Using Transfer Learned Features of Annotated Articles to Detect Fake News with ML Models

Rushabh Phadkule
University of Dayton
phadkuler1@udayton.edu

Nileena Thomas
University of Dayton
thomasn10@udayton.edu

Saeedeh Shekarpour
University of Dayton
sshekarpour1@udayton.edu

Abstract

We are all overwhelmed by the amount of fake news and misinformation circulating in the digital world. The amount of consequences is substantial enough to take actions and find new methods to resolve this problem. This paper tries to introduce a new way to use the annotated data to detect subtle and underlying characteristics. The indicators for article credibility defined by a diverse coalition of experts are used as transfer learned features.

I. Introduction

Fake news has been one of the most hotly-debated socio-political issues of recent years. Websites which deliberately published hoaxes and misleading information surfaced across the internet. As stated in Pew Research[1] 2019, Nearly seven-in-ten U.S. adults (68 Percent) say made-up news and information greatly impacts Americans' confidence in government institutions, and roughly half (54 Percent) say it is having a major impact on our confidence in each other.

In recent times, amidst pandemic, misinformation and speculation about Covid-19 have flooded digital media. Also due to recent developments in machine learning and artificial intelligence, the models that detect fake news have advanced, but so have the models that can fool the state-of-the-art soft-wares. So, we thought it was necessary to introduce some features which are subtle to machines, but apparent to humans. Therefore, we have used human annotated data. Annotated data-set of 40 articles of varying credibility annotated with indicators by 6 trained annotators using specialized platforms by Credibility Coalition[2].

[1] https://pewrsr.ch/3fzCgi4

[2] https://credibilitycoalition.org/research/

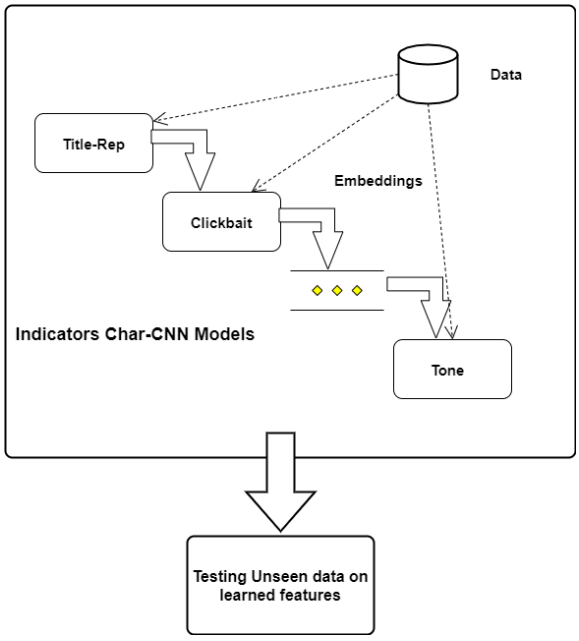

Figure 1: *Model Architecture of Transfer Learning of Indicators*

II. Literature Review

We heavily relied on the information in the paper titled: A Structured Response To Misinformation: Defining And Annotating Credibility Indicators In News Articles by authors: Amy X. Zhang, Aditya Ranganathan, Sarah Emlen

Metz, Scott Appling, Connie Moon Sehat, Norman Nick B. Adams, Emmanuel Vincent, Jennifer 8. Lee, Martin Robbins Factmata London, Sandro Hawke, David Karger, An Xiao Mina . The article presents an initial set of indicators for article credibility defined by a diverse coalition of experts. These indicators originate from both within an article's text as well as from external sources or article metadata. A dataset of 40 articles of varying credibility annotated with our indicators by 6 trained annotators.the authors describe a set of initial indicators for article credibility, grouped into content signals, that can be determined by considering the content of an article, as well as context signals, that can be determined through consulting external sources or article metadata.

These indicators were iteratively developed through consultations with journalists, researchers, platform representatives The authors found domain experts, such as scientists or industry practitioners, to score each article on a 5 point scale. This gold standard is an overall rating of the credibility of the article.

This standard metric was then used to compare how much the annotators' assessments of overall article credibility agreed with domain experts' assessments. Standard metrics were used to compare: how much annotators agreed with one another, and how much the annotators' assessments of overall credibility agreed with domain experts' assessments. After aggregating annotations, they determined correlation with domain expert scores, and multiple linear regression performed.

III. Dataset

Credibility Coalition effort presents an initial set of indicators for article credibility defined by a diverse coalition of experts. A dataset of 40 articles was annotated with indicators by trained annotators. Indicators for article credibility, grouped into content signals, that can be determined by considering the content of an article, as well as context signals, that can be determined through consulting external sources or article metadata.

We focused our research on Content Indicators, because we interpreted that content (i.e title, tone, fallacies, etc) of the article is more revealing than context (i.e No. of Ads, "Spammy" Ads, Originality).

