# UKParl: A Data Set for Topic Detection with Semantically Annotated Text

**Federico Nanni**, Mahmoud Osman, Yi-Ru Cheng, Simone Paolo Ponzetto and Laura Dietz

# My Research

Post-Doc in **computational social science** at the University of Mannheim, working on natural language processing applications in political science.

Focus:

1. Topic detection
2. Collection building
3. Text scaling

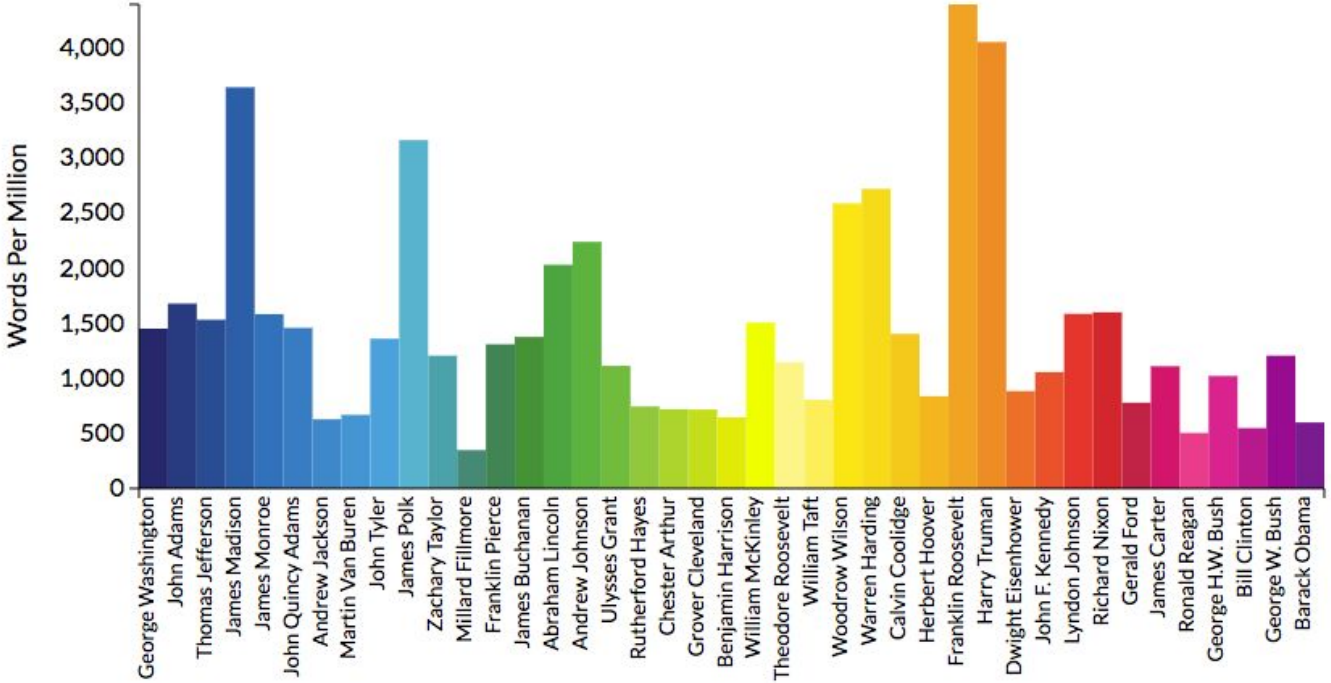
# My Research

Post-Doc in **computational social science** at the University of Mannheim, working on natural language processing applications in political science.

Focus:

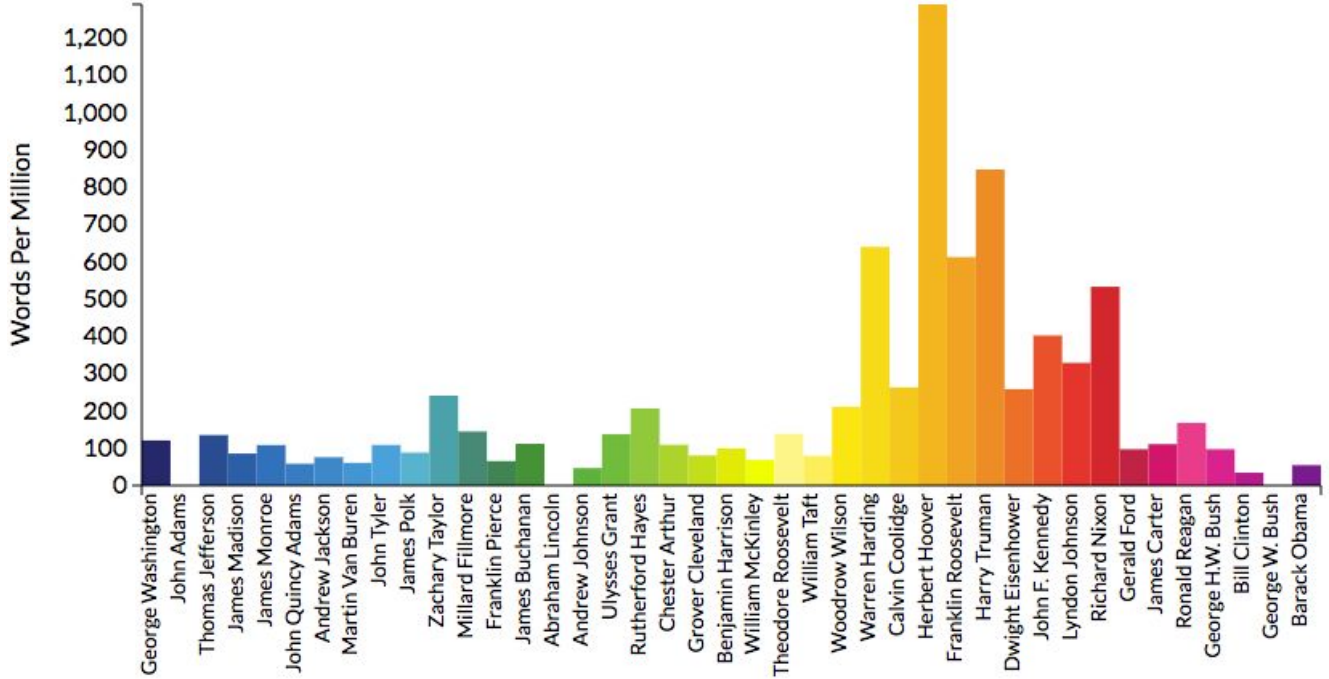
1. **Topic detection**
2. Collection building
3. Text scaling

# Topic War in State of the Union Addresses



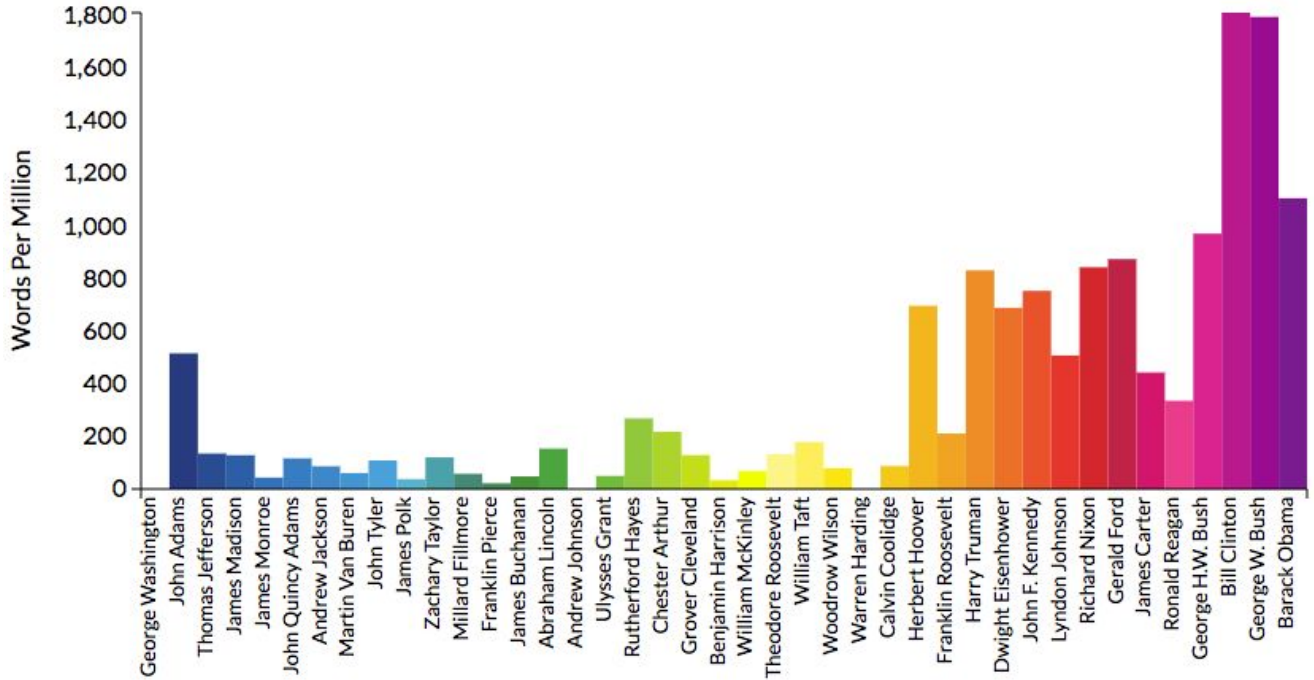

From The Atlantic (01/2015)

