Building a Corpus for Hyperpartisan News Detection

Master’s Thesis by Payam Adineh

Referees:
Prof. Dr. Benno Stein
Prof. Dr. Siegmund

Advisors:
Prof. Dr. Martin Potthast
Johannes Kiesel
Motivation
Fake News

An 'extremely credible source' has called my office and told me that @BarackObama's birth certificate is a fraud.
9:23 PM - Aug 6, 2012
❤️ 21K  ❱ 32.7K people are talking about this
Partisan News

- Ideologically biased creation and distribution of news by authors and organizations.
- Hyperpartisanship has similar meaning, and is used especially when news producers manipulate the coverage of the reality with a tendency to the right or left wing political parties.

**After each school shooting!**

- Change gun control laws
- Arm teachers with guns
Partisan News

- Ideologically biased creation and distribution of news by authors and organizations
- Hyperpartisanship has similar meaning, and is used especially when news producers extremely manipulate the coverage of the reality with a tendency to the right or left wing political parties.

“grab them by the p*ssy”

Example of sexual harrasment
Partisan News

- “Journalism in the United States was born partisan and remained, for much of its history, loud, boisterous and combative”. Mitchel Stephens, a journalism professor
- “Political conflicts spill from political aspects of our lives to nonpolitical aspects of our lives. In the modern era, we view party identity as something akin to gender, ethnicity or race.” Sean Westwood
- “The more partisan we become, the more emotionally we react to normal political events. The angrier the electorate, the less capable we are of finding common ground on policies, or even of treating our opponents like human beings.” Lilliana Mason
Partisan News

- “Journalism in the United States was born partisan and remained, for much of its history, loud, boisterous and combative”. **Mitchel Stephens**, a journalism professor
- “Political conflicts spill from political aspects of our lives to nonpolitical aspects of our lives. In the modern era, we view party identity as something akin to gender, ethnicity or race.” **Sean Westwood**
- “The more partisan we become, the more emotionally we react to normal political events. The angrier the electorate, the less capable we are of finding common ground on policies, or even of treating our opponents like human beings.” **Lilliana Mason**

How we can take an action against this phenomenon in the field of computer science?
Article Collection

News Producers Discovery

Article URL Collection
How we can find the articles?

- Find news producers.

What kind of news producers?

- Cover the United States news
- Must have a political section
- Must own a Facebook page or a working sitemap
BuzzFeed Partisan News Sites List

- BuzzFeed published article about ecosystem of partisan websites
- This list contains 677 political news websites
- Each website has following information:
  - URL
  - Political Category (Left, Right)
  - Facebook Id

<table>
<thead>
<tr>
<th>Sites</th>
<th>Left</th>
<th>Right</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number</td>
<td>178</td>
<td>499</td>
<td>677</td>
</tr>
<tr>
<td>Duplicate</td>
<td>9</td>
<td>39</td>
<td>48</td>
</tr>
<tr>
<td>Unavailable</td>
<td>31</td>
<td>140</td>
<td>171</td>
</tr>
<tr>
<td>Facebook</td>
<td>120</td>
<td>231</td>
<td>351</td>
</tr>
<tr>
<td>Sitemap</td>
<td>18</td>
<td>89</td>
<td>107</td>
</tr>
</tbody>
</table>
Media Biased Fact Check

- A fact-checking website which labels websites according to their political orientation
- Media Biased Fact Check contains 1310 political news websites
- They labeled websites in extreme left, left-center, least biased, right-center, and extreme right categories
- Focused on a variety of topics like politics, economy, and sport, to name but a few
- We found those which have political section and own a sitemap

<table>
<thead>
<tr>
<th></th>
<th>Left</th>
<th>Leftcenter</th>
<th>Least</th>
<th>Rightcenter</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Subjects</td>
<td>290</td>
<td>447</td>
<td>362</td>
<td>211</td>
<td>1310</td>
</tr>
<tr>
<td>Political</td>
<td>12</td>
<td>49</td>
<td>28</td>
<td>20</td>
<td>109</td>
</tr>
</tbody>
</table>
Article Collection

- News Producers Discovery
- Article URL Collection
Facebook Pages

- Due to the popularity of this platform, most of the news producers tried to absorb a new audience as well as keep the current ones by using this medium.
- News producers not only publish the news on their websites, but also they share a link to their new articles on Facebook.
- We used Facebook Graph API to retrieve all the posts published by news producers.
Sitemap

