Talos Targets Disinformation with Fake News Challenge Victory

This post was authored by Sean Baird with contributions by Doug Sibley and Yuxi Pan

Executive Summary

For the past several months, the problem of "fake news" has been abuzz in news headlines, tweets, and social media posts across the web. With historical roots in information warfare and disinformation, "fake news" is a different kind of cyber-threat affecting people all around the globe. Using advanced machine learning and artificial intelligence technology, Talos researchers set their sights on this different kind of cyber-threat and beat out over 80 registered teams worldwide to claim first place in the Fake News Challenge.

Context

Background

While there has been significant media coverage regarding fake news in the recent months, the modern fake news problem is rooted in a long history of information operations and disinformation campaigns.

In a very in-depth paper about the topic, Facebook defines information operations "[...] as actions taken by organized actors...to distort domestic or foreign political sentiment, most frequently to achieve a strategic and/or geopolitical outcome [...]" and classifies "false news" as a useful tool in the information operations toolkit. The paper specifies that "false news" is "[...] news articles that

purport to be factual, but which contain intentional misstatements of fact with the intention to arouse passions, attract viewership, or deceive."

Alternately, the Wired article announcing the victory describes fake news in a simpler manner as "[...] made-up news stories created to convert social media shares into page views, ad dollars, and perhaps even political traction."

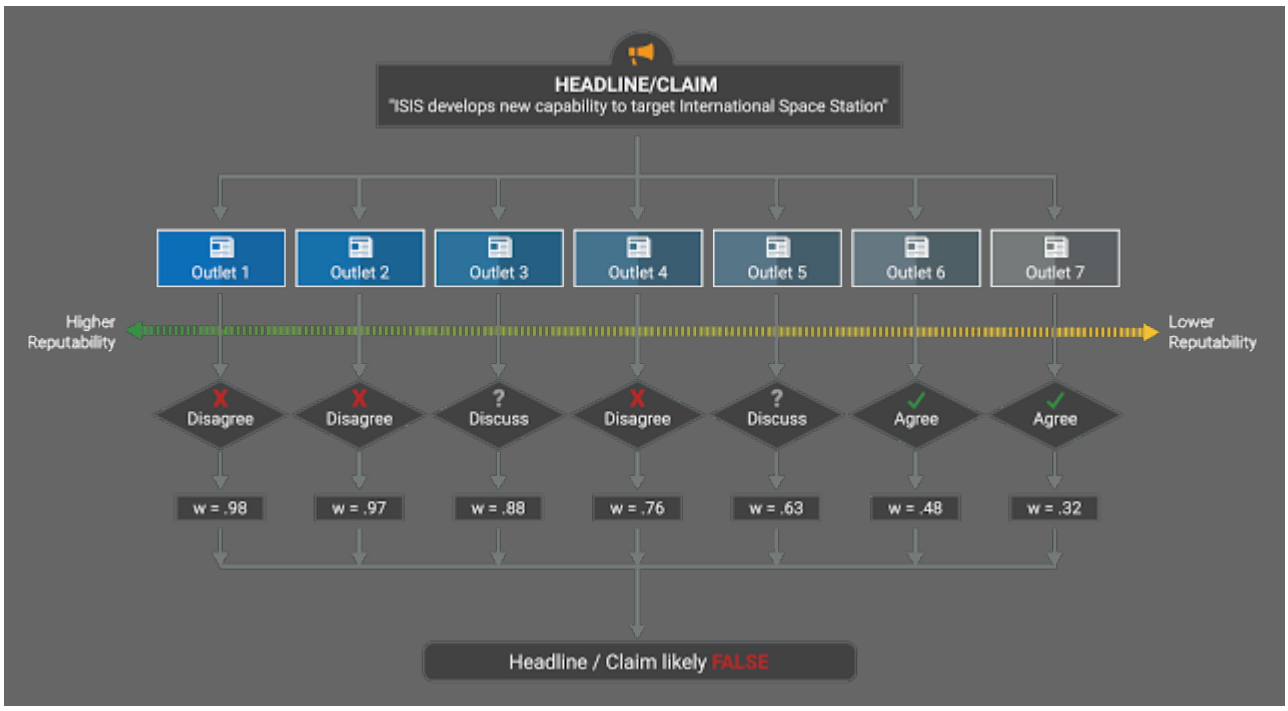
Clearly, this is a difficult problem to solve in cyberspace, especially in a world where technology and social media can help amplify these stories to a much broader audience. This prompted several researchers in academia and industry to create the Fake News Challenge (FNC). The self described goal of the FNC is to "[...] address the problem of fake news by organizing a competition to foster development of tools to help human fact checkers identify hoaxes and deliberate misinformation in news stories."

