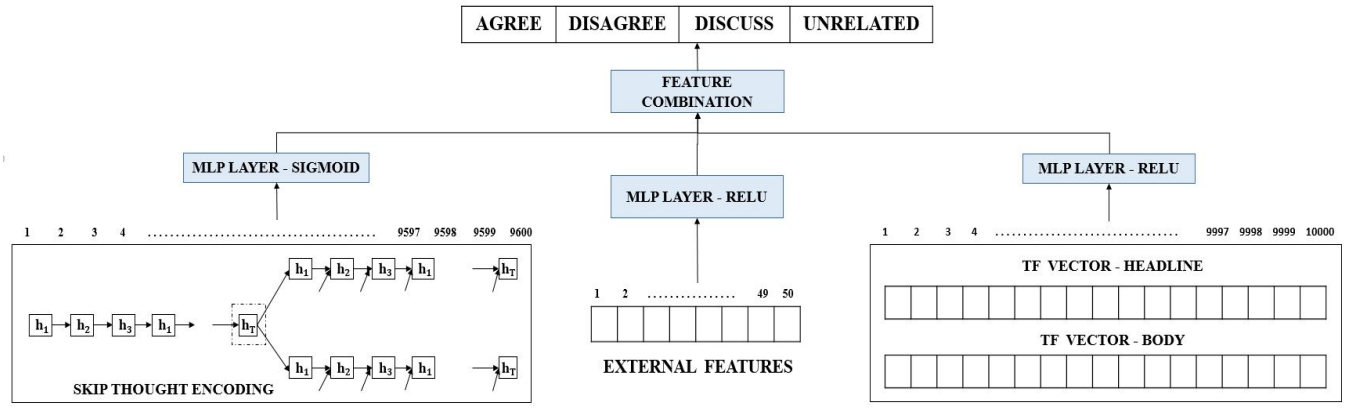

Figure 1: Combining the neural, statistical and external features using deep MLP.

2 RELATED WORK

From an NLP perspective, researchers have studied numerous aspects of credibility of online information. For example, [4] applied the time-sensitive supervised approach by relying on the tweet content to address the credibility of a tweet in different situations. [5] used LSTM in a similar problem of early rumor detection. Other work on rumor detection includes [2, 10, 28]. In an another work, [6] aimed at detecting the stance of tweets and determining the veracity of the given rumor with convolution neural networks. A submission [3] to the SemEval 2016 Twitter Stance Detection task focuses on creating a bag-of-words auto encoder, and training it over the tokenized tweets.

In their work, [19] experimented on four basic models on which the final result was evaluated: Bag Of Words (BOW), basic LSTM, LSTM with attention and conditional encoding LSTM with attention (CEA LSTM). Similarly, The work by [20], focuses on generating lexical and similarity features using (TF-IDF) representations of BOW. Another team, [27], combined multiple models in an ensemble providing 50/50 weighted average between deep convolution neural network and a gradient-boosted decision trees. In a similar attempt, a team [1] concatenated various features vectors and passed it through an MLP model.

3 DEEP LEARNING ARCHITECTURES

To predict the stance for a given sample in FNC-1[1] dataset, a multi-channel deep neural network can be used to encode a given headline-body pair, which can be classified into one of the four stances. This is achieved by using a multi channel convolution neural network with $softmax$ layer at the output. Similarly, instead of using the convolution and pooling layers, LSTM and GRU can be used to encode the headline-body pairs. The LSTMs and GRUs encode the given sequence of words into fixed length vector representation which can be used to score the relevance of headline-body pair.

We experiment with some of the deep architectures that have been shown to be effective for non-factoid based question answering [11, 23].

4 PROPOSED IDEA

The *unrelated* headline-body pairs in the FNC-1 dataset are created by randomly assigning a news body to the given headline. This type of data augmentation has been successfully used in NLP problems such as non-factoid question answering where it results in reasonable performance by the deep learning models [16, 24]. However, in the case of FNC-1 challenge, the *agree*, *disagree*, and *discuss* headline-body pairs are relatively smaller in quantity than the *unrelated* stance. This bias leads to a uneven distribution of dataset across the four classes, with the *unrelated* category being the least interesting. Interestingness of a headline-body pair is evaluated in terms of information it contains.

