Bauhaus-Universität Weimar

Building a Corpus for Hyperpartisan News Detection

Master's Thesis by Payam Adineh

Referees:

Prof. Dr. Benno Stein

Prof. Dr. Siegmund

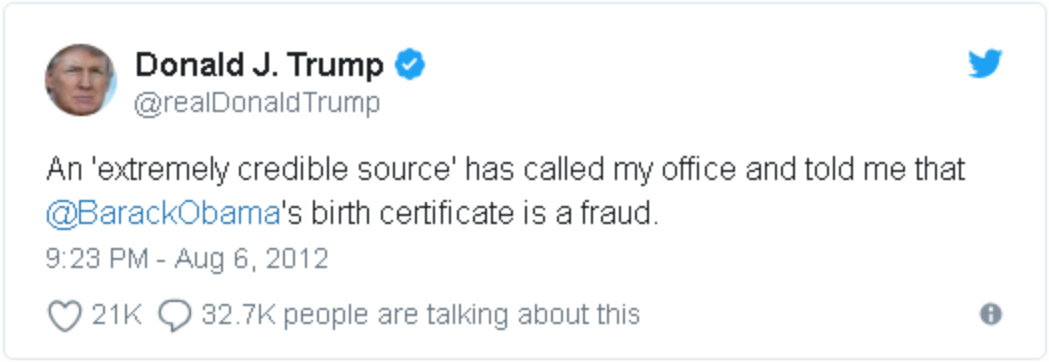
Advisors:

Prof. Dr. Martin Potthast

Johannes Kiesel

Motivation

Fake News


Donald J. Trump
@realDonaldTrump
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


Melania Trump Bans Monsanto Products from the White House
Mandalay Bay Security Guard "vanished"
FBI Raids CDC for Data on Vaccines & Autism
Mass Grave of Dozens of Tortured Black Men Found in Deceased KKK Leader's Estate
President Trump Closed US Airspace for 4th of July
Ohio Cemetery Exhuming Bodies of Confederate Soldiers
Florida's Sheriff's Office Will Not Escort Miami Dolphins Until They Stand For National Anthem
Queen Elizabeth Removes Obamas From Royal Wedding Guest List

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.

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

Locker room talk

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 pehonomenom in the filed 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

Media Bias Fact Check

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

Sites	Left	Right	Total
Total Number	178	499	677
Duplicate	9	39	48
Unavailable	31	140	171
Facebook	120	231	351
Sitemap	18	89	107

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

	Left	Leftcenter	Least	Rightcenter	Total
All Subjects	290	447	362	211	1310
Political	12	49	28	20	109

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

Sitemap	Last Modified
https://www.snopes.com/post-sitemap.xml	2016-01-20 09:12 -08:00
https://www.snopes.com/post-sitemap2.xml	2016-07-07 15:25 -07:00
https://www.snopes.com/post-sitemap3.xml	2016-09-06 09:34 -07:00

Filtering keyword

```
▼<url>
  ▼<loc>
      https://www.reviewjournal.com/news/politics-and-government/nevada/heller-to-introduce-bill-to-speed-removal-of-immigrant-gang-members/
    </loc>
    <lastmod>2018-02-08T19:18:20+00:00</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.7</priority>
  </url>
```

Keyword to filter URLs

```
▼<url>
  ▼<loc>
      https://www.mercurynews.com/2018/02/08/mccain-staying-away-from-dc-over-flu-concerns/
    </loc>
    <changefreq>monthly</changefreq>
    <priority>0.7</priority>
    <lastmod>2018-02-08T22:45:12+00:00</lastmod>
  ▼<n:n>
    ▼<n:publication>
        <n:name>The Mercury News</n:name>
        <n:language>en-US</n:language>
      </n:publication>
    ▼<n:genres>
        Health, Nation &amp;amp; World, News, Politics, Midday Wire, National News, Senate, U.S. Congress, World News
      </n:genres>
```

Tag and Value to filter URLs

BuzzFeed and Media Bias Fact Check Dissagreement

News Producers	MBFC	BuzzFeed
Consortiumnews.com	Least	left
Amgreatness.com	Rightcenter	Right
Rare.us	Rightcenter	Right
Theamericanconservative.com	Rightcenter	Right
Washingtonexaminer.com	Rightcenter	Right
Liberalmountain.com	Leftcenter	Left
Billmoyers.com	Leftcenter	Left
Mintpressnews.com	Leftcenter	Left
Secondnexus.com	Leftcenter	Left

URLs Statistics

	Left	Leftcenter	Least	Rightcenter	Right	Total
BuzzFeed Facebook	169,091	-	-	-	254,673	423,764
BuzzFeed Sitemap	70,279	-	-	-	508,929	579,208
MBFC Sitemap	628,638	1,392,187	750,872	420,218	-	3,191,915
Total	868,008	1,392,187	750,872	420,218	763,602	4,194,887

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

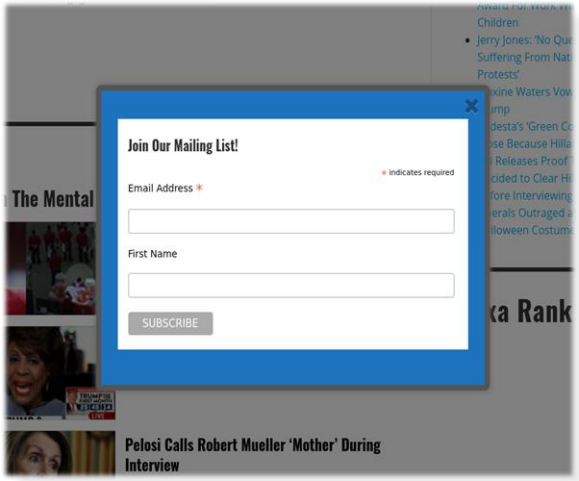
Problem	Affected Producers
Modal Windows	140
Down	68
Share Others	12
Read More	3
Homepage	3
Captcha	2

