A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features

Tanik Saikh[1(✉)], Amit Anand[2], Asif Ekbal[1], and Pushpak Bhattacharyya[1]

[1] Indian Institute of Technology Patna, Bihta, India
{1821cs08,asif,pb}@iitp.ac.in
[2] Indian Institute of Information Technology Kalyani, Kalyani, India
amitanand@iiitkalyani.ac.in

Abstract. The phenomenal growth in web information has nourished research endeavours for automatic fact checking, or fake news and/or misinformation detection. This is one of the very emerging and challenging problems in Natural Language Processing (NLP), Machine Learning (ML) and Data Science. One such problem relates to estimating the veracity of a news story, which is a complex and deep problem. The very recently released Fake News Challenge Stage 1 (FNC-1) dataset introduced the benchmark FNC stage-1: stance detection task. This task could be an effective first step towards building a robust fact checking system. In this paper, we correlate this stance detection problem with Textual Entailment (TE). We present the systems which are based on statistical machine learning (ML), Deep Learning (DL), and a combination of both. Empirical evaluation shows encouraging performance, outperforming the state-of-the-art system.

Keywords: Fake news · Stance detection · Deep learning · Machine learning · Textual entailment

1 Introduction

In recent years, people are very communicative with the advent of the Internet. A lot of communications and conversations are happening through text, image, audio and video etc. This generates a lot of data everyday. The proliferation of these data/information in social media, online news feeds and tweets etc. demand for checking the truthfulness of these data/information. It is a tedious job even for the human being to do it manually. Hence, it is imperative to build the automated system which should be able to perform the tasks of detecting fake or misinformation, false claim detection, judging the veracity of a textual content made by a person etc.

Detecting veracity of information is a very challenging and demanding problem in Artificial Intelligence (AI), difficult even for a human being to understand the news contents all the time. Lately, [12] organized a shared task to

E. Métais et al. (Eds.): NLDB 2019, LNCS 11608, pp. 345–358, 2019.
https://doi.org/10.1007/978-3-030-23281-8_30

investigate how AI and Natural Language Processing (NLP) techniques could be promoted to combat fake news, entitled as Fake News challenge stage-I (FNC-I): Stance Detection. It could be a valuable first step towards helping human fact checkers to identify the false claims. Basically, to check the veracity of a claim/headline/report, it is important to see what other news agencies are saying about that particular claim/headline/report. There are multiple reportings available for a particular claim/headline/report produced by the different news agencies. Sometimes the document (body texts) agrees/supports the claim, sometimes it contradicts, sometimes discusses, or sometimes it remains completely unrelated to the claim. This is called stance, i.e. the relation between the *headline* and the *body* text. This is exactly what is defined in the dataset released in the shared task, FNC-I. The dataset contains <*Headline, Body Text, Stance*> triples. An example from the dataset is shown in Table 1. For this experiment, we assume the titles as claim/fact and the documents related to a particular title as body text. So if a particular title generally agrees with one and/or many of the body texts, then that particular title/claim could be most probably legitimate, otherwise, if there is no supporting body text to that claim, then that claim might be most probably fake. In this way, we can detect the truthfulness of a claim/report through stance detection. The shared task gained a lot of responses, with 50 teams from both academia and industry submitted their systems. Briefly, input to the system is a claim and the output corresponds to determining whether it is fake or genuine. We pose the problem as a classification problem, i.e. stance classification. The problem is conceptually very similar to a very well-known problem in NLP, namely TE [9] or Natural Language Inference (NLI) [3,15,16]. The definition of which is as follows: Given two pieces of texts, one is the *Premise(P)* and the other one is *Hypothesis(H)*, the system has to decide whether H is the logical consequence of P or not and/or H is true in every circumstance (possible world) in which P is true. 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*. We correlate the problem of stance detection to TE as follows: If a body text entails a claim, then it corresponds to actually support or agree or discuss; if it contradicts, then it corresponds to refute/disagree and if it does not provide any information related to the claim then it is completely unrelated (to the claim). We propose two approaches which are based on *viz. i. Statistical/Traditional ML and ii. DL*. The first approach makes use of a conventional set of features which are typically used for the task of TE. The second approach is an end-to-end deep learning approach and is based on the prior work [20]. We consider their model as the baseline in our experiments. The task described in [6] has shown how external knowledge could be helpful for DL based NLI models. Motivated by this we incorporate the ML features into our proposed DL architecture.

