# Bidirectional LSTM Based on POS tags and CNN Architecture for Fake News Detection

*

Manoj Kumar Balwant
*Dept. of Computer Science*
*Uttar Pradesh Rajarshi Tandon Open University*
Prayagraj, India
balwantuoh89@gmail.com

***Abstract*—Fake news generally on social media spreads very quickly and this brings many serious consequences. Traditional lexico-syntactic based features have limited success to detect fake news. Majority of fake news detection techniques are tested on small dataset containing limited training examples. In this work, we evaluate our architecture on Liar-Liar dataset which contain 12836 short news from different sources including social media. The proposed architecture incorporates POS (part of speech) tags information of news article through Bidirectional LSTM and speaker profile information through Convolutional Neural Network. The results show that the resulting hybrid architecture significantly improves detection performance of Fake news on Liar Dataset.**

***Index Terms*—Fake News, Long Short Term Memory, POS tags, Convolutional Neural Network, Liar Liar Dataset**

## I. Introduction

Fake news is intentionally and verifiably false news articles that mislead readers [1]. This is a narrow definition of fake news which contains two key features: authenticity and intent. First, fake news contains false information that can be verified. Second, fake news is created to mislead consumers. A Broader definition of fake news focuses on the either authenticity or intent of the news content. Some research papers consider satire news as fake news whose contents are false even though they are entertainment-oriented [2], [3]. While, other literature directly treats deceptive news that contains serious fabrications and hoaxes as fake news [4]. In this research, we use the narrow definition of fake news: a news article that is intentionally and verifiably false. This definition eliminates ambiguities between fake news and other related concepts such as satire news, rumours, conspiracy theories which, misinformation that is created unintentionally and hoaxes that are only motivated by fun or to scam targeted individuals.

Fake news exists from long time since news began to circulate through printing press in 1439. Fake news generally spreads by traditional print, broadcast Media and online social media. These news are often due to unethical practices which involve paying reporters for stories. Fake news often fabricates headlines to increase its readership and to mislead people. Many times, news stories are over exaggerated and sensationalized to spread misleading and frightening rumours. Most often, It is propagated to damage a person or a agency for financial or political gain. Journalism is often driven by advertising revenue which depends on circulations and subscriptions for their news articles. Nowadays, online advertising revenue becomes main source of income for many websites. Publishing false stories attract viewers which benefits advertisers and increase their popularity.

In the early 90s, when the Internet was opened to public, it was just a mean to access information. Over the time, Internet grows very rapidly with tons of unmanageable information coming daily. It also allows anyone to easily spread unwanted and misleading information. A 2018 study at Oxford University has found that during 2016 U.S. presidential election, Trump's supporters consume largest volume of fake news on Twitter and Facebook. "Just how partisan is Facebook's fake news? We tested it" reported that there are higher sharing of fake news than legitimate news on Facebook which may be possible because fake news are often fit to expectations and are more interesting than legitimate news. According to a 2017 BuzzFeed article [5], a fake story about a rape festival in India, has helped to generate over $ 250,000 to a non-profit organization GiveIndia.

Nowadays, people often read from various social media like Facebook, Twitter, WhatsApps, Instagram and many more rather than publication. These news are generally diverted from their context. This results in spreading of misinformation very quickly and brings many serious consequences. For example, a 2008 hoax claiming that Apple CEO Steve Jobs had suffered a serious heart attack which led to the companys stock price falling by 10 % [6]. On November 8, 2016, Indian government introduced a 2,000 rupees currency note in an effort of demonization of 500, 1000 rupees notes. Fake news went viral on various social media including WhatsApp that the new 2000 rupees note will come with spying technology containing nano GPS chip that would be capable of tracking note 120 meter below earth. However, later finance minister has dismissed such rumours [7].

Usually, the news from Facebook, Twitter trends to be short which challenge the current fake news detection system. Even, it is difficult for human beings to differentiate between areal

news and a fake news. Human beings are able to identify real and fake news with only 50-60% success rates [8]. So, developing automated fake news detection system is extremely important.