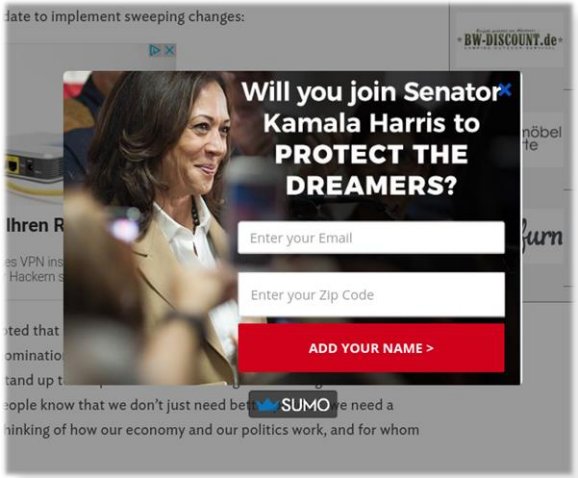
Checking the Producers and Archiving Tool

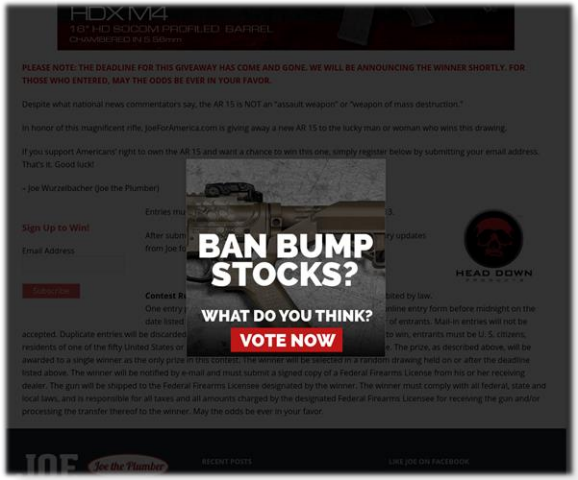
- Archived three random articles from each producer (567 total) to find the possible errors

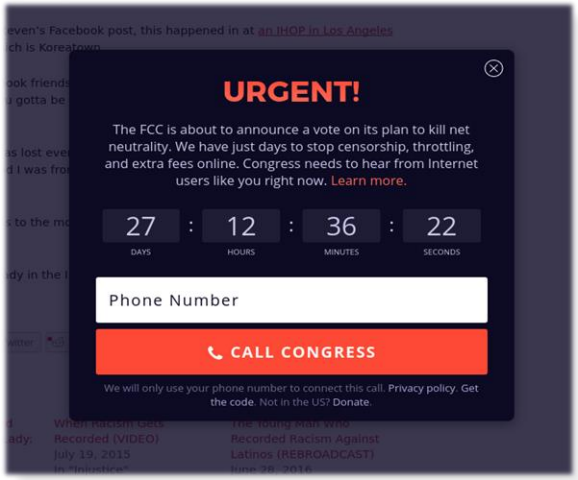
Problem	Affected Producers
Modal Windows	140
Down	68
Share Others	12
Read More	3
Homepage	3
Captcha	2

