Comparing Topics and Ideology in Online News:
Social Media Data and Web Browsing Histories

Iulia Cioroianu

University of Bath
Brexit referendum campaign - February-June 2016.
Online information availability vs. exposure

- Brexit referendum campaign - February-June 2016.
- How do political topics in the news shared on Twitter differ from those that people read, based on individual web browsing histories?
Online information availability vs. exposure

- Brexit referendum campaign - February-June 2016.
- How do political topics in the news shared on Twitter differ from those that people read, based on individual web browsing histories?
- How does the ideological leaning of news shared on Twitter differ from those that people read?
Online information availability vs. exposure

- Brexit referendum campaign - February-June 2016.
- How do political topics in the news shared on Twitter differ from those that people read, based on individual web browsing histories?
- How does the ideological leaning of news shared on Twitter differ from those that people read?
- Are the networks of news exposure different?
Online information availability vs. exposure

- Brexit referendum campaign - February-June 2016.
- How do political topics in the news shared on Twitter differ from those that people read, based on individual web browsing histories?
- How does the ideological leaning of news shared on Twitter differ from those that people read?
- Are the networks of news exposure different?
  - What about echo chambers?
- Overview of the methods used to collect and process the data used in the project.
ExpoNET Project

Survey  Clickstream  Twitter
ExpoNET Project
ExpoNET Project

Survey
- Collecting and processing linked survey and clickstream data
- Cleaning clickstream

Clickstream
- Cleaning clickstream

Twitter
- Collecting tweets
- Processing tweets

URLs
- Links to text
  - Text to numbers
- URLs to nodes and edges
ExpoNET Project

Flowchart:

- Survey
  - Collecting and processing linked survey and clickstream data
  - Comparing news consumption
  - Comparing topics

- Clickstream
  - Cleaning clickstream
  - Links to text: Text to numbers
  - Comparing networks

- Twitter
  - Collecting tweets
  - Processing tweets
  - URLs to nodes and edges
ExpoNET Project

Survey
Collecting and processing linked survey and clickstream data
Cleaning clickstream
Comparing news consumption

Clickstream

Twitter
Collecting tweets
Processing tweets

URLs
Links to text
Text to numbers
URLs to nodes and edges
Comparing networks
Comparing topics
ICM Unlimited Reflected Life Panel
3 waves: February, April, June 2016
1154 respondents
Underrepresents retired respondents (15% to 30%), younger (avg age < 4 yrs), similar on women/men, regionally representative.
ICM Unlimited Reflected Life Panel
3 waves: February, April, June 2016
1154 respondents
Underrepresents retired respondents (15% to 30%), younger (avg age < 4 yrs), similar on women/men, regionally representative.
900 respondents with clickstream/browsing history
2 weeks per wave
Cleaning web history data

https://mvt.api.bbc.com/buckets?activate=false
http://www.bbc.co.uk/news/components?alternative=loading=true&%5Bmost-popular%5D%5Bid%5D=comp-most-popular&%5Bmost-popular%5D%5Bopts%5D=assetId%5D=36616028&%5Bmost-popular%5D%5Bopts%5D=loading_strategy%5D=include_content&%5Bmost-popular%5D%5Bopts%5D=position_info%5D=instanceNo%5D=1&batch%5Bmost-popular%5D%5Bopts%5D=position_info%5D=positionInRegion%5D=4&batch%5Bmost-popular%5D%5Bopts%5D=position_info%5D=blastInRegion%5D=true&batch%5Bmost-popular%5D%5Bopts%5D=position_info%5D=blastOnPage%5D=true&batch%5Bmost-popular%5D%5Bopts%5D=position_info%5D=column%5D=secondary_column
http://www.bbc.co.uk/news/uk-politics-36615028
http://www.bbc.co.uk/news/components?alternative=loading=true&batch%5Bfrom-other-news-sites%5D%5Bid%5D=comp-from-other-news-sites&batch%5Bfrom-other-news-sites%5D%5Bopts%5D=assetId%5D=36616028&batch%5Bfrom-other-news-sites%5D%5Bopts%5D=loading_strategy%5D=include_content&batch%5Bfrom-other-news-sites%5D%5Bopts%5D=position_info%5D=instanceNo%5D=1&batch%5Bfrom-other-news-sites%5D%5Bopts%5D=position_info%5D=positionInRegion%5D=7&batch%5Bfrom-other-news-sites%5D%5Bopts%5D=position_info%5D=lastInRegion%5D=true&batch%5Bfrom-other-news-sites%5D%5Bopts%5D=position_info%5D=blastOnPage%5D=false&batch%5Bfrom-other-news-sites%5D%5Bopts%5D=position_info%5D=column%5D=primary_column&batch%5Bfrom-other-news-sites%5D%5Btemplate%5D=%2Fcomponent%2Ffrom-other-news-sites
EU referendum: The result in maps and charts

How did the Leave camp clinch victory in the referendum on the UK’s membership of the EU after a very closely fought contest?