The first iteration of the challenge (FNC-1), which lasted from December 1, 2016 until June 2, 2017 focused solely on stance detection, a crucial first step in helping to detect fake news.

FNC-1: Stance Detection

While actual truth-labeling is a hefty task, rife with political and technical issues, stance detection is a potential first step toward a more robust solution. Dean Pomerleau, one of the organizers of the challenge, explained in a Mediashift interview that "[...] the goal [of stance detection] is to determine which has the best argument, not just which is the most popular or widely cited or read, the way a search engine does."

In the context of the FNC, stance detection can be defined as labeling the relationship an article body has to its headline/claim -- specifically, whether the body agrees with, disagrees with, or discusses the headline/claim or whether the body is completely unrelated. Thus, the four possible outputs of a stance detection system should be "agree," "disagree," "discuss," and "unrelated." An example of how stance detection could be implemented in a broader fake news detection system is available in the figure below:



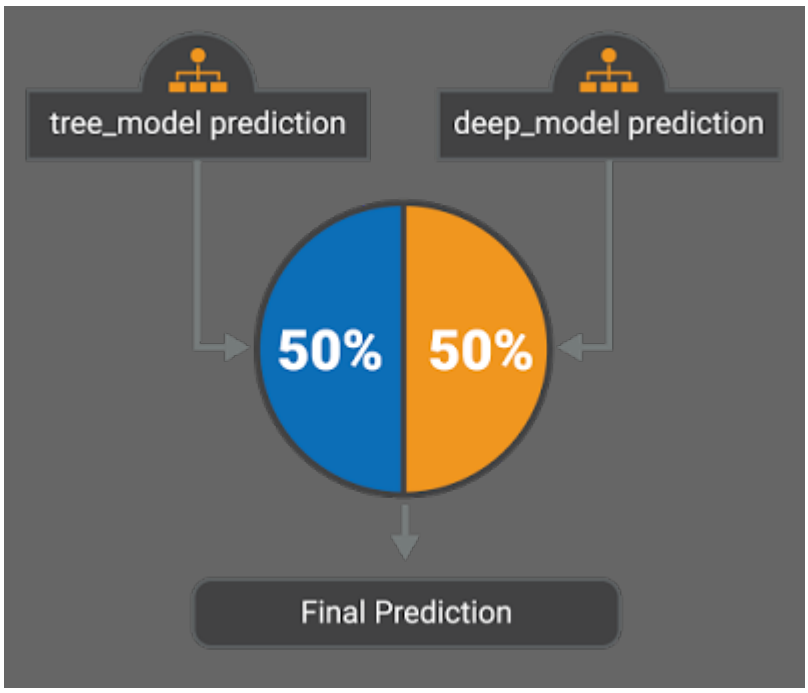
Stance detection's role in fake news detection

Always interested in a challenge, the FNC piqued the interest of Talos researchers who chose the team name "SOLAT IN THE SWEN" as a clever anagram of their true affiliation. Immediately, these researchers began development work on various models and solutions in their spare time -- models which would eventually net them a first place victory.

SOLAT IN THE SWEN - Talos's covert team name

Our Solution

One of the goals of this challenge was "[...] to explore how artificial intelligence technologies, particularly machine learning and natural language processing, might be leveraged to combat the fake news problem [...]." Because of this, team SOLAT IN THE SWEN decided to test how various cutting-edge machine learning techniques performed. After successfully implementing several different models, the team found that their results were best when combining multiple models in an ensemble. The team's final submission was an ensemble based on an 50/50 weighted average between gradient-boosted decision trees and a deep convolutional neural network. The full code can be found on the Talos GitHub, open sourced with an Apache 2.0 license.