Modal Windows Examples

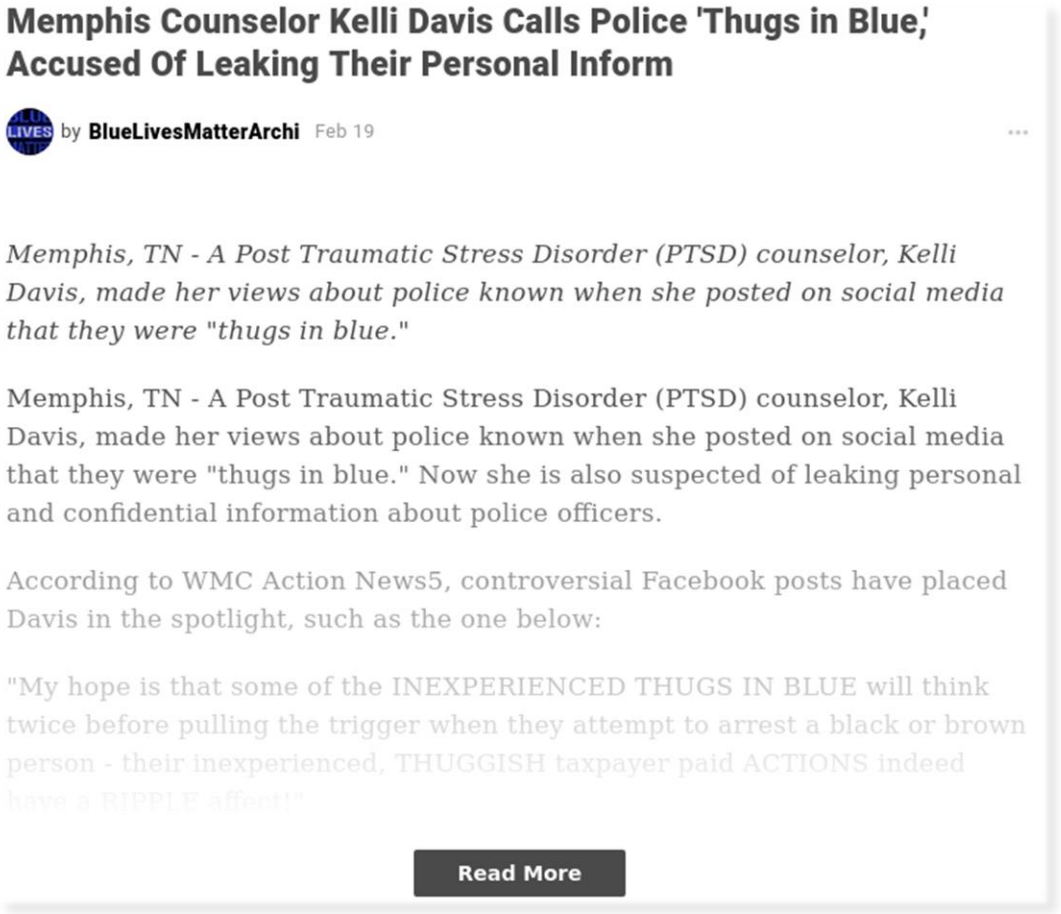








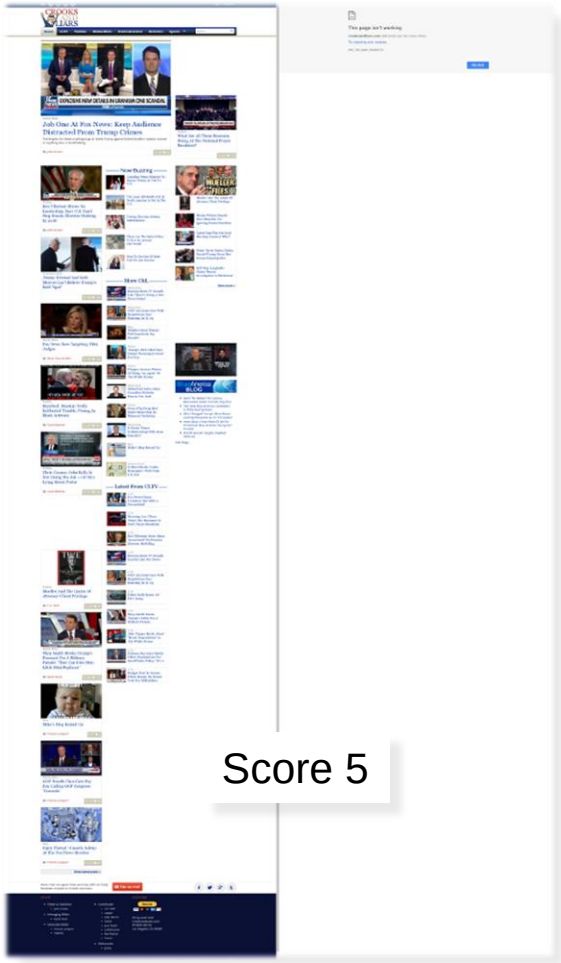
Read More example



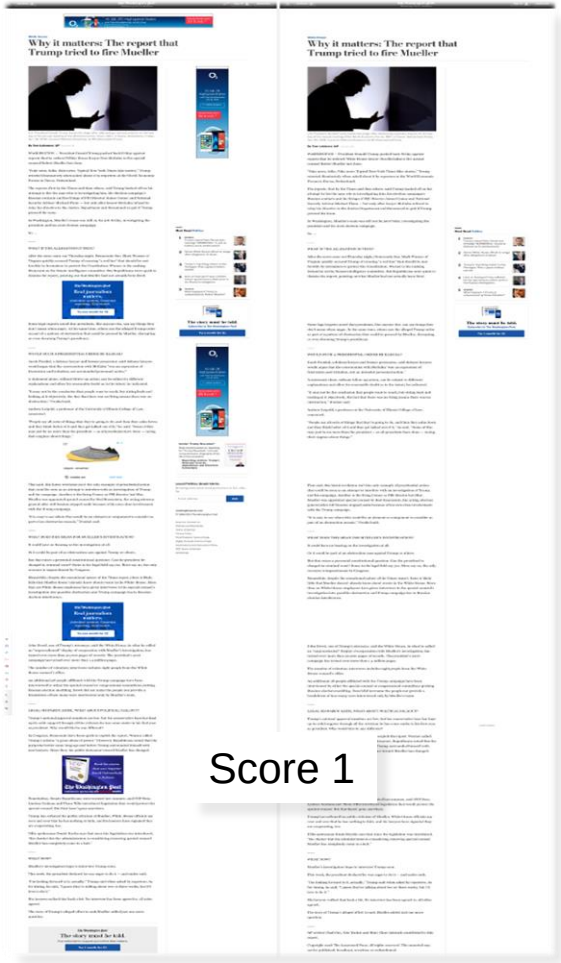
Example to Remove Modal Window

```
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



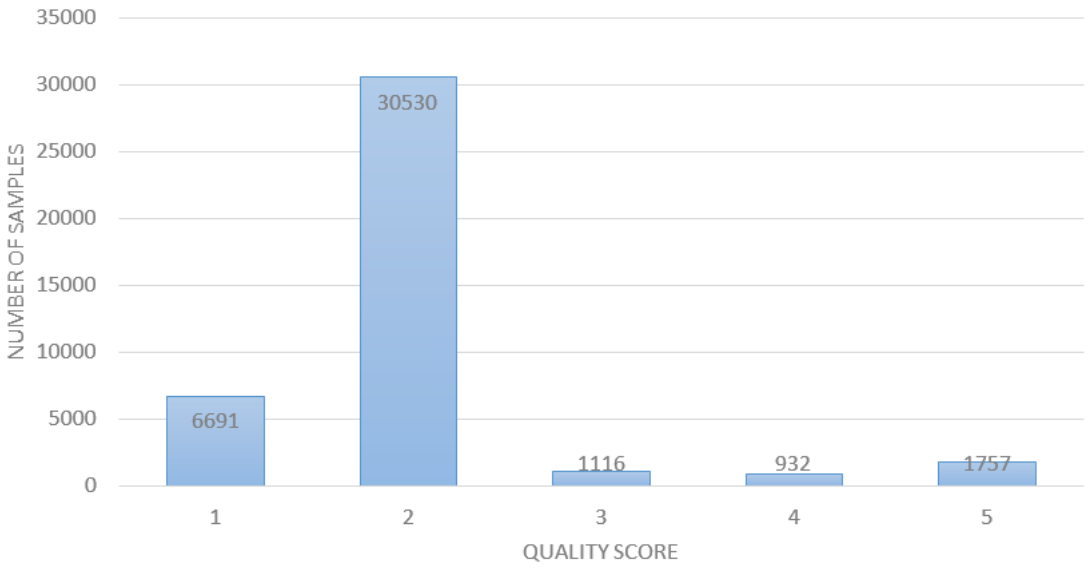
Score 1

Archive Quality Check

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

	Producers	Articles
Total	475	44,806
At Least One Successful Archive	450	38,309

Archive Quality Score Chart



Archive Quality Check Discovered Problems

Problem	Number of Producers
Modal Windows	33
Park Domain	25
Reproduction Problem	8
