A Deep Transfer Learning Approach for Fake News Detection

Tanik Saikh*, Haripriya B†, Asif Ekbal* and Pushpak Bhattacharyya*
Department of Computer Science and Engineering, Indian Institute of Technology Patna*
Bihta, Patna, India*
Department of Computer Science and Engineering, Indian Institute of Information Technology Senapati†
Manipur, India†
Email: {tanik.srf17,asif,pb}@iitp.ac.in*, haripriya@iiitmanipur.ac.in†

***Abstract*—Fake or incorrect or miss-information detection has nowadays attracted attention to the researchers and developers because of the huge information overloaded in the web. This problem can be considered as equivalent to lie detection, truthfulness identification or stance detection. In our particular work, we focus on deciding whether the title of a news is consistent with its body text- a problem equivalent to fake information identification. In this paper, we propose a deep transfer learning approach where the problem of detecting title-body consistency is posed from the viewpoint of Textual Entailment (TE) where the title is considered as a hypothesis and news body is treated as a premise. The idea is to decide whether the body infers the title or not. Evaluation on the existing benchmark datasets, namely Fake News Challenge (FNC) dataset (released in Fake News Challenge Stage 1 (FNC-I): Stance Detection) show the efficacy of our proposed approach in comparison to the state-of-the-art systems.**

***Index Terms*—Text Entailment, Title-Body Consistency, Stance Detection, Fake News, Deep Transfer Learning**

I. Introduction

The online platforms like social media websites, e-commerce sites of products and services, blogs, online forums and discussion forums etc. are very much attached today with our day-to-day lives. A large volume of textual contents are generated daily from these sources. This information can be effectively utilized to build models for any applications related to Artificial Intelligence (AI), Natural Language Processing (NLP) and Machine Learning (ML). As the sources are diverse in nature and large in number, studying the credibility and authenticity of information is a crucial step. There are multiple reports available for a particular event and the vice-versa. Different news agencies also produce different reports on a particular topic. A legitimate report should be consistent with its title. In order to judge the truthfulness of a particular event/claim it is necessary to observe what other agencies are saying on that particular topic. So to justify the truthfulness of a particular fact/claim, it is necessary to judge the consistency between the fact/claim and the body texts related to that very topic. This could be a vital module for robust fake news detection system - the task of which is to identify whether the news is genuine or fake.

In this paper, we tackle the problem of fake news detection through stance detection. Stance is basically an article's response to a title/headline/claim. The response could be in any of the followings: *Agree, Disagree, Discuss and Unrelated*. It is one of the fundamental approaches for fake news detection. We make use of this stance detection to combat fake news. We can detect fake information/claims through stance as follows: suppose a person claims like *"Barak Obama is not born in United States"* or *"The pope has a new baby"*, we can take that claim and search for many news articles with respect to that subject. If we have many reputable (and well-known) sources which all *Agree* with this claim, then we can say that the particular claim is most probably true.

More concisely, the task can be defined as: Given a claim and a body of text (like news article), the system has to decide whether the body of the text generally *Agree*, *disagree*, *neutral* or is completely *Unrelated* to the claim. This problem is typically called a stance detection problem. Hence, detecting the truthfulness of a particular information/claim, and the consistency of that particular information with its' context is a challenging task for fake news detection. According to [1] fake news is *"made-up stories with an intention to deceive"*. Basically, the task of fake news detection is to estimate the probability of a piece of text being fake. The problem of fake information detection has been viewed from the different perspectives, *viz.* (i). determining whether the textual content of a news article is true or not, and (ii). evaluating the intrinsic prejudice of a written text. In our current work, we use the setup which consists of News Title (NT), News Body (NB) and their stance (relation). We make use of the benchmark dataset which is released as a part of the Fake News Challenge [2] [1].

We pose the task of stance detection as equivalent to consistency detection between the NT and the NB which is conceptually very similar to a very popular task in NLP, namely Natural Language Inference (NLI) [3] or Textual Entailment (TE) [4], [5]. The definition is as follows: Given two pieces of texts, one being the *Premise(P)* and the another one is the *Hypothesis(H)*, the system has to decide whether

- H is the logical consequence of P or not.
- H is true in every circumstance (possible world) in which P is true.

