Title

Transforming Deep Learning to Harness the Interpretability of Shallow Models: An Interactive End-to-End System

Team Composition

Texas A&M University (TAMU), Xia (Ben) Hu, Principal Investigator, hu@cse.tamu.edu
TAMU, Shuiwang Ji
University of Florida, Eric Ragan

Problem Statement

With the prevalence of web applications, it has become much easier for the public to get access to various news items through different channels, such as news websites and social media. However, this kind of convenience also brings about the wide spread of fake news at the same time. Fake news is detrimental to both individuals and society, and it is usually generated as hoaxes to mislead public through false or biased information. In 2016, the term "Fake News" was even shown as the word of the year by Macquarie Dictionary. Though its increasing importance, effective fake news detection is still considered to be a challenging task due to the following two aspects. First, fakeness of news could come from diversified perspectives, which is beyond the boundaries of traditional textual analysis. For example, a fake news item may result from its contexts or speaker, and it may not directly relate to its content information. The source of fakeness is highly dependent on each specific instance. Second, fakeness detection results need further explanation, which is important and necessary for users' final decisions. The explanation could provide evidences on model predictions and further help users better understand why certain news items are classified as fake news.

Considering these two challenges, TAMU has designed and implemented an explainable system, named XFake, on fake news detection. In particular, three different frameworks (MIMIC, ATTN, and PERT) are designed to analyze news items from different perspectives. MIMIC aims to judge each news item from its attribute information, which may include news statement, speaker, etc. ATTN's goal is to investigate each news statement through its semantic meaning. PERT is employed to study the linguistic features of news statements for detection. All three designed frameworks are self-explainable, where relevant explanations could be extracted for interpreting detection results. The system provides supporting examples for each news instance by a corresponding matching algorithm, and visualization of detection results for user-friendly interaction. Both supporting examples and visualization further assist users to understand why the system made a certain prediction on a given news. Overall, the XFake system is capable of not only giving prediction scores on news fakeness, but also providing relevant explanation as prediction evidences. With the aid of XFake, users can identify the news reliability.

Approach and Technique

The overall idea of the MIMIC framework is to mimic the performance of deep neural networks with the forest-based models. That is, the deep model is regarded as the teacher model; and the forest model is used as the student model. By calculating the node importance of each attribute in the mimic forest, the significance of each attribute in the input instance can be obtained. Further, through the constructed forest, users can better understand what happened in the prediction model, which provides users an overall transparency towards the model itself. The ATTN framework aims to build an interpretable deep learning model to deal with fake news detection. Several techniques are employed in the model, such as word2vec pre-trained word embedding, 1D convolutional network, and self-attention mechanism. The self-attention mechanism is employed because it can capture the long-range relationships between

Approved for Public Release, Distribution Unlimited

words. The mechanism enables the system to obtain input dependent weights to realize interpretation. Furthermore, by applying different kernel sizes for the 1D convolutional network, interpretation results can be obtained based on one-gram, two-grams, or three-grams. The PERT framework aims to extract linguistic features from the statement of a news and then make predictions using the XGBoost classifier. The perturbation-based method provides explanation for each prediction. The goal for this part is to explain the fake news detection method using interpretable linguistic features.

