Report : Where is your Evidence: Improving Fact-checking by Justification Modeling

1. Problem Statement

Fake news statements are typically intended to deceive or mislead readers. For example, in the past, fake news articles were written to increase profit by directing web traffic to “Clickbait” content, which was often designed to go viral by targeting controversial topics. The majority of people nowadays consume their content online, and as a consequence, many rely on social media as their news source. Non-expert users generally have difficulty determining the veracity of news articles and claims, making social media fertile ground for the spread of fake news, which has emerged as a new type of cybersecurity threat. As a result, it has become critical to provide tools for automated detection of fake news.

William Yang Wang’s paper **"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection** introduced a large dataset of claims from POLITIFACT, associated metadata, and for each claim and the verdict class for the development of such automatic models. The paper **Where is your Evidence: Improving Fact-checking by Justification Modeling** introduces the LIAR-PLUS dataset. The paper shows that modeling extracted justification in conjunction with the claim (and metadata) provides a significant improvement regardless of the machine learning model used (feature-based or deep learning) - both in a binary classification task (true, false) and in a six-way classification task (pants on fire, false, mostly false, half-true, mostly true, true).

2. State of the art Solutions available

LSTMs and transformer-based language models are viewed as state-of-the-art techniques available for Natural Language Process tasks.

Long Short-Term Memory (LSTM) networks are a modified version of recurrent neural networks, which makes it easier to remember past data in memory. Yet the information at the very beginning of the sequence can still be forgotten and important phrases or information at the beginning of a statement may be overlooked. Bidirectional LSTM (also known as Bi-LSTM) solves this problem. In parallel with the main LSTM, a second LSTM is run which reads the sequence from the end. Thus, the model will treat all words equally, regardless of their location. The paper introducing the LIAR PLUS dataset suggests two techniques of modeling the claim along with justification using Bi-LSTMs. In the first one, the justification is just concatenated to the statement and passed to a single BiLSTM. In the second one, a dual/parallel architecture is used where one BiLSTM reads the statement, and another one reads the justification (architecture denoted as P-BiLSTM).

The BERT model uses the Transformer architecture with multi-headed attention - essentially producing multiple sets of representations of the input sequence, each one encoding a different characteristic of the input. The pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks. Pre-trained

representations reduce the need for many heavily engineered task-specific architectures. BERT has made significant breakthroughs in various natural language understanding tasks and is hence tested as a state-of-the-art transformer-based technique as part of our implementation.

3. Research Gaps and our ideas

Extraction of evidence/justification for such a data mining task would usually require human-written articles or extracting information from sources whose authenticity is known or verified by human intervention making it difficult to make the complete task autonomous. Additionally, the simple method used for extraction of the justification provided by humans in the fact-checking articles may lead to noisy data on both extremes where either the claim may just be repeated or evidence is completely missed out on. Implementing a better method to extract justifications is part of the future work that needs to be done and currently seems to be out of the scope of our study.

Another issue is a robust method to model justification and metadata along with the statement. We tried out various approaches for this, but a significant increase in performance was not observed when using the architectures and methods specified in the paper. This may be due to:

a. Nonoptimal way of concatenating/combining justification, metadata, and other additional information along with the statement.
b. Bad tuning of hyperparameters of the model.

A recent approach modeled each input feature (statements, justifications, subject, etc) using BiLSTMs. Then, they are passed through a convolutional layer followed by a max-pooling layer. These features were later concatenated to form a single fully connected layer. Then, depending upon the classification required, it was either passed through a sigmoid or softmax activation function to get the final outputs. The above model gave us the highest accuracy among all.

For the BERT-based approach, we separately pass the statement, justification, and metadata (string concatenation of the attributes representing metadata) to BERT. The outputs are then concatenated and then each element is added with the credit score, a derived attribute that represents the previous track record of the news source. This tensor is passed into a fully connected layer which then gives us our output for the binary and six-way classification tasks respectively. It should be noted that here we have used BERT as a feature extractor instead of finetuning it for our specific task. We did so as we lacked sufficient GPUs to run the training in parallel, resulting in a longer training time than anticipated. Nonetheless, this model produces promising results on both tasks.

Submitted by:
Nalin Deepak : 2018A7PS0223P
Kalit Naresh Inani : 2018A7PS0207P
Yash Bansal : 2019A7PS0484P
Harsh Mahajan : 2019A7PS0036P