The uneven distribution of FNC-1 dataset thwarts the performance of deep learning architectures introduced in Section 3. Deep learning models are dependent on a huge training corpus (few million headline-body pairs) in order to identify such nuances in patterns. The FNC-1 dataset, though the largest publicly available dataset on stance detection, does not satiate this criteria. For this reason, we introduce a much simpler strategy that consists of heavy use of feature engineering.

4.1 Neural Embeddings

We use skip-thought vectors which encodes sentences to vector embedding of length 4800 (shown in Figure 1). We follow the work of [15, 22] and compute two features from the skip-thought embeddings. These features have been shown to be effective in evaluating contextual similarity between sentences. The task of stance detection is analogous to the computation of contextual similarity between two sentences - headline and its body. We speculate that the features introduced by [15, 22] should be effective for stance detection as well. Given the skip-thought encoding of news and headline as u^{news} and v^{head}, we compute two features

[1] http://www.fakenewschallenge.org/
https://competitions.codalab.org/competitions/16843#results

Hyperparameter	Skip-thought	External Features	TF-IDF Vectors
MLP layers	2	1	2
MLP neurons	500 ; 100	50	500 ; 50
Dropout	0.2 ; -	-	0.4 ; -
Activation	sigmoid ; sigmoid	relu	relu ; relu
Regularization	L2 - 0.00000001 ; -	-	L2 - 0.00005 ; -
MLP Layers	1		
MLP neurons	4		
Activation	Softmax		
Optimizer	Adam		
Learning rate	0.001		
Batch size	100		
Loss	Cross-entropy		

Table 1: Values of hyper-parameters. The first half of the table shows the parameters used in architectures for extracting individual features. The second half shows the parameter setting of the feature combination layer that is shown in Figure 2.

$$feat_1 = u^{news}.v^{head} \quad (1)$$

$$feat_2 = |u^{news} - v^{head}| \quad (2)$$

where $feat_1$ is the component-wise product and $feat_2$ is the absolute difference between the skip-thought encoding of news and headlines. Both of these features results in a 4800 dimensional vector each.

4.2 Statistical Features

We capture the statistical information from the text to vectors with the help of BOW, TF-IDF and n-grams models. We follow the work of [20] and [8], and produce the following vectors for each headline-body pair

(1) 1-gram TF vector of the headline.
(2) 1-gram TF vector of the body.

This gives us a vector of 5000 dimension each. We concatenate both of the TF vectors and pass it to a MLP layer (as shown in Figure 2).

4.3 External Features

The external features include feature engineering heuristics such as number of similar words in the headline and body, cosine similarity between vector encodings of headline-body pairs, number of n-grams matched between the pairs, etc. We leveraged ideas for computing the external features from the baseline and add some extra features, which includes

(1) Number of characters n-grams match between the headline-body pair, where $n = 2, \cdots, 16$.
(2) Number of words n-grams match between the headline-body pair, where $n = 2, \cdots, 6$.
(3) Weighted TF-IDF score between headline and its body using the approach mentioned in [26].
(4) Sentiment difference between the headline-body pair, also termed as polarity and is computed using lexicon based approach.

All the external features adds up to a 50-dimensional feature vector and is passed to a MLP layer similar to neural and statistical features.

5 EXPERIMENTATIONS

5.1 Dataset Description

We use the dataset provided in the FNC-1 challenge which is derived from the Emergent Dataset [12], provided by the fake news challenge administrators. The former consist of 49972 tuple with each tuple consisting of a headline-body pair followed by a corresponding class label *stance* of either *agree*, *disagree*, *unrelated* or *discuss*. Word counts roughly ranges between 8 to 40 for headlines and 600 to 7000 for article body. The distribution of FNC-1 dataset is as follows: 73.13 % *unrelated* pairs, 17.82 % *discuss*, 7.36 % *agree*, and 1.68 % *disagree* pairs.

The final results are evaluated over a test dataset provided by fake news organization consisting of 25413 samples.

5.2 Baselines methods

Organizers of FNC-1 have provided a baseline model that consists of a gradient-boosting classifier over n-gram subsequences between the headline and the body along with several external features such as word overlap, occurrence of sentiment using a lexicon of highly-polarized words (like *fraud* and *hoax*). Following the work of [19], we introduce three new baselines for the FNC-1 dataset: word2vec+external features baseline, skip-thought baseline, and TF-IDF baseline. All these baselines focuses on performance of neural, statistical, and external features, when used individually.