Contributions of our current work are two-fold, *viz* (i). We relate the problem to TE and propose various ML based models. We exploit the TE-based features and show the effect of TE for stance classification and further for fake news

Table 1. Headline and text snippets from documents and respective stances from the FNC training dataset

Headline: Hong Kong protesters go Ferguson style: 'Hands up, don't shoot'	
Stance	Body text
Agree	Hong Kong protesters have "emulated" the Ferguson gesture in their recent protests
Disagree	Photographs of Hong Kong protests have been discussed in the context of Ferguson....
Discuss	HONG KONG—Thousands of pro-democracy demonstrations in Hong Kong have....
Unrelated	A Russian fisherman says that Justin Bieber saved his life...

detection. (ii). We merge the ML feature values and the features extracted from the DL network, and feed into a feed-forward neural network. In this way we provide the external knowledge to neural network based model. This system outperforms the state-of-the art reported in the literature for the problem on this particular dataset. The paper is organized as follows. Section 2 describes brief overview of the related works followed by proposed methodologies (Sect. 3), dataset (Sect. 4), the experiments, results along with proper analysis (Sect. 5), and conclude (Sect. 6).

2 Related Work

Automatic fake news detection has recently gained attention to the researchers and developers. The papers [7,26] defined fact checking problem and they correlated this problem with the problem of TE. We also correlate, and make use of different TE based features. The work defined in [27] 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 [11] also posited a novel dataset called *Emergent*, which was driven from the digital Journalism project, namely Emergent [22]. They additionally proposed a logistic regression model for the stance detection, where features are extracted from the headline and news body pairs. The dataset that we employ in this experiment is an extended version of this Emergent dataset.

The task defined in [1] made use of conditional encoding network with two Bi-LSTMs to detect stance of tweets with some targets. They nurtured two separate LSTM networks, one for the tweet and another one for the target. The first hidden state of the LSTM for the target was initialized with the final hidden state of the LSTM for the tweet. The work described in [19] also utilized the stance detection dataset. They proposed four models which are based on *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 yielded the highest result among the results produced by

all these 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). They reported the highest accuracy of 80.8%.

The task defined in [23] presented a novel hierarchical attention model for stance detection. Especially they fostered a model to represent the document and their linguistic features with attention technique. Additionally, on the top of document representation, they made use of attention mechanism to estimate the importance of different linguistic features and learnt overlapping attention between the document and the linguistic information. The work described in [12] performed deep analysis of the three best participating systems of FNC-1. They showed that, the class wise and macro-averaged F1 score is the best way for validating the model for stance detection, as the shared task's standard evaluation metric is severely affected by the imbalanced class distribution of the dataset. We also followed these two metrics in addition to the standard metric provided by fake news challenge to evaluate our systems. Apart from these, the tasks on stance detection for fake news detection which made use of Fake news dataset could be found in [12,14,17,18]. It has been studied in other languages too like Arabic which could be found in [10].

3 Proposed Method

As stated earlier, We use both traditional supervised Machine learning and the deep learning approaches.

3.1 Feature Based Machine Learning Approach

We propose a supervised machine learning approach based on Support Vector Machine (SVM) [5,24] and Multilayer Perceptron (MLP) [2,8] to detect the stance between the headline and the body text. This model aims to develop a machine learning based system where different TE-based features are employed. The features include *Synonyms, Antonyms, Hypernyms, Hyponyms, Overlapping Tokens, Longest Common Overlap, Modal verbs, Polarity, Numerals, Named Entities, and Cosine Similarity*. The following points elaborate all these features.

Synonyms: Presence of synonymous words in two pieces of text snippets reveal that they are semantically similar, like *X bought Y* implies *X acquired Z% of the Y's shares*, because *acquire* is the synonym of *bought*. For each word in title, we search for the synonym of that particular word in the body text. If it is present then the feature value of "1" is assigned otherwise "0".