## II. Related Work

Part-of-speech (POS) tagging is a fundamental step in most Natural Language Processing. It assigns each word in a text a proper POS tag. The Part-of-speech (nouns, verbs, adjectives, adverbs, etc.) gives a number of informations about a word such as syntactic categories of words (nouns, verbs, adjectives, adverbs, etc.) and similarities and dissimilarities etc. [9] experimented with a large corpus of Twitter data based on POS Tags and observes differences in distribution of positive, negative and neutral texts. Some POS tags are strong indicator of emotional text. They reports that neutral text often contains more proper and common noun (NNS,NP,NNS), while positive or negative text tends to contain personal pronoun (PP,PP$). Positive or negative texts in the corpus usually contain first person author or second person audience (VBP). Theses texts trends to contain more simple past tense (VBD) instead of past participle(VBN). They also contain base form of verb with frequent use of Modal verb (MD). While, neutral text often contains verb in third person (VBZ). Author also notices that, Superlatives adjectives (JJS) are frequently used to express emotions and opinions while, comparative adjectives (JJR) are used to provide information or stating facts. Positive text contains superlative adverb (RBS) like most, best. Negative text more often contains verbs in past tense(VBN,VBD) to express some loss or disappointment. Order sensitive sequential model Recurrent Neural Network is capable of processing sequential input of arbitrary length that makes it a natural choice for such tasks.

These days, CNNs have become a modern standard baseline model similar to SVM and logistic regression. Convolutional neural networks (CNN ) have shown remarkable performance in many computer vision tasks including visual question answers, image classification, image captioning. In natural language processing, CNNs have also shown impressive results on sentence classification [10], [11], [12]. Much of the work with deep learning have involved CNN on top of word embeddings. Word Embeddings are the vector representation of words where each words is represented with a vector of real numbers. Each real number in the vector may correspond to different features which allows rich feature representation of a word. [13] propose Recurrent Neural Network based model that combines text, response, and source characteristics of news for a more accurate prediction. [10] has achieved state of art results on several datasets with a simple CNNs model with one layer only. William Yang Wang proposed a hybrid model based on [10] cnn model and LSTM for incorporating both text news and metadata information(including profile information). They used pre-trained word embedding word 2vec from Google news [14] for text embedding. [15] experimented with different set of features in Twitter data, and claims that prior polarity of part of speech tags are most important features in sentiment analysis.

Recently, RNN based on lstm model emerges as a popular architecture due to its representational order and ability to capture long range dependence. It has been applied to a variety of sequence modelling, classification task, machine translation etc. Some of the notable achievements include machine translation [16], [17], speech recognition [18], image caption generation [19], visual question answering, handwriting recognition [20] and many more. [21] have proposed lexico-syntactic features driven from context free Grammar (CFG) parse tree. [22] Proposed deception detection on short text using SVM on different set of features including POS and production rules derived from Probabilistic Context Free Grammar (PCFG) trees. In addition to linguistic features derived from content of news articles, additional context features are also useful in determining credibility of any news [1], [23]. Contextual features can be the speaker profile informations such as party affiliation, location of speech, job title, credit history as well as topic are also useful to determine credibility of any news. [24] has used speaker profile information to design a new benchmark dataset for fake news detection and purposes a hybrid cnn model which uses speaker profile informations along with the text news. [25] also uses speaker profile informations and proposes an attention based LSTM model. In this work, we will present a hybrid architecture based on Bidirectional LSTM and Convolutional Neural Network to incorporate both news content and user profile information. A overview of the proposed architecture is shown in Figure 1

