A Stylometric Inquiry into Hyperpartisan and Fake News

Martin Potthast*, Johannes Kiesel†, Kevin Reinartz†, Janek Bevendorff†, Benno Stein†
*Leipzig University, †Bauhaus-Universität Weimar

webis.de

ACL, July 16th, 2018
Fake news is not our friend.

We’re committed to reducing its spread; so we’re working with more fact-checkers globally, improving our technology, and giving you background information on the articles in your News Feed.

Find out more: fb.me/changesAU

facebook
...it is very possible that those sources don't exist but are made up by fake news writers. #FakeNews is the enemy!
What are Fake News?

Disinformation displayed as news articles
What are Fake News?

Disinformation displayed as news articles

7 TYPES OF MIS- AND DISINFORMATION

**SATIRE OR PARODY**
No intention to cause harm but has potential to fool

**MISLEADING CONTENT**
Misleading use of information to frame an issue or individual

**IMPOSTER CONTENT**
When genuine sources are impersonated

**FABRICATED CONTENT**
New content is 100% false, designed to deceive and do harm

**FALSE CONNECTION**
When headlines, visuals or captions don’t support the content

**FALSE CONTEXT**
When genuine content is shared with false contextual information

**MANIPULATED CONTENT**
When genuine information or imagery is manipulated to deceive

Image: Claire Wardle, First Draft
What are Fake News?

Disinformation displayed as news articles

7 TYPES OF MIS- AND DISINFORMATION

SATIRE OR PARODY
No intention to cause harm but has potential to fool

MISLEADING CONTENT
Misleading use of information to frame an issue or individual

IMPOSTER CONTENT
When genuine sources are impersonated

FABRICATED CONTENT
New content is 100% false, designed to deceive and do harm

FALSE CONNECTION
When headlines, visuals or captions don’t support the content

FALSE CONTEXT
When genuine content is shared with false contextual information

MANIPULATED CONTENT
When genuine information or imagery is manipulated to deceive

Image: Claire Wardle, First Draft
A Stylometric Inquiry into Hyperpartisan “News” and “News” in False Context and/or with Content that is Impostered, Manipulated, and/or Fabricated

Martin Potthast*, Johannes Kiesel†, Kevin Reinartz†, Janek Bevendorff†, Benno Stein†
*Leipzig University, †Bauhaus-Universität Weimar

webis.de
ACL, July 16th, 2018
The Political Spectrum

The left-right political spectrum is a system of classifying political positions, ideologies and parties. Left-wing politics and right-wing politics are often presented as opposed, although either may adopt stances from the other side. [Wikipedia]
The Political Spectrum

The left-right political spectrum is a system of classifying political positions, ideologies and parties. Left-wing politics and right-wing politics are often presented as opposed, although either may adopt stances from the other side. [Wikipedia]
The left-right political spectrum is a system of classifying political positions, ideologies and parties. Left-wing politics and right-wing politics are often presented as opposed, although either may adopt stances from the other side. 

Partisan: someone with a psychological identification with one major party.
The Political Spectrum

The left-right political spectrum is a system of classifying political positions, ideologies and parties. Left-wing politics and right-wing politics are often presented as opposed, although either may adopt stances from the other side. [Wikipedia]

Partisan: someone with a psychological identification with one major party. [Wikipedia]

News media reporting on politics can be aligned on this spectrum as well.

We are observing an increasing number of hyperpartisan news publishers.
Fake News and Hyperpartisan News

Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate

A BuzzFeed News analysis found that three big right-wing Facebook pages published false or misleading information 38% of the time during the period analyzed, and three large left-wing pages did so in nearly 20% of posts.

By Craig Silverman (BuzzFeed Founding Editor, Canada), Lauren Strapagiel (BuzzFeed Staff), Hamza Shaban (BuzzFeed News Reporter), Ellie Hall (BuzzFeed News Reporter), Jeremy Singer-Vine (BuzzFeed News Reporter)
Why are Fake News Published by Hyperpartisan Pages?

