CS 224U Experimental Protocol: Fake News Detection

Meng Tang
Stanford University
mengtang@stanford.edu

Alex Tsun
Stanford University
alextsun@stanford.edu

David Xue
Stanford University
dxue@stanford.edu

1. Hypothesis

Our exploration in this project is guided by the following hypotheses:

1. Natural language processing (NLP) models can find subtle language patterns based solely on statements, without any contextual knowledge base, to detect *inauthenticity*.

2. Deep learning architectures for detecting fake news using only statements can be improved by combining contextual metadata with deep learning language features.

3. Jointly modelled statements and contextual metadata will significantly improve the performance on fake news detection.

2. Dataset

2.1. LIAR

For our project, we are using the LIAR [2] dataset, which collects around 12.8K labeled short statements from POLITIFACT.COM. We select the dataset because besides contextual statements, it contains extra metadata such as the speaker of the statements as well as other information, which is helpful to fulfill our exploration on the third hypothesis mentioned above. However, the dataset does has drawbacks with a limited number of data, which, unfortunately, is shared by most available datasets for fake news detection, due to the difficulties embedded in labeling news authenticity of news. Although more qualified labeled benchmark datasets seem urgently needed for fake news detection, we are focusing on exploring the models that are effective for fake news detection using the LIAR dataset.

Specifically, the dataset contains 6 different labels ranging across different levels of truth: "pants-fire", "false", "barely-true", "half-true", "mostly-true", and "true". Since the bins seem a bit too fine-grained, and may be too close, we merge the first three labels into "false" and the last three into "true", for binary classification. The dataset consists of a predetermined training set of around 10,000 statements, and a validation and test set around 1300 statements. Each row contains a statement, the speaker, their job, state, party, and other features that can be used in addition to the statement itself.

The dataset is unfortunately a bit small for deep learning approaches, but the benefit is that we can also try new models and do hyperparameter search more efficiently.

3. Metrics

We plan to tackle this problem in a few ways.

1. **Classification (Categorical and Binary):** We plan to use standard metrics such as precision, recall, and F1-Score. We also plan to visualize and examine the confusion matrix.

2. **Regression:** We were going to convert the labels to real numbers between 0 and 1. 0 being "pants-fire", "0.2" being "false", ..., and "1" being true. We could then measure mean squared error, or round our predictions and treate it as binary classification. In this case, we can use the same metrics as above. We plan will report this metric in future work.

4. Models

1. **Logistic Regression**
   For our baseline, we did binary classification using logistic regression. To featurize our data, we fed the statement into a sentence2vec model to get a 100-dimensional real vector. In addition to the statement data, we also included some auxiliary data such as the the number of "pants-fire" statements the speaker made, the number of "false" statements the speaker made, etc. (included in the dataset). We however did not include features such as speaker, party, and context, as they were additional text features that needed to be vectorized (context) or binarized (finite set of speakers and poltical parties).

2. **Deep Learning: LSTM, CNN**
   We implemented two deep learning approaches,

specifically LSTM- and CNN-based models. For the LSTM-based model, news text was fed to the LSTM and the output was added to a condensed representation of the metadata. For the CNN-based model with 128 filters each of size 2,5 and 8. Metadata was added in a condensed form just like the LSTM model. We used categorical cross-entropy loss.

Two options for word representations were considered static GloVe embeddings taken from the Wikipedia (6 Billion token dataset) and non static embeddings initialized by the model and learned during the training time. Word embeddings (both static and non static) were 100 dimensional.

3. **BERT Model**
Bidirectional Encoder Representations from Transformers (BERT) [1] has proven to be a conceptually simple while empirically powerful model on different natural language processing tasks including the sequence-level classification problem. The key differentiation of BERT is that it enables bidirectional training of a transformers and achieved a better sense of language context than left-to-right or combined shallow left-to-right and right-to-left training.

As far as we are concerned, BERT has not been applied for fake news detection problem so far. Compared to traditional feature-based approaches such as combinations of word or sentence embedding and deep neural networks, BERT may lead to better fake news detection results when only purely contextual statements are available. Therefore, it is of our interest to explore the performance of BERT on the fake news detection problem.