[1] http://www.fakenewschallenge.org/

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

For example, *P: "John's assassin is in jail"* entails *H: "John is dead"* and *P: "Mary shifted to France three years back."* entails *H: "Mary lives in France"*. Indeed, in both the above examples *H* is the logical consequence of *P*. On the other hand, *P: "Mary lives in Europe"* does not entail *H: "Mary lives in US"*. Obviously, *H* does not have any logical consequence of *P* in this example. In addition to this, the classes of stance detection problem are quite similar with the classes as what defined in the benchmark NLI dataset i.e. Stanford Natural Language Inference (SNLI) [3] [2]. Three broad applications areas of TE/NLI are found namely: (i). direct application of trained NLI models. (ii). NLI as a evaluation task for new ML methods and (iii). NLI as a pre-training task in transfer learning. A few direct applications are found in Fact Extraction and Verification (FEVER) shared task [6], [7], multi-hop reading comprehension tasks [8], generating video caption [9] and long form text [10]. The tasks of [11]–[14] make use of various entailment corpora, especially SNLI and Multi-NLI, as the benchmark datasets. The tasks defined in [15]–[18] applied transfer learning based approaches where a neural network was trained on the SNLI corpus. These tasks have shown considerable improvement in the target tasks. The underlying task is inspired by the third application (i.e. the transfer learning approach [19], [20]). We also train a deep neural network models on large NLI corpus (i.e. SNLI) and apply the training model on our target task, i.e. stance classification. Our task is also an example of application of NLI/TE.

Hence, we formulate the problem of stance classification with respect to TE/NLI. We take the concept of TE/NLI to detect the consistency between the NT and NB. We make use of this semantically enriched large SNLI corpus to utilize the notion of TE/NLI in our task, as this is a widely recognized benchmark corpus for the NLI/TE problem. The size of this corpus is very large compared to the FNC on which our systems were evaluated. Both the corpus are conceptually similar with a title, body and appropriate stance. So we adopt a transfer learning strategy to solve our target task, i.e. title-body consistency detection (stance classification for fake news detection).

We train our proposed system with this SNLI dataset, store and apply the knowledge gained to our target task (i.e. title body consistency detection problem). The key contributions and/or characteristics of the current work can be summarized as follows:

- We want to leverage the notion of TE for fake news detection. To the best of our knowledge there is no prior work in this line.
- We use deep transfer learning based approaches, where the model trained on a large structured SNLI dataset is adopted for fake news detection. It is to be noted that using this technique we mitigate the problem of data scarcity for the target task.

[2]https://nlp.stanford.edu/projects/snli/

Rest of the paper is organized as follows. At first we discuss the related work in Section II. We describe our methodology in Section III. This section comprises of problem definition, proposed approaches. In Section IV, we describe the dataset used, experimental setup, results obtained, and comparison with the existing state-of-the-art models followed by error analysis. Section V concludes the paper with pointers to future research directions.

II. Related work

Automatic fake news detection has recently gained attention to the researchers and developers. The tasks defined in [21], [22] describe the fact checking problem, and they correlated this with the problem of TE. The work of [23] first released a large dataset for fake news detection and proposed a hybrid model to integrate the statement and speaker's meta data and performed classification. The task of [24] also posited a novel dataset called *Emergent*, which was driven from the Digital Journalism Project, namely *Emergent* [25]. They additionally proposed a feature based logistic regression model for the stance detection.

The task defined in [26] employed conditional encoding network with two Bi-LSTMs to detect stance of tweets with some targets. The work described in [27] utilized the stance detection dataset. They proposed four models which are based on the *Bag of word (BoW), basic LSTM, LSTM with attention, and condition encoding LSTM with attention* and showed that the model with condition encoding LSTM with attention mechanism yields the highest performance among all the models, which demonstrated the efficiency of attention technique in extracting from a long sequence (news body) of information relevant to a small query (article title).

[28] defined a corpus which combines *stance detection, stance rationale, relevant document retrieval, and fact checking* tasks. Apart from these tasks on stance detection for fake news detection which made use of Fake news dataset could be found in [29]–[31]. The works [32] solved the problem which had been defined in SemEval-2017 Task 8. It has been studied in other languages too, like Arabic [33]. [31] performed a rigorous analysis of FNC-I and top three participating systems. This task could be considered as reproduction study of FNC-I. Error analysis of this study shows that the existing models mostly rely on the lexical overlap for stance classification. This study also concluded that stance detection problem is really very challenging tasks and suggested that more advanced machine learning approaches (having deeper semantic understanding) are essential to combat with this problem. This study tries to fill this gap.