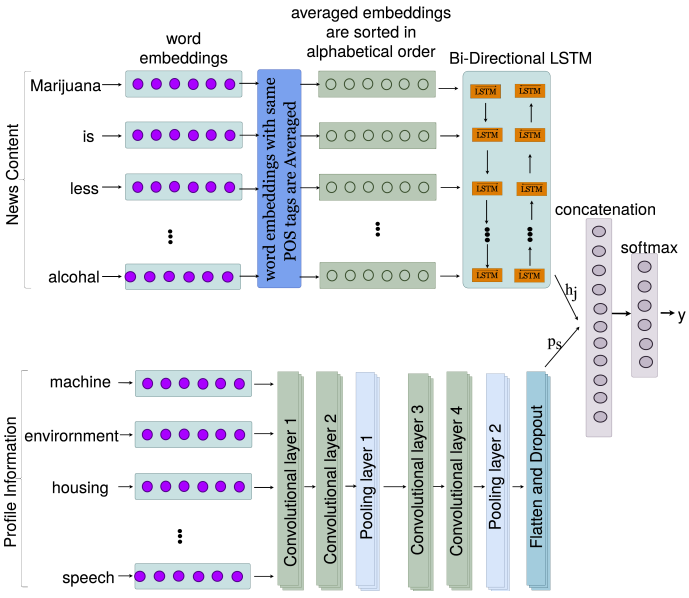


Fig. 1. An overview of the proposed architecture

## III. Long Short-Term Memory Networks

A recurrent neural network (RNN) processes an input sequence by updating a same weight matrix through a recursive

function call. The recursive function call at each time step takes current input and previous hidden state. At each time step t, the RNN hidden state $h_t$ is a function of current input $x_t$ and its previous hidden state $h_{t-1}$. The input $x_t$ can be a vector representation of $t^{th}$ word in the input sequence. The hidden state $h_t \in \mathbb{R}^d$, is representation of sequence of words up to time t. Usually, the RNN transition function is a affine transformation followed by a non-linearity function.

$$h_t = tanh(Wx_t + Uh_{t-1} + b)$$

Unfortunately, a major problem with RNN is that during training the gradient grows or decays exponentially. This is exploding or vanishing gradient problem which is reported by [26], [27]. This makes RNN difficult to remember long range dependence in an input text. This problem is addressed by LSTM architecture [28] by introducing a memory cell that preserves state over long time step. We can define the LSTM as a collection of vectors $\in \mathbb{R}^d$ at reach time step: an input gate $i_t$, a forget gate $f_t$, an output gate $o_t$, a memory cell $c_t$ and a hidden state $h_t$. Each entry in theses vectors are in range from [0, 1] which is computed by following equations:

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}),$$

$$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}),$$

$$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}),$$

$$u_t = tanh(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}),$$

$$c_t = i_t * u_t + f_t * c_{t-1},$$

$$h_t = o_t * tanh(c_t),$$

Where $x_t$ is input at time step t, $\sigma$ is sigmoid function, * is element wise multiplication. Intuitively, the input gate $i_t$ controls how much each unit is updated. The forget gate $f_t$ controls the extent to which information is forgotten from previous memory cell ($c_{t-1}$). The output gate $o_t$ controls which information should be going to next hidden state. The Hidden state $h_t$ describes value of these vectors (gates) which varies for each time step. The LSTM learns to represent information over multiple time steps.

There are two commonly used variant of LSTM: bidirectional LSTM and multilayer LSTM. A bidirectional LSTM [18] consists of two LSTM units: ones for left-to-right and other for right-to-left in an input sequence. The first LSTM unit runs on input sequence while other LSTM unit runs on reverse input sequence. These units run in parallel. The hidden state at time step t of the bidirectional LSTM is concatenation of forward and backward hidden states. This architecture allows hidden states to capture future information. In, Multilayer LSTM architecture [18], [17], [29]the hidden state $h_t$ of one LSTM unit at layer l is fed as input to another LSTM unit at layer l+1. The key idea for this architecture is to capture long range dependency in the input sequence. These two LSTM architecture can further be combined as a Multilayer bidirectional LSTM [18].

## IV. CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural network was initially proposed for image classification task which gave remarkable performance in object detection tasks. It is similar to conventional neural network which contains a sequences of operations. Usually, these operations includes: convolution and pooling. For natural language processing tasks [10], has proposed a variant of CNN which involves 1D convolution and pooling operation. Consider a sequence of words $w_{i:n} = w_1, w_2, w_3...w_n$. Each word wi is associated with an embedding vector of dimension d. A 1d convolution of width k moves along the words in the sentence like a sliding window of size k i.e each window contains k words. Each window of k words . The convolution filter or kernel is applied to each window. For example, consider a $i_{th}$ window of words $w_i, w_{i+1}...w_{i+k}$ This $i_{th}$ window is represented by $x_i = [w_i, w_{i+1}, ...w_{i+k}] \in \mathbb{R}^{d*k}$. If, a convolutional filter v of size $\mathbb{R}^{d*k}$ is applied to $i^th$ window, it generates a scalar feature $c_i$.