Antonyms: This is also a vital feature for detecting TE, which is a pervasive form of entailment trigger, where a word is replaced by it's antonym. Sentences like *T: "Oil price is surging"* does not imply *T: "Oil price is falling down."*. The feature value is computed in the reverse direction to what was followed in the synonym feature.

Hypernyms: Sometime certain concepts are generalized from one text to another, which leads to entailment. Like *T: "Beckham plays football."* entails H: "Beckham plays game.". So if there was *football* in headline and *game* in the body then we assign "1" otherwise "0".

Hyponyms: It is also observed that sometimes concepts are specialized, which, in turn, lead to entailment. Like *T: "Reptiles have scale."* entails H: "Snakes have scale.". So if Hyponyms of a word in title is present in body text, then the value of "1" is assigned, otherwise "0".

Overlapping Tokens: Overlapping tokens between two comparing text snippets can help in deciding entailment. The number of overlapping tokens between the headlines and body texts become the feature value of this feature.

Longest Common Overlap: Longest matching between two texts also matters a lot in taking the decision of Entailment. The value of this feature is computed as the maximum overlapping length between two pair of texts normalized by the number of words present in the body text.

Modal Verbs: It represents the presence of modal auxiliary verbs (like: can, should, must etc) which denote the possibility or necessity and sometimes lead to wrong entailment. Like *T: "The govt. may approve anti-corruption bill."* does not entails H: "The govt. approved anti corruption bill.". This feature is important for predicting the classes (like agree and discuss) between title and body text pairs. So, if it is present in any of the title or body text then the value of "0" is assigned and if it is present or absent in both the headline and body text then the value of "1" is assigned.

Polarity Features: These features determine whether the fact asserted or it's negation is going to occur, like (not, never, deny etc) are the polarity features. If we fully rely on lexical matching, the presence of negation word might cause problem in taking the decision for entailment. For example, *T: "The watchman denied that he was sleeping."* does not entail H: "The watchman was sleeping.". We compute this feature's value following the procedure as described in [21] for computing this polarity feature value.

Numerals: In some cases certain level of numeric calculation affect the entailment decision. Like *T: "3 men and 2 women were found dead in the apartment."* entails H: "5 people were found dead in apartment.". We assign the value of "1", if we found such matching, otherwise "0" is assigned.

Named Entity Information: Named Entities (NEs) (like, person, location, organization) between two text snippets sometime affect in entailment decision. We search for any matching pair of NEs between the headline and body text. A value of "1" is assigned if NEs match, otherwise a value of "0" is assigned.

Cosine Similarity: This is very popular and a benchmark similarity metric, widely used among the researchers over the years to find similarity between two pieces of texts. It could be a feature for entailment also. We pass headline and body separately to Universal Sentence Encoder (USE). USE produce vector

representation of headline and body. We compute the cosine similarity between these two vectors and assign as the value of this feature.

We apply different classifiers like SVM and MLP. The results obtained using these classifiers are shown in the results and discussion section (i.e in Sect. 5).

