A Deep Learning Approach for Automatic Detection of Fake News

Tanik Saikh[1]   **Arkadipta De**[2]   **Asif Ekbal**[1]   **Pushpak Bhattacharyya**[1]

Department of Computer Science and Engineering[1]
Indian Institute of Technology Patna[1]
Department of Computer Science and Engineering[2]
Government College of Engineering and Textile Technology, Berhampore[2]
{tanik.srf17, asif, pb}@iitp.ac.in[1]
de.arkadipta05@gmail.com[2]

Abstract

Fake news detection is a very prominent and essential task in the field of journalism. This challenging problem is seen so far in the field of politics, but it could be even more challenging when it is to be determined in the multi-domain platform. In this paper, we propose two effective models based on deep learning for solving fake news detection problem in online news contents of multiple domains. We evaluate our techniques on the two recently released datasets, namely *Fake-News AMT and Celebrity* for fake news detection. The proposed systems yield encouraging performance, outperforming the current hand-crafted feature engineering based state-of-the-art system with a significant margin of 3.08% and 9.3% by the two models, respectively. In order to exploit the datasets, available for the related tasks, we perform cross-domain analysis (i.e. model trained on FakeNews AMT and tested on Celebrity and vice versa) to explore the applicability of our systems across the domains.

1 Introduction

In the emergence of social and news media, data are constantly being created day by day. The data so generated are enormous in amount, and often contains miss-information. Hence it is necessary to check it's truthfulness. Nowadays people mostly rely on social media and many other online news feeds as their only platforms for news consumption (Jeffrey and Elisa, 2016). A survey from the *Consumer News and Business Channel (CNBC)* also reveals that more people are rely on social media for news consumption rather than news paper [1]. Therefore, in order to deliver the genuine news to such consumers, checking the truthfulness of such online news content is of utmost priority to news industries. The task is very difficult for a machine as even human being can not understand news article's veracity (easily) after reading the article.

Prior works on fake news detection entirely rely on the datasets having satirical news contents sources, namely *"The Onion" (Rubin et al., 2016)*, fact checking website like *Politi-Fact (Wang, 2017), and Snopes (Popat et al., 2016)*, and on the contents of the websites which track viral news such as *BuzzFeed (Potthast et al., 2018)* etc. But these sources have severe drawbacks and multiple challenges too. Satirical news mimic the real news which are having the mixture of irony and absurdity. Most of the works in fake news detection fall in this line and confine in one domain (i.e. politics). The task could be even more challenging and generic if we study this fake news detection problem in multiple domain scenarios. We endeavour to mitigate this particular problem of fake news detection in multiple domains. This task is even more challenging compared to the situation when news is taken only from a particular domain, i.e. uni-domain platform. We make use of the dataset which contained news contents from multiple domains. The problem definition would be as follows:

Given a *News Topic* along with the corresponding *News Body Document*, the task is to classify whether the given news is *legitimate/genuine* or *Fake*. The work described in Pérez-Rosas et al. (2018) followed this path. They also offered two novel computational resources, namely *FakeNews AMT* and *Celebrity news*. These datasets are having triples of *topic, document and label (Legit/Fake)* from multiple domains (like *Business, Education, Technology, Entertainment and Sports* etc) including *politics*. Also, they claimed that these datasets focus on the deceptive properties of

---

[1]https://www.cnbc.com/2018/12/10/social-media-more-popular-than-newspapers-for-news-pew.html

online articles from different domains. They provided a baseline model. The model is based on Support Vector Machine (SVM) that exploits the hand-crafted linguistics features. The SVM based model achieved the accuracies of 74% and 76% in the FakeNews AMT and Celebrity news datasets, respectively. We pose this problem as a classification problem. So the proposed predictive models are binary classification systems which aim to classify between fake and the verified content of online news from multiple domains. We solve the problem of multi-domain fake news detection using two variations of deep learning approaches. The first model (denoted as Model 1) is a Bi-directional Gated Recurrent Unit (BiGRU) based deep neural network model, whereas the second model (i.e. Model 2) is *Embedding from Language Model (ELMo)* based. It is to be noted that the use of deep learning to solve this problem in this particular setting is, in itself, very new. The technique, particularly the word attention mechanism, has not been tried for solving such a problem. Existing prior works for this problem mostly employ the methods that make use of handcrafted features. The proposed systems do not depend on hand crafted feature engineering or a sophisticated NLP pipeline, rather it is an end to end deep neural network architecture. Both the models outperform the state-of-the-art system.