Gateway Error	3
Archiving Problem	2
Captcha	1
Sitemap Issue	1

Archive Quality Check Discovered Problems

Problem	Number of Producers
Modal Windows	33
Park Domain	25
Reproduction Problem	8
Gateway Error	3
Archiving Problem	2
Captcha	1
Sitemap Issue	1

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 the 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

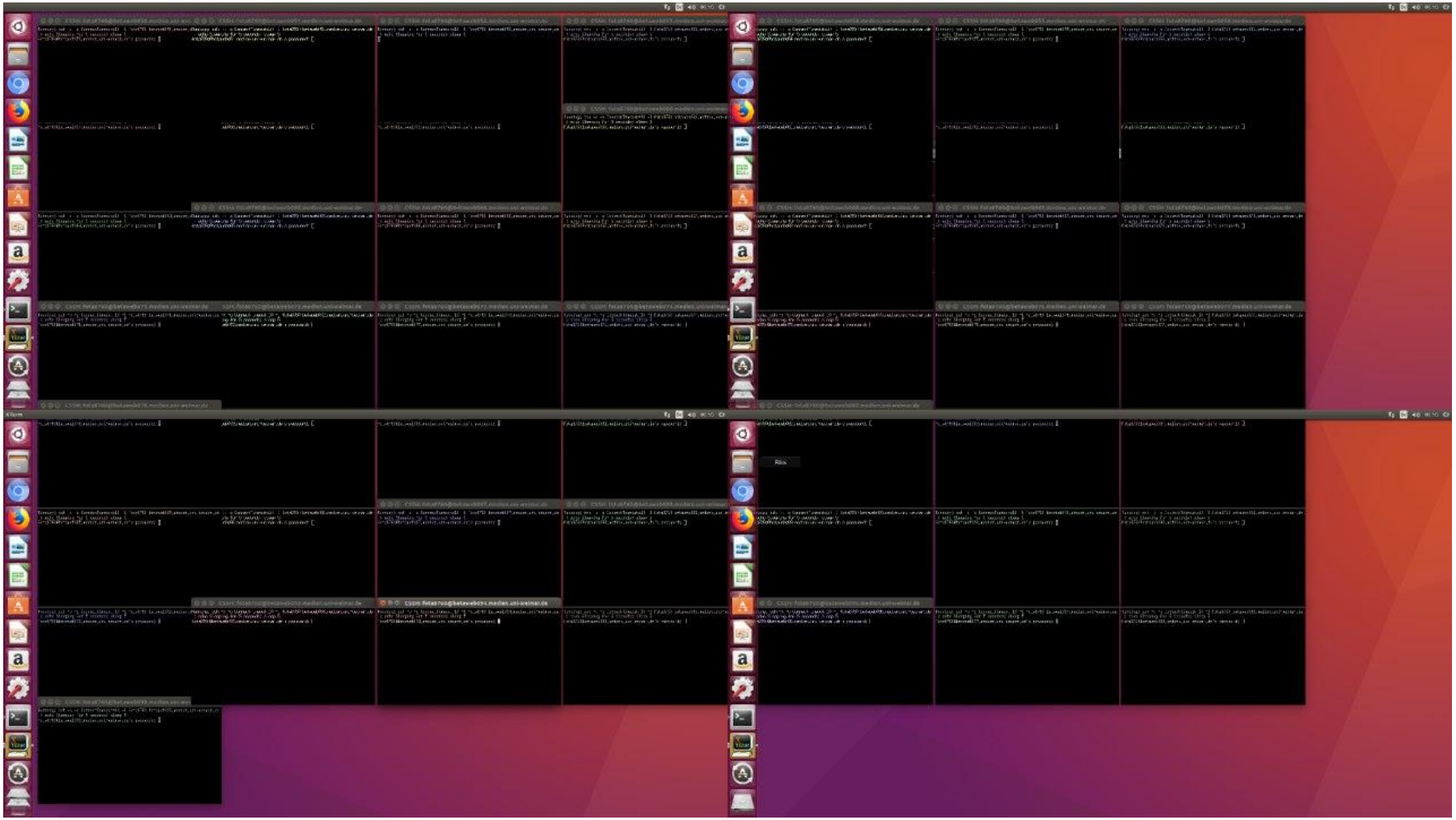
Automatic Archive Quality Assurance

- 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

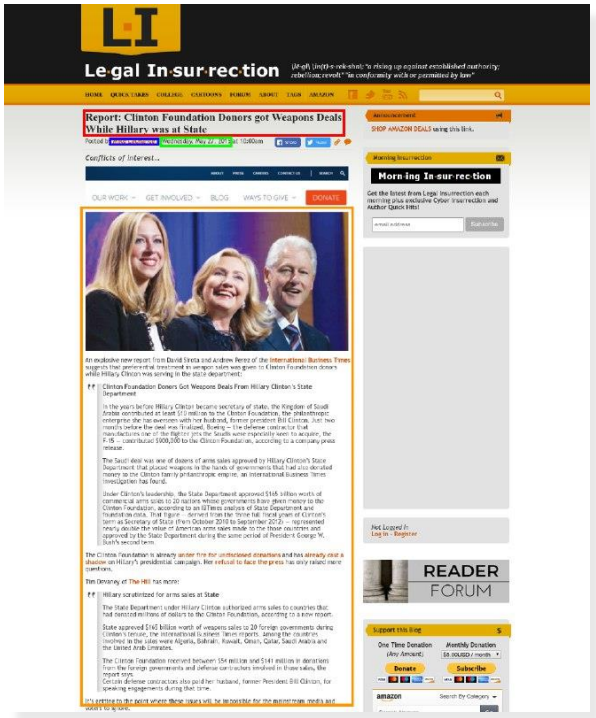
- We have HTML pages, but we need the main parts

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