Figure 1. XFake system overview

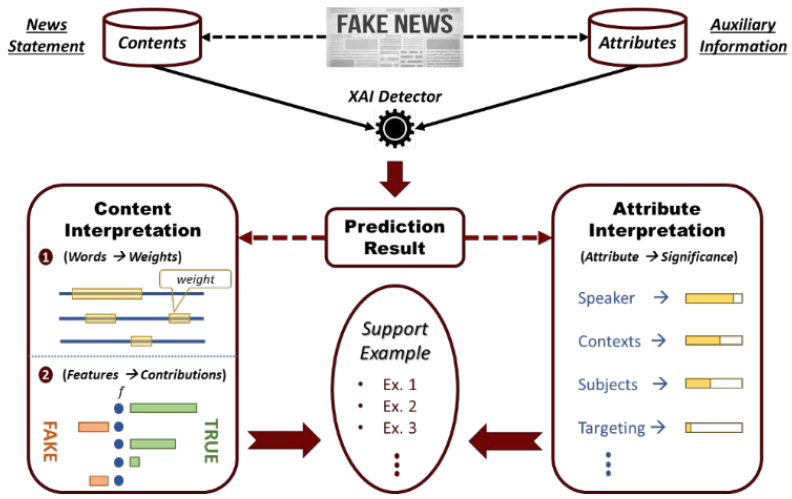
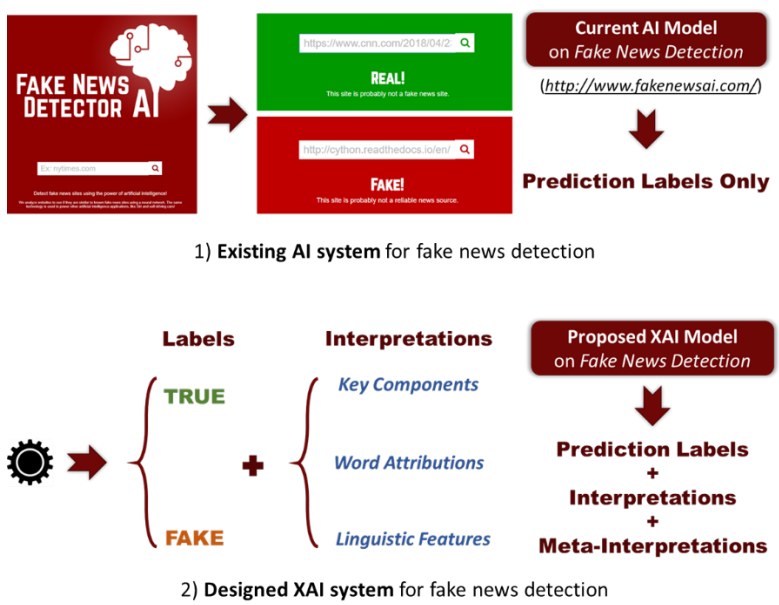


Figure 2. Comparison of AI system and XAI system on fake news detection. 1) Conventional AI for fake news; 2) TAMU designed XAI system for fake news



Key Objectives and Current Status

TAMU hypothesized that XAI system users would have an improved mental model of the system, and higher user-machine task performance. Because they included different levels of explanation, they

Approved for Public Release, Distribution Unlimited

hypothesized that the highest level of explanation detail would adversely impact participant performance.

TAMU designed an experiment to test different types and levels of explanation detail in order to evaluate the effects on participant performance and model understanding. The results of initial studies supported the hypothesis that additional explanation information can improve understanding of the model. However, this result did not support the hypothesis that added explanation supports improved human task performance; however, this is logical due to the difficulty of the task. The error estimation results did not support the hypothesis that the highest level of explanation detail would be detrimental to understanding; but the time results did indicate that the highest level of detail did take longer for participants to process.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
User's mental model	To verify if explanations can increase user understanding of how the system works	Additional explanation information can improve understanding of the model
User-machine task performance	To verify if users can achieve higher performance while working with an XAI system vs. a non-explainable system	Additional explanation may or may not support improved human task performance

	Numerical Value		Hypothesis Results	User Machine Task
	With Explanation	**Without Explanation**		
User-Machine Task Performance	65%	64%	No significant improvement among conditions	Detecting fake news among other news

	Numerical Value		Hypothesis Results	Mental Model Task
	With Explanation	**Without Explanation**		
Mental Model Understanding	83%	72%	Significant improvement, $p = 0.001$ (for the best case)	Predicting model's response

	Numerical Value		Hypothesis Results	Satisfaction Measure
	With Explanation	**Without Explanation**		
Explanation Satisfaction	N/A	N/A	Significant, $p < 0$	Likert scale

	Size	Length	Type
User Study	147	2-steps study; average study time of 10 minutes Each participant performed 16 news fact checking on each step Controlled study had six conditions	AMT crowdsourcing study Between-subject study

