

# Tweet Stance Detection Using an Attention based Neural Ensemble Model

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono

Department of Computer Science and Engineering

Toyohashi University of Technology, Toyohashi, Aichi, Japan

{aymun, nowshed}@kde.cs.tut.ac.jp, and aono@tut.jp

## Abstract

Stance detection in twitter aims at mining user stances expressed in a tweet towards a single or multiple target entities. To tackle this problem, most of the prior studies have been explored the traditional deep learning models, e.g., LSTM and GRU. However, in compared to these traditional approaches, recently proposed densely connected Bi-LSTM and nested LSTMs architectures effectively address the vanishing-gradient and overfitting problems as well as dealing with long-term dependencies. In this paper, we propose a neural ensemble model that adopts the strengths of these two LSTM variants to learn better long-term dependencies, where each module coupled with an attention mechanism that amplifies the contribution of important elements in the final representation. We also employ a multi-kernel convolution on top of them to extract the higher-level tweet representations. Results of extensive experiments on single and multi-target stance detection datasets show that our proposed method achieves substantial improvement over the current state-of-the-art deep learning based methods.

## 1 Introduction

Tweet stance detection is the task of automatically determining the stance of a tweet whether the tweet is in favor of, against, or none towards a target (Mohammad et al., 2017). We can consider it as a sub-domain of sentiment analysis. However, the goal of sentiment analysis is to classify the polarity of a tweet sentiment based on its contents, whereas identification of stance is dependent on the specific target. For example, Figure 1 depicts the stance of a tweet towards different targets.

Stance detection in twitter poses unique challenges to the research community since tweets are short and informal user-generated text, which usually tend not to follow the grammatical rules.

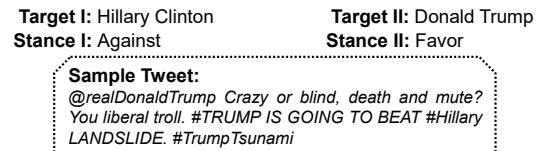


Figure 1: Example of target-specific stance detection.

Moreover, tweets contain plenty of idiosyncratic abbreviations as well as other twitter specific syntaxes such as #hashtags and emoticons. To address the challenges of stance detection in twitter, Mohammad et al. (2016) presented a tweet stance detection task that focused on a single target in SemEval-2016. Top performing systems in this task proposed several deep learning based approaches by using CNN (Wei et al., 2016), RNN (Zarrella and Marsh, 2016), and so on.

Later, (Du et al., 2017) utilized the target-augmented embeddings in an attention based neural network, whereas (Zhou et al., 2017) proposed an attention mechanism at the semantic level in the bidirectional GRU-CNN structure to perform target-specific stance detection on tweets. More recently, (Dey et al., 2018) proposed a two-phase LSTM based model with attention and (Wei et al., 2018b) proposed an end-to-end neural memory model via target and tweet interactions.

By considering the dependency of related targets, Sobhani et al. (2017) introduced a multi-target stance detection (MTSD) task and proposed an attentive encoder-decoder network to capture the dependencies among stance labels regarding multiple targets. Later, (Wei et al., 2018a) proposed a dynamic memory-augmented network that utilized a shared external memory to capture and store multi-targets stance indicative clues.

However, most of the related work of tweet stance detection explored the traditional deep learning models in their methods. In this paper,

we propose a neural ensemble method that combines the attention based state-of-the-art densely connected Bi-LSTM and nested LSTMs models with the multi-kernel convolution in a unified architecture. Experimental results on both the single and multi-target benchmark stance detection datasets demonstrate the efficacy of our method over the state-of-the-art deep learning based methods discussed above.

The rest of the paper is structured as follows: In **Section 2**, we introduce our proposed stance detection framework. **Section 3** includes experiments and evaluations as well as the comparisons with the state-of-the-arts to show the effectiveness of our proposed method. Some concluded remarks and future directions of our work are described in **Section 4**.

## 2 Proposed Stance Detection Framework

In this section, we describe the details of our proposed neural ensemble model (PNEM) for twitter stance detection. Figure 2 depicts an overview of our proposed framework.

