Outline & Summary of Project Goals and Deliverables

News webpage reliability study

Summary

I propose to complete an analysis comparing inaccurate news stories to accurate ones. This analysis will be done with both available databases and by scraping articles linked to by fact checking websites. Statistics will be presented in charts summarizing these findings. Clusters of sites sharing articles on the same topic will be created and presented. If time permits, a field will be available to the user to input a URL and receive a chart comparing statistics to the averages of accurate and inaccurate stories. An artificial intelligence will be created to predict the likeliness that a queried news source is accurate.

Goals & Deliverables

-Analysis will be done comparing stories considered true and stories considered false

- Pareto-like charts will be created for the most common words for true and false articles
- Similar charts will be created for sentiment
- Clusters of news sites will be created showing sites that post similar articles
- Charts for similar properties (e.g. occurrences of specific words or sentiments) will be created
- Charts showing differing properties will be created
- If time permits, the analysis will include other properties such as political leaning or countries most reported on
- A field will be provided for entering a URL to produce such charts on demand

- Machine learning will be implemented to predict whether an article is true/false, left/right leaning, etc.
  - A neural net can be implemented to determine the truthfulness of an article which can be provided via a URL. The last neuron will be a sigmoid, giving the probability of truth
  - A support vector machine is also applicable

- A number of clustering algorithms can be used to classify articles. These can be used in either an unsupervised manner, or by including the 'truth' value (1 or 0) as a dimension and clustering with an unknown article with truth value of .5

Completion Plan

- Fact checking articles such as those posted on PolitiFact and Snopes will be scraped using tools such as Scrapy.
  - For PolitiFact, for example, "pants on fire" and "True" rated articles will be used as two sets for analysis and for training the neural net.
  - For sites such as PolitiFact, similar articles making the same claim can be found under "Sources". These sources will be analyzed for the "True" set and "pants on fire" set
- These sources will be analyzed with tools such as WordNet, SentiWordNet or Google Brain. A tool such as Docker can be used to integrate such apps and web-based queries are also possible.
  - The number of occurrences of each synset (synonyms) will be collected for each source article
  - The most common synsets for either set will be determined.
  - The sentiment (positivity, negativity, objectivity) can be determined with available tools.
  - Additional properties based on the synset used, such as a score for "accusatory" or "laudatory" can be created by this analysis.
- A neural net, or other machine learning technique, can be implemented to provide a prediction for whether a given news source is likely reliable or unreliable
  - For a neural net, for example, the input layer could be occurrences of synsets, frequency of words (bag of words approach), sentiment, other sources with similar articles or other properties. The final neuron could be a sigmoid with "true" represented by an output of 1 and 0 representing "false". The output of the neural net would be something similar to a probability.
  - The hyperspace of occurrences of synsets and other parameters can be used for a support vector machine approach.

- A summary analysis comparing characteristics of false and true articles from news sites will be produced
  - Pareto-like histograms will be created for most common words, sentiments and other properties for true and untrue articles
  - Clusters of sites that post similar stories can be displayed
  - If time permits, reliability of sites might be analyzed statistically, or by machine learning for new sites by scraping them to analyze the reliability of recent articles
  - If time permits, properties exclusive of true and false can also be used as a basis for analysis. For example, the same graphs as described above can be created for left and right leaning websites. Other correlations between reliability and political leaning or other properties can be inferred.

Preliminary Results

- A dense neural net has been trained to distinguish between reliable and unreliable articles based on the words used
  - The neural net scores an accuracy of 99.9% on the training data and 95.4% on the test data
  - The architecture of the net is given as follows:
    - Three layers, input dimensions of 1000 x 65 neurons (relu) x 65 neurons (relu) x 1 neuron (sigmoid)
    - Dropout of .5 at each layer
    - L1 regularization on each layer (kernel regularization, lambda = .01)
    - AdaDelta optimizer used with binary cross-entropy for the loss function and accuracy as the metric
    - Net is trained with 2156 examples and validated with 414 examples (50% reliable/50% unreliable)
- Two methods of reducing the dimensionality of input data for methods other than neural networks were used
  - Each training example (dimensions of 1000) were reduced to two dimensions by two methods
    - Principal component analysis

    - A custom metric where the word most common in reliable articles compared to unreliable is given a weight of 500, second most common 499, etc. and most common in unreliable articles compared to reliable ones is given a weight of -500, second most common -499 etc. Weights were multiplied by scaled frequency of corresponding word in article.

- A support vector machine gives an accuracy of 95.9 % (training data)
- A single layer net was also written to test that the premise of using a neural net for this purpose in this manner was valid. As expected, the weights corresponding to inputs for words more common in unreliable articles were negative. Similarly, the weights corresponding to inputs for words more common in reliable articles, were positive.
- There are multiple ways in which clustering algorithms can be used to classify data. They can be used in an unsupervised manner to identify populations or they can be used with their truth label as a dimension and an unknown datum with a truth label of .5 to classify data
    - Methods giving accuracy of higher than 90% for data decomposed by PCA or the custom method (or both) are:
        - Agglomerative clustering (94.0%, 94.39%)
        - Spectral Clustering (91.4%)
        - Kmeans (91.3%)
- Successful predictive strategies could be used together to give a combined classification, for example, by voting (similar to a random forest) or by entering their outputs into another neural net