Approved for Public Release, Distribution Unlimited

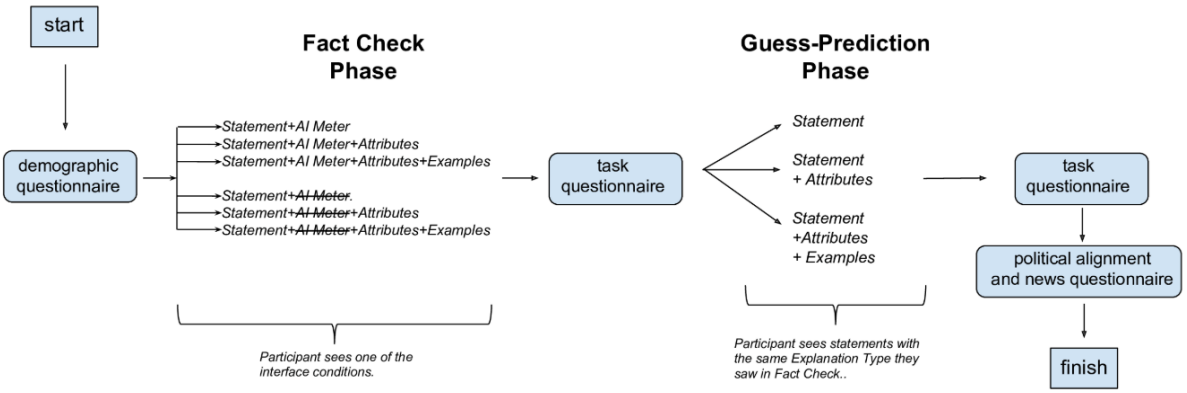
Test and Evaluation

Participants were instructed at different stages during the study. They were presented with text instructions and screenshots of components of the XAI system specific to their experimental conditions.

There were two phases. The first phase, "Fact Check", asked participants to rate whether news was more likely to be true or false. They would use the XAI system (if present) to help them with this task. The second phase, "Prediction-guess", asked participants to estimate the recommendation of the XAI system. The overall recommendation was hidden, but different levels of XAI could be used to help participants guess.

Figure 3 shows an overview of the study procedure and how the tasks were influenced by the experimental conditions. While there are two main phases, questionnaires were also used that asked participants to answer questions about their demographics, political leanings, news consumption, and experiences using the XAI system.

Figure 3. Overview of the study procedure and task phases



In the Fact Check Phase, participants used the XAI system to review news statements (one at a time) along with the corresponding output and explanations per the assigned experimental condition. Participants were tasked with assessing whether the news statement was true or false; they were free to decide whether or not to agree with the AI classification. After they evaluated a statement, the system showed the correct answer before allowing them to proceed to the next news statement.

In the Guess Prediction Phase, following the Fact Check Phase, participants viewed 16 news statements, except their task was now to predict the classification output of the XAI system. In other words, rather than predicting the actual statement veracity, the goal was to predict how the model would assess the statement. This task was done as a way to evaluate participant understanding of the model based on their ability to infer its output for a new set of inputs. Questionnaires were used between main phases to collect information about the participants and their experiences.

Mohseni, Sina, Eric Ragan, and Xia Hu. *Open Issues in Combating Fake News: Interpretability as an Opportunity.* Submitted to IJCAI 2019 (survey track).

Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan. *A Survey of Evaluation Methods and Measures for Interpretable Machine Learning.* arXiv preprint arXiv:1811.11839 (2018).

Approved for Public Release, Distribution Unlimited

Mohseni, Sina, and Eric Ragan. *Combating Fake News with Interpretable News Feed Algorithm.* arXiv preprint arXiv:1811.12349 (2018).

Mohseni, Sina, and Eric D. Ragan. *A human-grounded evaluation benchmark for local explanations of machine learning.* arXiv preprint arXiv:1801.05075 (2018).

Linder, Rhema, Mohseni, Sina, Yang, Fan, Kumar, Shiva, Eric D. Ragan and Xia Hu. *How Level of Explanation Detail Affects Human Performance in Interpretable Intelligent Systems: A Study on Explainable Fake News Detection* (ongoing submission).

For additional information:

https://datamllab.github.io/XAI_TAMU/

https://www.youtube.com/watch?v=U7Lrb5baeYs&t

https://youtu.be/Y_E0K-M3ZNs

http://csedatasrv.cs.tamu.edu:3000/

https://rhema.github.io/demos/xai-mturk/vdemo2x3/?condition=2x3-on-attr&godmode=1

Approved for Public Release, Distribution Unlimited