Extracting Topic-Specific Ideological Positions from News Articles*

Iulia Cioroianu†, Susan Banducci‡, Zoltan Szlavik§

28th August 2018

1 Introduction

Media exposure is a central concept in the social sciences, and previous research showed that the information individuals are exposed to has the potential to influence their political beliefs and attitudes (Prior 2013). In order to understand change (but also stability) in opinions and behaviour, it is therefore necessary to understand individual patterns of news consumption.

However, the information exposure process has several characteristics which make it difficult to capture and analyse. First, it is a continuous process. Political information comes as a constant stream of news, facts and opinions, which when taken individually may not have a strong impact, but which may have considerable impact on political attitudes and beliefs when taken together and analysed as a dynamic process. Second, selectivity in news and information exposure makes it difficult to disentangle the causal pathways between individuals' political positions and the information that they choose to consume, or to which they are being exposed to. People choose to consume content that fits with their prior beliefs, immersing themselves into "echo chambers" (Sunstein, 2009). They opt for partisan sources over sources that offer a variety of perspectives (Mutz, 2001) and seek out websites that match their pre-held beliefs and promote their ideological views (Bimber and Davis, 2003). These tendencies seem to be exacerbated by the online environment (Quattrociocchi et al., 2016). Third, aside from self-selection effects, content availability and individual access to a variety of sources and perspectives is constrained by search engines, news aggregators and social media sites that personalise the user experience and promote information and opinions which conform to, and reinforce opinions, creating 'filter bubbles' (Dillahunt et al. 2015).

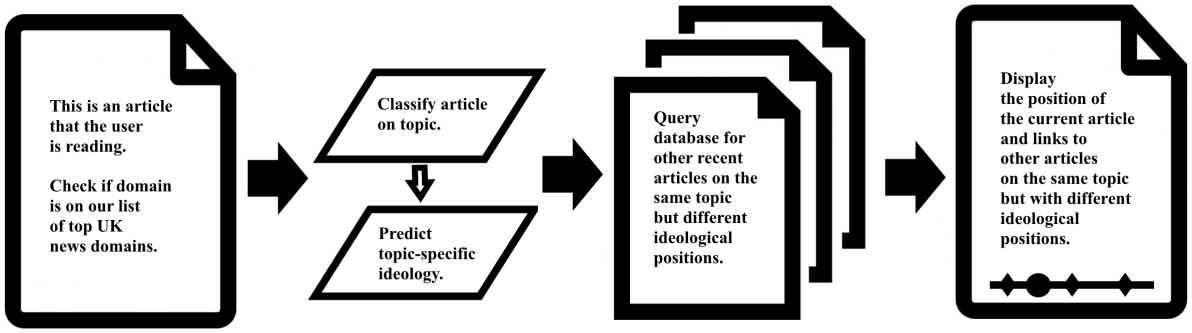
*Paper prepared for presentation at the Annual Meeting of the American Political Science Association, Boston MA, September 2018. This work has been supported by ESRC This work was supported by the Economic and Social Research Council ES/N012283/1 "Measuring Information Exposure in Dynamic and Dependent Networks (ExpoNet), an ESRC IAA UKRI Innovation Fellowships and an IBM Faculty Award. We weould like to thank Mustafa Tekman and Ioanna Birmpa for valuable research assistance.

†Postdoctoral Fellow, University of Exeter. Corresponding author, i.cioroianu@exeter.ac.uk

‡Professor, University of Exeter.

§Lead, Centre for Advanced Studies, IBM Netherlands.

Figure 1: UNBias pipeline



So even if some users may want to broaden their views, they may not have the chance to do so. This is especially problematic in light of normative theories of democracy, which suggest that its functioning depends on citizens' access to a variety of sources of information (Downs 1957).

In order to address these challenges, we would have to capture exposure to political information as it happens, and be able to analyse it as a dynamic process. At the same time, in order to establish the direction of causal relations, we would want to emulate an experimental design in which we offer individuals the choice to read news articles that can broaden their ideological perspective, and to which they may not be exposed otherwise. This would allow us to analyse selective exposure, but also identify the factors associated with the willingness to expand one's political views. To this end, the EXPONet team at the University of Exeter and the IBM Netherlands Center for Advanced Studies are working together to develop a news reading application and the methods and algorithms that underpin it.