# Topic Employment in State of the Union Addresses




From The Atlantic (01/2015)

# Topic Health in State of the Union Addresses




From The Atlantic (01/2015)

# Issue

Collections of political speeches often **do not provide topic annotation** at document level.

# Methods for Topic Detection

Mostly unsupervised, based on **Latent Dirichlet Allocation**.

# Methods for Topic Detection

Mostly unsupervised, based on **Latent Dirichlet Allocation**.

Results often are:

1. Hard to interpret
2. Difficult to evaluate

# A New Data Set

Collected all speeches from the UK House of Commons (2013-2016), organized under fine-grained topics by the curators.

# A New Data Set

Collected all speeches from the UK House of Commons (2013-2016), organized under fine-grained topics by the curators.

-> The entire **Hansard Corpus** is way larger, containing nearly every speech given in the British Parliament since 1803.

# A New Data Set

Collected all speeches from the UK House of Commons (2013-2016), organized under fine-grained topics by the curators.

We have:

- Aligned each topic (when possible) with the related Wikipedia entity

# A New Data Set

Collected all speeches from the UK House of Commons (2013-2016), organized under fine-grained topics by the curators.

We have:

- Aligned each topic (when possible) with the related Wikipedia entity

Topic: **Brexit** ->

Brexit

From Wikipedia, the free encyclopedia

**Brexit** (/ˈbrɛksɪt, ˈbrɛgzɪt/) is the prospective withdrawal of the United Kingdom (UK) from the European Union (EU).

In a referendum on 23 June 2016, 51.9% of the participating UK electorate voted to leave the EU, out of a turnout of 72.2%. On 29 March 2017, the UK government invoked Article 50 of the Treaty on the European Union. The UK is thus due to leave the EU on 29 March 2019.[1]

Prime Minister Theresa May announced that the UK would not seek permanent membership of the single market or the customs union after leaving the EU[2][3] and promised to repeal the European Communities Act of 1972 and incorporate existing European Union law into UK domestic law.[4] A new government department, the Department for Exiting the European Union (DExEU), was created in July 2016, with Eurosceptic David Davis appointed its first Secretary of State. Negotiations with the EU officially started in June 2017.

The UK joined the European Communities (EC) in 1973, with membership confirmed by a referendum in 1975. In the 1970s and 1980s, withdrawal from the EC was advocated mainly by Labour Party members and trade union figures. From the 1990s, the main advocates of withdrawal were the newly founded UK Independence Party (UKIP) and an increasing number of Eurosceptic Conservative Party members.

There is strong agreement among economists and a broad consensus in existing economic research that Brexit is likely to reduce UK's real per-capita income in the medium and long-term.[5][6] Studies on effects that have already materialised since the referendum show annual losses of £404 for the average British household and a

Part of a series of articles on the
United Kingdom
in the
European Union

[show]

# A New Data Set

Collected all speeches from the UK House of Commons (2013-2016), organized under fine-grained topics by the curators.

We have:

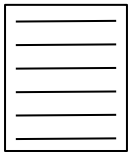
- Aligned each topic (when possible) with the related Wikipedia entity
- Entity-linked the entire dataset, using 

# A New Data Set

Collected all speeches from the UK House of Commons (2013-2016), organized under fine-grained topics by the curators.

We have:

- Aligned each topic (when possible) with the related Wikipedia entity
- Entity-linked the entire dataset, using TAG·ME