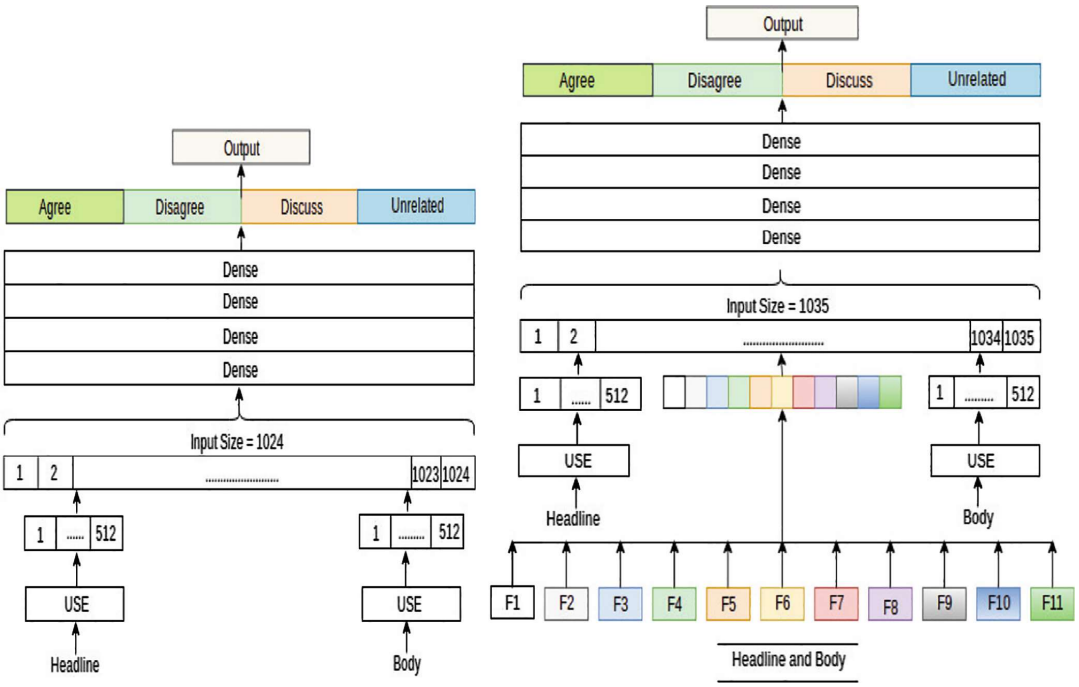
3.2 Deep Learning Based Approach

We propose two DL based approaches. One is based on the model defined in [20]. The difference from our propose model is in the representation layer. We apply the universal sentence encoder (USE) [4] to obtain the representations of titles and body texts, whereas they utilized Term Frequency-Inverse Document Frequency (tf-idf) for the same purpose. The another one is based on the first one but incorporated with ML based features values. The USE comes into two variants one exploiting the Transformer [25] architecture and the other one is based on the Deep Average Network (DAN) [13]. We make use of the Transformer based USE because it is observed that transfer learning from the transformer based sentence encoder performs better than transfer learning from the DAN encoder.

This model utilize the encoding sub-graph of the transformer architecture to produce the sentence/document's embedding. This kind of sub-graph provides context aware representation of words in a sentence by utilizing attention without hampering the ordering and the identity of other words. To obtain the fixed length sentence encoding vector, element-wise sum of the representations of each word is taken into account, which is further normalized by the square root of the length of the sentence.

The headline and body pairs are given to USE, which produces the representations for both headline and body, but separately. These representations are concatenated and subjected as inputs to feed-forward neural networks (dense layers) with ReLU activation function. Four such layers have been used, and this decision was taken in an empirical manner. We perform the experiments by taking the different number of layers. We obtain the highest performance with four layers. The outputs obtain from the fourth layer are given to a final layer with softmax activation function for final prediction. This layer predicts the class having the highest probability score. Architecture of the proposed model is shown in Fig. 1(a).

We modify our first approach to offer the second one. We incorporate the features values used in ML approach in the representation layer, as shown in the Fig. 1(b). We concatenate these values (computed for 11 features) with the representations obtained for headline and body from USE.



(a) Architectural diagram of the propose first DL system with Universal Sentence Encode

(b) Architectural diagram of the propose second DL system augmented with different ML features

Fig. 1. The architecture of the proposed two systems

4 Data

We make use of the benchmark dataset released in the shared task FNC-I for fake news detection through stance detection. The key statistics of the dataset are shown in Table 2. The dataset is highly imbalanced. So the task organizers[1] provide a standard metric to mitigate this problem. The metric is a weighted based evaluation system which comprises of two levels. In the first level, 25% weight is given for classifying headline and body text as related or unrelated and in the second level, 75% weight is given for classifying related pairs as agrees, disagrees, or discuss. The justification behind this is: classifying agrees, disagrees, or discusses is more difficult and relevant to fake news detection rather than just classifying headline–body pairs as related and unrelated.

Table 2. Number of instances, distribution of classes and average length of title and body in training and test set of FNC-1 dataset

Dataset	Example pairs	Classes				Avg. Length	
		Unrelated	Discuss	Agree	Disagree	Body	Title
Training	49972	0.73131	0.17828	0.0736012	0.0168094	369	11
Testing	25413	0.722032	0.17466	0.074833	0.027427	347	11

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

5 Experiments, Results and Discussions

In a nutshell, we perform three sets of experiments. The following subsections show the experimental procedures and results obtained.

