CookieBox - Fake News Classifier

Ruj Haan, George Boktor, and Nibraas Khan (Advisor: Dr. Sal Barbosa)

Department of Computer Science, Middle Tennessee State University, 37132

Abstract

In today's political climate, there is an overwhelming sense of confusion and deception when considerations are made for the divisive and agenda-driven nature of modern media. A campaign to slander and devalue political opponents, countries, organizations, and affiliations on both sides of the political spectrum makes it hard for the unsuspecting consumer of information to find reliable, untinged sources. What CookieBox tries to do is allow users to gauge the accuracy of a piece of written media using a simple online interface (accuracy, in this case, being defined as the grading a piece receives based on factuality, written bias, and the general tendency to disseminate false information). Under the hood, CookieBox employs an ensemble of methods, ranging from rule-based approaches using word classifications and knowledge bases, to machine learning, primarily deep learning techniques such as Autoencoding and Convolutional Neural Networks, in the hopes to create a consistent, reliable accuracy grading that users can use to accredit, or discredit, a piece of media.

Introduction

It can be difficult for the unsuspecting consumer of media to recognize whether or not the information they're reading is biased or not. The most important factors for gauging the unreliability of information are as follows:

- SOURCE OF INFORMATION
  - Low credibility sources:
    - The Onion
    - 70 News
    - Bients.com
- QUALITY OF INFORMATION
  - Grammar
  - Quality of Thought
  - Word Choice
- ACCURACY OF INFORMATION
  - Withdrawal of pertinent facts
  - Inaccurate representation of data/facts
  - Inaccurate characterization of events
  - Biased reporting of events/data

We propose a fake news classifier that employs an ensemble of methods to address most of these issues when it comes to quantifying the accuracy of a piece of media. From the user-end, a reader only needs to supply the link to an article to which they want to assign an accuracy grading. **Figure 1** illustrates how that article will be processed and what methods we will use to assemble a final percentage, or grading, to assign the given article/written media.

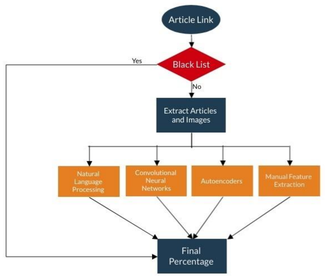

Figure 1:
Graphic representation of how an article is pipelined through the CookieBox classifier.

Methods

- NATURAL LANGUAGE PROCESSING (NLP)
  - Knowledge Graphs
    - The model extracts the main points from the article, then it gathers other articles related to the topic to create a knowledge graph.
    - Knowledge Graph: Graph representation of relationships between different Named Entities that we extract using our natural language tool kits. Entities are expressed as nodes and the relationships (Attributes) between those entities are expressed as edges. The utility of knowledge graphs is having a queryable reservoir of knowledge that we can use as a source of truth when analyzing a provided article. **Figure 2** illustrates the basic structure of a simple group of nodes and edges.
  - Word Embeddings
    - We are working with several machine learning algorithms including neural networks, and these algorithms have a lot of difficulty working with text. In order to overcome this obstacle we are using wording embeddings created through Glove and Word2Vec.
    - In essence, word embeddings are just distributed representations or vectors of words. For example, the word "king" is translated to a vector of N numbers. Certain arithmetic can also be performed on text converted to word embeddings such as "king" - "crown" = "man".
    - We are working with two word embedding tools: Word2Vec and Glove
      - Word2Vec
        - This is a statistical method for learning word embedding using local usage context. Meaning, the context of a word is defined by the words in the window around it.
        - There are two learning models within Word2Vec: Continuous Bag-of-Words that works by predicting the current word based on its context and Continuous Skip-Gram Model that learns by predicting the surrounding words given the current context.
      - Glove
        - This is an extension of to the Word2Vec model which combines the local context based approach and global text statistics.

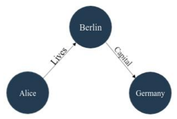

Figure 2:
Basic model of how relationships are expressed in the Knowledge Graph. Alice (N.E.) lives (Attr.) in Berlin (N.E.) which is the Capital (Attr.) of Germany (N.E.).

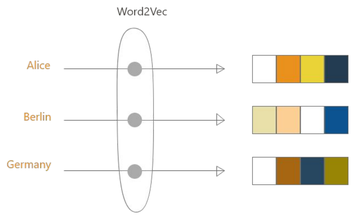

Figure 3:
Example of Word2Vec converting between text to a vector.