III. Methodology

In this section, at first we define the problem and then describe various deep learning based models that we propose.

A. Problem Definition

Given a title and its supporting documents (news body text), the system has to determine its stance, i.e, *Agree, Disagree,*

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Discuss or Unrelated. So the overall input and output of the system will be as follows:
Input: A claim, and its supporting document.
Output: Agree, Disagree, Discuss and Unrelated.
We pose this problem as classification problem.

B. Proposed Approaches

We propose deep transfer learning approach to solve the underlying problem of detecting title-body consistency. We detect the truthfulness of a claim using stance detection. The dataset contains title and body pairs with its labels/stance (*Agree, Disagree, Discuss, Unrelated*). We develop two transfer learning based methods, one is trained on the SNLI corpus and tested on FNC (i.e. Model I); and the other one is trained on the combination of both the SNLI and FNC, and then tested on the FNC test set (i.e. Model II).

First Model: Our first model (i.e. Model - I) has two versions, one is based on Bi-LSTM [34] and second one is with Bi-LSTM followed by max-pooling layer as the sentence encoder. The paper [35] showed that Bi-LSTM network with max-pooling is the best sentence encoder by exploring various architectures. The input dataset is primarily reshaped into three dimensions and given as input to the Bi-LSTM layer. We make use of fastText word vectors [36] [3] to obtain the embedding of each word contained in the dataset. The vector representations of sequence of words contained in NT/H is S_1 = $\{x_1, x_2, x_3....x_N\}$ and NB/P is S_2 = $\{y_1, y_2, y_3....y_N\}$ are given to two separate Bi-LSTM networks. At a particular time-stamp t, the memory c_t and the hidden state h_t are updated with the help of the following equations:

$$f_t = \alpha(W_f[a^{<t-1>}, x^t] + b_f) \tag{1}$$

$$f_u = \alpha(W_u[a^{<t-1>}, x^t] + b_u) \tag{2}$$

$$f_o = \alpha(W_o[a^{<t-1>}, x^t] + b_o) \tag{3}$$

Then the output a_t as follows:

$$C_t = f_u * C^{N<t>} + f_t * C^{N<t-1>} \tag{4}$$

Bi-LSTM learns the sequence of words in two directions, one from beginning to end and the other from end to beginning. We take the last representation of each direction as follows.

$$a_t = f_o * C^{<t>} \tag{5}$$

We concatenate the last representations of the both the forward and backward LSTM as shown in Equation 6.

$$y_t = \overrightarrow{a_t} \oplus \overleftarrow{a_t} \tag{6}$$

This y_t is obtained for each NB/P and NT/H. We consider y_t for NB/P as y_t^P and similarly for NT/H as y_t^H. They are concatenated further as below in Equation 7.

$$y_f = y_t^P \oplus y_t^H \tag{7}$$

This representation is further passed into four stacked feed forward neural network as follows. $Y_1 = F(W_1 * y_f + b_1)$, $Y_2 = F(W_2 * Y_1 + b_2)$, $Y_3 = F(W_3 * Y_2 + b_3)$ and $Y_4 = F(W_4 * Y_3 + b_4)$ where F is a ReLU activation function.

[3]https://fasttext.cc/docs/en/english-vectors.html

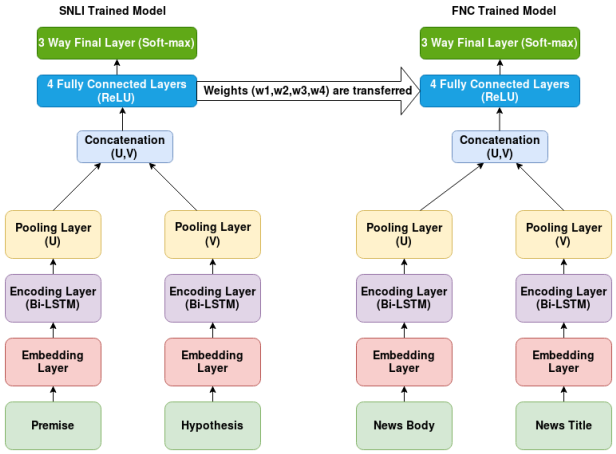

Fig. 1. Architecture of the Proposed Model - I.