```
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"
}
```

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

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

Number of Articles With Different Minimum Words Threshold

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

Average Length of Articles in Words

Average Length	Left	Leftcenter	Least	Rightcenter	Right	Total
Min. 2 words	867	549	593	750	570	691
Min. 50 words	893	580	637	775	612	691
Min. 100 words	924	608	676	844	648	732
Min. 150 words	965	643	723	930	690	782

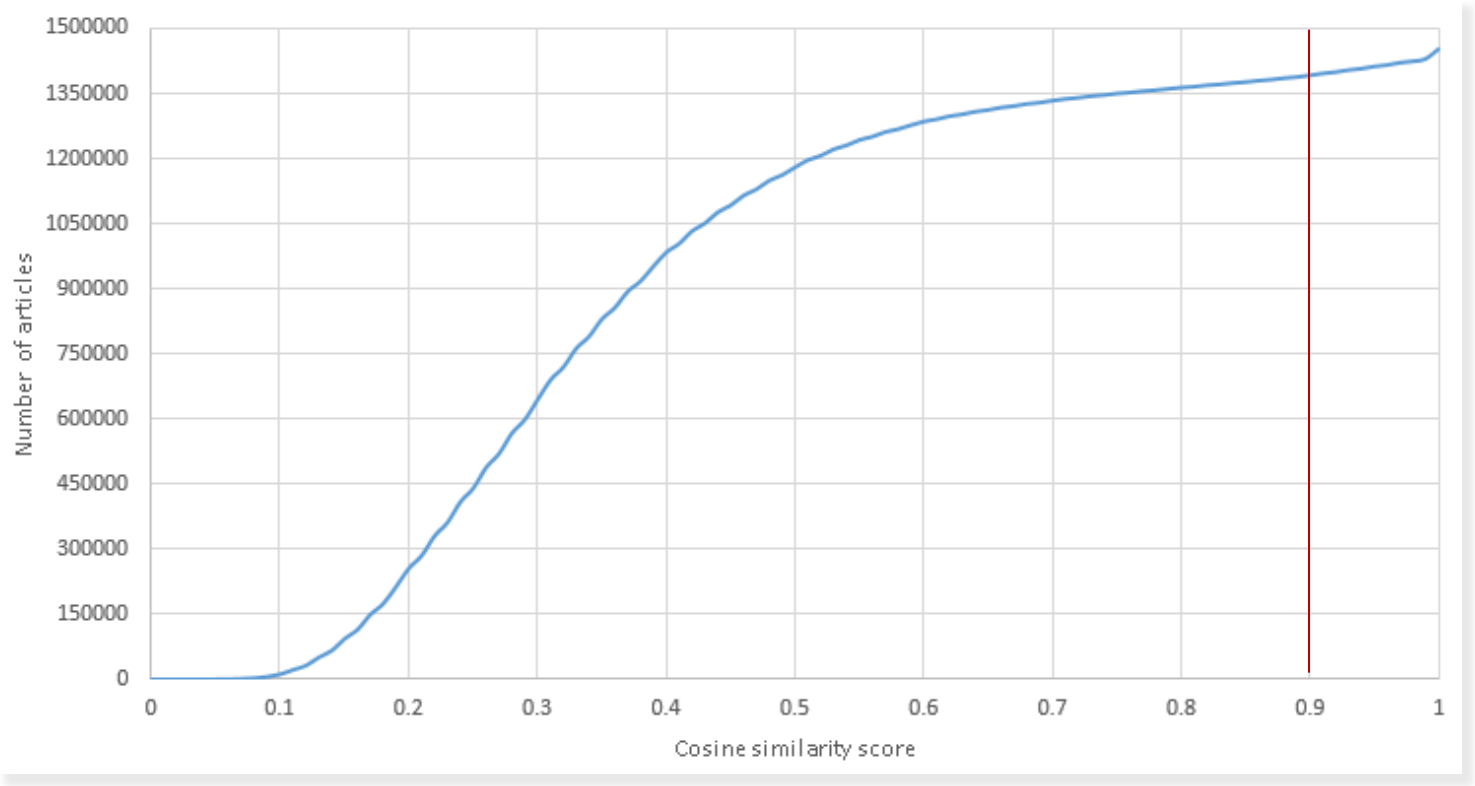
Analysis

- 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