- A sitemap is a systematic and hierarchical view of the website in an XML format
- It usually contains all the links which is published by a website
- We only need the political news, so we filtered the links’ list
Sitemap - Filtering Sitemap index and Sitemap URL

<table>
<thead>
<tr>
<th>Sitemap</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://www.snopes.com/post-sitemap.xml">https://www.snopes.com/post-sitemap.xml</a></td>
<td>2016-01-20 09:12:08:00</td>
</tr>
<tr>
<td><a href="https://www.snopes.com/post-sitemap3.xml">https://www.snopes.com/post-sitemap3.xml</a></td>
<td>2016-09-06 09:34:07:00</td>
</tr>
</tbody>
</table>

Filtering keyword

Keyword to filter URLs

```
<url>
<lastmod>2018-02-08T19:18:20+00:00</lastmod>
<changefreq>monthly</changefreq>
<priority>0.7</priority>
</url>
```

Tag and Value to filter URLs

```
<url>
<loc>https://www.mercurynews.com/2018/02/08/mccain-staying-away-from-dc-over-flu-concerns/</loc>
<lastmod>2018-02-08T22:45:12+00:00</lastmod>
<pub date="2018-02-08T22:45:12+00:00">The Mercury News</pub>
<language>en-US</language>
```
BuzzFeed and Media Bias Fact Check Dissagreement

<table>
<thead>
<tr>
<th>News Producers</th>
<th>MBFC</th>
<th>BuzzFeed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consortiumnews.com</td>
<td>Least</td>
<td>left</td>
</tr>
<tr>
<td>Amgreatness.com</td>
<td>Rightcenter</td>
<td>Right</td>
</tr>
<tr>
<td>Rare.us</td>
<td>Rightcenter</td>
<td>Right</td>
</tr>
<tr>
<td>Theamericanconservative.com</td>
<td>Rightcenter</td>
<td>Right</td>
</tr>
<tr>
<td>Washingtonexaminer.com</td>
<td>Rightcenter</td>
<td>Right</td>
</tr>
<tr>
<td>Liberalmountain.com</td>
<td>Leftcenter</td>
<td>Left</td>
</tr>
<tr>
<td>Billmoyers.com</td>
<td>Leftcenter</td>
<td>Left</td>
</tr>
<tr>
<td>Mintpressnews.com</td>
<td>Leftcenter</td>
<td>Left</td>
</tr>
<tr>
<td>Secondnexus.com</td>
<td>Leftcenter</td>
<td>Left</td>
</tr>
</tbody>
</table>
## URLs Statistics

<table>
<thead>
<tr>
<th>Platform</th>
<th>Left</th>
<th>Leftcenter</th>
<th>Least</th>
<th>Rightcenter</th>
<th>Right</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BuzzFeed Facebook</td>
<td>169,091</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>254,673</td>
<td>423,764</td>
</tr>
<tr>
<td>BuzzFeed Sitemap</td>
<td>70,279</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>508,929</td>
<td>579,208</td>
</tr>
<tr>
<td>MBFC Sitemap</td>
<td>628,638</td>
<td>1,392,187</td>
<td>750,872</td>
<td>420,218</td>
<td>-</td>
<td>3,191,915</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>868,008</td>
<td>1,392,187</td>
<td>750,872</td>
<td>420,218</td>
<td>763,602</td>
<td>4,194,887</td>
</tr>
</tbody>
</table>
Corpus Construction

- Archiving
- Distributed Storage
- Distributed Archiving
- Main Content Extraction
- Corpus Formatting
Select Archiving Tool

- HTML content is not enough to get a content of a web page
- A lots of external sources such as JavaScript files, CSS, multimedia files
- Sometimes content loads by client-side scripts
- Web ARChive is the format we used to archive web pages
- We employed Webis Web Archiver
Checking the Producers and Archiving Tool

- Archived three random articles from each producer (567 total) to find the possible errors

<table>
<thead>
<tr>
<th>Problem</th>
<th>Affected Producers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modal Windows</td>
<td>140</td>
</tr>
<tr>
<td>Down</td>
<td>68</td>
</tr>
<tr>
<td>Share Others</td>
<td>12</td>
</tr>
<tr>
<td>Read More</td>
<td>3</td>
</tr>
<tr>
<td>Homepage</td>
<td>3</td>
</tr>
<tr>
<td>Captcha</td>
<td>2</td>
</tr>
</tbody>
</table>
Checking the Producers and Archiving Tool