This method comprises of two models: one trained on the SNLI corpus and the weights (like W_1, W_2, W_3 and W_4 along with their biases) of the feed forward network (dense layers) layers of this model are transferred to the second model. The second model is being trained and tested on the FNC. Here, we use weights of the SNLI trained model and initialize the dense layers of the second model which are being trained and tested on FNC. The main motto of this process is that, the weights from the SNLI model will be updated in the FNC model while feeding them into the second model. This way, we apply the semantics learned from the SNLI corpus to detect stances *(Unrelated, Discuss, Agree, or Disagree)* of an unknown example pair. The architecture of this approach is shown in Figure 1.

Second Model : In our second model (i.e. Model-II), we use Bi-LSTM with and without max-pooling for extracting features from the input sentences. At first we build the models following the same process as we described in Model-I. This method comprises of three models: the first one is trained on the SNLI corpus which is having four dense layers. The weights (i.e. W_1, W_2 along with their biases) of the first two lower layers are saved and transferred to the third model. We transfer from the two lower layers, as the SNLI corpus is very large, the vanishing gradient problem in the lower layers will be less compared to the upper layers. The second model is trained on the FNC which is also having four dense layers. Here, we transfer the weights (i.e. W_3, W_4 along with their biases) of the upper two dense layers as shown in the Figure 2 to the dense of the third model. Here, we transfer the weights from the upper two layers, because the training corpus (i.e. FNC) is comparatively less in size, the vanishing gradient problem in the upper layers will be less as compared to the lower layers. There is a final layer with soft max activation

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

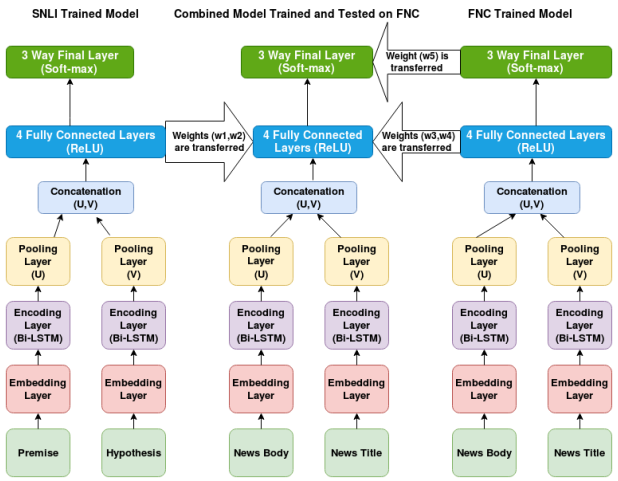

Fig. 2. Architecture of the Proposed Model - II

function like below.

$$Y_5 = f(W_5 * Y_4 + b_5) \tag{8}$$

where f is a soft-max function.

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{9}$$

Here i is for each of the classes.
We transfer the weights (i.e. W_5 along with the bias b_5) of this 4-way final (soft-max activation) layer of the second model also to the third model's final layer as the actual classification has 4 classes which is performed in the third model. The justification of transferring this final layer's weights is empirical. Using these weights we test the third model on the test set of FNC. Following this strategy we leverage the benefits of both SNLI and FNC which are further utilized to train and test an empty third model (it was initially empty, after completion of first and second models' training, it's weights are initialised as mentioned earlier). The final model assigns a label to each unknown example pair in one of the following four classes (*"Agree","Disagree","Discuss","Unrelated"*). We compare the predicted label of each instance with the gold label of that instance to obtain the accuracy. The architecture of this approach is shown diagrammatically in Figure 2.

IV. Experiments, Results and Discussions

In this section we describe the two kinds of datasets that we used, experimental procedures, results obtained and the discussions followed by error analysis.

A. Datasets

There are two kinds of datasets utilized for this experiment. One is SNLI and the another is the FNC. The FNC is derived from the Emergent dataset [24]. The various statistics including training and test set distribution of both the datasets are shown in Table I, Table II and Table III.