Number of Articles / Cosine Similarity Score

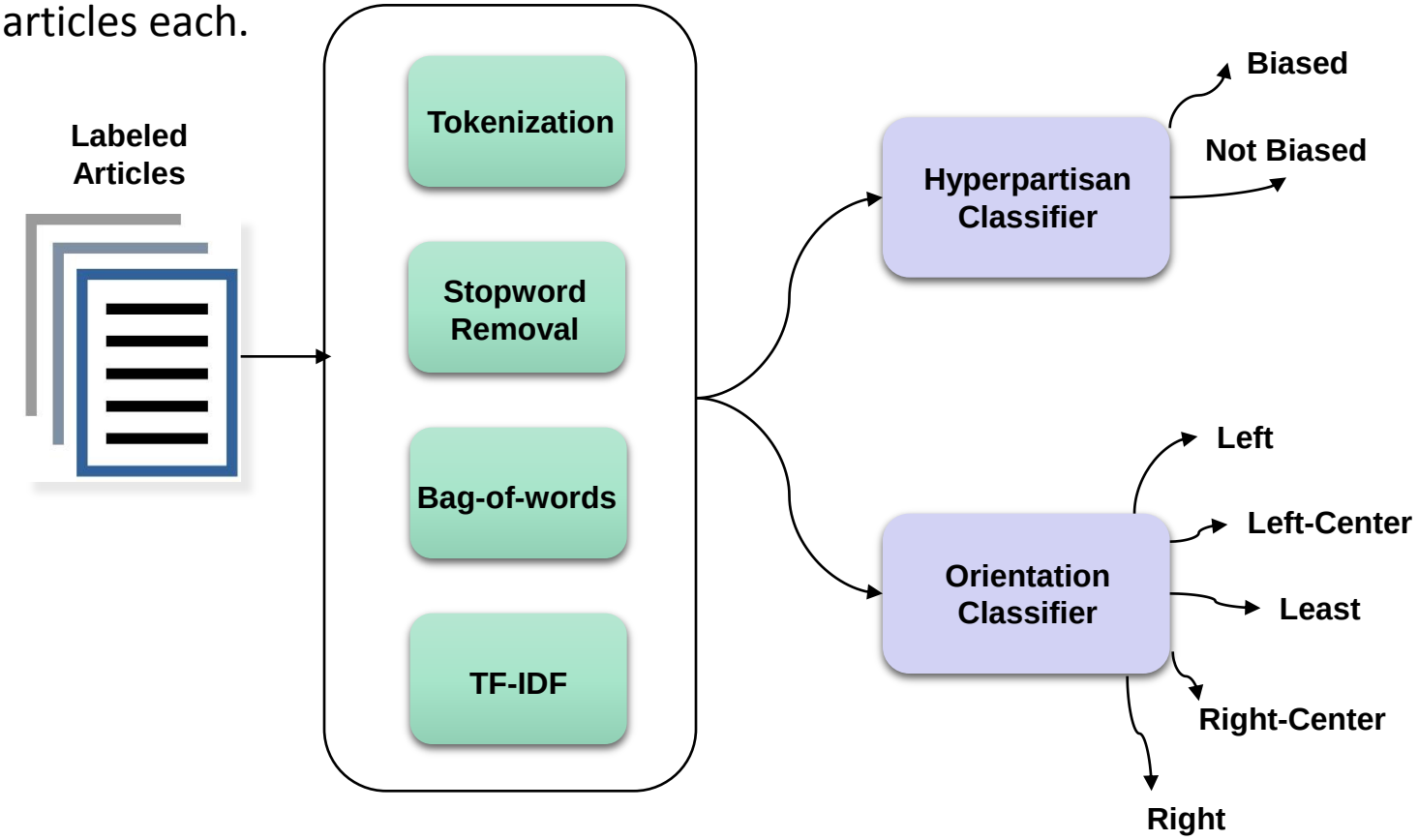


- The first priority is to keep the article which was published most recently
- Delete one randomly if no publishing date exist

Analysis

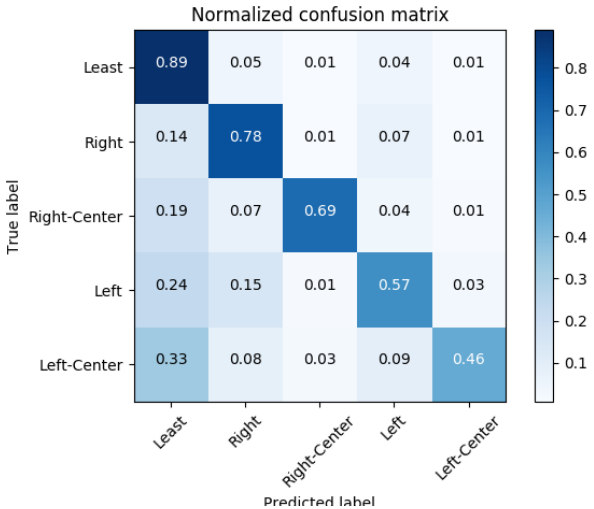
- Corpus Statistics
- Duplicate Article Detection
- Classification Experiment

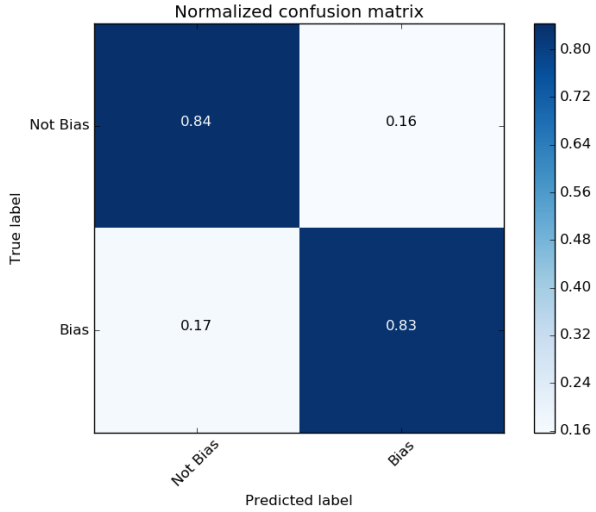
- Two different pipelines to detect hyperpartisan and articles' orientation
- To balance the corpus for hyperpartisan detection, we used 250,000 articles from each left and right wings, in addition to other bias categories we also include 167,000 articles each.