- Archived three random articles from each producer (567 total) to find the possible errors

<table>
<thead>
<tr>
<th>Problem</th>
<th>Affected Producers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modal Windows</td>
<td>140</td>
</tr>
<tr>
<td>Down</td>
<td>68</td>
</tr>
<tr>
<td>Share Others</td>
<td>12</td>
</tr>
<tr>
<td>Read More</td>
<td>3</td>
</tr>
<tr>
<td>Homepage</td>
<td>3</td>
</tr>
<tr>
<td>Captcha</td>
<td>2</td>
</tr>
</tbody>
</table>
Modal Windows Examples
Memphis Counselor Kelli Davis Calls Police 'Thugs in Blue,' Accused Of Leaking Their Personal Inform

Memphis, TN - A Post Traumatic Stress Disorder (PTSD) counselor, Kelli Davis, made her views about police known when she posted on social media that they were "thugs in blue."

Memphis, TN - A Post Traumatic Stress Disorder (PTSD) counselor, Kelli Davis, made her views about police known when she posted on social media that they were "thugs in blue." Now she is also suspected of leaking personal and confidential information about police officers.

According to WMC Action News5, controversial Facebook posts have placed Davis in the spotlight, such as the one below:

"My hope is that some of the INEXPERIENCED THUGS IN BLUE will think twice before pulling the trigger when they attempt to arrest a black or brown person - their inexperienced, THUGGISH taxpayer paid ACTIONS indeed have a RIPPLE affect!"
Example to Remove Modal Window

domain = uri.getHost();
switch (domain) {
  case "americasfreedomfighters.com":
    code = "document.querySelector('#modaal-close').click();";
    break;
  case "angrypatriotmovement.com":
    code = "document.querySelector('#revexitcloseme').click();";
    break;
}

Archive Quality Check

Score 5

Score 1
Archive Quality Check

- Archived and reproduced 100 random articles from each news producer
- Automatic and manual archive quality check

<table>
<thead>
<tr>
<th></th>
<th>Producers</th>
<th>Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>475</td>
<td>44,806</td>
</tr>
<tr>
<td>At Least One Successful Archive</td>
<td>450</td>
<td>38,309</td>
</tr>
</tbody>
</table>
Archive Quality Score Chart
Archive Quality Check Discovered Problems

<table>
<thead>
<tr>
<th>Problem</th>
<th>Number of Producers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modal Windows</td>
<td>33</td>
</tr>
<tr>
<td>Park Domain</td>
<td>25</td>
</tr>
<tr>
<td>Reproduction Problem</td>
<td>8</td>
</tr>
<tr>
<td>Gateway Error</td>
<td>3</td>
</tr>
<tr>
<td>Archiving Problem</td>
<td>2</td>
</tr>
<tr>
<td>Captcha</td>
<td>1</td>
</tr>
<tr>
<td>Sitemap Issue</td>
<td>1</td>
</tr>
</tbody>
</table>
# Archive Quality Check Discovered Problems

<table>
<thead>
<tr>
<th>Problem</th>
<th>Number of Producers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modal Windows</td>
<td>33</td>
</tr>
<tr>
<td>Park Domain</td>
<td>25</td>
</tr>
<tr>
<td>Reproduction Problem</td>
<td>8</td>
</tr>
<tr>
<td>Gateway Error</td>
<td>3</td>
</tr>
<tr>
<td>Archiving Problem</td>
<td>2</td>
</tr>
<tr>
<td>Captcha</td>
<td>1</td>
</tr>
<tr>
<td>Sitemap Issue</td>
<td>1</td>
</tr>
</tbody>
</table>
Corpus Construction

- Archiving
- Distributed Storage
- Distributed Archiving
- Main Content Extraction
- Corpus Formatting
How to Store the Data?