TABLE I
Statistics of SNLI dataset

Dataset	# of pairs
Training	550152
Development	10000
Test	10000
Sentence Length (mean token count):	
Premise	14.1
Hypothesis	8.3

TABLE II
Statistics of the FNC datasets. TBP: Number of Title–Body pairs, ABL: Average Body Length, ATL: Average Title Length

Dataset	Training Set			Test Set		
	TBP	ABL	ATL	TBP	ABL	ATL
FNC	49971	369	11	25413	347	11

TABLE III
Distribution of classes in training and test set of FNC

Dataset	**Example Pairs**	**Classes**			
		Unrelated	**Discuss**	**Agree**	**Disagree**
Training	49972	0.73	0.17	0.07	0.016
Test	25413	0.72	0.17	0.07	0.027

The distribution in Table III shows that the dataset is fully biased towards Unrelated class. The average length differences (approx. 358-336) between NB and NT show that, predicting stances between these two texts is really very challenging task.

B. Implementation

As already mentioned, we employ two kinds of corpus, namely SNLI and FNC. SNLI dataset consists of P and H pairs and labels (Entailment, Contradiction, and Neutral) corresponding to that pairs. First of all, the dataset is converted into two lists, one containing all the sentences pairs and other one which is having the labels. The labels are converted into integers and then to one-hot encoded vectors using the respective functions available within the scikit learn ML package [4]. The categorical data is not operable by many deep learning (DL) algorithms. They require all inputs and outputs variables to be in numerical form. So the labels are firstly converted into the integer encoding. For categorical variables where no ordinal relationship exists, the integer encoding is not enough. Therefore, it is converted into one-hot vector representation. Next, the sentences are processed through a NLTK tokenizer [5] which transforms sentences to a sequences of words.

Most of the modern sophisticated ML techniques rely on the vector representation of words. We apply pre-trainted fastText word vector method for the purpose. These vectors carry the hidden information of a language like, word analogy or

[4]http://scikit-learn.org/stable/
[5]https://www.nltk.org/api/nltk.tokenize.html

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

semantic. We take vector representation of each word and create the embedding matrix. The embedding matrix is then given to our proposed model as input with a batch size of 32. We take the last hidden representations obtained from forward and backward pass of Bi-LSTM model of a particular sentence and then concatenate those two representations to obtain the whole sentence representation. The obtained representation is considered as output of Bi-LSTM, which is considered as sentence vector representation. This representation is further fed into another layer, namely max-pool layer. This layer reduces the vector dimension, which is considered to be the more condensed feature representation. The outputs from this layer are further passed through multiple stacked layers of feed forward neural network (four dense layers with Relu activation function) for learning. Finally, we put a final layer with soft-max activation function to obtain the classification output. The implementation of the proposed networks are performed in Python Keras library [6] platform.

C. Results and Discussion

We tackle this problem with two approaches as described in the previous sections. These approaches are Transfer Learning Trained with SNLI (Model-I), Transfer Learning Trained with both the SNLI and FNC (Model-II).

The first model yields 76.37% and 72.82% accuracies by two variants, namely Bi-LSTM and Bi-LSTM with max-pooling, respectively. The two variants of the second model yields better classification accuracy compared to the previous method. The Bi-LSTM and Bi-LSTM with max-pooling produce the accuracy of 84.35% and 90.20%, respectively.

The Table III depicts that most of instances of FNC are having *Unrelated* class. FNC-I organizers have come up with a two levels weighted based scoring system.[7] They proposed this metric to tackle the data imbalance problem of having large number of unrelated examples. In first level 25% score weight is given for classifying NT and NB as related (combination of agrees, disagrees, and discusses) or unrelated, and in the second level 75% score weight is given for classifying related pairs as agrees, disagrees, or discusses. Less weight is given for classifying *related/unrelated* as it is trivial and less relevant for fake news detection. On the other hand, more weight is given in the evaluation scoring system for stance classification (i.e. classifying as agrees, disagrees or discuss). This is non-trivial and very relevant for fake news detection. We can say this score as FNC-1.

We also follow the guideline and compute that scores (FNC-1) for our proposed systems. We compute the overall F1 (i.e. F1) and class-wise F1 score (for Agree, Disagree, Discuss and Unrelated) to see the proposed models' efficacy in different modalities of evaluation. All these results and comparison with the existing scores are shown in Table IV.