Our models were ensembled with a 50/50 weighted average

Deep Learning Approach

The first model used by the team applies several different neural networks used in deep learning. This model applies a one-dimensional convolutional neural net (CNN) on the headline and body text, represented at the word level using the Google News pretrained vectors. CNNs allow for efficient, effective parallel computation while performing The output of this CNN is then sent to an multi-layer perceptron (MLP) with 4-class output -- “agree,” “disagree,” “discuss,” and “unrelated” -- and trained end-to-end. The model was regularized using dropout (p=.5) in all convolutional layers. All hyperparameters of this model were set to sensible defaults, however, they were not further evaluated to find better choices.

Diagram outlining our deep learning model

The architecture of this model was selected due to its ease of implementation and fast computation since we can rely on convolutions instead of recurrence. Judging from the relative strength of this model convolutions seem to be able to capture a wide variety of topics; however, the model is limited in that it only gets to observe the text once. A potential extension to this model would be to include some sort of attention mechanism with recurrence after the convolutions which would allow the model query specific aspects of the headline/body after receiving a general summary from the CNN.

Gradient-Boosted Decision Trees (GBDT) Approach

The other model employed in the ensemble is a Gradient-Boosted Decision Trees (GBDT) model.

This model inputs few text-based features derived from the headline and body of an article, which are then fed into Gradient Boosted Trees to predict the relation between the headline and the body.