UNBias (Understanding News Bias) is a browser extension that allows users to evaluate the ideological leaning of the articles they read, estimate how left-leaning or right-leaning their overall media exposure is, and broaden their ideological perspective on different topics. It uses the text of the article that the user is currently reading, classifies it into one of 11 categories that correspond to salient political issues using supervised multi-class classification, computes its topic-specific ideological position, and displays this position on screen, along with links to other recent news articles which are on the same topic, but cover a broader range of ideological perspectives (Fig. 1). The extension also allows researchers to conduct surveys and experiments which will provide a rich source of data for studying the political information exposure process, the relation between individual preferences and article selection, and individual response to being offered a diverse set of views on different topics.

This paper presents the main text classification and ideological scaling methods that underpin the UNBias app, comparing several common approaches and evaluating their effectiveness. We evaluate these methods on a large corpus of articles which were collected around the 2015 UK General elections (Stevens et al., 2017), as well as a corpus of more recent articles collected between April and June 2018 using Watson Discovery News. Our goal is to develop a robust and efficient method that takes news articles as input, identifies if their main topic is one of eleven pre-defined political topics, and calculates their topic-specific

ideological position.

2 Background

2.1 Media ideology

Media slant has been previously studied using text analysis, network analysis and survey methods. Groseclose and Milyo (2005) use speeches by politicians in the US Congress to form a training set, which they use to scale US media outlets by the frequency of mentions of political think tanks. Gentzkow and Shapiro (2010) build upon this approach, but instead of mentions of think tanks, they count mentions of key political phrases, and measure the slant of 433 US daily newspapers. Gentzkow and Shapiro (2011) and Flaxman et al. (2016) rely on a similar method for classifying the ideological orientations of online domains - they take the average ideological orientations of those visiting them. Barbera and Sood (2016) present a method of scaling media sources on the same scale as politicians and voters which is based on follower relations on Twitter.

A common feature of all of these studies is the fact that they focus on extracting ideological positions for the outlets as a whole, not individual articles within an outlet. However, to be able to measure the effect of ideology in the information exposure process we need to compute ideology at the article level. Individuals consume news stories, and the assumptions that we would have to make about the ideological content of articles, that of the outlet, and the process through which the user choose which articles to read within an outlet in order to use outlet ideology to evaluate exposure are not realistic. Therefore, we need to identify methods that can extract ideology at the article level.

2.2 Extracting ideological positions from text

Several approaches have been used to extract ideological positions from political texts (for an overview, see Grimmer and Stewart, 2013; Gentzkow et al. 2017; and Perry and Benoit, 2018). The most common ones are Wordscores (Laver et al. 2003) and Wordfish (Slapin and Proksch, 2008). Wordscores is a supervised text scaling method, which starts with a set of documents with known positions - the reference texts. It assumes that every document has a policy position score, and that the score of a document is the average of the scores of its words, and calculates the probability of document i upon observing word w_j. Unseen, or virgin texts are scored based on the average scores of the words they contain. The method makes no functional or distributional assumptions, is easy to implement, and has been shown to work well in different contexts (Klemmensen et al. 2007). But it is highly sensitive to the choice and quality of the reference texts and may perform poorly in other contexts.

Wordfish is an unsupervised method for estimating latent text-specific traits, which does not need any anchoring or reference texts to scale documents. Instead, it relies on a statistical model of word counts. The model assumes that each word j from document i, W_{ij}, is drawn from a Poisson distribution with rate λ_{ij}, where $\lambda_{ij}$is modeled as a function of document length α_{ij}, the frequency of word j, ψ_{ij} and the degree to which the word captures the underlying ideology space , $\beta_{ij}$and the document's underlying position θ_{ij}, $\lambda_{ij} = exp(\alpha_{ij} +$

$\psi_{ij} + \beta_{ij} + \theta$). A Poisson regression model estimates both document and word parameters. It start by guessing the parameters. It then assumes that the document parameters are correct, and fits a Poisson regression model for the word parameters. It then assumes that the fitted word parameters are correct, and fits a Poisson regression model for the document parameters. The process is repeated until convergence. The disadvantage of this method is that we cannot be sure that the estimated latent trait corresponds to the quantity of interest, in this case ideology.