The Leave campaign triumphed right across England and Wales, winning in large northern cities including Sheffield, the Welsh valleys, across the Midlands including Birmingham, and the south and west of England.

The Leave share of the vote mapped

Find the result in your area

Enter a postcode, council name or NF constituency

Top Stories

Zimbabwe's president has defied calls to resign, and his party summons its MPs to discuss impeachment.

1 hour ago

Blow for Merkel as German talks collapse

26 minutes ago

Charles Warner dies after decades in jail

3 hours ago

Features

Charles Warner: Strong leader of down south

Germany odd: What are the options?

Find the result in your area

Enter a postcode, council name or NF constituency
List of UK news domains

- Amazon Alexa Top Sites (alternative: SimilarWeb)
  - Expert coders checked the list.

- Restrict clickstream links to news domains
  - 460 news domains

Identify news article pages on these domains (vs. ads, widgets, images, videos, etc.)

- Regular expressions
  - `http://www.bbc.co.uk/news/\d{8}$`

- Restrict data to news articles

26,781 unique news URLs in the clickstream data
Collecting Twitter data

The method you choose depends on:
- Your programming skills or willingness to develop these skills
- The characteristics of the data you want to collect
- Your budget

Twitter APIs
- Some programming skills required
- Many available packages in Python (Yweepy) and R (TwitteR)
- Flexible but constraints on data collected.

Web scraping
- More advanced programming
- More flexible, you can get more data
- Still have to follow Twitter rules of service

Commercial or free software
- Chorus, NodeXL, Voson, etc.
- Easy access
- Some include data analysis options
- Less flexible
- Same constraints as the API

Purchasing data from Twitter
- Convenient but expensive
Processing social media data

```
{
  "created_at": "Thu Apr 06 15:24:15 +0000 2016",
  "id": 85006145121695744,
  "id_str": "85006145121695744",
  "text": "Support our campaign!
  https://t.co/XweGnmxIP",
  "user": {
    "id": 2244994945,
    "id_str": "2244994945",
    "screen_name": "Campaigner",
    "followers_count": 47684,
    "friends_count": 1524,
    "favourites_count": 251,
    "statuses_count": 3121
  },
  "entities": {
    "urls": [{
      "indices": [32, 52],
      "url": "http://t.co/iOwBrTZR",
      "display_url": "youtube.com/watch?v=oHg5SJ...",
      "expanded_url": "http://www.campaign.com"
    }]
  }
}
```

- Working with .json files.
  - Hierarchical data.
  - More efficient.
  - Scales up.

Data storage
- Relational databases
  - MySQL
- Non-relational databases
  - MongoDB

Extracting relevant information
- Text
- User information
- Lists of friends and followers
- Links shared

1.6 million tweets with links to news domains
ExpoNET Project

Clickstream

Twitter

Resolving urls

These are the same:
- Goo.gl/C8Snv6
- http://www.bbc.co.uk/news/politics/uk_leaves_the_eu

Python requests, urllib, urllib

Extracting text, title, author, etc. from article pages

Web scraping
- Python Scrapy, BeautifulSoup, Selenium

Boilerplate/content extraction tools
- Diffbot, Dragnet, Skin Tech
Topic models

Uncovering hidden thematic structures in documents.

Latent Dirichlet Allocation (LDA)

- Documents are a mixture of topics.
- Topics generate words based on their probability distribution.
- Algorithm:
  - Determine number of words in document.
  - Determine mixture of topics in document.
  - Based on the topics' multinomial distribution, assign words to documents.