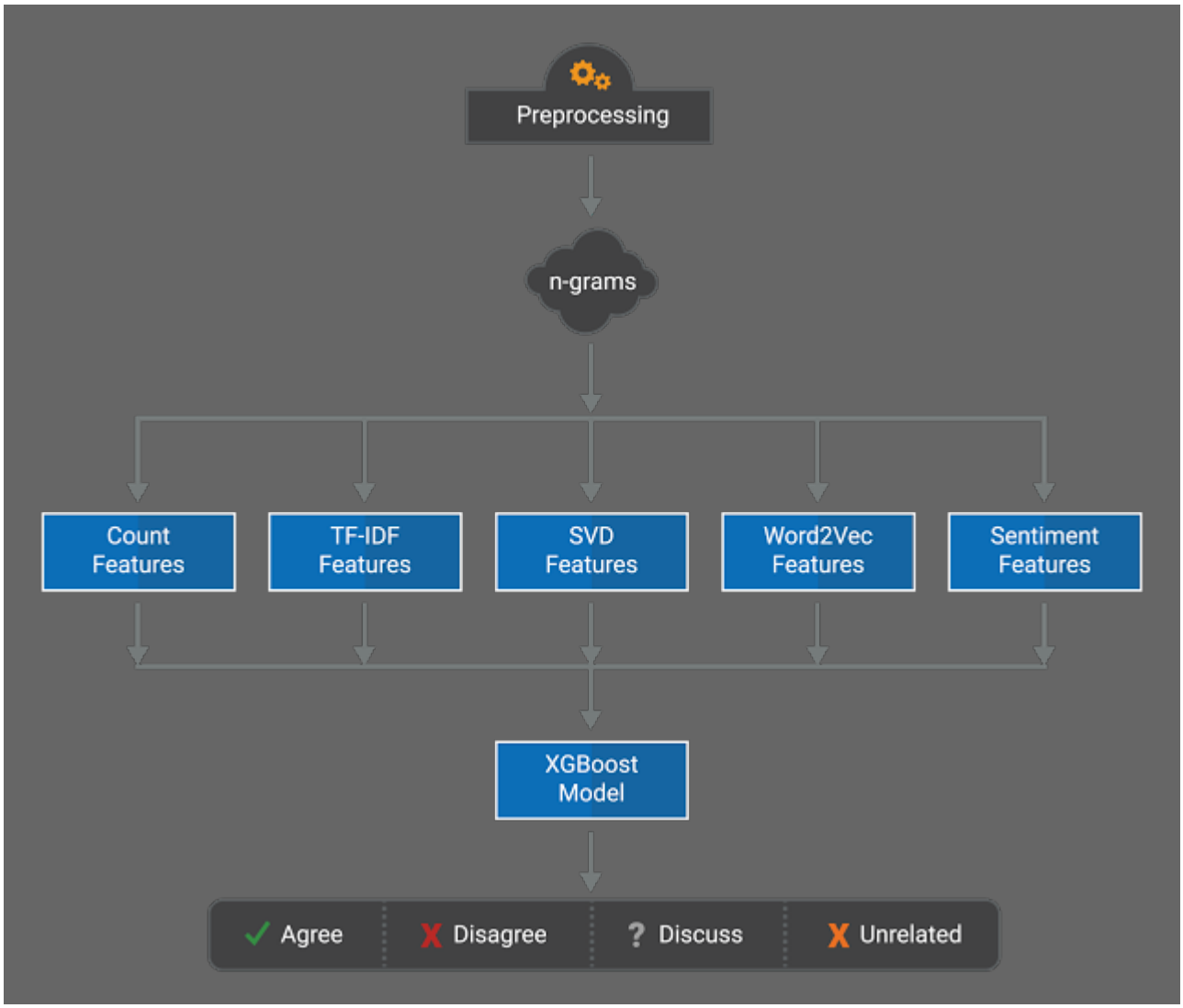


Diagram outlining our decision tree model

After exploring the dataset, a few features that are likely to be informative of headline/body relationships became obvious -- for example:

- The number overlapping words between the headline and body text;
- Similarities measured between the word count, 2-grams and 3-grams; and
- Similarities measured after transforming these counts with term frequency-inverse document frequency (TF-IDF) weighting and Singular Value Decomposition (SVD).

Using these features, it is not necessary to use a powerful and expressive model to learn the complex mapping from these features to the stance label.

For this, Gradient-Boosted Decision Trees were chosen because of the model's robustness with regard to the different scales of our feature vectors. Specifically, no normalization is needed and it

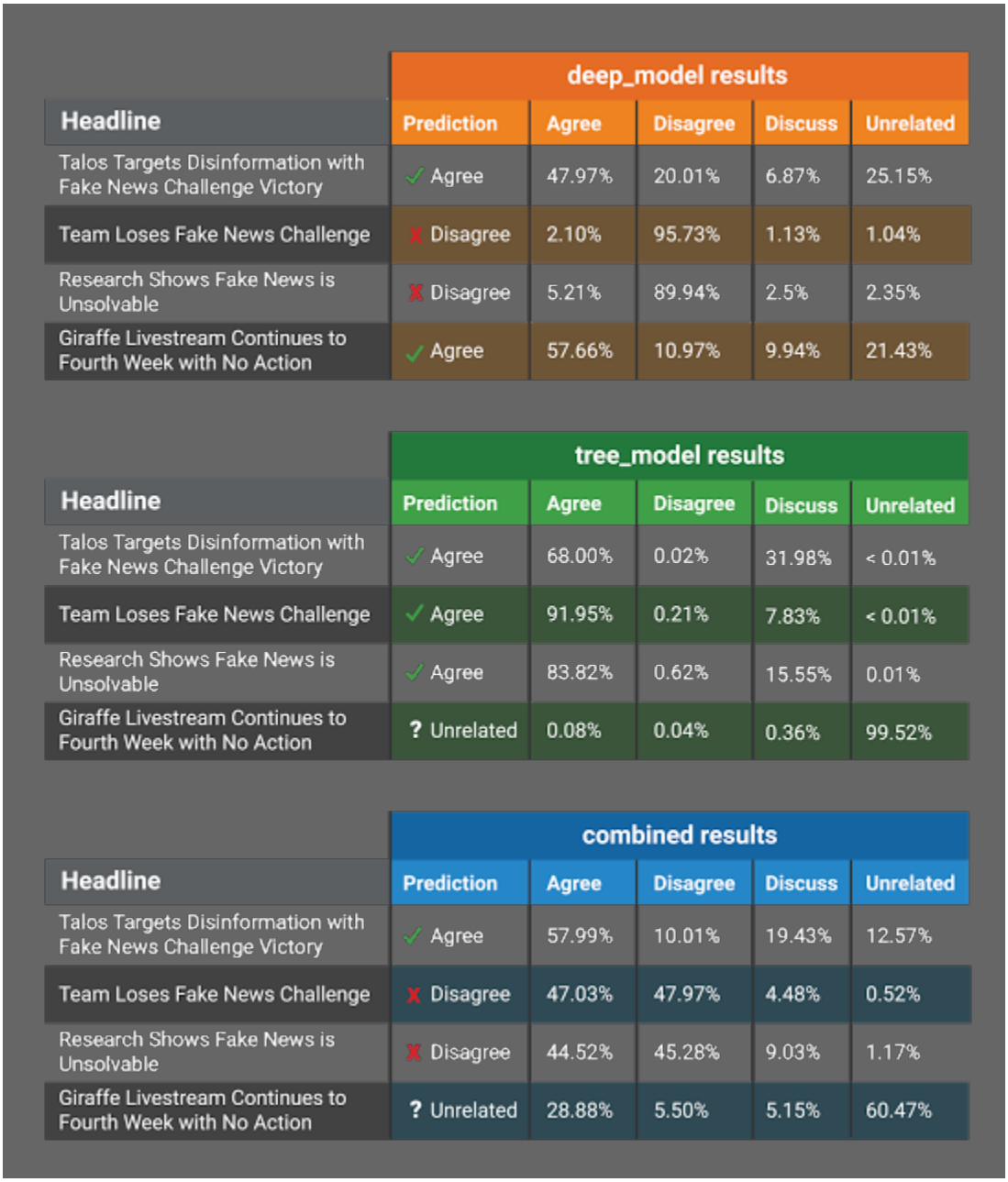
can be regularized in several different ways to avoid overfitting. Furthermore, XGBoost is a very efficient, open-source implementation that was easily applied to the handcrafted features.