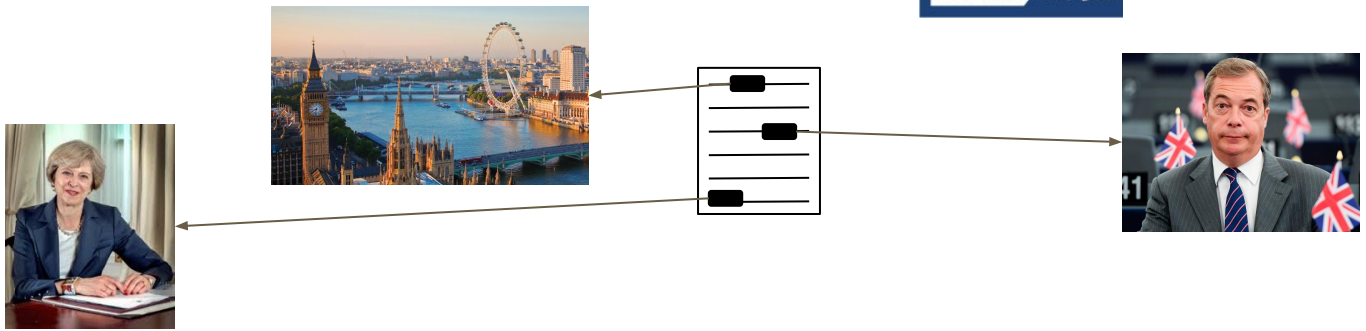


# A New Data Set

Collected all speeches from the UK House of Commons (2013-2016), organized under fine-grained topics by the curators.

We have:

- Aligned each topic (when possible) with the related Wikipedia entity
- Entity-linked the entire dataset, using TAG·ME



# A New Data Set

Collected all speeches from the UK House of Commons (2013-2016), organized under fine-grained topics by the curators.

| Session | # Speech | # Topic | # Token | # Entity |
|---|---|---|---|---|
| 2013-14 | 23,935 | 2,343 | 175,604 | 72,791 |
| 2014-15 | 19,439 | 1,987 | 166,777 | 72,248 |
| 2015-16 | 26,605 | 1,923 | 169,119 | 74,678 |
| Total | 69,979 | 5,634 | 354,403 | 125,886 |

# A New Data Set

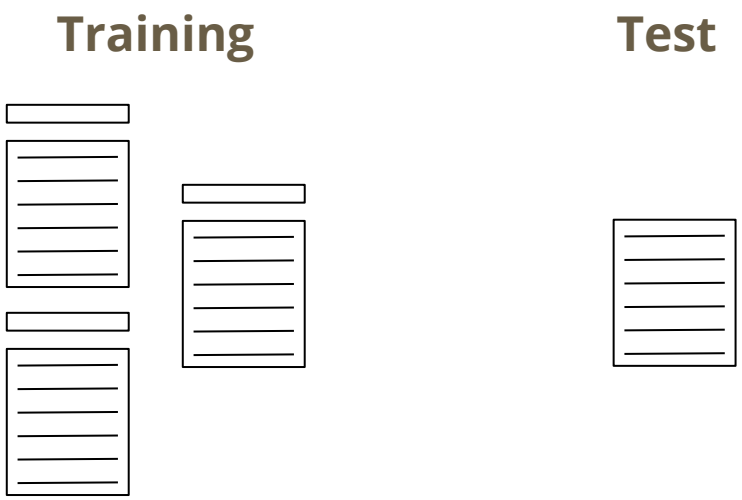
Collected all speeches from the UK House of Commons (2013-2016), organized under fine-grained topics by the curators.

| Session | # Speech | # Topic | # Token | # Entity |
|---|---|---|---|---|
| 2013-14 | 23,935 | 2,343 | 175,604 | 72,791 |
| 2014-15 | 19,439 | 1,987 | 166,777 | 72,248 |
| 2015-16 | 26,605 | 1,923 | 169,119 | 74,678 |
| Total | 69,979 | 5,634 | 354,403 | 125,886 |

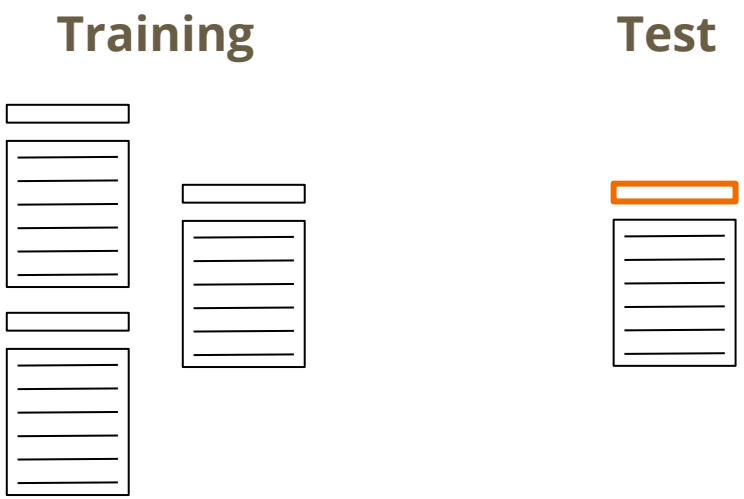
The first version of the dataset is available here:
https://federiconanni.com/ukparl/

