FAKE OR NOT?

Classification and Visualization of News Articles

Mansi Arora
MS Analytics
marora33@gatech.edu

Rishi Bhatia
MS Analytics
rbhatia37@gatech.edu

Michael Cho
MS Analytics
mcho89@gatech.edu

Jeh Lokhande
MS Analytics
jlokhande3@gatech.edu

Taylor Million
MS Computer Science
tmillion3@gatech.edu

Nabila Usmani
MS Analytics
nusmani6@gatech.edu

Summary

Fake news spreads misinformation and leads to biased opinions in society. The existing approach of human fact checking to identify fake news is expensive, cumbersome and erroneous. This project aims to identify fake news articles based on a machine learning model and visualize related topics using Google Trends. The idea is to not only classify news articles, but also provide intuition to the user as to why an article was classified as fake or real.

Approach

Our approach began with data collection, cleaning and integration. The raw data was cleaned by removing unwanted characters using **regex**.

Feature Extraction

Hand-crafted features were extracted from the data which included number of special characters, number of first person words, average word length, etc. Each word in the article text was converted to its word-vector using **word2vec** and **GloVe**.

Classification Model

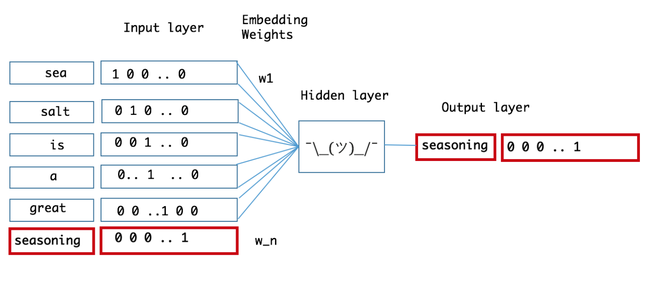
We used an SVM model as our baseline, and experimented with different kinds of deep learning models such as LSTMs, RNNs and seq2seq. Currently, a seq2seq model is used with 2 hidden layers.

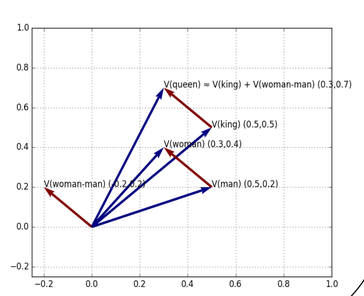
Visualization

The visualization consists of 2 pages – a user input for the article URL, and a results dashboard. Flask was used as a webserver to route webpage traffic, after which, a python scraper retrieves the data. Results are displayed using a keen.io dashboard template. The dashboard displays an intuitive explanation of the features using plotly and Google Trends of related topics using pytrends.

word2vec

word2vec is a deep learning algorithm that transforms words into vectors, so that words with similar meaning end up being close to each other. It allows us to use vector arithmetics to work with analogies, for example king – man + queen = woman.

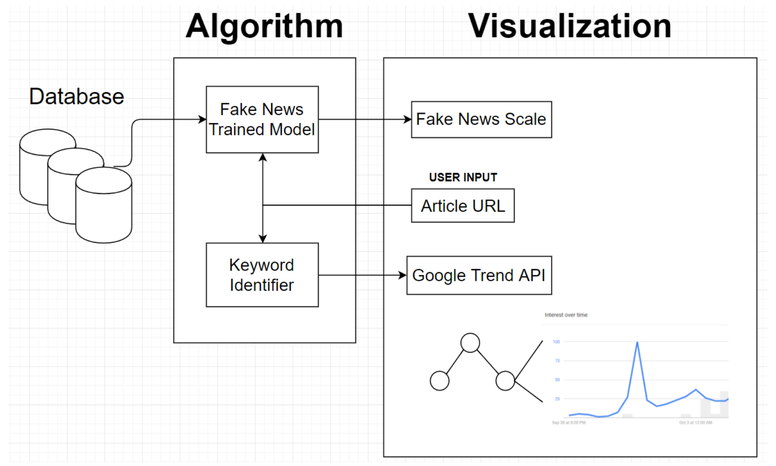


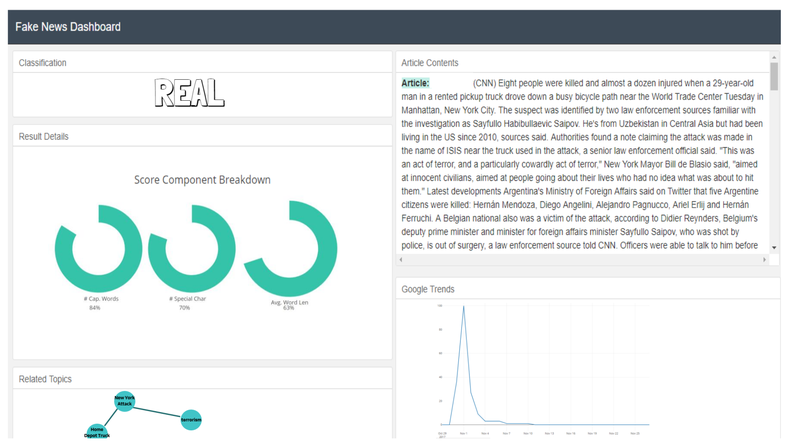


Data

Data was obtained mainly from two sources: Kaggle's 'fake news dataset containing news articles tagged as fake and Signal Media One's dataset containing credible news articles. A Python script was developed to scrape newspaper articles from any website for testing purposes. The final dataset contained of approx. 27,000 data points of articles with their meta data.

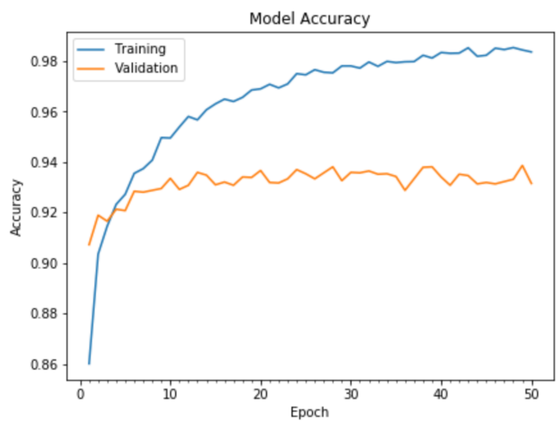
Data pipeline and visualization prototype





Results

The baseline SVM model reported an accuracy of 64% on out of sample news articles, as compared to the deep learning model which resulted in an accuracy of 93%. We benchmarked our model's accuracy with contemporary models being used for fake news classification. The figure below shows how the model accuracy varies for training and validation sets over epochs. The interactive visualization component of our tool is a key differentiating factor as that provides the user with the intuition to why the news was classified as fake/real.



References

[1] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation

[2] Samir Bajaj. 2017. "The Pope Has a New Baby!" Fake News Detection Using Deep Learning. (2017). https://web.stanford.edu/class/cs224n/reports/2710385.pdf

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. (2003). http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf

[4] Justin Chiu, Ajda Gokcen, Wenyi Wang, and Xiaohua Yan. 2013. Classification of Fake and Real Articles Based on Support Vector Machines. (2013). http://www.cs.cmu.edu/~xiaohuay/papers/report_11761.pdf