A number of other machine learning approaches could be used to predict the ideological score of each article. We follow the approach of Rennie and Srebro (2005), and Pedregosa-Izquierdo (2016) and compare the performance of a multi-class classifier against that of simple regression, as well as ordinal logistic regression with different loss functions.

2.3 News bias apps

Several news bias and fact checking apps and websites have been developed recently, such as Media Bias/Fact Check (2018) and AllSides (2018). However, none of these evaluate media bias or slant at the article level. Most of the apps focus on media sources based in, or popular in the United States, and assume that ideology is uni-dimensional. However, publications such as Financial Times or the Economist differ in their positions on economic and social issues. Both of them adopt conservative positions on economic issues, but liberal positions on social issues. The problem is exacerbated in countries with multiple salient cross-cutting ideological dimensions and social cleavages, and for niche publications, which focus on a particular topic of interest. UNBias attempts to address these issues by providing topic-specific ideological estimates at the article level.

3 Data and methods

3.1 Corpus

The first dataset of articles used in this analysis is based on the Stevens et al. (2017) collection of media data around the 2015 UK general election. The collection includes articles from national and regional newspapers[1] in England, Scotland, and Wales which were published during the February 1st, 2015 to May 28th, 2015 period. The articles were collected through the Nexis UK service for academics. This initial collection of newspaper content resulted in 338,923 stories, the vast majority of which were not related to politics or the election. To identify the stories focusing primarily on the election, we trained a Stochastic Gradient Descent classifier on a random sample of 11,000 news stories coded on whether the primary topic was related to the election or not. Overall, the classification performance was very high, with an average F1-score of 0.96 based on 10-fold cross-validation. The final dataset consist of 23.000 articles about politics and the 2015 general elections.

[1]The Daily Telegraph, The Financial Times, The Guardian, The Independent, The Times, Daily Express, Daily Mail, Daily Mirror, Daily Star, Sun, Daily Record, The Scotsman, Evening Standard, Birmingham Mail, Western Morning News, Western Mail (Wales) and Yorkshire Evening Post.

The second dataset was collected using Watson Discovery News, between April 2018 and June 2018. Articles were collected using lists of keywords associated with each topic (based on an LDA model on the 2015 data), and the search was restricted to popular UK political news domains. To identify these domains we followed the procedure detailed in Banducci et al. (2017), which starts from a list of the most popular news domains, as identified by Alexa, an Amazon service that ranks websites by traffic and classifies them into multiple categories based on content. Within the "News" category, we selected websites which were places in the "Newspapers", "Analysis and Opinion", "Breaking News", "Current Events", "Extended Coverage", "Internet Broadcasts", "Magazines and E-zines", "Journalism" and "Weblogs" sub-categories, and selected the top 400 domains in each news category, as well as the top domains categorized by UK region. The total number of news domains considered was 4,179. These domains were further pruned to eliminate those which were never visited by a panel of ICM Unlimited survey respondent for which clickstream data was also collected before the 2015 elections. We also eliminated domains which only included weather and other procedural articles (such as traffic information, sports results, TV programming guides, stocks monitoring pages), news aggregation websites (such as Google and Yahoo News, Flipboard, etc.), videos without an attached article or description, guides and how-to pages (recipes, reviews, self-diagnosis, travel guides, etc.) and those that were not visited at all by UK users in a which left a total number of 480 news domains. A total of 3500 articles were collected using this method, through Watson Discovery News.