Logistic Regression

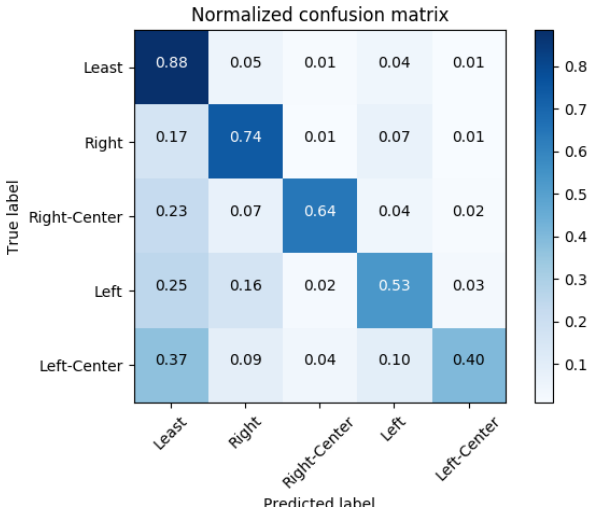
	f1	Precision	Recall	Accuracy
Orientation	0.7495	0.7600	0.7522	0.7522
Hyperpartisan	0.8507	0.8498	0.8492	0.8508

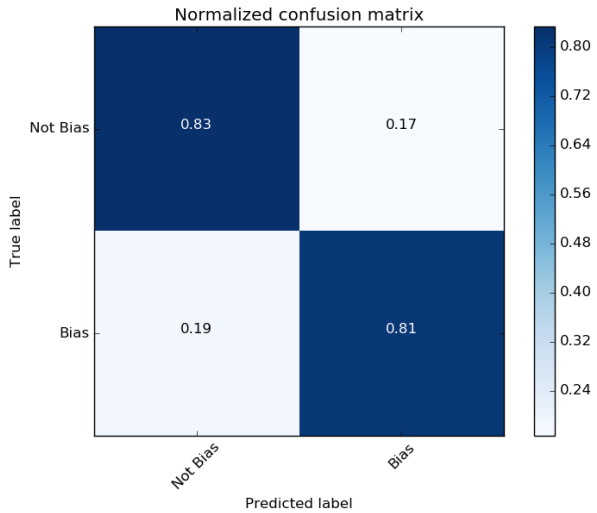




Logistic Regression – TF-IDF

	f1	Precision	Recall	Accuracy
Orientation	0.6732	0.7110	0.6807	0.6807
Hyperpartisan	0.8239	0.8236	0.8231	0.8236



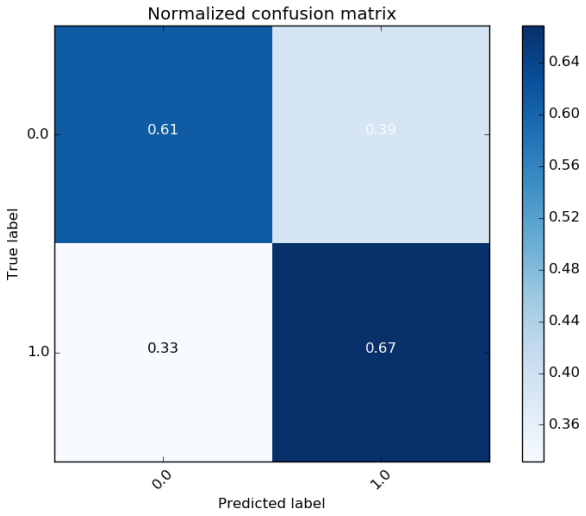


Experiment on SemEval 2019 Corpus

- 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

	f1	Precision	Recall	Accuracy
Hyperpartisan	0.6397	0.6405	0.6400	0.6400