- Archiving web pages is a very time-consuming process
- Hardware failures are unavoidable
- We need to store the data in a reliable storage
- Hadoop Distributed File System
  - Distributed file system
  - Designed to store large data sets reliably
- We have huge number of files
- Need a to store them in a format which is convenient for both storage and future computation
- MapFile is chosen format
  - It is similar to SequenceFile, with the random access feature
- In the end we stored 64TB of information on HDFS
Corpus Construction

- Archiving
- Distributed Storage
- Distributed Archiving
- Main Content Extraction
- Corpus Formatting
- We have 3,651,229 news articles to crawl
- With single machine and single thread: 600 pages per day
- It means 6,000 days to crawl all the pages
- We employed 80 machines from Betaweb cluster
- Around 45,000 articles per machine
Task Control Automation

- 80 machines is so difficult to control manually
- We develop Archiving baseline with three subtask
  1. Archiving and Reproduction
  2. Archive Quality Check
  3. MapFiles Creation
Archiving and Reproduction

- We used the Webis Commands' Task Manager
- To speed up the archiving process we run 3 threads on each machine

Automatic Archive Quality Assurance

- Find new done tasks every 24 hours
- Copy archive and reproduction images in two different folders
- Run archive quality check

MapFiles Creation

- Recognize verified tasks by archive quality assurance every 24 hours
- Create a Mapfile for each 1000 archive articles
- Copy the mapfiles to HDFS
Corpus Construction - Distributed Archiving
Corpus Construction

- Archiving
- Distributed Storage
- Distributed Archiving
- Main Content Extraction
- Corpus Formatting
We have HTML pages, but we need the main parts

Main part of the HTML page

- **Headline**: Title of the article
- **Content**: Main content of article
- **Author**: The person who wrote the article
- **Date**: The date of publication
Writing the Wrappers

- 383 news producers, 383 different layout
- Query selectors for each producer to find required information
Wrapper Example

```java
wrappers.addWrapper(new UriSpecificWrapper(  // This wrapper will only be used for leftvoice.org  // optional second parameter for path matching  new UriPredicate("leftvoice.org"),  // This wrapper  new SelectorBasedWrapper()  .withTag(Wrapper.TAG_TITLE, "article div.header-articulo h1")  .withTag(Wrapper.TAG_AUTHOR, "article div.autor-articulo a")  .withTag(Wrapper.TAG_DATE, "article div.row:nth-of-type(2) div.col-md-12 span")  .withTag(Wrapper.TAG_CONTENT, new String[] {  "article div.articulo p",  })));
```
Corpus Construction

- Archiving
- Distributed Storage
- Distributed Archiving
- Main Content Extraction
- Corpus Formatting
Corpus Formats

- JSON lines
- XML (SemEval 2019 task)

Fields Description:
- **id**: Article id which is a number with 7 digits.
- **published-at**: A date in ISO 8601 format in which a news producer published the article.
- **title**: Headline of the article.
- **content**: The field content is used to store the main content of the article.
- **bias**: This field indicates the orientation of an article which can be "left", "left-center", "least", "right-center", or "right".
- **hyperpartisan**: Hyperpartisan is a boolean value. It is "true" when bias value is "left" or "right", otherwise the value of this field is "false".
- **url**: In this field the URL of the article is stored.
- **author**: Name of the author of the article can be found in this field.
JSON Lines

{
    "id":"0000001",
    "published-at":"YYYY-MM-DD",
    "title":"Headline",
    "content":"Main Content",
    "hyperpartisan":"true",
    "bias":"right",
    "url":"URL",
    "author":"Author"
}

Corpus Construction - Corpus Formatting
XML

Article Instance:
<?xml version="1.0" encoding="UTF-8" standalone="no"?><articles>
    <article id="0000001" published-at="YYYY-MM-DD" title="HeadLine">
        Content
    </article>
</articles>

Ground Truth:
<?xml version="1.0" encoding="UTF-8" standalone="no"?><articles>
    <article id='0000001' hyperpartisan='true' bias='right' url='URL'
        labeled-by='publisher'/>
    <article id='0000002' hyperpartisan='true' bias='least' url='URL'
        labeled-by='publisher'/>
</articles>
Analysis

- Corpus Statistics
- Duplicate Article Detection
- Classification Experiment
### Number of Articles With Different Minimum Words Threshold

<table>
<thead>
<tr>
<th>Articles</th>
<th>Left</th>
<th>Leftcenter</th>
<th>Least</th>
<th>Rightcenter</th>
<th>Right</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>311,511</td>
<td>837,422</td>
<td>551,673</td>
<td>335,552</td>
<td>442,461</td>
<td>2,478,620</td>
</tr>
<tr>
<td>Min. 2 words</td>
<td>262,345</td>
<td>228,467</td>
<td>459,324</td>
<td>272,239</td>
<td>362,164</td>
<td>1,583,173</td>
</tr>
<tr>
<td>Min. 50 words</td>
<td>253,311</td>
<td>214,598</td>
<td>423,675</td>
<td>262,078</td>
<td>334,413</td>
<td>1,488,075</td>
</tr>
<tr>
<td>Min. 100 words</td>
<td>243,863</td>
<td>203,245</td>
<td>396,570</td>
<td>238,297</td>
<td>313,334</td>
<td>1,395,309</td>
</tr>
<tr>
<td>Min. 150 words</td>
<td>232,717</td>
<td>189,475</td>
<td>365,429</td>
<td>212,733</td>
<td>290,031</td>
<td>1,289,785</td>
</tr>
</tbody>
</table>
### Number of Articles With Different Minimum Words Threshold