5.1 ML Approach

In this experiment, We make use of 11 different features. We extract features values from headline and body text. We concatenate all these values, and given to classifier for classification. We make use of different classifiers and perform experiments. We obtain the remarkable results with Support Vector Machine (SVM) and Multi-layer Perceptron (MLP). We compute the FNC score using the evaluation metric provided in Fake News Challenge Competition. We obtain the FNC score of 72.13 and 56.04 for MLP and SVM, respectively. SVMs are well known good performer for two-class classification problem, even if it plays with a multi-class problem, it assumes the problem as two class problem. As our problem is a multi-class problem, this might be the reason for the poor performance of SVM compared to MLP. Results are shown in Table 4. Due to space constraints we are unable to show the confusion matrices for all of our proposed models. However, we show the confusion matrix for the best performing model.

Sensitivity Analysis of the Features: We perform feature ablation study to understand the contribution of each feature. The F1 scores are obtained by removing one feature after another. Results are shown in the Table 3. It shows that cosine similarity followed and Named Entities (because news titles/documents are full of different names) are the most contributing features in our experiment.

Table 3. Feature sensitivity analysis and effect of each feature on F1

Features removed	F1	Increment/decrement
None	**0.4777**	0
Synonyms	0.4757	−0.0020
Antonyms	0.4756	−0.0021
Longest common overlap	0.4679	−0.0098
Hypernym	0.4701	−0.0076
Hyponym	0.4724	−0.0053
NER	0.4653	−0.0124
Modality	0.4731	−0.0046
Overlapping tokens	0.4729	−0.0048
Numerals	0.4700	−0.0077
Polarity	0.4763	−0.0014
Cosine similarity	0.4364	**−0.0413**

5.2 Deep Learning

We propose two models which utilize the DL platform. The first one is based on USE and another one is where we incorporate the ML features values into USE based Model.

Universal Sentence Encoder Model: All the modern ML techniques fully rely on the vector representation of words, phrases and sentences. We obtain the embedding of title and body by utilizing transformer based USE. It takes lowercased Pen Tree Bank (PTB) tokenized[2] string of any length as input and produces the representation of fixed (512) dimensional embedding vector as output. We concatenate the representations of title and body text. The concatenated vector further send to four feed forward neural network layers. The representation obtained from the fourth feed forward neural network is further fed into a final layer for classification. The final layer predicts appropriate labels (Agree, Disagree, Discuss and Unrelated) having the maximum probability score. The architecture of this approach is shown in Fig. 1(a). We obtain the FNC score of 76.9 in this experiment.

Universal Sentence Encoder Model Incorporated with ML Features: In this experiment we inject the ML based features in the previous model. We concatenate the 11 features values with the vector representation for headline and body text. So the representation become a vector of 1035 dimension. This representation is further subjected as input to four feed forward neural network layers, placed one after another. The output obtained from the fourth feed forward neural network is given to a final layer with softmax activation function for final prediction. The architecture of this model is shown in the Fig. 1(b). We obtain the FNC score of 82.54 in this experiment.

Hyperparameters: We tune the hyperparameters in this experiment and mark the results and freeze the model having the hyperparameters which produces the best result. For example, the hidden layer size is tuned from 64 units to 256 units, batch size input from 64 to 256, dropout from 0.2 to 0.3. For all the experiments Rectified linear Unit (ReLU) activation function is used in all the feed-forward neural networks. The loss function and optimizer are cross entropy and ADAM respectively. The training iterations i.e. epoch was 50 for all the experiments and also we used checkpoint, to check the model's accuracy get increased or not, if it get increased only then the weights get updated. The final layer for the output prediction is with softmax activation function.

5.3 Comparison with the State of the Art and Other Prior Models

We perform an exhaustive comparison with previous three best participating systems on this dataset. The comparison is shown in Table 4. Apart from the FNC, we also compute the performance of our model using different modalities of evaluation metrics like *"overall F1"*, *"FNC"*, *"per class F1"* (for Agree, Disagree,

[2] https://nlp.stanford.edu/software/tokenizer.shtml.