# Benchmark: Topic Classification

# Benchmark: Topic Classification



# Benchmark: Topic Classification



# Benchmark: Topic Classification

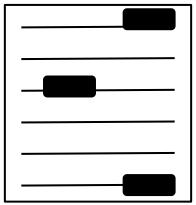
| Doc. Representation | Classifier | Topic Prediction | | | |
|---|---|---|---|---|---|
| | | Macro | | | Micro |
| | | P | R | $F_1$ | $F_1$ |
| TF-IDF (words) | NB<br>NearestCentroid<br>*k*-NN<br>SVM | | | | |
| TF-IDF (entities) | NB<br>NearestCentroid<br>*k*-NN<br>SVM | | | | |
| Word embeddings | NB<br>NearestCentroid<br>*k*-NN<br>SVM | | | | |
| Entity embeddings | NB<br>NearestCentroid<br>*k*-NN<br>SVM | | | | |

# Benchmark: Topic Classification

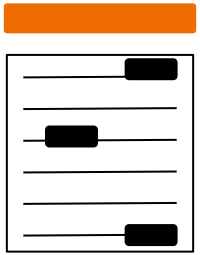
| Doc. Representation | Classifier | Topic Prediction | | | |
|---|---|---|---|---|---|
| | | Macro | | | Micro |
| | | P | R | $F_1$ | $F_1$ |
| TF-IDF (words) | NB | 0.18 | 0.14 | 0.15 | 0.17 |
| | NearestCentroid | **0.52** | **0.49** | **0.50** | **0.46** |
| | $k$-NN | 0.41 | 0.42 | 0.41 | 0.42 |
| | SVM | 0.49 | 0.39 | 0.43 | 0.44 |
| TF-IDF (entities) | NB | 0.10 | 0.09 | 0.09 | 0.10 |
| | NearestCentroid | **0.30** | **0.30** | **0.30** | **0.28** |
| | $k$-NN | 0.22 | 0.23 | 0.22 | 0.24 |
| | SVM | 0.27 | 0.25 | 0.25 | **0.28** |
| Word embeddings | NB | 0.31 | 0.28 | 0.29 | 0.24 |
| | NearestCentroid | 0.33 | **0.33** | **0.33** | 0.33 |
| | $k$-NN | 0.26 | 0.27 | 0.26 | 0.29 |
| | SVM | **0.36** | 0.31 | **0.33** | **0.38** |
| Entity embeddings | NB | 0.16 | 0.18 | 0.16 | 0.15 |
| | NearestCentroid | 0.23 | 0.23 | 0.23 | 0.21 |
| | $k$-NN | 0.17 | 0.18 | 0.17 | 0.20 |
| | SVM | **0.27** | **0.25** | **0.25** | **0.28** |

# Benchmark: Topic Ranking

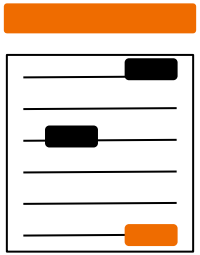
# Benchmark: Topic Ranking



# Benchmark: Topic Ranking



# Benchmark: Topic Ranking



Only 22% of the documents mention the topic in the content.

# Benchmark: Topic Ranking

| | MAP | P@1 |
|---|---|---|
| Baseline (Random) | | |
| Entity frequency<br>Entity TF-IDF | | |
| Centroid (embeddings) | | |
| Position (doc. order)<br>Position + frequency<br>Position + TF-IDF<br>Position + centroid | | |

# Benchmark: Topic Ranking

| | MAP | P@1 |
|---|---|---|
| Baseline (Random) | 0.13 | 0.04 |
| Entity frequency | **0.37** | **0.24** |
| Entity TF-IDF | **0.37** | **0.24** |
| Centroid (embeddings) | 0.20 | 0.10 |
| Position (doc. order) | 0.23 | 0.09 |
| Position + frequency | 0.24 | 0.14 |
| Position + TF-IDF | 0.22 | 0.11 |
| Position + centroid | 0.22 | 0.12 |

# Conclusions and Next Steps

UKParl: a data set for supporting the **evaluation** of supervised and unsupervised **topic detection methods** on parliamentary speeches.

# Conclusions and Next Steps

UKParl: a data set for supporting the **evaluation** of supervised and unsupervised **topic detection methods** on parliamentary speeches.

In the near future we will:

1. Extend it diachronically
2. Offer manual alignment between topics and Wikipedia pages
3. Expand the set of tested baselines
4. Align with other resources

# Questions?

Federico Nanni

Data and Web Science Group

University of Mannheim

federico@informatik.uni-mannheim.de