5.3 Evaluation metrics

The FNC-1 dataset shows a heavy bias towards unrelated headkdline-body pairs. Recognizing this data bias and the simpler nature of the *related/unrelated* classification problems, the organizers of FNC-1 introduced the following weighted accuracy score as their final evaluation metric.

$$Score_{FNC} = 0.25 * Accuracy_{Unrelated} + 0.75 * Accuracy_{Agree,Disagree,Discuss} \quad (3)$$

We use the $Score_{FNC}$ as the main evaluation criteria while comparing the proposed model with other related techniques.

Method	$Score_{FNC}$	Agree	Disagree	Discuss	Unrelated	Overall
FNC-1 baseline	75.20	9.09	1.00	79.65	97.97	85.44
Word2vec + External Features	75.78	50.70	**9.61**	53.38	96.05	82.79
Skip-thought baseline	76.18	31.8	0.00	81.20	91.18	82.48
TF-IDF baseline	81.72	44.04	6.60	81.38	97.90	88.46
SOLAT in the SWEN [27]	82.05	58.50	1.86	76.18	98.70	89.08
Athene [1]	81.97	44.72	9.47	80.89	**99.25**	**89.50**
UCL Machine Reading [20]	81.72	44.04	6.60	81.38	97.90	88.46
Chips Ahoy! [21]	80.12	55.96	0.28	70.29	98.98	88.01
CNN	60.91	35.89	2.10	46.77	88.47	74.84
biLSTM	63.11	38.04	4.59	58.13	78.27	69.88
biLSTM + Attention	63.17	58.74	0.03	63.48	77.49	73.27
CNN + biLSTM	64.95	**74.09**	2.46	57.85	74.87	72.89
Proposed	**83.08**	43.82	6.31	**85.68**	98.04	89.29

Table 2: Performance of different models on FNC-1 Test Dataset. The first half of the table shows the baselines, followed by the top-4 submissions, and different architectures used in our work. Column 2-5 shows the class-wise accuracy in % while the last column shows the overall accuracy.

5.4 Results

The results on FNC-1 test dataset are shown in Table 2. The FNC-1 baseline achieves a score of 75.2 which is better than the performance of all deep architectures introduced in Section 3. The FNC-1 baseline is comprised of training gradient tree classifier on the hand crafted features (described in Section 5.2). Provided the simplicity of this baseline, it is indeed remarkable to achieve such a high score. The FNC-1 baselines achieves *approx* 7% higher class-wise accuracy on *unrelated* stance as compared to skip-thought baseline, whereas the latter receiving a higher $Score_{FNC}$. Skip-thought baselines achieves a higher accuracy on *agree* and *discuss* than the *unrelated* stance.

Since the interestingness of *agree* and *discuss* is higher than the *unrealted* stance, therefore, skip-thought achieves a higher $Score_{FNC}$. This also explains the reason for the introduction of new scoring criterion by the FNC organizers (see Section 5.3). Finally, the $Score_{FNC}$ by skip-thought, external features, and TF-IDF baselines are higher than the FNC-1 baseline. Therefore, our speculation to combine these three baselines models, is guaranteed to achieve a higher score on $Score_{FNC}$ evaluation metric. Moreover, all the baselines achieves very low or zero score on the *disagree* stance. Therefore, apart from the $Score_{FNC}$, the class-wise performance is worth considering as a performance criterion.

The performance of top-4 teams that participated in FNC-1 are shown in the middle part of Table 2, with *SOLAT in the SWEN* [27] winning the challenge achieving a score of 82.05. All the teams achieved higher score and class-wise accuracy on all stances except for the *disagree* stance. This should be a concern, since the importance of *disagree* is equivalent to the *agree* and *discuss* stance. We observed that the news pairs in the *disagree* category are not only very few, but also consists of divergent news articles. This is one of the reason for poor performance of most of the deep models, including the top teams, on identifying *diagree* stance.

From Table 2, it is evident that the overall accuracy achieved by the proposed model is slightly lower than [1], although the proposed model outperformed all the other techniques by a clear margin (in terms of $Score_{FNC}$). The possible reason for this deviation is that the [1] gives more focus to the classification of *unrelated* stances rather than the rest, which is the reason for highest overall accuracy. Since *unrelated* stances are of least interest to us, this results in lower $Score_{FNC}$.

Finally, a confusion matrix is given in Table 3 that provides in-detail analysis of the performance of our approach.

