MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection

Guido Zarrella and **Amy Marsh**

Problem and Motivation:
This work was submitted by two researchers in MITRE corporation, as a SemEval-2016 Task 6. The Problem was to Detect Stance in the Tweets. Twitter and other social media are platforms where people express their opinion or stance (implicitly or explicitly) towards a topic. In stance detection they attempt to measure how a writer's opinion is expressed in spontaneous, unstructured messages rather than the explicit prompts of formal opinion votes. The challenges arise when the stance are formulated figuratively, language which will be difficult for machines to unfold. It is even more difficult for the machine to understand rhetoric statements which portray sarcasm, irony, analogy and metaphor. Where a human reader would include personal experience and knowledge to infer the broader context.

Proposed Method:
The authors approach is to solve the problem by employing a recurrent neural network which initialized from pre-trained features learned in repeated attempts to encode corpus using very less external supervision. The pre-training is done using weak supervision, meaning the data which was trained on contained no labels so that the model automatically learns the useful representation of words and input sequences. If the weights of a already trained model is obtained somehow, the model can be initialized with this weight and fine tuned with the current data to tackle the problem. In that way, the new data required to solve this problem would be less. That is the main idea of transfer learning. However the training in the later stage does require labels in this case. On the whole, with minimal effort, the task can be completed. It is said that, this method scored the highest in the SemEval-2016 Task 6-A.

Evaluation:
Figure 1 represent the sequential workflow to solve the stance detection problem. Inputs are the tokens encoded in one-hot fashion. A sequence of these inputs are given to the 256-dimension embedding layer which in turn feeds into a recurrent layer containing Long Short Term Memory(LSTM). The output of this layer is connected to Rectified Linear Units(ReLU) with 90% drop out. Finally, this is given to 3 dimensional softmax layer to represent the classes, FAVOUR, AGAINST, or NONE.

Author's do not mention any baseline in the paper. But similar papers in the same task mention SVM classifier with n-gram features as the baseline. The system depicted in the paper was evaluated using **F1 score**. Data set used, was obtained from the SemEval task itself.

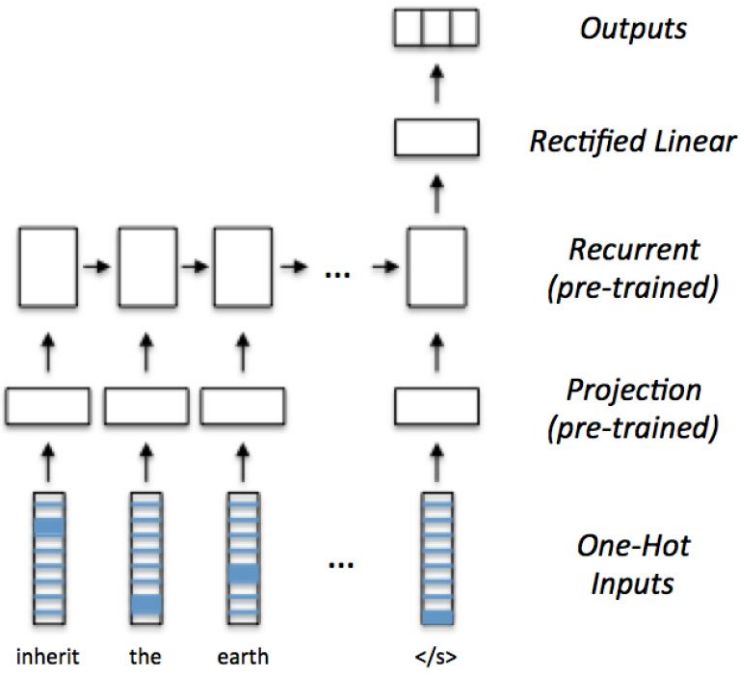

Figure 1: Workflow graph for stance detection

Results :

The system experimented achieved an average F1 score of 67.8 on the FAVOR and AGAINST classes. This was the best result obtained in the SemEval task. When cross validation was used with the system, the average F1 score of 71.1 was obtained. According to the author's observation, Majority class outperformed the corresponding minority class. So to quantify this observation, the author finds the correlation score. There was positive correlation of 0.67 between F1 score for a class and the number of training examples representing each class.

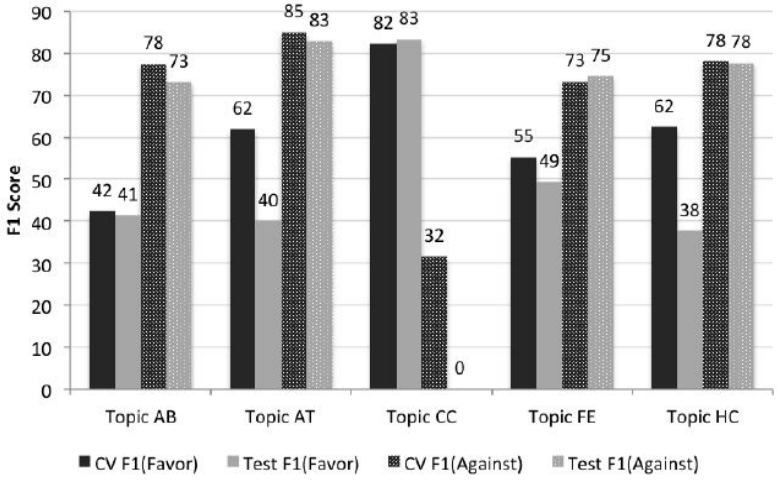

Figure 2: F1 scores for each topic and class on both cross-validation and test conditions

Conclusion:
The author concludes by saying that the system automatically determines stance of a tweet with limited training data, as it made use of transfer learning. Also mentions that the results obtained during hashtag prediction and skip-gram task are used as feature for this stance detection problem. Where in, domain relevant hashtags enhanced the performance of the system. As a future step, author mentions about investigating other techniques for identifying relevant hashtags, with goal to maximise the diversity of opinions represented in the dataset.

My opinion and Insights:
After reading the paper, I had to comment on the way the paper was written. Paper is short in simple words, direct to the point, describing the implementing strategies. But, I feel it was sluggish. There is no, clear reasoning as to why, some parameters are chosen. A comparison of different values for the parameter would have given more insight. There is no mentioning of any such analysis made in the experiment section. Moreover there is no comparison made with baseline. As a consequence the paper has not many comparison tables or figures. The results shown are only for the current model for 5 different categories in the data set. The abbreviations used in the graph are not explained, it becomes hard to understand what values is for which category. Besides, I came to conclusion that graph is articulating about the categories performance only after going through another paper of the same challenge. Figure 2, shows performance of this model on 5 categories in the dataset. But the abbreviation for the categories are not defined nor it is given in the task description on the SemEval website.

The Dataset looks imbalanced. In climate change category, only 4% of tweets are AGAINST, compared to 54% in FAVOR. Author could have oversampled the data in the initial stage itself and scrutinized the performance of the system.

The author mentions that with minimum data, this task can be performed. But again, the initial pretraining has to be done with huge amount of this same(or similar tweet) data only. If we use a question answering dataset, this model might not perform so well. Hence, on the whole to get good performance author is using large dataset.

The ultimate merit of this paper is that, Idea of transfer learning was successfully implemented showing that, transfer learning can be used effectively for text analytics as well.