<table>
<thead>
<tr>
<th>Articles</th>
<th>Left</th>
<th>Leftcenter</th>
<th>Least</th>
<th>Rightcenter</th>
<th>Right</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>311,511</td>
<td>837,422</td>
<td>551,673</td>
<td>335,552</td>
<td>442,461</td>
<td>2,478,620</td>
</tr>
<tr>
<td>Min. 2 words</td>
<td>262,345</td>
<td>228,467</td>
<td>459,324</td>
<td>272,239</td>
<td>362,164</td>
<td>1,583,173</td>
</tr>
<tr>
<td>Min. 50 words</td>
<td>253,311</td>
<td>214,598</td>
<td>423,675</td>
<td>262,078</td>
<td>334,413</td>
<td>1,488,075</td>
</tr>
<tr>
<td>Min. 100 words</td>
<td>243,863</td>
<td>203,245</td>
<td>396,570</td>
<td>238,297</td>
<td>313,334</td>
<td>1,395,309</td>
</tr>
<tr>
<td>Min. 150 words</td>
<td>232,717</td>
<td>189,475</td>
<td>365,429</td>
<td>212,733</td>
<td>290,031</td>
<td>1,289,785</td>
</tr>
</tbody>
</table>
### Average Length of Articles in Words

<table>
<thead>
<tr>
<th>Average Length</th>
<th>Left</th>
<th>Leftcenter</th>
<th>Least</th>
<th>Rightcenter</th>
<th>Right</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. 2 words</td>
<td>867</td>
<td>549</td>
<td>593</td>
<td>750</td>
<td>570</td>
<td>691</td>
</tr>
<tr>
<td>Min. 50 words</td>
<td>893</td>
<td>580</td>
<td>637</td>
<td>775</td>
<td>612</td>
<td>691</td>
</tr>
<tr>
<td>Min. 100 words</td>
<td>924</td>
<td>608</td>
<td>676</td>
<td>844</td>
<td>648</td>
<td>732</td>
</tr>
<tr>
<td>Min. 150 words</td>
<td>965</td>
<td>643</td>
<td>723</td>
<td>930</td>
<td>690</td>
<td>782</td>
</tr>
</tbody>
</table>
Analysis

- Corpus Statistics
- Duplicate Article Detection
- Classification Experiment
Finding Similarity Between Articles

- Employed part of Text Reuse Analysis Pipeline
- we used a representation method based on TF-IDF weighting scheme to each document in the dataset in a set of feature vectors
- we perform a pairwise cosine similarity comparison on all the possible pair of articles
The first priority is to keep the article which was published most recently.
Delete one randomly if no publishing date exist.
Analysis

- Corpus Statistics
- Duplicate Article Detection
- Classification Experiment
- Two different pipelines to detect hyperpartisan and articles' orientation
- To balance the corpus for hyperpartisan detection, we used 250,000 articles from each left and right wings, in addition to other bias categories we also include 167,000 articles each.
Logistic Regression

<table>
<thead>
<tr>
<th></th>
<th>f1</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orientation</td>
<td>0.7495</td>
<td>0.7600</td>
<td>0.7522</td>
<td>0.7522</td>
</tr>
<tr>
<td>Hyperpartisan</td>
<td>0.8507</td>
<td>0.8498</td>
<td>0.8492</td>
<td>0.8508</td>
</tr>
</tbody>
</table>
## Logistic Regression – TF-IDF

<table>
<thead>
<tr>
<th></th>
<th>f1</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orientation</td>
<td>0.6732</td>
<td>0.7110</td>
<td>0.6807</td>
<td>0.6807</td>
</tr>
<tr>
<td>Hyperpartisan</td>
<td>0.8239</td>
<td>0.8236</td>
<td>0.8231</td>
<td>0.8236</td>
</tr>
</tbody>
</table>