3.2 Selecting major political topics

We selected major topics of interest based on their importance and ranking in the following types of data:

a. the political manifestos of major UK political parties from 2015 and 2017. The section headers of the Conservative, Labour, Liberal Democratic, SNP and UKIP manifestos were compared, and the topics that were addressed in most manifestos were selected.

b. voter answers to the question "What is the single most important issue facing the country at the present time?" from the British Election Study waves 5-13 (2015-2017) were analysed, and the most common answers were included in our short list of topics;

c. the results of a LDA model on our corpus of election articles.

Combining these sources we ended up with 11 topics which were popular in party manifestos, among voters and in the news articles: Brexit and the EU, Budget, Crime and Policing, Economy, Education, Foreign policy, Housing, Immigration, Jobs, NHS and Welfare.

3.3 Selecting documents for initial human coding

While some of the topics, such as the economic ones were over-represented in our 2015 corpus, others, such as Crime and Policing were underrepresented. To ensure we were coding enough sentences on each topic, we oversampled the training set texts by using a 3-step process. First, we compiled lists of the most relevant keywords by topic from a Latent Dirichlet Allocation (LDA) model on the entire election corpus. The lists were pruned based on the words' contribution to the topic, relative to its contribution to other topics, and human coder

judgment. Finally, we queried the database for the words in the list and selected the top 500 articles for each topic category for human coding.

For the analysis presented in this paper, we focus on 12 of the major UK newspapers, while keeping the pre and post-election period the same as the one for the entire collection (1 February to 28 May 2015). The Telegraph, Daily Mail and the Sun endorsed the Conservative Party and have a long tradition in the last electoral cycles of strong endorsement of Conservative policies while the Guardian and the Mirror had endorsed Labour with similar historical support. The Independent officially endorsed the coalition but have no historical ties to either party.

The 2017 collection was designed to include an approximately equal number of articles per topic, using similar keyword queries as above, which resulted in an average of 320 articles per topic.

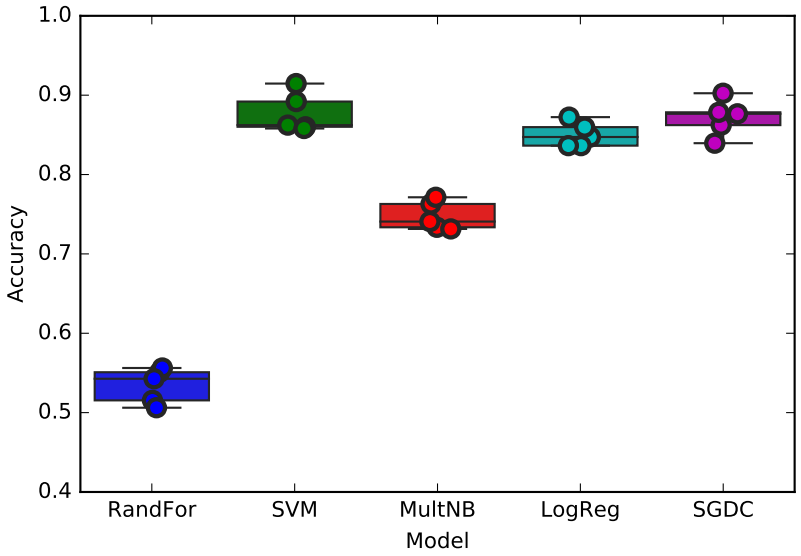
3.4 Human coding by topic

Coding news articles on CrowdFlower raises specific challenges. One of them is having a great degree of variation between the lengths of the newspaper articles, which can range from less than a hundred words to thousands of words. The required amount of attention from coders differs in this case, while payment stays the same. To overcome this issue, we hypothesize that the main topic of an article is preserved if the article is summarized using automatic text summarizing. We test this hypothesis using a small initial set of 550 articles (50 per topic) which were expert coded for topic. We summarized the articles into short texts with an average of either 100 or 300 words using the Gensim (Radim and Sojka, 2010) implementation of the TextRank algorithm (Mihalcea and Tarau, 2004) and compared coder agreement across the two types of sets, as well as with the the expert coding. The results showed that the two codings were highly correlated, with a correlation of more than 0.80 for each of the topics. Somewhat surprisingly, shorter texts were slightly more correlated with the expert coding than longer ones. We are therefore confident that summarized articles preserves the most important topic in the newspaper articles, and we designed a multi-label coding task for CrowdFlower on summaries of our articles with an average of 150 words. Coders were allowed to select any of 11 topic categories for an article, as well as a category for "none". Table 1 reports coder agreement for each of the 11 topics, and overall. Low levels of agreement for the Jobs and Welfare categories seem to be driven by the tendency to confuse the two, as well as the fact that articles on these topics also often have Budget or Economy as additional topics. Overall however, coder agreement levels are good for a multi-label classification task.