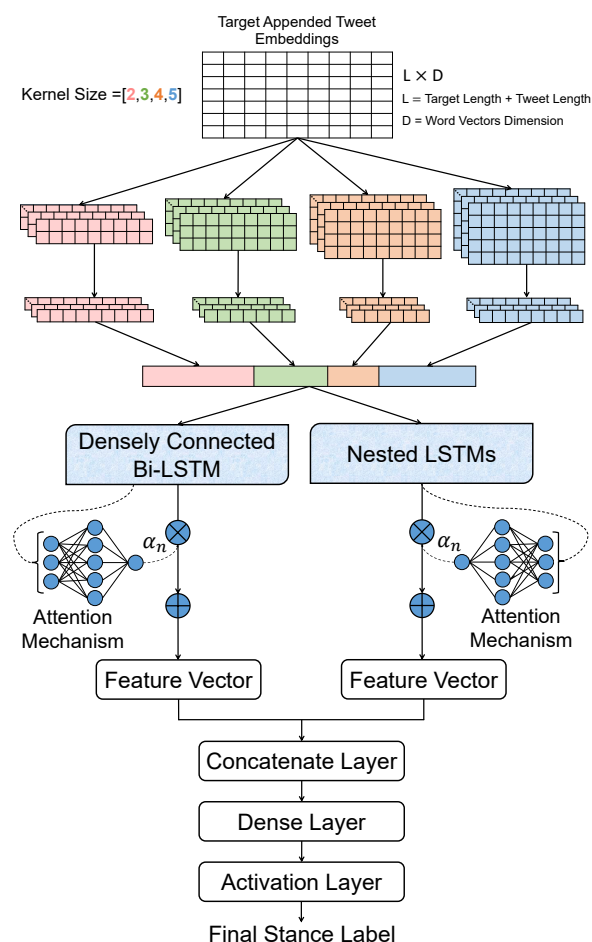


Figure 2: Proposed stance detection framework.

At first, we utilize the multi-kernel convolution filters to extract higher-level feature sequences from the target appended tweet embeddings. These feature sequences are fed into the attention based densely connected Bi-LSTM and nested LSTMs to learn long-term dependencies. Final representations of these modules are concatenated and pass to the stance prediction module to determine the stance label. Next, we describe each component elaborately.

**Embedding Layer:** Prior works already established the significance of target information for stance detection. To integrate the target information, we generate a unified word vector matrix by concatenating the vector representations of the target and tweet. The dimensionality of the matrix will be  $L \times D$ , where length  $L$  is the sum of the target length and tweet length, and  $D$  denotes the word-vector dimension. We utilize a pre-trained word embedding model for obtaining the vector representation of words.

**Multi-Kernel Convolution:** In our multi-kernel convolution, we adopt the idea proposed by (Kim, 2014) to extract the higher-level features. The input of this module is the target appended tweet matrix generated in the embedding layer. We then perform the convolution on it by using a filter. We apply multiple convolutions based on four different kernel sizes, i.e., the size of the convolution filters: 2, 3, 4, and 5. After performing convolutions, each filter generates the corresponding feature maps and a max pooling function is then applied to generate a univariate feature vector. Finally, the feature vectors generated from each kernel are concatenated to form a single high-level feature vector.

**Densely Connected Bi-LSTM:** With the emerging trend of deep learning, LSTM based models are the most popular for sequential tasks. Recently, the densely connected structure of LSTM models gets attention among the researchers (Li et al., 2018; Wu et al., 2017) that enable the effective connection from lower to upper layers features without any loss of information on lower-layer features thus alleviate the vanishing-gradient and overfitting problems effectively.

In our proposed framework, we utilize the densely connected Bi-LSTM (Ding et al., 2018) (DC-Bi-LSTM) model. A DC-Bi-LSTM model consists of multiple Bi-LSTM layers, where representation of each layer is estimated by concatenat-

ing its hidden states and all the preceding layers' hidden states. Hence, for the first Bi-LSTM layer, the input is a higher-level features sequences generated from multi-kernel convolution (MKC) and the output is  $\{h^1 = h_1^1, h_2^1, \dots, h_s^1\}$ . For the second Bi-LSTM layer, the input is the concatenation of higher-level feature sequences from MKC and the output from first Bi-LSTM layer to generate the corresponding output. Rest of the layers are processed accordingly. We can define the above process as follows:

$$\begin{aligned} h_t^l &= [\overrightarrow{h_t^l}; \overleftarrow{h_t^l}], h_t^0 = \text{MKC feature sequence}, \\ \overrightarrow{h_t^l} &= \text{lstm}(\overrightarrow{h_{t-1}^l}, M_t^{l-1}), \\ \overleftarrow{h_t^l} &= \text{lstm}(\overleftarrow{h_{t+1}^l}, M_t^{l-1}), \\ M_t^{l-1} &= [h_t^0; h_t^1; \dots; h_t^{l-1}]. \end{aligned}$$

Therefore, from a  $L$  layer DC-Bi-LSTM model, the output is  $\{h^L = h_1^L, h_2^L, \dots, h_s^L\}$ .