<table>
<thead>
<tr>
<th><strong>MISINFORMATION MATRIX</strong></th>
<th>SATIRE OR PARODY</th>
<th>FALSE CONNECTION</th>
<th>MISLEADING CONTENT</th>
<th>FALSE CONTEXT</th>
<th>IMPOSTER CONTENT</th>
<th>MANIPULATED CONTENT</th>
<th>FABRICATED CONTENT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POOR JOURNALISM</strong></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TO PARODY</strong></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>TO PROVOKE OR TO ‘PUNK’</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PASSION</strong></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>PARTISANSHIP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PROFIT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>POLITICAL INFLUENCE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PROPAGANDA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Image: Claire Wardle, First Draft
## Why are Fake News Published by Hyperpartisan Pages?

### Misinformation Matrix

**FirstDraft**

<table>
<thead>
<tr>
<th></th>
<th>Satire or Parody</th>
<th>False Connection</th>
<th>Misleading Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor Journalism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To Parody</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>To Provoke Or To 'Punk'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partisanship</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political Influence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propaganda</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Image:** Claire Wardle, First Draft

---

@KieselJohannes

14
Fake News Detection

Knowledge-based
- Requires political knowledge base
- Unavailable ahead of time
- We cannot trust the web

Context-based
- Limited to social media platforms
- Part of damage already done

Style-based
- Allows for pre-posting check
- Real-time reaction possible
- Hard to mask
- But are style differences sufficient?

Taxonomy of Approaches

Knowledge-based (also called fact checking)
- Information retrieval
- Semantic web / LOD
- Social network analysis
- Deception detection
- Text categorization

Information retrieval
- Etzioni et al., 2018
- Magdy and Wanas, 2010
- Ginsca et al., 2015
- Wu et al., 2014
- Ciampaglia et al., 2015
- Shi and Weninger, 2016

Semantic web / LOD
- Long et al., 2017
- Mocanu et al., 2015
- Acemoglu et al., 2010
- Kwon et al., 2013
- Ma et al., 2017
- Volkova et al., 2017
- Budak et al., 2011
- Nguyen et al. 2012
- Derczynski et al., 2017
- Tambuscio et al., 2015

Social network analysis
- Wei et al., 2013
- Chen et al., 2015
- Rubin et al., 2015
- Wang et al., 2017
- Bourgonje et al., 2017
- Afroz et al., 2012
- Badaskar et al., 2008
- Rubin et al., 2016
- Yang et al., 2017
- Rashkin et al., 2017
- Horne and Adali, 2017
- Pérez-Rosas et al., 2017
Fake News and Hyperpartisan News  Corpus Construction
## Orientation

<table>
<thead>
<tr>
<th>Publisher</th>
<th>true</th>
<th>mix</th>
<th>false</th>
<th>n/a</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Center</strong></td>
<td>806</td>
<td>8</td>
<td>0</td>
<td>12</td>
<td>826</td>
</tr>
<tr>
<td>ABC News</td>
<td>90</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>95</td>
</tr>
<tr>
<td>CNN</td>
<td>295</td>
<td>4</td>
<td>0</td>
<td>8</td>
<td>307</td>
</tr>
<tr>
<td>Politico</td>
<td>421</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>424</td>
</tr>
<tr>
<td><strong>Left-wing</strong></td>
<td>182</td>
<td>51</td>
<td>15</td>
<td>8</td>
<td>256</td>
</tr>
<tr>
<td>Addicting Info</td>
<td>95</td>
<td>25</td>
<td>8</td>
<td>7</td>
<td>135</td>
</tr>
<tr>
<td>Occupy Democrats</td>
<td>59</td>
<td>25</td>
<td>7</td>
<td>0</td>
<td>91</td>
</tr>
<tr>
<td>The Other 98%</td>
<td>28</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td><strong>Right-wing</strong></td>
<td>276</td>
<td>153</td>
<td>72</td>
<td>44</td>
<td>545</td>
</tr>
<tr>
<td>Eagle Rising</td>
<td>106</td>
<td>47</td>
<td>25</td>
<td>36</td>
<td>214</td>
</tr>
<tr>
<td>Freedom Daily</td>
<td>49</td>
<td>24</td>
<td>22</td>
<td>4</td>
<td>99</td>
</tr>
<tr>
<td>Right Wing News</td>
<td>121</td>
<td>82</td>
<td>25</td>
<td>4</td>
<td>232</td>
</tr>
<tr>
<td><strong>Σ</strong></td>
<td>1264</td>
<td>212</td>
<td>87</td>
<td>64</td>
<td>1627</td>
</tr>
</tbody>
</table>