The pre-training procedure conducted by the BERT authors on the combination of BooksCorpus (800M words) and English Wikipedia (2500M words) provides useful pretrained BERT weights that can be used for fine tuning specific tasks. Therefore, a natural baseline of BERT application in this study will be fine-tuning BERT for fake news detection. During fine-tuning process for sentence-level classification task, the final hidden state $C \in \mathbb{R}^H$ corresponding to the [CLS] word embedding will be taken to obtaine a summarized representation of the input sentence, where H is the hidden dimension. And a new classification layer with weights $W \in \mathbb{R}^{K \times H}$ (K denotes the number of classes) is added and the probability for different classes will be computed by $P = \text{softmax}(CW^T)$. The pretrained BERT weights as well as the newly added W will be fine-tuned together during the training. We have been focusing on adapting the existing BERT model (Pytorch version) for the fake news detection problem on both binary and multi-class classification.

5. General Reasoning

Modeling contextual contextual information, through pretrained word embeddings and metadata, is crucial to successfully detecting statements of *in-authenticity*. The experiments conducted on the LIAR dataset with different NLP models using or not using metadata can reveal the importance of NLP architectures as well as the contextual metadata for fake news detection.

We first applied logistic regression as a baseline approach by featurizing the statement using sentence2vec based on GloVe word embeddings. Then we concatenate some (but not all) of the other contextual features. Deep learning has proven effective in modeling word embeddings, but it is unclear how to successfully incorporate sparse contextual metadata (e.g. reputation score, state, venue, etc.). As a state-of-the-art fine-tuning model, BERT demonstrates superior performance on fake news detection than traditional models in cases where only statements are used. The performances of those models combined with extra metadata will be explored in the next step.

6. Preliminary Results

1. **Logistic Regression:** Using logistic regression, we were able to achieve a classification accuracy of 0.5888, and an F1 score of 0.681. Without the extra few contextual features, we get an accuracy of 0.5319 and an F1 score of 0.647. This is basically random guessing - statements alone do not contain enough information to distinguish truth or falsity. With just the contextual features, we get 0.566 accuracy and 0.694 F1 score.

2. **Deep Learning: LSTM, CNN:** For the LSTM- and CNN-based approaches, were able to achieve an binary classification training and validation accuracy scores of 0.7035 and 0.6285, respectively. Our macro-F1 score on validation was 0.6339. We found that the LSTM models tended to overfit easily. Dropout was used for regularization.

3. **BERT Model:** We are able to achieve an binary classification validation accuracy of 0.681 and macro-F1 score 0.687, which improves over the results obtained from both logistic regression and LSTM/CNN, when only statement sentences are used for the classification.

7. Discussion / Progress

So far, we have implemented different models including logistic regression combined with sen2vec, LSTM/CNN

combined with GloVE as well as BERT on the fake news detection problems using only the statement sentences available, and concluded that the NLP models are helpful for fake news detection even only statements are available by pushing the fake news detection accuracy above 0.5, which is roughly guess for human when no knowledge base or other information is available. And this proves our first hypothesis that those NLP models are able to identify language patterns to help identify fake news.

Also, by comparing the results achieved by different NLP models using only statement sentences, we showed that the improved model architectures can lead to the improved performance on fake news detection.

Furthermore, the incorporation of metadata into the prediction process was conducted for the logistic regression part where the counts are also used as extra features for prediction. Metadata was difficult to categorize because of the diverse information present. We tried to quantify different meta tags into discrete classes and a symbolic 'rest' class that allowed us to condense the values into a fixed (limited) set. We did observe the accuracy improvement when the extra metadata are introduced. And this lead us to further explore ways to combine metadata features with LSTM/CNN and BERT models in our future work.

8. Future Work

We plan to understand the examples that we get wrong, and trying to find a common factor which leads to this. Fundamentally, we want develop an intuition for the weaknesses and strengths of different modeling approaches, and use this to inform our hypothesis and future joint approaches.

We seek to find possibly better ways to combine the information. For example, since sentence2vec was trained using deep learning approaches, but the final classifier was not, we could see if the contextual features could be added earlier on in our pipeline.

We need to binarize the remaining categorical features, and also get embeddings for the context (free-form text). We will experiment with methods to incorporate contextual information better than just concatenating it to the word embedding models. Also maybe learning some sort of reliability score, instead of separate counts of how many times the speaker was in each class (part of the data). Basically a nonlinear function of these 5 count scores, which could help inform decisions.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] W. Y. Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.