Mallet, Python Gensim, R Quanteda.
<table>
<thead>
<tr>
<th>id</th>
<th>label</th>
<th>top probability words</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>0.018**&quot;uk&quot;+0.015**&quot;growth&quot;+0.014**&quot;brexit&quot;+0.012**&quot;year&quot;+0.010**&quot;economy&quot;+0.010**&quot;uncertainty&quot;+0.009**&quot;referendum&quot;+0.008**&quot;term&quot;+0.008**&quot;economic&quot;+0.008**&quot;may&quot;+0.007**&quot;investment&quot;+0.007**&quot;cent&quot;</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.032**&quot;young&quot;+0.029**&quot;university&quot;+0.026**&quot;people&quot;+0.020**&quot;women&quot;+0.019**&quot;students&quot;+0.015**&quot;research&quot;+0.014**&quot;professor&quot;+0.013**&quot;school&quot;+0.011**&quot;education&quot;+0.011**&quot;eu&quot;+0.010**&quot;student&quot;+0.010**&quot;science&quot;+0.009**&quot;study&quot;</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.036**&quot;eu&quot;+0.027**&quot;immigration&quot;+0.025**&quot;uk&quot;+0.020**&quot;ireland&quot;+0.015**&quot;migration&quot;+0.014**&quot;people&quot;+0.011**&quot;citizens&quot;+0.011**&quot;british&quot;+0.011**&quot;work&quot;+0.011**&quot;migrants&quot;+0.011**&quot;irish&quot;+0.010**&quot;northern&quot;+0.009**&quot;living&quot;</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.027**&quot;security&quot;+0.021**&quot;russia&quot;+0.021**&quot;nato&quot;+0.017**&quot;military&quot;+0.015**&quot;eu&quot;+0.014**&quot;defence&quot;+0.014**&quot;putin&quot;+0.013**&quot;russian&quot;+0.011**&quot;foreign&quot;+0.010**&quot;former&quot;+0.010**&quot;army&quot;+0.009**&quot;china&quot;+0.008**&quot;us&quot;+0.008**&quot;world&quot;</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.044**&quot;prices&quot;+0.044**&quot;house&quot;+0.034**&quot;property&quot;+0.032**&quot;london&quot;+0.031**&quot;housing&quot;+0.026**&quot;market&quot;+0.023**&quot;price&quot;+0.019**&quot;homes&quot;+0.017**&quot;mortgage&quot;+0.016**&quot;buyers&quot;+0.013**&quot;average&quot;+0.013**&quot;home&quot;+0.012**&quot;new&quot;+0.012**&quot;demand&quot;</td>
</tr>
<tr>
<td>id</td>
<td>label</td>
<td>top probability words</td>
</tr>
<tr>
<td>----</td>
<td>-------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>0</td>
<td>?</td>
<td>0.016**&quot;uk&quot;+0.015**&quot;growth&quot;+0.014**&quot;brexit&quot;+0.012**&quot;year&quot;+0.010**&quot;economy&quot;+0.010**&quot;uncertainty&quot;+0.009**&quot;referendum&quot;+0.008**&quot;term&quot;+0.008**&quot;economic&quot;+0.008**&quot;may&quot;+0.007**&quot;investment&quot;+0.007**&quot;cent&quot;</td>
</tr>
<tr>
<td>1</td>
<td>?</td>
<td>0.032**&quot;young&quot;+0.029**&quot;university&quot;+0.026**&quot;people&quot;+0.020**&quot;women&quot;+0.019**&quot;students&quot;+0.015**&quot;research&quot;+0.014**&quot;professor&quot;+0.013**&quot;school&quot;+0.011**&quot;education&quot;+0.011**&quot;eu&quot;+0.010**&quot;student&quot;+0.010**&quot;science&quot;+0.009**&quot;study&quot;</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>0.036**&quot;eu&quot;+0.027**&quot;immigration&quot;+0.025**&quot;uk&quot;+0.020**&quot;ireland&quot;+0.015**&quot;migration&quot;+0.014**&quot;people&quot;+0.011**&quot;citizens&quot;+0.011**&quot;british&quot;+0.011**&quot;work&quot;+0.011**&quot;migrants&quot;+0.011**&quot;irish&quot;+0.010**&quot;northern&quot;+0.009**&quot;living&quot;</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>0.027**&quot;security&quot;+0.021**&quot;russia&quot;+0.021**&quot;nato&quot;+0.017**&quot;military&quot;+0.015**&quot;eu&quot;+0.014**&quot;defence&quot;+0.014**&quot;putin&quot;+0.013**&quot;russian&quot;+0.011**&quot;foreign&quot;+0.010**&quot;former&quot;+0.010**&quot;army&quot;+0.009**&quot;china&quot;+0.008**&quot;us&quot;+0.008**&quot;world&quot;</td>
</tr>
<tr>
<td>4</td>
<td>?</td>
<td>0.044**&quot;prices&quot;+0.044**&quot;house&quot;+0.034**&quot;property&quot;+0.032**&quot;london&quot;+0.031**&quot;housing&quot;+0.026**&quot;market&quot;+0.023**&quot;price&quot;+0.019**&quot;homes&quot;+0.017**&quot;mortgage&quot;+0.016**&quot;buyers&quot;+0.013**&quot;average&quot;+0.013**&quot;home&quot;+0.012**&quot;new&quot;+0.012**&quot;demand&quot;</td>
</tr>
<tr>
<td>id</td>
<td>label</td>
<td>top probability words</td>
</tr>
<tr>
<td>----</td>
<td>-----------</td>
<td>---------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>0</td>
<td>economy</td>
<td>0.016**&quot;uk&quot;+0.015**&quot;growth&quot;+0.014**&quot;brexit&quot;+0.012**&quot;year&quot;+0.010**&quot;economy&quot;+0.010**&quot;uncertainty&quot;+0.009**&quot;referendum&quot;+0.008**&quot;term&quot;+0.008**&quot;economic&quot;+0.008**&quot;may&quot;+0.007**&quot;investment&quot;+0.007**&quot;cent&quot;</td>
</tr>
<tr>
<td>1</td>
<td>education</td>
<td>0.032**&quot;young&quot;+0.029**&quot;university&quot;+0.026**&quot;people&quot;+0.020**&quot;women&quot;+0.019**&quot;students&quot;+0.015**&quot;research&quot;+0.014**&quot;professor&quot;+0.013**&quot;school&quot;+0.011**&quot;education&quot;+0.011**&quot;eu&quot;+0.010**&quot;student&quot;+0.010**&quot;science&quot;+0.009**&quot;study&quot;</td>
</tr>
<tr>
<td>2</td>
<td>immigration</td>
<td>0.036**&quot;eu&quot;+0.027**&quot;immigration&quot;+0.025**&quot;uk&quot;+0.020**&quot;ireland&quot;+0.015**&quot;migration&quot;+0.014**&quot;people&quot;+0.011**&quot;citizens&quot;+0.011**&quot;british&quot;+0.011**&quot;work&quot;+0.011**&quot;migrants&quot;+0.011**&quot;irish&quot;+0.010**&quot;northern&quot;+0.009**&quot;living&quot;</td>
</tr>
<tr>
<td>3</td>
<td>defence</td>
<td>0.027**&quot;security&quot;+0.021**&quot;russia&quot;+0.021**&quot;nato&quot;+0.017**&quot;military&quot;+0.015**&quot;eu&quot;+0.014**&quot;defence&quot;+0.014**&quot;putin&quot;+0.013**&quot;russian&quot;+0.011**&quot;foreign&quot;+0.010**&quot;former&quot;+0.010**&quot;army&quot;+0.009**&quot;china&quot;+0.008**&quot;us&quot;+0.008**&quot;world&quot;</td>
</tr>
<tr>
<td>4</td>
<td>housing</td>
<td>0.044**&quot;prices&quot;+0.044**&quot;house&quot;+0.034**&quot;property&quot;+0.032**&quot;london&quot;+0.031**&quot;housing&quot;+0.026**&quot;market&quot;+0.023**&quot;price&quot;+0.019**&quot;homes&quot;+0.017**&quot;mortgage&quot;+0.016**&quot;buyers&quot;+0.013**&quot;average&quot;+0.013**&quot;home&quot;+0.012**&quot;new&quot;+0.012**&quot;demand&quot;</td>
</tr>
</tbody>
</table>