**Nested LSTMs:** The nested LSTMs (NLSTMs) architecture (Moniz and Krueger, 2017) creates a temporal hierarchy of memories that achieved significant improvement over the single-layer or stacked LSTM architectures to learn longer-term dependencies.

In NLSTMs, the LSTM memory cells have access to their inner memory, where they can selectively read and write relevant long-term information. While the value of the outer memory cell in the LSTM is estimated as  $c_t^{outer} = f_t \odot c_{t-1} + i_t \odot g_t$ , memory cells of the NLSTMs use the concatenation ( $f_t \odot c_{t-1}, i_t \odot g_t$ ) as input to an inner LSTM (or NLSTM) memory cell, and set  $c_t^{outer} = h_t^{inner}$ . Therefore, in compared to the LSTM and stacked LSTM, the inner memories of NLSTMs operate on longer time-scales and effectively capture the context information from the input texts.

**Feed-Forward Attention:** Recently, the attention mechanism has been introduced in the neural network models for effectively modeling the long-term dependencies by enabling the model to learn what to attend based on the input text (Vaswani et al., 2017). In order to amplify the contribution of important elements in the final representation of both the DC-Bi-LSTM and NLSTMs module, we employ a feed-forward attention mechanism (Raffel and Ellis, 2015) to aggregate all the hidden states according to their relative importance weight.

**Stance Prediction and Model Training:** We concatenate the final tweet representation from the attention based DC-Bi-LSTM and NLSTMs module and pass it to a fully connected softmax layer for stance detection. We consider cross-entropy as the loss function and train the model by minimizing the error, which is defined as:

$$E(x^{(i)}, y^{(i)}) = \sum_{j=1}^k 1\{y^{(i)} = j\} \log(y_j^{\sim(i)})$$

where  $x^{(i)}$  is the training sample with its true label  $y^{(i)}$ .  $y_j^{\sim(i)}$  is the estimated probability in  $[0, 1]$  for each label  $j$ .  $1\{condition\}$  is an indicator which is 1 if true and 0 otherwise. We use the stochastic gradient descent (SGD) to learn the model parameter and adopt the Adam optimizer (Kingma and Ba, 2014).

## 3 Experiments and Evaluations

### 3.1 Model Configuration

In the following, we describe the set of parameters that we have used in our proposed neural ensemble model (PNEM) during experiments. We used the 300-dimensional fastText embedding model pre-trained on Wikipedia with skip-gram (Bojanowski et al., 2017) to initialize the word embeddings in the embedding layer. For the multi-kernel convolution, we employed 4 kernel sizes (2,3,4,5), and the number of filters was set to 600. In our model, DC-Bi-LSTM module contains 5 layers and NLSTMs module contains 2 layers. We trained all models for max 45 epochs with a batch size of 32 and an initial learning rate 0.001 by Adam optimizer.  $L2$  regularization with a factor of 0.01 was applied to the weights in the softmax layer. In this paper, we reported the results based on these settings. Unless otherwise stated, default settings were used for the other parameters.

To preprocess the data, we removed the stop words based on NLTK's standard stoplist, special characters removal, and performed hashtag segmentation according to (Baziotis et al., 2017).

### 3.2 Single Target Stance Detection

**Dataset and Setup:** To validate the effectiveness of our proposed method for the single target stance detection, we made use of a widely popular benchmark twitter dataset used in the SemEval-2016 Task 6-A (Mohammad et al., 2016). The training set consists of 2914 tweets and the test set consists

of 1249 tweets relevant to 5 targets. Each tweet was annotated as *Favor*, *Against* or *None* towards the specific target.

Following the SemEval-2016 Task 6-A benchmark, we employed the macro-average of F1-score for the *Favor* and *Against* stance classes as the evaluation measure. To estimate it, we used the evaluation script provided by the organizer.

**Results and Analysis:** We divided the whole dataset across targets and trained the model accordingly. We used 5% of the training samples as the validation set. The summarized experimental results of our proposed neural ensemble model (PNEM) on single target stance detection are presented in Table 1.