- CONVOLUTIONAL NEURAL NETWORKS (CNN)
  - CNNs are a class of Neural Networks, which are commonly used to analyze visual images. This method works by encoding pixel values from an input image into a matrix. The input will go through different layers of the model including convolution, Relu, Max-pool, and an input layer, similar to **Figure 4**.
    - Convolution layer: This layer computes the output of neurons that are connected to local regions in the input.
    - Relu layer: This layer applies the activation function.
    - Max-Pool layer: A pooling layer is a layer added after the convolutional layer, used for ordering layers within a CNN that may be repeated one or more times in a given model. Max-pool calculates the maximum value in each patch.
    - The input layer holds raw pixel values of the image.
  - Most articles have images that are related to the topic of the article, most discredited articles use irrelevant or poor quality images. We are using a pretrained CNN model to evaluate the article based on its constituent images.

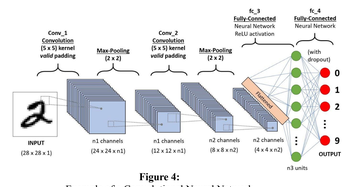

Figure 4:
Example of a Convolutional Neural Network learning how to understand handwritten digits

- AUTOENCODERS (AU)
  - An Autoencoder is an unsupervised learning technique that uses backpropagation to minimize the difference between the inputs and the outputs.
  - The purpose of this algorithm is for dimensionality reduction. In other words, the algorithm is able to take complex data with a plethora of features and condense it down to the most important features.
  - In our work, we are using the encoder to understand what makes a fake news article fake.
    - For the input, the model is fake news articles converted to vectors using Word2Vec.
    - The compressed representation becomes the features of the fake articles.
    - The output is the same as the inputs.
  - When the model is pretrained, we can pass the news article that we get from the user through the model and see how well it is able to reconstruct it. If there is high reconstruction, then the article has many features that fake articles have, and if the reconstruction is low, the model does not have many features in common.

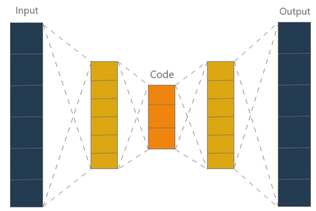

Figure 5:
An example of Autoencoder encoding the input layer down to the code section and decoding it back out to the output section.

- MANUAL FEATURE EXTRACTOR
  - Fake articles can also contain certain features such as certain kinds of syntax or word choice.
  - We can manually figure out what these features and see how many of these features the test article has in common.

Conclusion

The CookieBox fake news classifier is a working title for the public interface people will use to independently verify the veracity of any news article they provide. With the use of an ensemble of Machine Learning and Natural Language Processing algorithms, our model will be able to assign an accuracy grading. Our model's prediction heavily depends on the sources of text we use for pre-training and the sources we use when collecting data for our knowledge graphs. While the algorithm will be able to generate a grading that corresponds to the accuracy of an article, the quality of the scoring is limited by the quality of the data with which we train the classifier. For success, the data needs to be valid and legitimate.

Future Work

There are many parts of our algorithm that can be pre-trained and useable the moment the user requests an article for evaluation. The models that we are working with need to be updated often to retain accuracy as new kinds of fake news circulate. This means that the pre-trained models also need a way to be replaced with updated models. Along with the update of the models already in the algorithm, we hope to integrate more models into our ensemble. With more methods to scrutinize articles, the algorithm will be able to predict their accuracy with more and more accuracy.

There are many avenues of research that this work hints towards, and many of them can be extremely fruitful.

References

- Lin, Yankai, et al. "Learning entity and relation embeddings for knowledge graph completion." Twenty-ninth AAAI conference on artificial intelligence. 2015.
- Srinivasa-Desikan, Bhargav. Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd, 2018.
- Lawrence, Steve, et al. "Face recognition: A convolutional neural-network approach." IEEE transactions on neural networks 8.1 (1997): 98-113.
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
- Rong, Xin. "word2vec parameter learning explained." arXiv preprint arXiv:1411.2738 (2014).
- Chen, Xi, et al. "Variational lossy autoencoder." arXiv preprint arXiv:1611.02731 (2016).
- Lewis, David D. "Feature selection and feature extraction for text categorization." Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992.
- Howard, Jeremy, and Sebastian Ruder. "Fine-tuned language models for text classification." arXiv preprint arXiv:1801.06146 (2018): 1-7.
- Wang, Yasi, Hongxun Yao, and Sicheng Zhao. "Auto-encoder based dimensionality reduction." Neurocomputing 184 (2016): 232-242.
- Zabalza, Jaime, et al. "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging." Neurocomputing 185 (2016): 1-10.

Contact Information

Ruj Haan: gm3g@mtmail.mtsu.edu
George Boktor: gsb3c@mtmailmtsu.edu
Nibraas Khan: nak2z@mtmail.mtsu.edu
Dr. Sal Barbosa: sal.barbosa@mtsu.edu