Discuss and Unrelated). The DL model augmented with TE based features i.e. the third one has achieved the highest FNC score which outperforms the state-of-the-art reported in the literature by the FNC score of 0.5 margin. This model also beats the official baseline provided by the shared task organizers and also the score of the system [20] which we assumed as the baseline in this experiments. The result of this system is shown in the 3rd row (UCLMR system) in all formats. We also obtain the overall F1 score of 63.6%, and also the F1 score of 61.1% for agree class which is the highest among all the prior models. We also obtain the highest F1 score of 59.54% in disagree class with SVM classifier which is also the highest F1-score among all the previous system's score. However, we are not able to overcome the performance of human which is shown in row no 12 of the Table 4. This indicates there are lots of room that are available for improvement. The first participating system obtained an FNC score of 0.8204. The system is an ensemble of two 2D CNNs on word embedding of headline and body respectively. The resulting output is then fed into an MLP of three hidden layers and a decision tree based system composition of 5 features. Our two deep learning systems are based on the UCLMR system [20] with some modifications *viz: i. at the representation layer and ii. at hidden layer (that model was one feed-forward neural network, and we have four)*. In the third model, in addition to these we inject TE based ML features.

Table 4. The prior six best results and the results obtained by our proposed models on the dataset

SN	System	FNC-1	F1	Agree	Disagree	Discuss	Unrelated
Previous Models							
1	**TALOSCOMB(TREE+CNN)**	0.8204	0.582	0.539	0.035	0.760	0.994
2	**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	**SVM**	0.5604	0.4150	0.0073	**0.5954**	0.1084	0.9489
8	**MLP**	0.7213	0.4777	0.3462	0.0	0.6328	0.9315
9	**Univ_Sen_Enc**	0.769	0.570	0.436	0.187	0.712	0.944
10	**Univ_Sen_Enc_Features**	**0.8254**	**0.636**	**0.611**	0.214	0.746	0.972
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

5.4 Error Analysis

Every system has some pros and cons. Our system has some disadvantages too. We perform error analysis of our best performing system. We take miss-classified

Table 5. Confusion matrix obtained by the best performing DL approach on the test set

	Agree	Disagree	Discuss	Unrelated
Agree	1162	55	590	96
Disagree	233	149	258	57
Discuss	804	154	3323	180
Unrelated	92	33	395	17829

instances into account. We make a rigorous analysis of those instances and try to analysis why our model fails. The Table 5 shows the confusion matrix.

Our observations could be as follows:

• The dataset is enriched with Named Entities, phrasal verbs, and Multi-word expressions. The bodies are having multiple number of repetitive words, and sentences too which we need to take care separately in future. • The length variation between the title and the body is very high. • It is observed that the model is performing badly where headlines and body texts are of question answer type, i.e. Headline is question and the body text explaining it like answer. We need to investigate this in future.

6 Conclusion and Future Work

Detection of misinformation/fake news and fact checking is a very challenging and utmost task these days to mankind. In this paper, we try to mitigate this problem. The dataset released in Fake News Challenge for detecting fake news through stance detection serves this purpose. We relate this problem to TE as they are conceptually similar. We offer the systems which are based on ML, DL and combination of both. In ML, we foster the different TE-based features apply to different classifiers (SVM and MLP), and obtain remarkable results. In DL, we pose two models, one is USE based and the other one is the modified version of the USE model but augmented with TE based features. We make use of different performance measures i.e. *FNC*, *overall F1*, *per class F1 score* etc. Our proposed model outperforms the state-of-the-art system in *FNC* and *F1 score*, and *F1 score of Agree class* by the third DL model i.e. the model augmented with TE features. The system also outperforms the state-of-the-art *F1 score of Disagree class* by our SVM based model. In future we would like to: • enrich the propose models by incorporating many more lexical/syntactic/semantic based features and address the issues raised by the proposed models. • do more in-depth and rigorous error analysis of the previous three best participating systems to get more insights. • incorporate the external knowledge (i.e. world knowledge) into the existing system.