2 Related Work

A sufficient number of works could be found in the literature in fake news detection. Nowadays the detection of fake news is a hot area of research and gained much more research interest among the researchers. We could detect fake news at two levels, namely the conceptual level and operational level. Rubin et al. (2015) defined that conceptually there are three types of fake news: *viz i. Serious Fabrications ii. Hoaxes and iii. Satire*. The work of Conroy et al. (2015) fostered linguistics and fact checking based approaches to distinguish between real and fake news, which could be considered as the work at conceptual level. Chen et al. (2015) described that fact-checking approach is a verification of hypothesis made in a news article to judge the truthfulness of a claim. Thorne et al. (2018) introduced a novel dataset for fact-checking and verification where evidence is large Wikipedia corpus. Few notable works which made use of text as evidence can be found in (Ferreira and Vlachos,

2016; Nie et al., 2018).

The Fake News Challenge [2] organized a competition to explore, how artificial intelligence technologies could be fostered to combat fake news. Almost 50 participants were participated and submitted their systems. Hanselowski et al. (2018) performed retrospective analysis of the three best participating systems of the Fake News Challenge. The work of Saikh et al. (2019) detected fake news through stance detection and also correlated this stance classification problem with Textual Entailment (TE). They tackled this problem using statistical machine learning and deep learning approaches separately and with combination of both of these. This system achieved the state of the art result.

Another remarkable work in this line is the verification of a human- generated claim given the whole Wikipedia as evidence. The dataset, namely (*Fact Extraction and Verification (FEVER)*) proposed by Thorne et al. (2018) served this purpose. Few notable works in this line could be found in (Yin and Roth, 2018; Nie et al., 2019).

3 Proposed Methods

We propose two deep Learning based models to address the problem of fake information detection in the multi-domain platform. In the following subsections, we will discuss the methods.

3.1 Model 1

This model comprises of multiple layers as shown in the Figure 1. The layers are *A. Embedding Layer B. Encoding Layer (Bi-GRU) C. Word level Attention D. Multi-layer Perceptron (MLP)*.

A. Embedding Layer: The embedding of each word is obtained using pre-trained fastText model[3](Bojanowski et al., 2017). FastText embedding model is an extended version of Word2Vec (Mikolov et al., 2013). Word2Vec (predicts embedding of a word based on given context and vice-versa) and Glove (exploits count and word co-occurrence matrix to predict embedding of a word) (Pennington et al., 2014) both treat each word as an atomic entity. The fastText model produces embedding of each word by combining the embedding of each character n-gram of that word. The model works better on rare words and also produces embedding for

---

[2]http://www.fakenewschallenge.org/
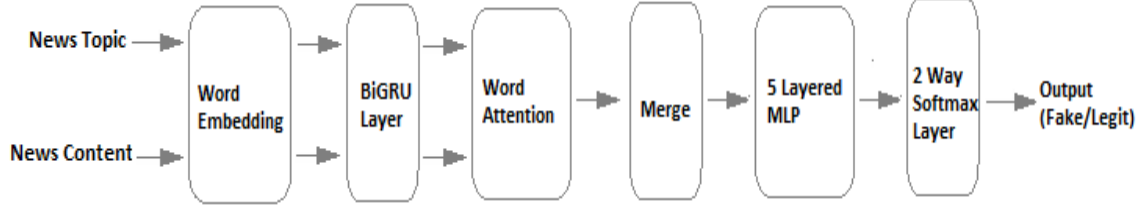[3]https://fasttext.cc/

231

Figure 1: Architectural Diagram of the Proposed First System

out-of-vocabulary words, where Word2Vec and Golve both fail. In the multi-domain scenario vocabularies are from different domains and there is a high chance of existing different domain specific vocabularies. This is the reason for choosing the fastText word vector method.

B. Encoding Layer: The representation of each word is further given to a bidirectional Gated Recurrent Units (GRUs) (Cho et al., 2014) model. GRU takes less parameter and resources compared to Long Short Term Memory (LSTM), training also is computationally efficient. The working principles of GRU obey the following equations:

$$z = \alpha(x_t U^z + s_{t-1} W^z) \qquad (1)$$

$$r = \alpha(x_t U^r + s_{t-1} W^r) \qquad (2)$$

$$h = tanh(x_t U^h + r_t \cdot s_{t-1} W^r) \qquad (3)$$

$$r = (1 - z) \cdot h + z \cdot s_{t-1} \qquad (4)$$

In equation 1, z is the update gate at time step t. This z is the summation of the multiplications of x_t with it's own weight U(z) and s_{t-1} (holds the information of previous state) with it's own W(z). A sigmoid α is applied on the summation to squeeze the result between 0 and 1. The task of this update gate (z) is to help the model to estimate how much of the previous information (from previous time steps) needs to be passed along to the future. In the equation 2, r is the reset gate, which is responsible for taking the decision of how much past information to forget. The calculation is same as the equation 1. The differences are in the weight and gate usages. The equation 3 performs as follows, i. multiply input x_t with a weight U and s_{t-1} with a weight W. ii. Compute the element wise product between reset gate r_t and s_{t-1}W. Then a non-linear activation function *tanh* is applied to the summation of i and ii. Finally, in the equation 4, we compute r which holds the information of the current unit. The computation procedure is
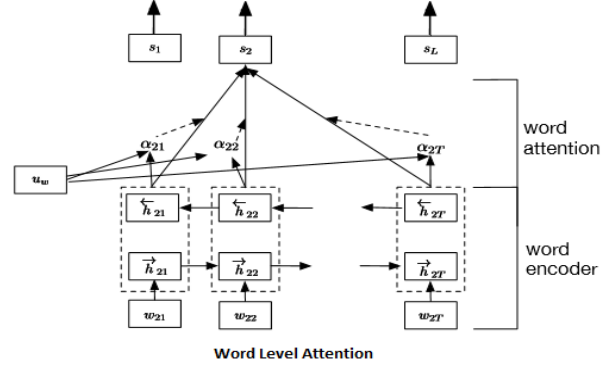

Figure 2: Word Level Attention Network

as follows: i. compute element-wise multiplication to the update gate z_t and $s_{(t-1)}$. ii. calculate element-wise multiplication to (1-z) with h. Take the summation of i and ii.

The bidirectional GRUs consists of the forward GRU, which reads the sentence from the first word (w_1) to the last word (w_L) and the backward GRU, that reads in reverse direction. We concatenate the representation of each word obtained from both the passes.

C. Word Level Attention: We apply the attention model at word level (Bahdanau et al., 2015; Xu et al., 2015). The objective is to let the model decide which words are importance compared to other words while predicting the target class (fake/legit). We apply this as applied in Yang et al. (2016). The diagram is shown in the Figure 2. We take the aggregation of those words' representation which are multiplied with attention weight to get sentence representation. We do this process for both the news topic and the corresponding document. This particular technique of the word attention mechanism, has not been tried for solving such a problem.

$$U_{i_t} = tanh(W_w h_{i_t} + b_w) \qquad (5)$$

232

$$\alpha_{i_t} = \frac{exp(u_{i_t}^T u_w)}{\sum_t exp(u_{i_t}^T u_w)} \qquad (6)$$

$$s_i = \sum_t \alpha_{i_t} h_{i_t} \qquad (7)$$

First get the word annotation h_{i_t} through GRU output and compute u_{i_t} as a hidden representation of h_{i_t} in 5. We measure the importance of the word as the similarity of u_{i_t} with a word level context vector u_w and get a normalized importance weight α_{i_t} through a softmax in 6. After that, in 7, we compute the sentence vector s_i as a weighted sum of the word annotations based on the weights α_{i_t}. The word context vector u_w is randomly initialized and jointly learned during the training process.

D. Multi-Layer Perceptron: We concatenate the sentence vector obtained for both the inputs. The obtained vector further fed into fully connected layers. We use 512, 256, 128, 50 and 10 neurons, respectively, for five such layers with ReLU (Glorot et al., 2011) activation in each layer. Between each such layer, we employ 20% dropout (Srivastava et al., 2014) as a measurement of regularization. Finally, the output from the last fully connected layer is fed into a final classification layer with softmax (Duan et al., 2003) activation function having 2 neurons. We use Adam (Kingma and Ba, 2014) optimizer for optimization.

3.2 Model 2

We propose another approach whose embedding layer is based on *Embedding for Language Model (ELMo)* (Peters et al., 2018) and the MLP Network, which is same as we applied in Model 1. The diagram of this model is shown in the Figure 3.