$$c_i = g(v * x_i) \in \mathbb{R}$$

Here, g is a nonlinear function which is generally reLu function. This filter is applied to each possible window of words which finally gives a sequence of features c.

$$c = [c_1, c_2...c_{n-k+1}] \in \mathbb{R}^{n-k+1}$$

In practice, generally multiple filters v1, v2,...vl are applied as shown in Figure 2 to produce output $r_i \in \mathbb{R}^{l*(n-k+1)}$. This is a output of ith layer of CNN. After that, a pooling operation is applied which is usually max- polling. The max-pooling operation extracts maximum values from a sliding window of a fixed size containing a sequence of features c. The idea here is to capture most important features from the sequence of features.

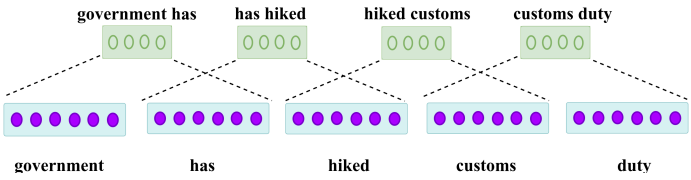


Fig. 2. Example of a sentence convolution with k=2 and dimensional output l=4.

## V. DATASET

Liar dataset [24] contains 12836 short news statements along with user profile information such as subject, context/venue, speaker, state, party, and prior history. The instances in the dataset come from different sources including interview, news release, political debate, Facebook post and tweets.

The dataset is labelled into 6 different class category: pants-fire, false, barely- true, half-true, mostly-true, and true. The labels in the dataset are well balanced where each label range from 2063 to 2638. While, the pant fire label are 1050. The Dataset covers a wide range of subjects/Topics including

| S.No | Statement | Subject | Speaker | Location of speech | Label |
|---|---|---|---|---|---|
| 1 | Says a U.S. Supreme Court justice suggested that some U.S. cases will be decided based on South African law | legal-issues,supreme-court | marco-rubio | a speech at the CPAC conference | False |
| 2 | Says Ohio budget item later signed into law by Gov. John Kasich requires women seeking an abortion to undergo a mandatory vaginal probe | abortion,pundits,women | rachel-maddow | a broadcast on MSNBC | pants-fire |
| 3 | Under Obama, home values in Florida collapsed, construction jobs were lost and the state had a high foreclosure rate | housing,message-machine-2012 | mitt-romney | a campaign ad | barely true |
| 4 | Under President Obamas health-care reforms, the premium for an average plan for a family didnt go down by $2,500 per year, its gone up about $2,500 per year | corporations,economy, health-care, public-health,workers | ron-johnson | an interview | half-true |
| 5 | Each year, 18,000 people die in America because they don't have health care. | health-care | hillary-clinton Presidential candidate | a speech in Des Moines, Iowa. | true |
| 6 | The number of illegal immigrants could be 3 million. It could be 30 million. | immigration | donald-trump President-Elect New York | a speech in Phoenix, Ariz. | pant on fire |
| 7 | Because of the federal health care law, 300,000 health plans canceled in Florida. | health-care | rick-scott Governor | a television ad from Let's Get to Work | barely-true |

TABLE I
SOME RANDOM NEWS FROM LIAR DATASET

most discussed subjects: economy, health-care, taxes, federal-budget, education, jobs, state-budget, candidates-biography, elections, and immigration. Some random news are shown in Table I.