**Normalized confusion matrix:**

```
Predicted Label

<table>
<thead>
<tr>
<th></th>
<th>Least</th>
<th>Right</th>
<th>Right-Center</th>
<th>Left</th>
<th>Left-Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Label</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Least</td>
<td>0.88</td>
<td>0.05</td>
<td>0.01</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Right</td>
<td>0.17</td>
<td>0.74</td>
<td>0.01</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>Right-Center</td>
<td>0.23</td>
<td>0.07</td>
<td>0.64</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>Left</td>
<td>0.25</td>
<td>0.16</td>
<td>0.02</td>
<td>0.53</td>
<td>0.03</td>
</tr>
<tr>
<td>Left-Center</td>
<td>0.37</td>
<td>0.09</td>
<td>0.04</td>
<td>0.10</td>
<td>0.40</td>
</tr>
</tbody>
</table>
```

**Normalized confusion matrix:**

```
Predicted Label

<table>
<thead>
<tr>
<th></th>
<th>Not Bias</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>True label</td>
<td>0.83</td>
<td>0.17</td>
</tr>
<tr>
<td>Not Bias</td>
<td>0.19</td>
<td>0.81</td>
</tr>
</tbody>
</table>
```

---

Bauhaus-Universität Weimar, November 9, 2018

Building a Corpus for Hyperpartisan News Detection

Payam Adineh
Experiment on SemEval 2019 Corpus

- Contains 1 million articles
- 800,000 articles for training and 200,000 articles for the test set
- The training set has 200,000 left, 400,000 least, and 200,000 right articles
- The test set has 50,000 left, 100,000 least, and 50,000 right articles
- We divided the data in the way that the training and test set share no news producers in common
- Task is to predict the extreme bias of the news articles
## Logistic Regression

<table>
<thead>
<tr>
<th></th>
<th>f1</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperpartisan</td>
<td>0.6397</td>
<td>0.6405</td>
<td>0.6400</td>
<td>0.6400</td>
</tr>
</tbody>
</table>

![Normalized confusion matrix](image)
Conclusion and Future Work

- Concentrate on a way to deal with ideological creation and contribution of political news.
- Create a hyperpartisan corpus 1.5 million news articles from 358 news producers as the first and the main contribution of this thesis.
- Develop a pipeline that can predict extreme bias of articles with accuracy of 0.85.
- So far 320 participant are working on our corpus on SemEval 2019 task.
- Further analysis of the data and topic classification in order to eliminate the non-political articles.
- There is always a chance to add new articles in the corpus.
- Use statistics we collected to detect duplicate and near-duplicate documents, and we can go further in this regard to detect possible text reuse in the corpus.
Thank you for your attention!
Mean Archive Quality Score of News Producers
Corpus Anchor Element Analysis

<table>
<thead>
<tr>
<th></th>
<th>Left</th>
<th>Leftcenter</th>
<th>Least</th>
<th>Rightcenter</th>
<th>Right</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>1,545,640</td>
<td>1,230,138</td>
<td>967,071</td>
<td>1,453,404</td>
<td>1,431,560</td>
<td>6,627,813</td>
</tr>
<tr>
<td><strong>Internal</strong></td>
<td>368,951</td>
<td>341,425</td>
<td>289,446</td>
<td>141,971</td>
<td>280,257</td>
<td>1,422,050</td>
</tr>
<tr>
<td><strong>External</strong></td>
<td>1,176,689</td>
<td>888,713</td>
<td>677,625</td>
<td>1,311,433</td>
<td>1,151,303</td>
<td>5,205,763</td>
</tr>
<tr>
<td><strong>Average per article</strong></td>
<td>7.45</td>
<td>5.72</td>
<td>4.57</td>
<td>3.04</td>
<td>4.21</td>
<td>4.43</td>
</tr>
</tbody>
</table>
Average Number of Internal and External Links per Producer
Top 20 News Producers Referenced by Other Producers

- nytimes.com
- washingtonpost.com
- upi.com
- reuters.com
- latimes.com
- tomdispatch.com
- breitbart.com
- dailycaller.com
- nbcnews.com
- thinkprogress.org
- businessinsider.com
- thedailylbeast.com
- washingtontimes.com
- motherjones.com
- realclearpolitics.com
- thenation.com
- mediامatters.org
- chicagotribune.com
- mediaite.com
- politifact.com

Values range from 0 to 140,000.