Embedding Layer: Embedding from Language Model (ELMo) has several advantages over the other word vector methods, and found to be a good performer in many challenging NLP problems. It has key features like i. Contextual i.e. representation of each word is based on entire corpus in which it is used ii. Deep i.e. it combines all layers of a deep pre-trained neural network and iii. Character based i.e. it provides representations which are based on character, thus allowing the network to make use of morphological clues to form robust representation of out-of-vocabulary tokens during training. The ELMO embedding is very efficient in capturing context. The multi-domain datasets are having different vocabularies and contexts, so

Dataset	# of Examples	Avg.words/sent	Words	Label
FakeNewsAMT	240	132/5	31,990	Fake
	240	139/5	33,378	Legit
Celebrity	250	399/17	39,440	Fake
	250	700/33	70,975	Legit

Table 1: Class Distribution and Word Statistics for Fake News AMT and Celebrity Datasets. Avg: Average, sent: Sentence

we make use of such a word vector representation method to capture the context. News topics and corresponding documents are given to Elmo Embedding model. This embedding layer produces the representation for news topic and news content.

After getting the embedding of the topic and the context, we merge them. The merged vector is fed into a five layers MLP (same as the previous model). Finally, we classify with a final layer having softmax activation function.

4 Experiments, Results and Discussion and Comparison with State-of-the-Art

Overall we perform four sets of experiments. In the following sub-sections we describe and analyze them one by one after the description of the datasets used.

Data: Prior datasets and focus of research for fake information detection are on political domain. As our research focus is on multiple domains, we foster the dataset released by Pérez-Rosas et al. (2018). They released two novel datasets, namely *FakeNews AMT* and *Celebrity*. The Fake News AMT is collected via crowdsourcing (Amazon Mechanical Turk (AMT)) which covers news of six domains (i.e. Technology, Education, Business, Sports, Politics, and Entertainment). The Celebrity dataset is crawled directly from the web of celebrity gossips. It covers celebrity news. The AMT manually generated fake version of a news based on the real news. We extract the data domain wise to get the statistics of the dataset. It is observed that each domain contains equal number of instances (i.e. 80). The class distribution among each domain is also evenly distributed. The statistics of these two datasets is shown in the following Table 1.

The news of the Fake News AMT dataset was obtained from a variety of mainstream news websites predominantly in the United States such as the ABCNews, CNN, USAToday, New York
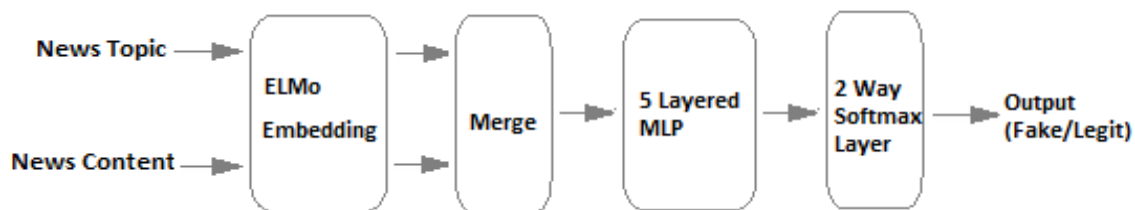
233

Figure 3: Architectural Diagram of the Proposed Second Model

Dataset	System	Model	Test Accuracy(%)
FakeNews AMT	Proposed	Model1	77.08
		Model2	83.3
	(Pérez-Rosas et al., 2018)	Linear SVM	74
Celebrity	Proposed	Model1	76.53
		Model2	79
	(Pérez-Rosas et al., 2018)	Linear SVM	76

Table 2: Classification Results for the FakeNews AMT and Celebrity News Dataset with Two Proposed Methods and Comparison with Previous Results

Training	Testing	Accuracy(%)
FakeNewsAMT	Celebrity	54.3
Celebrity	FakeNewsAMT	68.5

Table 3: Results Obtained in Cross-Domain Analysis Experiments on the Best Performing System.

Times, FoxNews, Bloomberg, and CNET among others.

Multi-Domain Analysis: In this section, we do experiments on whole Fake News AMT and Celebrity datasets individually. We train our models on the whole Fake News AMT and Celebrity dataset and test on the respective test set. As the datasets is evenly distributed between real and fake news item, a random baseline of 50% could be assumed as reference. The results obtained by the two proposed methods outperform the baseline and the results of Pérez-Rosas et al. (2018). The results obtained and comparisons are shown in the Table 2. Our results indicate this task could be efficiently handled using deep learning approach.

Cross-Domain Analyses: We perform another set of experiment to study the usefulness of the best performing system (i.e. Model2) across the domains. We train the model2, on FakeNews AMT and test on Celebrity and vice-versa. The results are shown in the Table 3. If we compare with the in domain results it is observed that there is a significant drop. This drop also observed in the work of Pérez-Rosas et al. (2018) in machine learning setting. This indicates there is a significant role of a domain in fake news detection, as it is established by our deep learning guided experiments too.