We accomplish a systematic comparison of our systems to the existing best systems. The approach based on Bi-LSTM with max-pooling of Model-II produces the best accuracy among the results obtained by the other two proposed models. The FNC-1, overall F1, Agree, Disagree and Discuss class' F1 produced by this system outperformed the existing system's results. In this model we incorporate the knowledge of semantically enriched corpus's (SNLI). Our second model's Bi-LSTM based approach yields the best F1 for Unrelated class. Our system outperforms the existing best system of [37] with a margin of 0.0269 in FNC-1 score. The system of [37] is a combination of statistical ML and DL based approach augmented with TE based features. However, the proposed systems are fully automatic end-to-end DL based approaches that avoid any hand-crafted feature engineering.

Talos Intelligence's SOLAT in the SWEN team [38] stood first in this competition and put a milestone on this dataset. They obtained an FNC-1 of 8204 on this dataset. Our fourth system performs better compared to this system also. The system of [38] comprises of two models: a gradient-boosted decision trees (TalosTree) and a deep convolutional neural (TalosCNN) network. TalosTree combines word2vec embedding along with word count, TF-IDF, sentiment, and singular-value decomposition features. TalosCNN is based various CNN models followed by three feed forward neural networks, followed by a soft-max for classification. This is also a combination of ML and DL approaches. The task of [39] was the second best system in the competition. Their model is also based on several hand-crafted features (unigrams, cosine similarity, latent Dirichlet allocation etc.) feeding into a Multi-layer Perceptron (MLP). The work proposed by UCL Machine Reading (UCLMR) team [40] was the third ranked model in the competition. The model is based on single hidden layer MLP. Term frequency vector, cosine similarity computed between the TF-IDF vectors of the NT and NB and two TF-IDF vectors etc. are the features for this model. As we can see all the previous models are hand-crafted feature engineering based approaches. The proposed approaches are fully automated deep neural network based approaches. The Table V shows the confusion matrix of our best performing system.

D. Error Analysis

From the Table V we extract the mis-classified instances. From these example pairs we try to analyze the cases where our proposed system failed. Below we show some of the errors that our system encounters.

- One of the main problems is the data imbalance problem. Maximum number of example pairs are having *"Unrelated"* class in the training set. So the models learn to predict an unknown example pair as *"Unrelated"* in most of the cases. In mis-classified instances maximum number of instances are wrongly predicted as *"Unrelated"*, even though they belong to the other classes in the gold label.
- The bodies are having multiple number of repetitive words and sentences. we believe these repetitive occurrences of such entries might have hampered the accuracies.

[6]https://keras.io/

[7]Please refer to evaluation section of http://www.fakenewschallenge.org/

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE IV
RESULTS OF THE EXISTING SYSTEMS AND THE PROPOSED SYSTEMS; SOTA: STATE-OF-THE-ART; COLUMNS NAMED WITH AGREE, DISAGREE, DISCUSS AND UNRELATED REPRESENT THEIR RESPECTIVE CLASS WISE F1 SCORE; F1: OVERALL F1

SN	System	FNC-1	F1	Agree	Disagree	Discuss	Unrelated
Existing Six SOTA Models and Results							
1	**ML_DL Combo**	0.8254	0.636	0.611	0.214	0.746	0.972
2	**TALOSCOMB(TREE+CNN)**	0.8204	0.582	0.539	0.035	0.760	0.994
	ATHENE	0.8197	0.604	0.487	0.151	**0.780**	**0.996**
3	**UCLMR**	0.8172	0.583	0.479	0.114	0.747	0.989
4	**featMLP**	0.825	0.607	0.530	0.151	0.766	0.982
5	**stackLSTM**	0.821	0.609	0.501	0.180	0.757	0.995
6	**MAJORITY VOTE**	0.394	0.210	0.0	0.0	0.0	0.839
Proposed Models							
7	**Model1_Bi-LSTM**	0.6405	0.5240	0.2748	0.4332	0.8270	0.5611
8	**Model1_Bi-LSTM_Max-Pooled**	0.6117	0.5102	0.2075	0.4691	0.9072	0.4569
9	**Model2_Bi-LSTM**	0.8072	0.6815	0.6095	0.3730	0.8810	**0.8626**
10	**Model2_Bi-LSTM_Max-Pooled**	**0.8523**	**0.7495**	**0.6946**	**0.5093**	**0.9548**	0.8393
Official Baseline and Human Performance							
11	**Official Baseline**	**0.7520**	X	X	X	X	X
12	**HUMAN UPPER BOUND**	0.859	0.754	0.588	0.667	0.765	0.997