Acknowledgments. Asif Ekbal acknowledges Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

1. Augenstein, I., Rocktäschel, T., Vlachos, A., Bontcheva, K.: Stance detection with bidirectional conditional encoding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, pp. 876–885. Association for Computational Linguistics (2016)
2. Becerra, R., Joya, G., García Bermúdez, R.V., Velázquez, L., Rodríguez, R., Pino, C.: Saccadic points classification using multilayer perceptron and random forest classifiers in EOG recordings of patients with ataxia SCA2. In: Rojas, I., Joya, G., Cabestany, J. (eds.) IWANN 2013. LNCS, vol. 7903, pp. 115–123. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38682-4_14
3. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 632–642. Association for Computational Linguistics (2015)
4. Cer, D., et al.: Universal sentence encoder for English. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, pp. 169–174. Association for Computational Linguistics (2018)
5. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27:1–27:27 (2011)
6. Chen, Q., Zhu, X., Ling, Z.H., Inkpen, D., Wei, S.: Neural natural language inference models enhanced with external knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2406–2417. Association for Computational Linguistics (2018)
7. Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F., Flammini, A.: Computational fact checking from knowledge networks. PLoS One **10**(6), e0128193 (2015)
8. Costa, W., Fonseca, L., Körting, T.: Classifying grasslands and cultivated pastures in the brazilian cerrado using support vector machines, multilayer perceptrons and autoencoders. In: Perner, P. (ed.) MLDM 2015. LNCS (LNAI), vol. 9166, pp. 187–198. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21024-7_13
9. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognising textual entailment challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 177–190. Springer, Heidelberg (2006). https://doi.org/10.1007/11736790_9
10. Darwish, K., Magdy, W., Zanouda, T.: 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, 31 July–03 August 2017, pp. 145–148 (2017)
11. Ferreira, W., Vlachos, A.: 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, pp. 1163–1168. Association for Computational Linguistics (2016)

12. Hanselowski, A., et al.: 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, pp. 1859–1874. Association for Computational Linguistics (2018)
13. Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, pp. 1681–1691. Association for Computational Linguistics (2015)
14. Thorne, J., Chen, M., Myrianthous, G., Pu, J., Wang, X., Vlachos., A.: Fake news stance detection using stacked ensemble of classifiers. In: Proceedings of the EMNLP Workshop on Natural Language Processing meets Journalism, Copenhagen, Denmark, pp. 80–83 (2017)
15. MacCartney, B., Grenager, T., de Marneffe, M.C., Cer, D., Manning, C.D.: Learning to recognize features of valid textual entailments. In: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference (2006)
16. MacCartney, B., Manning, C.D.: Natural logic for textual inference. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, RTE 2007, Stroudsburg, PA, USA, pp. 193–200. Association for Computational Linguistics (2007)
17. Mohtarami, M., Baly, R., Glass, J., Nakov, P., Màrquez, L., Moschitti, A.: 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, pp. 767–776. Association for Computational Linguistics (2018)
18. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. In: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, pp. 3391–3401. Association for Computational Linguistics (2018)
19. Pfohl, S., Triebe, O., Legros, F.: Stance detection for the fake news challenge with attention and conditional encoding (2017)
20. Riedel, B., Augenstein, I., Spithourakis, G.P., Riedel, S.: A simple but tough-to-beat baseline for the fake news challenge stance detection task. CoRR abs/1707.03264 (2017)
21. Saikh, T., Ghosal, T., Ekbal, A., Bhattacharyya, P.: Document level novelty detection: textual entailment lends a helping hand. In: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), Kolkata, India, pp. 131–140. NLP Association of India, December 2017
22. Silverman, C.: Lies, damn lies and viral content (2015)
23. Sun, Q., Wang, Z., Zhu, Q., Zhou, G.: Stance detection with hierarchical attention network. In: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, pp. 2399–2409. Association for Computational Linguistics (2018)
24. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995). https://doi.org/10.1007/978-1-4757-2440-0
25. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008 (2017)

26. Vlachos, A., Riedel, S.: Fact checking: task definition and dataset construction. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, MD, USA, pp. 18–22. Association for Computational Linguistics (2014)
27. Wang, W.Y.: "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), pp. 422–426. Association for Computational Linguistics (2017)