At first, we report the results based on the baseline, which is the combination of CNN and LSTM (Zhou et al., 2015) and obtained competitive performances on several text classification tasks. Next, we report the results of our proposed PNEM model. It showed that our PNEM method outperformed the baseline by a large margin. In order to estimate the effect of each component of our model, we performed the component ablation study on our proposed model. In this regard, we removed one component each time and repeated the experiment. From the results, it can be observed that when removing target embedding, multi-kernel convolution (MKC), attention mechanism (ATT), NLSTMs with corresponding attention (NLSTMs+ATT), and DC-Bi-LSTM with corresponding attention (DC-Bi-LSTM+ATT), the results decreased by 0.99%, 4.98%, 1.68%, 4.14%, and 1.11%, respectively. Thus deduced the importance of each of the component in our model.

Method	$F_{favor}$	$F_{against}$	$F_{avg}$
CNN+LSTM	59.36	74.93	67.15
<b>PNEM</b>	<b>66.56</b>	<b>77.66</b>	<b>72.11</b>
– Target Embedding	64.87	77.36	71.12
– MKC	60.49	73.78	67.13
– ATT	63.94	76.92	70.43
– (NLSTMs+ATT)	61.51	74.43	67.97
– (DC-Bi-LSTM+ATT)	64.47	77.52	71.00

Table 1: Comparative performance with different experimental settings for single target stance detection. The best results are highlighted in boldface.

Moreover, the comparative performance of our proposed model, PNEM with the state-of-the-art methods are presented in Table 2. It showed

that our model gained 3.13% and 4.29% improvement while compared with the SemEval-2016 baseline (SVM-ngrams) and the best performing system MITRE, respectively. Furthermore, in comparison with the related deep learning based methods, PNEM achieved at least 1.07% and at best 3.44% improvement. Overall, our proposed PNEM method greatly surpassed the previous works. The results show that attentive neural ensemble model could benefit stance detection.

### 3.3 Multi-Target Stance Detection

**Dataset and Setup:** In order to assess our method for the multi-target stance detection, we made use of a benchmark twitter dataset (Sobhani et al., 2017), where each tweet is annotated with two stance labels towards two targets of a pair. The tweets in this dataset are related to the 2016 US election. Overall, the dataset contains 4455 tweets for the three target-pairs including Hillary Clinton - Bernie Sanders, Hillary Clinton - Donald Trump, and Ted Cruz - Donald Trump. The train, development, and test set contains 3119, 446, and 890 tweets, respectively.

As the evaluation measure, we used a similar kind of approach reported in the dataset paper (Sobhani et al., 2017). Following this, the  $F_{avg}$  of each target is estimated according to the SemEval-2016 Task-A benchmark (Mohammad et al., 2016). Then, the  $F_{avg}$  of the two targets of a target-pair is averaged to estimate the average score of a target-pair. Finally, the average scores of all the target-pairs are averaged to estimate the overall score.

**Results and Analysis:** We divided the whole dataset based on each target-pairs. The dataset is then separated into two parts for each target in the target-pair. We trained and evaluated the model for each target in the target-pair and combined the results to estimate the overall performance.

We used a similar kind of baseline (CNN+LSTM) for comparison that we used in the single target stance detection as well as compared with the state-of-the-art deep learning based methods. The comparative results of our proposed PNEM model on multi-target stance dataset are presented in Table 3. It showed that our method surpassed the baseline by a large margin and gained 3.91% and 1.99% improvement over the state-of-the-art methods Seq2Seq and DMAN, respectively.



Method	Overall			Target				
	$F_{favor}$	$F_{against}$	$F_{avg}$	Atheism	Climate Change	Feminist Movement	Hillary Clinton	Legal. of Abortion
<b>PNEM</b>	<b>66.56</b>	<b>77.66</b>	<b>72.11</b>	67.73	44.27	<b>66.76</b>	60.28	64.23
<i>SemEval-2016 Baseline and Related Deep Learning based Methods</i>								
SVM-ngrams	62.98	74.98	68.98	65.19	42.35	57.46	58.63	66.42
MITRE	59.32	76.33	67.82	61.47	41.63	62.09	57.67	57.28
<i>n</i> -grams+embeddings	-	-	70.30	<b>68.30</b>	43.80	58.40	57.80	<b>66.90</b>
TGMN-CR	65.52	76.55	71.04	64.60	43.02	59.35	<b>66.21</b>	66.21
T-PAN	-	-	68.84	61.19	<b>66.27</b>	58.45	57.48	60.21
AS-biGRU-CNN	-	-	69.42	66.76	43.40	58.83	57.12	65.45
TAN	-	-	68.79	59.33	53.59	55.77	65.38	63.72