Real World Exercise

Some readers may be wondering what the output of our system looks like with real-world data. As a fun exercise, we ran the contents of a first draft of this blog post through our system with various headlines -- the real headline of the post, and a few others we made up for the sake of this activity. These headlines are:

- Talos Targets Disinformation with Fake News Challenge Victory (the real headline);
- Team Loses Fake News Challenge;
- Research Shows Fake News is Unsolvable; and
- Giraffe Livestream Continues to Fourth Week with No Action.

We were excited to see whether or not our models would be able to correctly detect the stance of our blog post with each of these headlines. The results can be found in the figure below:



deep_model results

Headline	Prediction	Agree	Disagree	Discuss	Unrelated
Talos Targets Disinformation with Fake News Challenge Victory	✓ Agree	47.97%	20.01%	6.87%	25.15%
Team Loses Fake News Challenge	✗ Disagree	2.10%	95.73%	1.13%	1.04%
Research Shows Fake News is Unsolvable	✗ Disagree	5.21%	89.94%	2.5%	2.35%
Giraffe Livestream Continues to Fourth Week with No Action	✓ Agree	57.66%	10.97%	9.94%	21.43%

tree_model results

Headline	Prediction	Agree	Disagree	Discuss	Unrelated
Talos Targets Disinformation with Fake News Challenge Victory	✓ Agree	68.00%	0.02%	31.98%	< 0.01%
Team Loses Fake News Challenge	✓ Agree	91.95%	0.21%	7.83%	< 0.01%
Research Shows Fake News is Unsolvable	✓ Agree	83.82%	0.62%	15.55%	0.01%
Giraffe Livestream Continues to Fourth Week with No Action	? Unrelated	0.08%	0.04%	0.36%	99.52%

combined results

Headline	Prediction	Agree	Disagree	Discuss	Unrelated
Talos Targets Disinformation with Fake News Challenge Victory	✓ Agree	57.99%	10.01%	19.43%	12.57%
Team Loses Fake News Challenge	✗ Disagree	47.03%	47.97%	4.48%	0.52%
Research Shows Fake News is Unsolvable	✗ Disagree	44.52%	45.28%	9.03%	1.17%
Giraffe Livestream Continues to Fourth Week with No Action	? Unrelated	28.88%	5.50%	5.15%	60.47%

Results of this blog post and various test headlines being evaluated by our system

As shown above, while neither the deep learning approach nor the GBDT approach had perfect accuracy, the combination of these two approaches with a 50/50 weighting detected the correct stance for each headline.

Conclusion

In the end, these innovative model implementations put Talos on top of the global leaderboard. While more research needs to be done, Talos's award-winning research on stance detection is an important first step toward tackling the problem of fake news and disinformation in the 21st century.

As the Fake News Challenge moves forward and the natural language processing community

continues to churn out cutting-edge research, Talos remains committed to continually forcing the bad guys to innov