Annotations provided by journalists at BuzzFeed
## Fake News and Hyperpartisan News

### Selected Results

<table>
<thead>
<tr>
<th>Orientation</th>
<th>Publisher</th>
<th>true</th>
<th>mix</th>
<th>false</th>
<th>n/a</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center</td>
<td>ABC News</td>
<td>90</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>295</td>
<td>4</td>
<td>0</td>
<td>8</td>
<td>307</td>
</tr>
<tr>
<td></td>
<td>Politico</td>
<td>421</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>424</td>
</tr>
<tr>
<td>Left-wing</td>
<td>Addicting Info</td>
<td>95</td>
<td>25</td>
<td>8</td>
<td>7</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>Occupy Democrats</td>
<td>59</td>
<td>25</td>
<td>7</td>
<td>0</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>The Other 98%</td>
<td>98</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>Right-wing</td>
<td>Eagle Rising</td>
<td>106</td>
<td>47</td>
<td>25</td>
<td>4</td>
<td>214</td>
</tr>
<tr>
<td></td>
<td>Freedom Daily</td>
<td>49</td>
<td>24</td>
<td>22</td>
<td>4</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Right Wing News</td>
<td>121</td>
<td>82</td>
<td>25</td>
<td>4</td>
<td>232</td>
</tr>
<tr>
<td></td>
<td>SUM</td>
<td>1264</td>
<td>212</td>
<td>87</td>
<td>64</td>
<td>1627</td>
</tr>
</tbody>
</table>

Fake News Detection

- Precision ≈ 42%
- Recall ≈ 41%

Annotations provided by journalists at BuzzFeed
## Fact-checking results

<table>
<thead>
<tr>
<th>Orientation</th>
<th>Publisher</th>
<th>true</th>
<th>mix</th>
<th>false</th>
<th>n/a</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Center</strong></td>
<td>ABC News</td>
<td>90</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>295</td>
<td>4</td>
<td>0</td>
<td>8</td>
<td>307</td>
</tr>
<tr>
<td></td>
<td>Politico</td>
<td>421</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>424</td>
</tr>
<tr>
<td><strong>Left-wing</strong></td>
<td>Addicting Info</td>
<td>95</td>
<td>25</td>
<td>8</td>
<td>7</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>Occupy Democrats</td>
<td>59</td>
<td>25</td>
<td>7</td>
<td>0</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>The Other 98%</td>
<td>98</td>
<td>28</td>
<td>1</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td><strong>Right-wing</strong></td>
<td>Eagle Rising</td>
<td>106</td>
<td>47</td>
<td>25</td>
<td>36</td>
<td>214</td>
</tr>
<tr>
<td></td>
<td>Freedom Daily</td>
<td>49</td>
<td>24</td>
<td>22</td>
<td>4</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Right Wing News</td>
<td>121</td>
<td>82</td>
<td>25</td>
<td>4</td>
<td>232</td>
</tr>
<tr>
<td></td>
<td><strong>Σ</strong></td>
<td>1264</td>
<td>212</td>
<td>87</td>
<td>64</td>
<td>1627</td>
</tr>
</tbody>
</table>

### Orientation Detection

- **Precision** ≈ 21%
- **Recall** ≈ 20%

- **Precision** ≈ 56%
- **Recall** ≈ 59%

Annotations provided by journalists at BuzzFeed
## Fake News and Hyperpartisan News