Content Indicators Include:

- Title Representativeness
- "Clickbait" Title
- Logical Fallacies
- Tone
- Inference
- Calibration of Confidence

Each indicator was extracted from the data-set according to each indicator question. Indicators were denoted by the elaborated questions. For example. Title Representativeness:–> Question: Does the title of the article accurately reflect the content of the article? Each indicator had scale of outputs. For example. Title Representativeness:–> • Somewhat Unrepresentative • Somewhat Representative • Completely Representative
The gold standard indicator is: Credibility. Which we also used to train our model.
We split the data according to each indicator. So we acquired different CSV files for all the indicators. This was further split into training and testing data by using python script.
Data Acquisition: Acquired JSON file of Content indicators from Credibility Coalition dataset from data.world platform. [Zhang, Ranganathan, and MetzZhang et al.2018]
Created a merged CSV file. For each indicator we created a different CSV files. (This will become apparent as to why we did this, in further slides) We also had to clean the data of off Unknown characters. (It's not deemed necessary when the data is massive. But we have limited data [40 articles] , so clean data is necessary for high dimension of data.)

IV. Methods

In the research paper Character-level Convolutional Networks for Text Classification [Zhang, Zhao, and LeCunZhang et al.2015]
they constructed datasets to show that

character-level convolutional networks could achieve state-of-the-art or competitive results. Comparisons are offered against traditional models such as bag of words, n-grams and their TFIDF variants, and deep learning models such as word-based ConvNets and recurrent neural networks.

That is why we used Character level CNN. Another reason why we used this model is that, as the data is less the dimensionality also reduces. So to overcome the limitation occurred by word grams and sentence formation, we used character level. Character level can learn higher level of dimensionality, even the subtle ones.

Model Architecture: We used a Char-CN Network with

- Embedding Layer
- Six convolutional layers, and 3 convolutional layers followed by a max pooling layer
- Two fully connected layer(dense layer in keras), neuron units are 1024.
- Output layer(dense layer). In this task, we set it 3.

Layer	Large Frame	Small Frame	Kernel	Pool
1	1024	256	7	3
2	1024	256	7	3
3	1024	256	3	N/A
4	1024	256	3	N/A
5	1024	256	3	N/A
6	1024	256	3	3

Figure 2: *Model Summary*

We thought it should be sparse enough to learn all the transfer learned features. Certain computations can be carried out more efficiently on them provided the matrix is sufficiently large and sparse. Our networks can leverage the efficiency gained from sparsity by assuming most connection weights are equal to 0. (i.e Drop out Value)

The model for each indicator was tuned to get better accuracy. Also keeping in mind that the model did not overfit (As the dataset was small). The hyper-parameters were changed but the model structure was same throughout.

V. Training

Each constructed model was trained on each indicator. The data was split 60-40. As the dataset is small we also had to give preeminence to Testing as well. As networks can even learn from small datasets, it is important to know if the learned network has network fine-tuned in a way that it can perform well on unseen data. Each indicator model transferred the learning through the previous saved model and weights. Then the transfer learned model was tested on Credibility Gold Standard. Also tested on Emergent Fact-Checking Website Dataset aquired from DISCOURSE PROCESSING LAB website.[3]

VI. Results

We tested the models as we trained them. In Figure 3 it's apparent as the features are transferred down the tree it has better performance.

We tested the models on multiple combinations of indicators on the Gold Standard (Credibility). Figure 4 indicates a few.

Content	loss	accuracy	val_loss	val_accuracy
title-rep	22.0834	63%	20.5782	56%
clickbait	10.1101	63%	10.2916	44%
cali-confider	5.8657	60%	5.7999	33%
log-fal	3.6615	31%	3.5019	33%
tone	2.5366	45%	12.3011	54%
credibility	1.7293	30%	1.6931	40%

Figure 3: *Model Training and Testing Accuracy and Loss with each transfer learned model*

The combination of models were tested on Emergent Dataset. We can see that the

Credibility	loss	accuracy	val_loss	val_accuracy
title-rep + clickbait	5.9768	34%	5.6702	40%
title-rep + clickbait + cali-confidence	3.6732	36%	3.5307	20%

Figure 4: *Model Training and Testing Accuracy and Loss on Credibility (The Gold Standard)*

[3] `http://fakenews.research.sfu.ca/#parseWebs`

model has more accuracy as more indicators are added to the combination. The Logical fallacies indication contributes to just 1 Percent increase in Accuracy.

Emergent	loss	accuracy
title-rep + clickbait	11%	19%
title-rep + clickbait + cali-confidence	6%	29%
title-rep + clickbait + cali-confidence + log-fa	4%	29%
title-rep + clickbait + cali-confidence + log-fa	2%	51%
all tranfer leaned	2%	52%

Figure 5: *Model Testing Accuracy and Loss on Emergent Dataset*

VII. Discussion

Being involved with the process some things came into the spot-light. Some interpretation from the results:

i. Subsection One

The accuracy decreases as transferring the learning to subsequent models. We think it's because of the correlations between the IRR metric used by the paper: A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles.

ii. Subsection Two

The models performed quite well on all transfer learned feature model. So maybe the features are quite distinct from each other.

VIII. Bibliographical References

References

[Zhang, Ranganathan, and MetzZhang et al.2018] Amy X. Zhang, Aditya Ranganathan, and Metz. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 603–612. https://doi.org/10.1145/3184558.3188731

[Zhang, Zhao, and LeCunZhang et al.2015] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. arXiv:cs.LG/1509.01626