	Agree	Disagree	Discuss	Unrelated	Overall
Agree	834	15	945	109	43.82
Disagree	208	44	328	117	6.31
Discuss	401	23	3825	215	85.68
Unrelated	22	12	325	17990	98.04

Table 3: Confusion matrix for proposed model on test data.

6 CONCLUSION

In this paper, we explored the benefit of incorporating neural, statistical and external features to deep neural networks on the task of fake news stance detection. The presented idea leverages features extracted using skip-thought embeddings, n-gram TF-vectors and several introduced hand crafted features.

We found that the uneven distribution of FNC-1 dataset undermines the performance of most deep learning architectures. The fewer training samples adds further to this aggravation. Furthermore, the introduced scoring function doesn't help in a fair evaluation. Creating a dataset for a complex NLP problems such as fake news identification is indeed a cumbersome task, and we appreciate the work by the FNC organizers, yet, a more detailed and elaborate dataset along with well defined scoring criterion should make this challenge more suitable to evaluate.

REFERENCES

[1] Benjamin Schiller Andreas Hanselowski, Avinesh PVS and Felix Caspelherr. 2017. Athenefnc. https://github.com/hanselowski/athene_system. (2017).
[2] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464* (2016).
[3] Isabelle Augenstein, Andreas Vlachos, and Kalina Bontcheva. 2016. USFD at SemEval-2016 Task 6: Any-Target Stance Detection on Twitter with Autoencoders.. In *SemEval@ NAACL-HLT*. 389–393.
[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. *Internet Research* 23, 5 (2013), 560–588.
[5] Tong Chen, Lin Wu, Xue Li, Jun Zhang, Hongzhi Yin, and Yang Wang. 2017. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection. *arXiv preprint arXiv:1704.05973* (2017).
[6] Yi-Chin Chen, Zhao-Yand Liu, and Hung-Yu Kao. 2017. IKM at SemEval-2017 Task 8: Convolutional Neural Networks for Stance Detection and Rumor Verification. *Proceedings of SemEval. ACL* (2017).
[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
[8] Richard Davis and Chris Proctor. 2017. Fake News, Real Consequences: Recruiting Neural Networks for the Fight Against Fake News. https://web.stanford.edu/class/cs224n/reports/2761239. (2017).
[9] Delip Rao Dean Pomerleau. 2017. Fake News Challenge. http://www.fakenewschallenge.org/. (2017).
[10] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972* (2017).
[11] Minwei Feng, Bing Xiang, Michael R Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 813–820.
[12] William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL.
[13] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5 (2005), 602–610.
[14] Hua He, Kevin Gimpel, and Jimmy J Lin. 2015. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks.. In *EMNLP*. 1576–1586.
[15] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. 3294–3302.
[16] Todor Mihaylov and Preslav Nakov. 2016. SemanticZ at SemEval-2016 Task 3: Ranking Relevant Answers in Community Question Answering Using Semantic Similarity Based on Fine-tuned Word Embeddings.. In *SemEval@ NAACL-HLT*. 879–886.
[17] Paul Neculoiu, Maarten Versteegh, Mihai Rotaru, and Textkernel BV Amsterdam. 2016. Learning Text Similarity with Siamese Recurrent Networks. *ACL 2016* (2016), 148.
[18] NYTimes. 2016. As fake news spreads lies, more readers shrug at the truth. https://www.nytimes.com/2016/12/06/us/fake-news-partisan-republican-democrat.html. (2016).
[19] Stephen Pfohl, Oskar Triebe, and Ferdinand Legros. 2017. Stance Detection for the Fake News Challenge with Attention and Conditional Encoding. (2017).
[20] Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *arXiv preprint arXiv:1707.03264* (2017).
[21] Jingbo Shang. 2017. Chips ahoy! at Fake News Challenge. https://github.com/shangjingbo1226/fnc-1. (2017).
[22] Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075* (2015).
[23] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108* (2015).
[24] Liu Yang, Qingyao Ai, Damiano Spina, Ruey-Cheng Chen, Liang Pang, W Bruce Croft, Jiafeng Guo, and Falk Scholer. 2016. Beyond factoid QA: Effective methods for non-factoid answer sentence retrieval. In *European Conference on Information Retrieval*. Springer, 115–128.
[25] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering.. In *EMNLP*. 2013–2018.
[26] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632* (2014).
[27] Sean Baird Yuxi Pan, Doug Sibley. 2017. Talos. http://blog.talosintelligence.com/2017/06/. (2017).
[28] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management* 54, 2 (2018), 273–290.