Multi-Domain Training and Domain-wise Test-

ing: There are very small number of examples pairs in each sub-domain (i.e. Business, Technology etc) in FakeNews AMT dataset. We combine the examples pairs of multiple domains/genres for cross corpus utilization. We train our proposed models on the combined dataset of five out of six available domains and test on the remaining one. This has been performed to see how the model which is trained on heterogeneous data react on the domain to which the model was not exposed at the time of training. The results are shown in *Exp. a* part of the Table 4. Both the models yield the best accuracy in the *Education* domain, which indicates this domain is open i.e. linguistics properties, vocabularies of this domain are quite similar to other domains. The models (i.e. Model 1 and 2) perform worst in the *Entertainment* and the *Sports*, respectively, which indicate these two domains are diverse in nature from the others in terms of linguistics properties, writing style, vocabularies etc.

Domain-wise Training and Domain-wise Testing: We also eager to see in-domain effect of our systems. The FakeNews AMT dataset comprises of six separate domains. We train and test our models, on each domain's dataset of Fake News AMT. This evaluates our model's performance domain-wise. The results of this experiment are shown in the *Exp. b* part of the Table 4. In this case both the models produce the highest accuracy in the *Sports* domain, followed by the *Entertainment*, as we have shown in our previous experiments that these two domains are diverse in nature from the others. This fact is established by this experiment too. Both the models produce the lowest result in the *Technology* and the *Business* domain, respectively.

Visualization of Word Level Attention: We take the visualization of the topic and the corresponding document at word level attention as shown in the Figure 4 and 5, respectively. The aim is to visualize the words which are assigned more weights during the prediction of the output class.

Domain	Exp. a		Exp. b	
	Model1	Model2	Model1	Model2
Business	74.75	78.75	63.56	68.56
Education	77.25	91.25	65.65	70.65
Technology	76.22	88.75	64.3	65.35
Politics	73.75	88.75	64.27	69.22
Entertainment	68.25	76.25	65.89	71.2
Sports	70.75	73.75	67.86	71.45

Table 4: Result of Exp. a (Trained on Multi-domain Data and Tested on Domain wise Data) and Exp. b (Trained on Domain wise Data and Tested on Domain wise Data)

Figure 4: Word Level Attention on News Topic

In these Figures, words with more deeper colour indicate that they are getting more attention. We can observe, the words *secretary*, *education* in 4 and *President*, *Donald* in 5 are the words having deeper colour, i.e. these words are getting more weight compared to others. These words are Named Entities (NEs). It could be concluded that NEs phrases are important in fake news detection in multi domain setting.

4.1 Error Analysis

We extract the mis-classified and also the truly classified instances produced by the best performing system. We perform a rigorous analysis of these instances and try to find out the pattern in the mis-classified instances and the linguistics differences between those two categories of instances. It is found that the model fails mostly in the *Entertainment* followed by the *sports* and the *Business* domain etc. To name a few, we are showing such examples which are actually "Legitimate", but predicted as "Fake" the Table 5 and which are actually "Fake", but predicted as "Legitimate" the Table 6. It is observed that both the topic and document are having ample number of NEs. It needs further investigation in this font.

5 Conclusion and Future work

In this article, we propose two deep learning based approaches to mitigate the problem of fake news detection from multiple domains platform. Antecedent works in this line pay attention on satir-

Figure 5: Word level Attention on News Document, A Part of it is Shown Due to Space Constraint.

ical news or made use of the content of the fact-checking websites, which was restricted to one domain (i.e. politics). To address these limitations, we focus to extend this problem into multi domain scenario. Our work extends the concept of fake news detection from uni-domain to multi-domain, thus making it more general and realistic. We evaluate our proposed models on the datasets whose contents are from multiple domains. Our two proposed approaches outperform the existing models with a notable margin. Experiments also reveal that there is a vital role of a domain in context of fake news detection. We would like to do more deeper analysis of the role of domain for this problem in future. Apart from this our future line of research would be as follows:

- It would be interesting for this work to encode the domain information in the Deep Neural Nets.

- BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019) embedding based model and make a comparison with fastText and ELMo based models in the context of fake news detection.

- Use of transfer learning and injection external knowledge for better understanding.

- Handling of Named Entities efficiently and incorporate their embedding with the normal phrases.