TABLE V
CONFUSION MATRIX FOR BI-LSTM-MAX-POOLING ENCODER IN METHOD - II

Label/Label	Agree	Disagree	Unrelated	Discuss
Agree	1322	98	330	153
Disagree	107	355	160	75
Unrelated	135	25	17520	669
Discuss	47	29	641	3747

- The biggest constraint is the length variation between the title and the body, which is very high. The body text is having more number of tokens compared to the title. We would keep this constraint in mind in the future work.
- It is to be noted that the corpus is having multiple numbers of Phrasal Verbs, Named Entities (NEs) and Multiword Expressions (MWEs), which need special modules to handle.

V. CONCLUSION AND FUTURE WORK

In this paper, we have investigated the role of Textual Entailment in fake information detection through stance detection. We have proposed various deep neural network models for solving the particular problem. The very first approach is composed of two models, one is trained on SNLI corpus, weights are transferred to another model which is trained and tested on FNC. Second approach is the combination of three models, one is trained on SNLI and another one is trained on Fake News Corpus (FNC), weights are transferred from these two models to another one which is trained and tested on FNC. Evaluation results show that the second approach with max-pooled layer is the best performing one. Our best model attains the state-of-the-art performance. Hence, we can draw the conclusion that indeed TE could be an effective way to handle the fake news detection problem. In future we would like to:

- incorporate the external knowledge (world knowledge) into the existing system.
- take care of NEs, MWEs and Phrasal Verbs present in the corpus in pre-processing module.
- take into account, the length difference between the headlines and bodies. We can introduce the concept of *Justification of name* in this regard.
- enrich the best performing model by incorporating the relevance score between the headline and body texts.
- apply attention model to address the length difference between the news body and title. Attention model will capture the important words of news body based on the corresponding title.

REFERENCES

[1] S. Tavernise, "As Fake News Spreads Lies, more Readers Shrug at the Truth," *New York Times*, vol. 6, 2016.

[2] D. R. Dean Pomerlau, "Post-facto Fake News Challenge."

[3] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A Large Annotated Corpus for Learning Natural Language Inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 632–642. [Online]. Available: http://aclweb.org/anthology/D15-1075

[4] B. MacCartney, "Natural Language Inference," in *Ph.D. thesis*. Stanford University, 2009.

[5] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL Recognising Textual Entailment Challenge," in *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, ser. MLCW'05. Southampton, UK: Springer-Verlag, 2006, pp. 177–190.

[6] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, and A. Mittal, "The Fact Extraction and VERification (FEVER) shared task," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 1–9. [Online]. Available: https://www.aclweb.org/anthology/W18-5501

[7] Y. Nie, H. Chen, and M. Bansal, "Combining Fact Extraction and Verification with Neural Semantic Matching Networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6859–6866.

[8] H. Trivedi, H. Kwon, T. Khot, A. Sabharwal, and N. Balasubramanian, "Repurposing Entailment for Multi-Hop Question Answering Tasks," 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun 2019, pp. 2948–2958.

[9] R. Pasunuru and M. Bansal, "Reinforced Video Captioning with Entailment Rewards," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen,

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Denmark: Association for Computational Linguistics, Sep 2017, pp. 979–985. [Online]. Available: https://www.aclweb.org/anthology/D17-1103

[10] A. Holtzman, J. Buys, M. Forbes, A. Bosselut, D. Golub, and Y. Choi, "Learning to Write with Cooperative Discriminators," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1638–1649. [Online]. Available: https://www.aclweb.org/anthology/P18-1152

[11] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kociský, and P. Blunsom, "Reasoning about Entailment with Neural Attention," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. [Online]. Available: http://arxiv.org/abs/1509.06664

[12] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A Decomposable Attention Model for Natural Language Inference," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2249–2255. [Online]. Available: https://www.aclweb.org/anthology/D16-1244

[13] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "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)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[15] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 670–680. [Online]. Available: https://www.aclweb.org/anthology/D17-1070

[16] S. Subramanian, A. Trischler, Y. Bengio, and C. J. Pal, "Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. [Online]. Available: https://openreview.net/forum?id=B18WgG-CZ