Table 2: (Single target) Comparative performance of our model against the SemEval-2016 official baseline (SVM-ngrams) (Mohammad et al., 2016) and related deep learning based methods including MITRE (SemEval-2016 best performing system) (Zarrella and Marsh, 2016), *n*-grams+embeddings (Mohammad et al., 2017), TGMN-CR (Wei et al., 2018b), T-PAN (Dey et al., 2018), AS-biGRU-CNN (Zhou et al., 2017), and TAN (Du et al., 2017). The best results are highlighted in boldface.

Method	Overall	Target-Pair		
		Clinton Sanders	Clinton Trump	Cruz Trump
CNN+LSTM	46.03	43.02	47.35	47.71
Seq2Seq	54.81	54.72	56.60	53.12
DMAN	56.73	56.25	<b>60.30</b>	53.64
<b>PNEM</b>	<b>58.72</b>	<b>57.74</b>	60.05	<b>58.36</b>

Table 3: (Multi-target) Comparative performance of our model against the baseline (CNN+LSTM) and state-of-the-art deep learning based methods including Seq2Seq (Sobhani et al., 2017) and DMAN (Wei et al., 2018a). The best results are highlighted in boldface.

## 4 Conclusion

In this paper, we proposed an attention based neural ensemble model for the target-specific tweet stance detection. The main contribution of our unified model is to learn the contextual information effectively which in turns improved the stance detection performance and outperformed the state-of-the-art deep learning based methods for both the single and multi-target stance detection benchmark datasets.

In the future, we have a plan to leverage external knowledge and generalize our model for target-independent stance detection in the same domain.

## Acknowledgments

The part of this research is supported by MEXT KAKENHI, Grant-in-Aid for Scientific Research (B), Grant Number 17H01746 and by grants from the KDDI Foundation.

## References

- Christos Baziotis, Nikos Pelekis, and Christos Douk-eridis. 2017. Datastories at SemEval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, pages 747–754, Vancouver, Canada. ACL.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)*, 5:135–146.
- Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical stance detection for twitter: A two-phase lstm model using attention. In *Proceedings of the 40th European Conference on Information Retrieval (ECIR)*, pages 529–536. Springer.
- Zixiang Ding, Rui Xia, Jianfei Yu, Xiang Li, and Jian Yang. 2018. Densely connected bidirectional lstm with applications to sentence classification. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 278–287. Springer.
- Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3988–3994.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. ACL.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

- Dexu Li, Yimin Chen, Mingke Gao, Siyu Jiang, and Chen Huang. 2018. Multimodal gesture recognition using densely connected convolution and blstm. In *Proceedings of the 24th International Conference on Pattern Recognition (ICPR)*, pages 3365–3370. IEEE.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*, pages 31–41.
- Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):26.
- Joel Ruben Antony Moniz and David Krueger. 2017. Nested lstms. In *Proceedings of the 9th Asian Conference on Machine Learning (ACML)*, pages 530–544. Springer.
- Colin Raffel and Daniel PW Ellis. 2015. Feed-forward networks with attention can solve some long-term memory problems. *arXiv preprint arXiv:1512.08756*.
- Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Volume 2, Short Papers*, pages 551–557. ACL.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008.
- Penghui Wei, Junjie Lin, and Wenji Mao. 2018a. Multi-target stance detection via a dynamic memory-augmented network. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1229–1232. ACM.
- Penghui Wei, Wenji Mao, and Daniel Zeng. 2018b. A target-guided neural memory model for stance detection in twitter. In *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. Pkudblab at SemEval-2016 task 6: A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval), NAACL-HLT*, pages 384–388.
- Chuhan Wu, Fangzhao Wu, Yongfeng Huang, Sixing Wu, and Zhigang Yuan. 2017. Thunngn at IJCNLP-2017 task 2: Dimensional sentiment analysis for Chinese phrases with deep lstm. *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP), Shared Tasks*, pages 47–52.
- Guido Zarrella and Amy Marsh. 2016. Mitre at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463.
- Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A c-lstm neural network for text classification. *CoRR*, abs/1511.08630.
- Yiwei Zhou, Alexandra I Cristea, and Lei Shi. 2017. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *Proceedings of the 18th International Conference on Web Information Systems Engineering (WISE)*, pages 18–32. Springer.