## VI. Proposed Model

Assume, a news article contains $w_i$ words, i $\in [0, l]$. The news article is padded wherever necessary to make exactly l words. In this work, NLTK POS tagger [30] is used to tag each word in a news content with one of the k tags. After pos tagging, each $w_i$ word is tagged with one of k tags. Each $w_i$ word in a given news is then represented by its word embedding $\overrightarrow{w_i}$. The word embeddings of remaining POS tags which do not appear in POS tags of the news content are made zero vector. The word embeddings of same pos tags are then averaged to get $T_j$ embeddings, j $\in [0, k]$. In this paper, each news content is represented by Tj embeddings each of 300 dimensional. The $t_j$ embeddings are sorted in alphabetical order before passing to the bi-directional LSTM. We use a pre-trained word vector trained on 100 billion words of Google News [31] and the bidirectional LSTM [18] to get a vector representation of a news content. The bidirectional LSTM contains the forward LSTM $\overrightarrow{f}$ which reads the word embeddings from $T_1$ to $T_k$ and a backward LSTM $\overleftarrow{f}$ which reads from $T_k$ to $T_1$. This way, we obtain a vector representation of any news content by concatenating the forward hidden state $\overrightarrow{h_j}$ and backward hidden state $\overleftarrow{h_j}$, i.e., $h_j = [\overrightarrow{h_j}, \overleftarrow{h_j}]$.

$$\overrightarrow{h_j} = \overrightarrow{LSTM(Tj)}, j \in [1, k],$$
$$\overleftarrow{h_j} = \overleftarrow{LSTM(Tj)}, j \in [k, 1],$$
$$h_j = [\overrightarrow{h_j}, \overleftarrow{h_j}],$$

The above feature vector $h_j$ is concatenated with feature vector $p_n$ (as shown in Figure 1). The feature vector $p_s$ is obtained through CNN using user profile information. For example, if a user profile information is $p_s$, s $\in [0, n]$ , it contains n different parameters such as subject, party and prior history. Since, each parameter $p_s$ may contains multiple words t $\in [0, m]$, they are padded to make exactly m words. Each word in $p_s^t$ is then represented by its word embedding $\overrightarrow{p_i^t}$. The concatenated feature vector is finally fed to a softmax classifier to get a class label.

## VII. Experimental setup

The proposed architecture is implemented by using Python and a deep learning library Keras. The CNN model encodes user profile information of the news content in the vector $p_J$. The CNN model consists 4 layers. The first two layers consists 64 filters of width 1*5. The next two layers consist of 128 filters each of 1*5 size. After two consecutive layers, a max-pooling operation is performed. The model is trained on the training dataset with 6 training epochs. For news contents, we have used a Bidirectional LSTM with 64 hidden unit and dropout of 0.5. The outputs of CNN and RNN area

merged before applying the softmax classifier. While training the network, 'adam' optimizer with 'categorical crossentropy' loss are used. All these parameters are used with their default settings. The model is trained on the training dataset with 6 training epochs with batch size of 32.

## VIII. Result and discussion

To validate the effectiveness of the proposed architecture, the CNN and LSTM architecture model are individually evaluated. Our implementation produces better result as compared to [24], [25], when considering news content without profile information. However, the third model which is based on the fusion of news content and user profile informations give overall better classification performance than the individual one. This hybrid model which is based on LSTM and CNN gives the performer gain of 3.3%. The hybrid architecture gives comparable result to [25] on the benchmark dataset . The results and comparisons are summarized in Table II. The first three models are the part of proposed framework which are evaluated against the baseline models individually to better understand the contribution of each individual component. Our CNN architecture based on user profile informations performs significantly better than our LSTM architecture based on news content. The next five methods shows results as reported by [24], [25].

| Methods | Features Name | Validation | Test |
|---|---|---|---|
| Bidirectional LSTM based on POS Tags [Proposed Architecture] | News content | 0.204 | 0.274 |
| CNN [Proposed Architecture] | Profile attributes | 37 | 38.9 |
| Hybrid model [proposed Architecture] | News content & Profile attributes | 40.7 | 42.2 |
| SVMs [William Yang Wang [24]] | News content | 0.258 | 0.255 |
| CNNs [William Yang Wang [24]] | News content | 0.26 | 0.27 |
| Hybrid CNNs [William Yang Wang [24]] | News content & Profile attributes | 0.247 | 0.27.4 |
| Base LSTM [Yunfei Long etl. [25]] | News content | 0.25 | 0.255 |
| LSTM [Yunfei Long etl. [25]] | News content & Profile attributes | 40.7 | 41.5 |

TABLE II
THE EVALUATION RESULTS OF PROPOSED MODEL ON THE LIAR DATASET.