[17] J. Phang, T. Févry, and S. R. Bowman, "Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks," *CoRR*, vol. abs/1811.01088, 2018. [Online]. Available: http://arxiv.org/abs/1811.01088

[18] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task Deep Neural Networks for Natural Language Understanding," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4487–4496. [Online]. Available: https://www.aclweb.org/anthology/P19-1441

[19] L. Y. Pratt, J. Mostow, and C. A. Kamm, "Direct Transfer of Learned Information among Neural Networks," in *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2*, ser. AAAI'91. AAAI Press, 1991, p. 584–589.

[20] S. Ruder, "Neural Transfer Learning for Natural Language Processing," Ph.D. dissertation, National University of Ireland, Galway, 2019.

[21] A. Vlachos and S. Riedel, "Fact Checking: Task Definition and Dataset Construction," in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics, 2014, pp. 18–22.

[22] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational Fact Checking from Knowledge Networks," *CoRR*, vol. abs/1501.03471, 2015.

[23] W. Y. Wang, ""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)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 422–426.

[24] W. Ferreira and A. Vlachos, "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*, San Diego, California, 2016, pp. 1163–1168.

[25] C. Silverman, "Lies, Damn Lies and Viral Content," 2015.

[26] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva, "Stance Detection with Bidirectional Conditional Encoding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, 2016, pp. 876–885.

[27] S. Pfohl, O. Triebe, and F. Legros, "Stance Detection for the Fake News Challenge with Attention and Conditional Encoding," 2017.

[28] R. Baly, M. Mohtarami, J. Glass, L. Màrquez, A. Moschitti, and P. Nakov, "Integrating Stance Detection and Fact Checking in a Unified Corpus," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 21–27. [Online]. Available: http://aclweb.org/anthology/N18-2004

[29] M. Mohtarami, R. Baly, J. Glass, P. Nakov, L. Màrquez, and A. Moschitti, "Automatic Stance Detection using End-to-End Memory Networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 767–776. [Online]. Available: http://aclweb.org/anthology/N18-1070

[30] G. M. J. P. X. W. James Thorne, Mingjie Chen and A. Vlachos., "Fake News Stance Detection using Stacked Ensemble of Classifiers." in *Proceedings of the EMNLP Workshop on Natural Language Processing meets Journalism*, Copenhagen, Denmark, 2017, p. 80–83.

[31] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych, "A Retrospective Analysis of the Fake News Challenge Stance-Detection Task," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 1859–1874. [Online]. Available: http://aclweb.org/anthology/C18-1158

[32] G. Gorrell, K. Bontcheva, L. Derczynski, E. Kochkina, M. Liakata, and A. Zubiaga, "Rumoureval 2019: Determining Rumour Veracity and Support for Rumours," *CoRR*, vol. abs/1809.06683, 2018. [Online]. Available: http://arxiv.org/abs/1809.06683

[33] K. Darwish, W. Magdy, and T. Zanouda, "Improved Stance Prediction in a User Similarity Feature Space," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, July 31 - August 03, 2017*, 2017, pp. 145–148. [Online]. Available: https://doi.org/10.1145/3110025.3110112

[34] A. Graves and J. Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM and other Neural Network Architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[35] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 670–680. [Online]. Available: http://aclweb.org/anthology/D17-1070

[36] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in Pre-Training Distributed Word Representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[37] T. Saikh, A. Anand, A. Ekbal, and P. Bhattacharyya, "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*. Springer, 2019, pp. 345–358.

[38] D. S. Sean Baird and Y. Pan, "Talos Targets Disinformation with Fake News Challenge Victory." in *https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html*, 2017.

[39] B. S. Andreas Hanselowski, Avinesh PVS and F. Caspelherr., "Description of the System Developed by Team Athene in the FNC-1, 2017." in

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

https://github.com/hanselowski/athene$_system/blob/master/system_description_athene.pdf$., 2017.

[40] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, "A Simple but Tough-to-Beat Baseline for the Fake News Challenge Stance Detection Task," *CoRR*, vol. abs/1707.03264, 2017. [Online]. Available: http://arxiv.org/abs/1707.03264

978-1-7281-6926-2/20/$31.00 ©2020 IEEE