3.5 Topic classification in the full corpus

In the UNBias browser extension, when a user opens a new page, the first thing the application does after grabbing the text from the page and checking if it is on one of the top UK news domains, is to classify it and assign it to one of the 11 topic categories, or to the "None" category. The classifier makes the assumption that each new article is assigned to a single category. This is a multi-class text classification problem. We tested and evaluated the performance of multiple classifiers which can handle multi-class supervised classification

Figure 2: Comparing classifier accuracy



problems. Figure 2 presents the average accuracy in 5-fold cross-validation for the following classifiers: (a) Random Forest, (b) linear Support Vector Machine, (c) multinomial Naive Bayes, (d) logistic regression, and (e) Stochastic Gradient Descent. Whereas in the corpus building phase the SGDC performed best for binary classification of news articles into election and non-election classes, the linear SVM overtakes it for the multi-class classification problem.

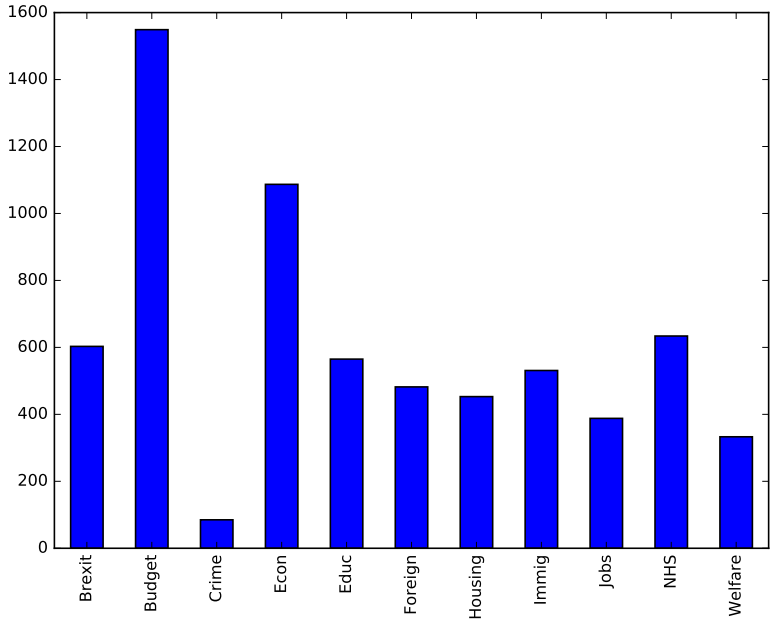
We therefore select the linear SVM classifier model to work with, and present its performance for each of the topics of interest in Table 2. The classifier performs very well overall, with F-1 scores close to, or above 0.90 for most topics. Recall is low however for three of the topics (Jobs, Crime and Brexit) so we inspect misclassifications using a confusion matrix based on a test size of 25% of the articles in Fig. 3. These Jobs, Crime and Brexit are the categories which are most commonly misclassified as "None". Since app users expect that not all the articles they are reading will be analysed, we are more concerned about precision than recall, so we are not overly concerned about losing some of the articles to the "None" category.

From a more substantive perspective, our classification exercise should be able to capture the ranking of topics in the 2015 election which corresponds to previous knowledge about the most relevant issues for voters and parties. Indeed, Figure 4 shows that the most discussed topics in the 2015 elections were the Budget and Economy, followed by the NHS, Brexit and Immigration and Education, which match voter concerns, but also media reports of the time.