## IX. Conclusion

In this paper, a hybrid architecture is proposed which is based on Bidirectional LSTM and Convolutional Neural Network to incorporate both news content and user profile information. The proposed bidirectional LSTM exploits POS tags information of news content and outperforms compared to state of art architecture when considering only news content information. Also, the proposed CNN architecture which encodes user profile information gives comparable results to state of art architecture. The experiment demonstrates that the proposed hybrid architecture performs better than individual architecture and it gives overall accuracy of 42.2%. There are several limitations and scope of improvement. First, the dataset contains 6 types of class levels which causes redundancy among class labels. For example, pants-fire, false, barely-true are almost same. This causes the architecture model difficult to learn to differentiate among them. Similarly, true and mostly-true class labels are barely differs from each other. Second, the two independent models ( i.e CNN and LSTM) can be combined in a more efficient way to give overall better performance. Because, the CNN model which exploits profile information alone gives an accuracy of 38.9 but still the hybrid model gives an overall accuracy of 42.2%. Third, although the LSTM model which exploits news content, gives increase in accuracy of only 27.4% but, the model finds difficulty in learning pattern from long news content.

## References


[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[2] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? using satirical cues to detect potentially misleading news," in *Proceedings of the second workshop on computational approaches to deception detection*, 2016, pp. 7–17.

[3] M. Balmas, "When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism," *Communication Research*, vol. 41, no. 3, pp. 430–454, 2014.

[4] V. L. Rubin, Y. Chen, and N. J. Conroy, "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*. American Society for Information Science, 2015, p. 83.

[5] BuzzFeed. (2013) An american website wrote a satirical article about an indian rape festival and many people thought it was real. [Online]. Available: https://www.buzzfeednews.com/article/tasneemnashrulla/an-american-website-wrote-a-satirical-article-about-an-india

[6] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading online content: Recognizing clickbait as false news," in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 2015, pp. 15–19.

[7] ZeeNews. (2016) Arun jaitley dismisses rumours of nano gps chip on rs 2000. but data show as many as cash fish catches have been done they had huge bundles of new currency_note. [Online]. Available: https://zeenews.india.com/personal-finance/arun-jaitley-dismisses-rumours-of-nano-gps-chip-on-rs-2000-note_1948129.html

[8] V. L. Rubin, "Deception detection and rumor debunking for social media," in *The SAGE Handbook of Social Media Research Methods*. SAGE, 2017, p. 342.

[9] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREc*, vol. 10, no. 2010, 2010, pp. 1320–1326.

[10] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[11] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[12] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," *arXiv preprint arXiv:1412.1058*, 2014.

[13] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 797–806.

[14] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[15] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 2011, pp. 30–38.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[17] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
[18] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 273–278.
[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
[20] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in neural information processing systems*, 2009, pp. 545–552.
[21] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 2012, pp. 171–175.
[22] V. Pérez-Rosas and R. Mihalcea, "Experiments in open domain deception detection," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1120–1125.
[23] S. Gottipati, M. Qiu, L. Yang, F. Zhu, and J. Jiang, "Predicting users political party using ideological stances," in *International Conference on Social Informatics*. Springer, 2013, pp. 177–191.
[24] W. Y. Wang, "" liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648*, 2017.
[25] Y. Long, Q. Lu, R. Xiang, M. Li, and C.-R. Huang, "Fake news detection through multi-perspective speaker profiles," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2017, pp. 252–256.
[26] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.
[27] Y. Bengio, P. Simard, P. Frasconi *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.
[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[29] W. Zaremba and I. Sutskever, "Learning to execute," *arXiv preprint arXiv:1410.4615*, 2014.
[30] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.
[31] T. Mikolov, K. Chen, G. S. Corrado, and J. A. Dean, "Computing numeric representations of words in a high-dimensional space," May 19 2015, uS Patent 9,037,464.