### Selected Results

<table>
<thead>
<tr>
<th>Orientation</th>
<th>Publisher</th>
<th>Fact-checking results</th>
<th></th>
<th></th>
<th></th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>true</td>
<td>mix</td>
<td>false</td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Center</strong></td>
<td>806</td>
<td>8</td>
<td>0</td>
<td>12</td>
<td></td>
<td>826</td>
</tr>
<tr>
<td><strong>ABC News</strong></td>
<td>295</td>
<td>4</td>
<td>0</td>
<td>8</td>
<td></td>
<td>307</td>
</tr>
<tr>
<td><strong>CNN</strong></td>
<td>421</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td></td>
<td>424</td>
</tr>
<tr>
<td><strong>Politico</strong></td>
<td>182</td>
<td>51</td>
<td>15</td>
<td>8</td>
<td></td>
<td>256</td>
</tr>
<tr>
<td><strong>Left-wing</strong></td>
<td>95</td>
<td>25</td>
<td>8</td>
<td>7</td>
<td></td>
<td>135</td>
</tr>
<tr>
<td><strong>Addicting In</strong></td>
<td>59</td>
<td>25</td>
<td>7</td>
<td>0</td>
<td></td>
<td>91</td>
</tr>
<tr>
<td><strong>Occupy Der</strong></td>
<td>98</td>
<td>28</td>
<td>1</td>
<td>0</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td><strong>Right-wing</strong></td>
<td>276</td>
<td>153</td>
<td>72</td>
<td>44</td>
<td></td>
<td>545</td>
</tr>
<tr>
<td><strong>Eagle Rising</strong></td>
<td>106</td>
<td>47</td>
<td>25</td>
<td>36</td>
<td></td>
<td>214</td>
</tr>
<tr>
<td><strong>Freedom Daily</strong></td>
<td>49</td>
<td>24</td>
<td>22</td>
<td>4</td>
<td></td>
<td>99</td>
</tr>
<tr>
<td><strong>Right Wing News</strong></td>
<td>121</td>
<td>82</td>
<td>25</td>
<td>4</td>
<td></td>
<td>232</td>
</tr>
<tr>
<td><strong>Σ</strong></td>
<td>1264</td>
<td>212</td>
<td>87</td>
<td>64</td>
<td></td>
<td>1627</td>
</tr>
</tbody>
</table>

### Hyperpartisanship Detection

- Precision ≈ 69%
- Recall ≈ 89%

Annotations provided by journalists at BuzzFeed
Fake News and Hyperpartisan News

How can it be that the alt left and the alt right cannot be distinguished from the mainstream, when both together (hyperpartisan news) can be?

Alt-left  Left  Center  Right  Alt-right

Liberal  Conservative

Hyperpartisan  Partisan  Partisan  Hyperpartisan
Fake News and Hyperpartisan News

How can it be that the alt left and the alt right cannot be distinguished from the mainstream, when both together (hyperpartisan news) can be?
Fake News and Hyperpartisan News

How can it be that the alt left and the alt right cannot be distinguished from the mainstream, when both together (hyperpartisan news) can be?

The horseshoe theory asserts that the alt left and the alt right, rather than being at opposite and opposing ends of a linear political continuum, in fact closely resemble one another, much like the ends of a horseshoe. [Wikipedia]
Horseshoe Validation Experiment I  Leave-out Classification

left-wing

center

right-wing
- **Classifier** is trained to distinguish left-wing and center articles
- Right-wing articles are used for testing
Horseshoe Validation Experiment I  Leave-out Classification

- **Classifier** is trained to distinguish left-wing and center articles
- Right-wing articles are used for testing
- Majority of right-wing articles are classified as left-wing rather than center

left-wing | center | right-wing
---|---|---
74% | 26% |
Horseshoe Validation Experiment I  Leave-out Classification

- **Classifier** is trained to distinguish left-wing and center articles
- Right-wing articles are used for testing
- Majority of right-wing articles are classified as left-wing rather than center

<table>
<thead>
<tr>
<th>left-wing</th>
<th>center</th>
<th>right-wing</th>
</tr>
</thead>
<tbody>
<tr>
<td>74%</td>
<td>26%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>left-wing</th>
<th>center</th>
<th>right-wing</th>
</tr>
</thead>
<tbody>
<tr>
<td>34%</td>
<td>66%</td>
<td></td>
</tr>
</tbody>
</table>
Horseshoe Validation Experiment II  Unmasking [Koppel/Schler 2004]
### Horseshoe Validation Experiment II

**Unmasking**

[Koppel/Schler 2004]

---

**Diagram**: Two figures A and B are connected by a line labeled with a question mark (\(\equiv\)). The figures depict paper documents, suggesting a validation or comparison process between the two sets of data.