- Using WordNet to retrieve connections between words on the basis of semantics in the news corpora (both topic and document of news) which may influence in detection of Fake News.

6 Acknowledgement

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representation (ICLR)*.

Domain	Topic	Content
Entertainment	Chris Pratt responds to body shamers telling him he's too thin	Big or small Chris Pratt has heard it all. These days the "Guardians of the Galaxy" star 37 is taking flak for being too thin but he's not taking it lying down. Pratt who has been documenting the healthy snacks he's eating while filming "Jurassic World 2" in a series of "What's My Snack" Instagram videos fired back – in his usual tongue-in-cheek manner – after some followers apparently suggested he looked too thin. "So many people have said I look too thin in my recent episodes of #WHATSMYSNACK he wrote on Instagram Thursday. Some have gone as far as to say I look 'skeletal.' Well just because I am a male doesn't mean I'm impervious to your whispers. Body shaming hurts."
Business	Banks and Tech Firms Battle Over Something Akin to Gold: Your Data	The big banks and Silicon Valley are waging an escalating battle over your personal financial data: your dinner bill last night your monthly mortgage payment the interest rates you pay. Technology companies like Mint and Betterment have been eager to slurp up this data mainly by building services that let people link all their various bank-account and credit-card information. The selling point is to make budgeting and bookkeeping easier. But the data is also being used to offer new kinds of loans and investment products. Now banks have decided they aren't letting the data go without a fight. In recent weeks several large banks have been pushing to restrict the sharing of this kind of data with technology companies according to the tech firms. In some cases they are refusing to pass along information like the fees and interest rates they charge. Both sides see big money to be made from the reams of highly personal information created by financial transactions.

Table 5: Examples of mis-classified instances from Entertainment and Business domain. Examples are originally "Legitimate" but predicted as "Fake".

Domain	Topic	Content
Sports	Slaven Bilic still has no support of West Ham's owners	"West Ham's owners have no faith in manager Slaven Bilic as his team won only six of their 11 games this year according to Sky sources. Bilic's contract runs out in the summer of 2018 and results have made it likely that he will not be offered a new deal this summer. Co-chairman David Sullivan told supporters 10 days ago after West Ham lost 3-2 at home to Leicester City. Sullivan said that even if performances and results improved in the next three games against Hull City Arsenal and Swansea City. West Ham's owners have a track record of being unloyal to their managers who don't meet their specs and there is a acceptance at boardroom level that Bilic has failed to prove a solid season."
Education	STEM Students Create Winning Invention	STREAMWOOD, Ill. (AP) – A group of Streamwood High School students have created an invention that is exciting homeowners everywhere - and worrying electricity companies at the same time. The kids competed in the Samsung Solve for Tomorrow contest, entering and winning with a new solar panel that costs about $100 but can power an entire home - no roof takeover needed! The contest won the state-level competition which encourages teachers and students to solve real-world issues using science and math skills; the 16 studens will now compete in a national competition and, if successful, could win a prize of up to $200,000.

Table 6: Examples of mis-classified instances from the Sports and Education domain. Examples are originally "Fake" but predicted as "Legitimate".

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. News in an Online World: The Need for an automatic crap detector. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic Deception Detection: Methods for Finding Fake News. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaibo Duan, S. Sathiya Keerthi, Wei Chu, Shirish Krishnaj Shevade, and Aun Neow Poo. 2003. Multi-Category Classification by Soft-max Combination of Binary Classifiers. In *Proceedings of the 4th International Conference on Multiple Classifier Systems*, MCS'03, pages 125–134, Berlin, Heidelberg. Springer-Verlag.

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*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gottfried Jeffrey and Shearer Elisa. 2016. News use Across Social Media Platforms 2016. In *In Pew Research Center Reports*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2018. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. *arXiv preprint arXiv:1811.07039*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6859–6866.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility Assessment of Textual Claims on the Web. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 2173–2178.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake News or Truth? using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. 2015. Deception Detection for News: Three Types of Fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, pages 83:1–83:4, Silver Springs, MD, USA. American Society for Information Science.

Tanik Saikh, Amit Anand, Asif Ekbal, and Pushpak Bhattacharyya. 2019. A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features. In *International Conference on Applications of Natural Language to Information Systems*, pages 345–358, Salford, UK. Springer.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

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, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. 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 (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural Image Caption Generation with Visual Attention. In *International conference on machine learning*, pages 2048–2057.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium. Association for Computational Linguistics.