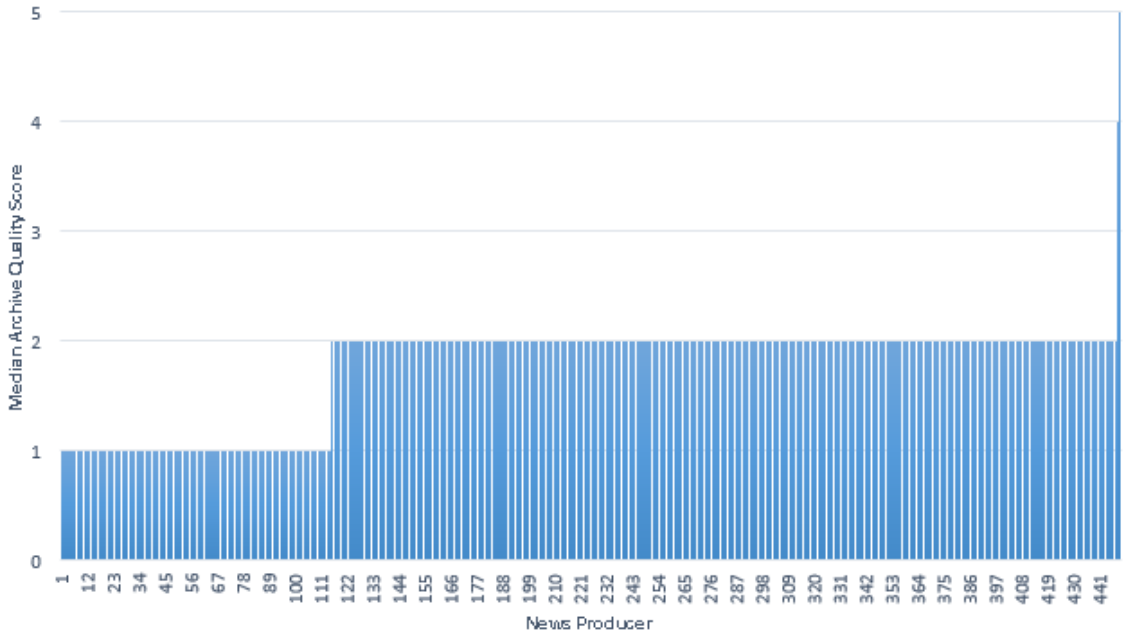


Conclusion and Future Work

- 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!

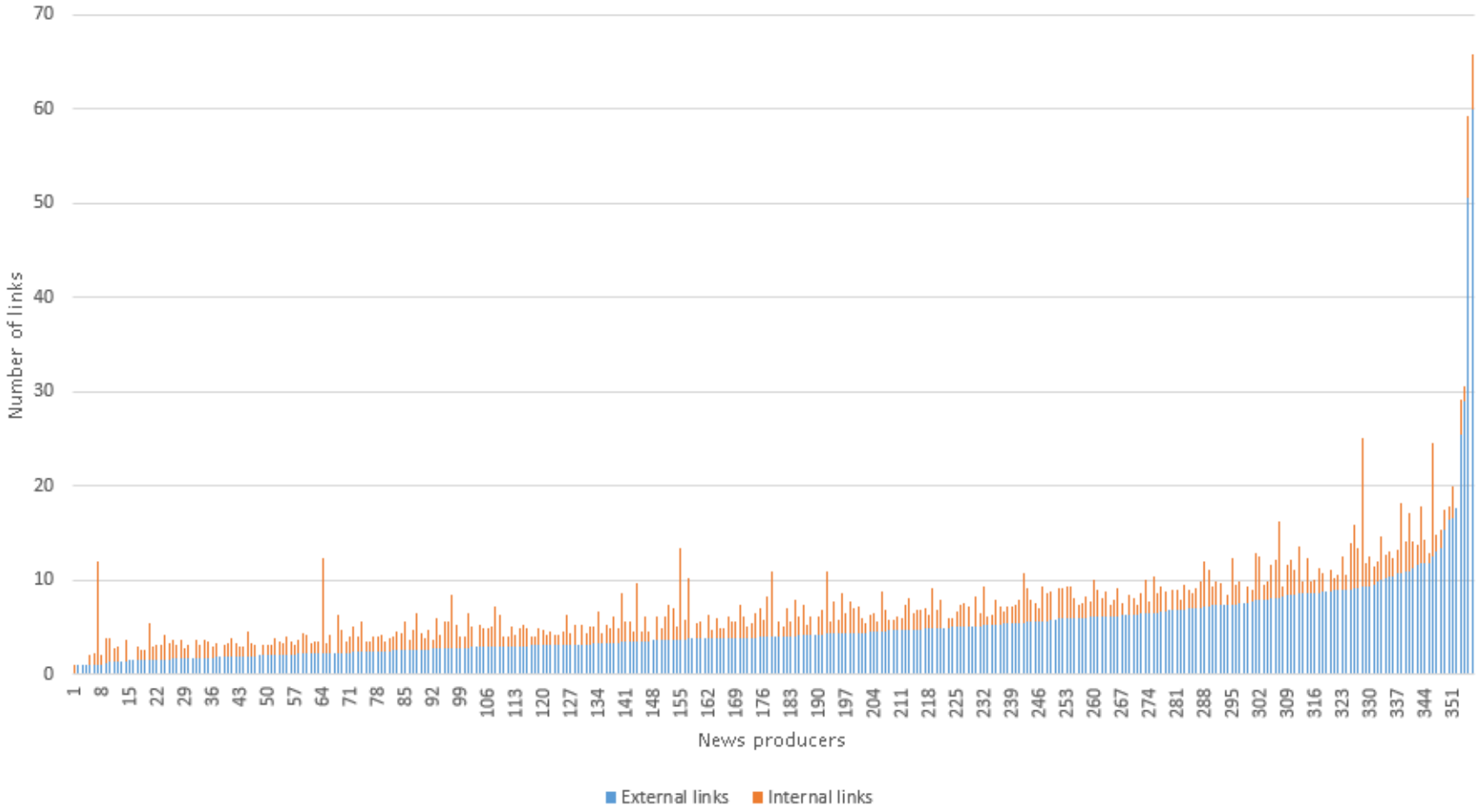
Mean Archive Quality Score of News Producers



Corpus Anchor Element Analysis

	Left	Leftcenter	Least	Rightcenter	Right	Total
Total	1,545,640	1,230,138	967,071	1,453,404	1,431,560	6,627,813
Internal	368,951	341,425	289,446	141,971	280,257	1,422,050
External	1,176,689	888,713	677,625	1,311,433	1,151,303	5,205,763
Average per article	7.45	5.72	4.57	3.04	4.21	4.43

Average Number of Internal and External Links per Producer



Top 20 News Producers Referenced by Other Producers