**Table**: A 3x3 table for each set A and B is shown, containing numerical values that likely represent data or results from the experiment.

---

**Note**: The table and diagram indicate a comparison or validation of data sets A and B, which may be related to the unmasking process described in the text.
Horseshoe Validation Experiment II  Unmasking  [Koppel/Schler 2004]
Horseshoe Validation Experiment II  Unmasking [Koppel/Schler 2004]

\[
\begin{array}{ccccccccc}
0.3 & 0.1 & 0.2 & 0.0 & 0.3 & 0.1 & 0.5 & 0.0 & 0.4 \\
0.2 & 0.2 & 0.3 & 0.1 & 0.2 & 0.2 & 0.1 & 0.3 & 0.4 \\
0.0 & 0.4 & 0.2 & 0.4 & 0.0 & 0.1 & 0.1 & 0.2 & 0.4 \\
0.1 & 0.0 & 0.2 & 0.5 & 0.1 & 0.1 & 0.4 & 0.2 & 0.1 \\
0.2 & 0.2 & 0.1 & 0.0 & 0.2 & 0.0 & 0.4 & 0.2 & 0.3 \\
0.0 & 0.2 & 0.1 & 0.0 & 0.2 & 0.0 & 0.2 & 0.3 & 0.2 \\
0.3 & 0.3 & 0.4 & 0.1 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 \\
0.1 & 0.3 & 0.4 & 0.2 & 0.1 & 0.2 & 0.1 & 0.1 & 0.2 \\
0.0 & 0.1 & 0.2 & 0.1 & 0.2 & 0.1 & 0.1 & 0.2 & 0.1 \\
0.3 & 0.3 & 0.2 & 0.1 & 0.1 & 0.0 & 0.2 & 0.3 & 0.1 \\
0.0 & 0.3 & 0.3 & 0.1 & 0.3 & 0.1 & 0.0 & 0.0 & 0.1 \\
\end{array}
\]
Horseshoe Validation Experiment II

Unmasking [Koppel/Schler 2004]
Horseshoe Validation Experiment II  Unmasking  [Koppel/Schler 2004]
Horseshoe Validation Experiment II  Unmasking  [Koppel/Schler 2004]
Typical learning characteristic for ...
Typical learning characteristic for ...
The typical learning characteristic can be learned. → “Meta Learning”

We apply Unmasking to distinguish style genres.
Horseshoe Validation Experiment II  Unmasking  [Koppel/Schler 2004]

![Graph showing normalized accuracy over iterations for different comparisons: mainstream vs left, mainstream vs right, and left vs right.](image_url)
Summary and Outlook

- Hyperpartisan news pages produce relatively many fake news articles
- Hyperpartisan news can be distinguished quite well based on style
- Style-based detection allows for real-time detection
  - Political extremism in news can be ousted or at least flagged

- The style of alt left and alt right news is very similar
- Linguistic evidence for the horseshoe theory of the political spectrum?
  - Large-scale analysis required
SemEval 2019 Task 4
Hyperpartisan News Detection
SemEval 2019 Task 4

Hyperpartisan News Detection

webis.de/events/semeval-19
Style Model

Features

- n-grams with \( n \in [1, 3] \) of characters, stop words, parts-of-speech
- 10 readability scores
- Dictionary features based on General Inquirer
- Ratios of quoted words, external links, number of paragraphs, and their average length

Feature selection

- Discard word features (n-gram features) occurring in less than 2.5% (10%) of documents

Training set

- Balancing using oversampling
- Publishers are not represented in both training and test set

Learning algorithm

- WEKA’s random forest with default parameters