Topic probability in each document.

For each topic, average probabilities across articles extracted from clickstream and twitter data.
Does the ideological leaning of articles read on Twitter and in clickstream data differ?
Beyond topics mentions: ideological position?

- Does the ideological leaning of articles read on Twitter and in clickstream data differ?
- Ideal points and spatial theory models
Beyond topics mentions: ideological position?

- Does the ideological leaning of articles read on Twitter and in clickstream data differ?
- Ideal points and spatial theory models
- Measuring ideological positions.
  - Correspondence analysis
  - Wordscores
  - Wordfish
Principal Component Analysis

- Dimension reduction method.
- Transform multiple possibly correlated features into a smaller set of composite variables.
  - Composite variables are orthogonal linear combinations of the features.
Principal Component Analysis

- Dimension reduction method.
- Transform multiple possibly correlated features into a smaller set of composite variables.
  - Composite variables are orthogonal linear combinations of the features.
- Components ordered by the amount of explained variance.
- Use first component for scaling purposes.
Principal Component Analysis

- Dimension reduction method.
- Transform multiple possibly correlated features into a smaller set of composite variables.
  - Composite variables are orthogonal linear combinations of the features.
- Components ordered by the amount of explained variance.
- Use first component for scaling purposes.
- Correspondence analysis - similar to principal component analysis but works for categorical variables.
Scaling model for estimating the positions of actors/news domain for dimensions that are specified a priori.

Supervised. Start with a set of documents with known positions.
Wordscores (Laver, Benoit and Garry, 2003)

- Scaling model for estimating the positions of actors/news domains for dimensions that are specified a priori.
- Supervised. Start with a set of documents with known positions.
- Every document has policy position score
Scaling model for estimating the positions of actors/news domains for dimensions that are specified a priori.

Supervised. Start with a set of documents with known positions.

Every document has policy position score

The score of a document is the average of the scores of its words
  - Use word frequency matrix
• Scaling model for estimating the positions of actors/news domains for dimensions that are specified a priori.
• Supervised. Start with a set of documents with known positions.
• Every document has policy position score
• The score of a document is the average of the scores of its words
  • Use word frequency matrix
• What is the probability of being in document $j$ upon observing $w_i$?
Wordscores (Laver, Benoit and Garry, 2003)

- Scaling model for estimating the positions of actors/news domains for dimensions that are specified a priori.
- Supervised. Start with a set of documents with known positions.
- Every document has policy position score
- The score of a document is the average of the scores of its words
  - Use word frequency matrix
- What is the probability of being in document $j$ upon observing $w_i$?
- To score new documents use average of contained words.
Wordfish (Slapin and Proksch, 2008)

- Poisson scaling model of one-dimensional document positions.
- Unsupervised. Poisson regression model, estimating both actor/document/news domain and word parameters.
• Poisson scaling model of one-dimensional document positions.
• Unsupervised. Poisson regression model, estimating both actor/document/news domain and word parameters.
• Start by guessing the parameters.
Wordfish (Slapin and Proksch, 2008)

- Poisson scaling model of one-dimensional document positions.
- Unsupervised. Poisson regression model, estimating both actor/document/news domain and word parameters.
- Start by guessing the parameters.
- Assume party parameters are correct, fit a Poisson regression model for the word parameters.
Wordfish (Slapin and Proksch, 2008)

- Poisson scaling model of one-dimensional document positions.
- Unsupervised. Poisson regression model, estimating both actor/document/news domain and word parameters.
- Start by guessing the parameters.
- Assume party parameters are correct, fit a Poisson regression model for the word parameters.
- Assume fitted word parameters are correct, fit a Poisson regression model for the party parameters.
Poisson scaling model of one-dimensional document positions.

Unsupervised. Poisson regression model, estimating both actor/document/news domain and word parameters.

Start by guessing the parameters.

Assume party parameters are correct, fit a Poisson regression model for the word parameters.

Assume fitted word parameters are correct, fit a Poisson regression model for the party parameters.

Repeat the process until convergence (..hopefully).
In general, produce similar results.
Comparing topics

Comparing Content

What do we learn about information exposure by comparing content/topics from news stories shared on Twitter to news stories viewed online?
Comparing ideological leaning
Comparing ideological content

Survey
- Collecting and processing linked survey and clickstream data
  - Cleaning clickstream
  - Links to text
  - Text to numbers

Clickstream
- Cleaning clickstream
  - Links to text
  - Text to numbers

Twitter
- Collecting tweets
  - Processing tweets
  - URLs to nodes and edges

Comparing networks

Comparing news consumption

Comparing topics
Figure: Comparing Networks

(a) Twitter
(b) Clickstream

But misleading: little overlap and very low modularity.
Conclusion

- Small differences between online and clickstream content when it comes to:
  - topics
  - ideology
  - networks

Little evidence of echo chambers.

But does this result hold when we further investigate each type of data?

See next talks!
Small differences between online and clickstream content when it comes to:
- topics
- ideology
- networks

Little evidence of echo chambers.
- But does this result hold when we further investigate each type of data?
- See next talks!