News Reliability Visualization

SARA KHEDR (PRESENTING MASTER’S PROJECT WORK)

FRONTIERS OF NETWORK SCIENCE: ASSIGNMENT 2
Problem History

Since the most recent election, we have seen a shift in the spread of ‘fake news’. These fake news stories were being repeatedly shared on social media (mainly Facebook) - helping increase traffic to dubious web news sources.

Throughout this, there were efforts to block and uncover this disinformation.

Google expands ‘Fact Check’ tool to flag up fake news in search results

Snopes

Hoaxy

Visualize the spread of claims and fact checking. Click here for a tutorial.
Problem Definition: Unreliability

Not all published truly made-up stories – some were clickbait-y, sensationalized, misleading, or even satirical.

Melissa Zimdar created a Google document, which went viral, categorizing news sources as: fake, satire, extreme bias, conspiracy, rumor mill, state news, junk science, hate news, etc.

My goal was to focus on a tool that could help users objectively analyze the reliability of a news source.
Dataset Collection

Goal: Create a wide-ranging, complete dataset of collected news articles from at least a 6 month time frame; used Melissa Zimdar’s list to maintain high variety

Details: 98 news sources, 121k articles, months April - October
Feature Extraction

Using article title and text, extracted **279** features mainly using **LIWC**

Also collected number of **Facebook** shares, comments, and reactions for each article

**LIWC** (Linguistic Inquiry Word Count Dictionary): uses set dictionaries to compute standardized feature values based on words within text

<table>
<thead>
<tr>
<th>Feature</th>
<th>Example Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Processes: Friends</td>
<td>pal, buddy, coworker</td>
</tr>
<tr>
<td>Affective Processes: Anxiety</td>
<td>nervous, afraid, tense</td>
</tr>
<tr>
<td>Cognitive Processes: Certainty</td>
<td>always, never</td>
</tr>
<tr>
<td>Cognitive Processes: Causation</td>
<td>because, effect, hence</td>
</tr>
<tr>
<td>Personal Concerns: Work</td>
<td>work, class, boss</td>
</tr>
<tr>
<td>Relativity: Time</td>
<td>hour, day, oclock</td>
</tr>
</tbody>
</table>
Data Issues/Considerations

In the smaller categories (i.e. junk science, conspiracy, hate, etc.), news sources didn’t publish articles as often. As a result, the number of articles in each classification was extremely unbalanced.

For analysis, I combined those categories into an **unreliable** classification, along with the **reliable** and **satire** classifications.

**Goal of Analysis:** Understand how to create a representative, accurate visual tool
Top Features

Reduced features to set of ‘important’ ones

Top features (based on random forest):

- Word Count
- Hear Words
- Adverbs
- Work Words
- Dashes
Visual Clustering

General separation of classified data

Changing this to source-level plotting instead of article-level reduces cluster spread
Classification Methods

**Goal**: Find methods for accurate classification within tool

- Naïve Bayes
- Decision Tree
- Random Forest

**Best Model**: Random Forest

**Overall Accuracy**: 75.5%
Turning Analysis into a Visual Tool

- Top 5 features initially shown
- Points can represent average, median, maximum
- User can select up to 4 features at a time
- User can select sources shown
- User can select date range

### Chart Settings

- **Show All Options**
- **Source Value:** Average
- **X-axis:** Word Count
- **Y-axis:** Hear Words
- ** Bubble Color:** Adverbs
- ** Bubble Size:** Work Words
- **Sources:** AP, Activist Post, Adding In
- **Date Range:** 2017-04-01 - 2017-10-31

### Data

<table>
<thead>
<tr>
<th>Total Sources</th>
<th>Total Articles</th>
<th>Most FB Shares</th>
<th>Most FB Comments</th>
<th>Most FB Reactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>92</td>
<td>121.0k</td>
<td>213.0k</td>
<td>185.0k</td>
<td>695.0k</td>
</tr>
</tbody>
</table>

### Source-level plotting
Note: Doesn’t have perfect separation, but point of tool is for an objective viewpoint – user can opt to get determined classifications if they choose.
Input Differences

Median

Average
Additional Visualization

Source specific page

Top 5 phrases extracted by month

Top 10 most shared articles

Source metadata

Activist Post

Type: Online
Country: United States
Founded: June 2010
Website: activistpost.com

Number of Articles per Month

Facebook Engagement

Lone Congresswoman Calls Out Trump For Insane Arms Deal With Human Rights Violator Saudi Arabia
Colossal Pedophile Ring Busts 800 Arrests 300 Kids Saved Corporate Media Ignores It
Forced Vaccinations Violate Constitutional Rights Sweden’s Parliament Declares
YouTube Begins Purging Alternative Media as the Deep State Marches Toward WW3
Julian Assange Just Destroyed The FBI CIA And Fake News In An Epic Tweetstorm
Children Now Face Fines And Arrest If They Don’t Get a Permit To Mow Grass For Money
Houston Prepares To Be Sprayed By Bee-Killing Chemicals To Fight Mosquitoes
Cops Detain Entire School Illegally Search/Grope 900 Kids Find NOTHING Parents Furious
Why Is ISIS Attacking The Philippines?
Kinder Morgan Paid $115,000 to Mass. State Police to Stop Protests Against Pipeline
Future Graph-Based Analysis

**Goal:** Discover which sources report on similar stories and where information originates from

**Nodes:** News sources

**Directed Edges:** $e=(u,v)$ if source $u$ and source $v$ have published an article with the same top phrase on the same day and source $u$ published before source $v$

**Edge Weight:** Similarity between the two articles’ text using TF-IDF and Cosine Similarity $[0,1]$

**Edge Metadata:** Top phrase

$$w_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right)$$

- $tf_{i,j}$ = number of occurrences of $i$ in $j$
- $df_i$ = number of documents containing $i$
- $N$ = total number of documents
Additional Future Work

> Collect information on whether source has published an article that has been proven to be fake in the past

> Integration with colleague's Reddit news toolkit
  ◦ Gives likelihood of an article being published in a set of subreddits (e.g. /news, /conspiracy, /thedonald)

> Publication as a dataset paper and potentially jointly demo paper
References