Table 1: Human coding of topics - CrowdFlower coder agreement

	Coder agreement	
	Mean	Std. Dev.
Brexit	.75	.14
Budget	.75	.14
Crime	.69	.13
Economy	.66	.20
Education	.72	.21
Foreign policy	.77	.17
Housing	.70	.21
Immigration	.74	.21
Jobs	.59	.21
NHS	.75	.22
Welfare	.50	.20
Overall	.67	.22

Figure 4: Predicted topic prevalence in GE2015 corpus



3.6 Human coding of ideological position

From the set of articles coded by topic, we selected those on which coder agreement was 100% (a total of 1500 articles), and coded them further on ideology. We created separate tasks for each of the 11 topics. Coders received the same summary of the articles as before, and had to rate the text on a 1 to 10 ideology scale with ends consisting of short statements outlining the most important stances associate with a left, and a right wing position on the

Table 2: Topic classification performance

	Average standard deviation		
	Precision	Recall	F1-score
Brexit	0.79	0.75	0.77
Budget	0.91	0.85	0.87
Crime	0.94	0.38	0.54
Economy	0.86	0.87	0.86
Education	0.87	0.99	0.92
Foreign policy	0.92	0.97	0.94
Housing	0.96	0.99	0.98
Immigration	0.98	1.00	0.96
Jobs	0.88	0.76	0.81
NHS	0.98	1.00	0.99
Welfare	0.95	0.89	0.92
Overall	0.88	0.87	0.87

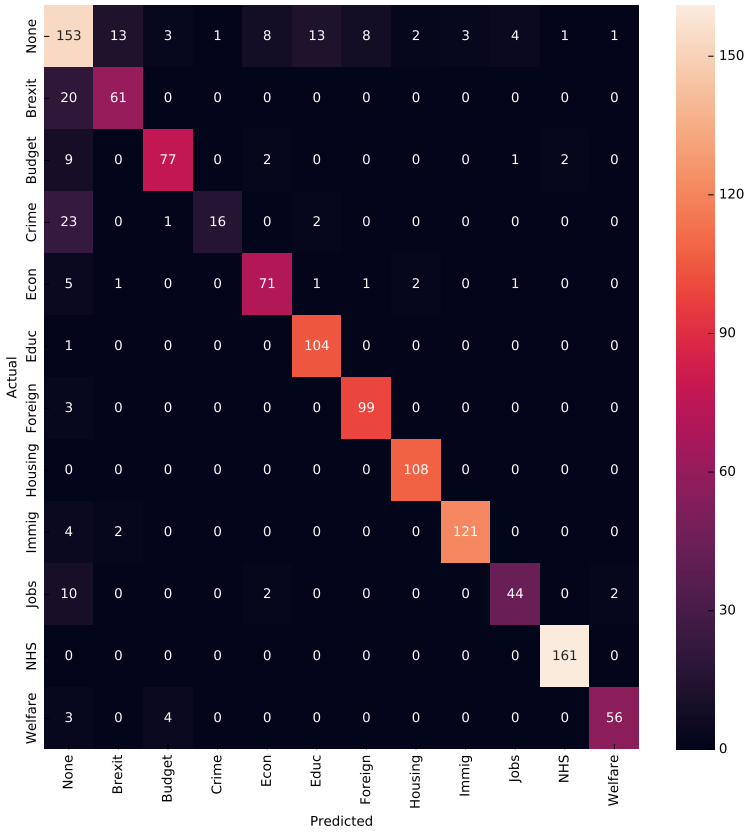
topic. The instructions provided a more detailed description of the meaning of left and right ideology in the context of each topic, as well as examples. The emphasis was placed not on coders being able to understand what left and right means for each topic, but on them being able to evaluate if the article content is more similar to the statements listed on the left, or those listed on the right. The level of difficulty associated with the task was relatively high, so we restricted it to experienced UK-based coders, and increased payment relative to the topic coding task by 50%. Coders also had the option to tick a box if they believed the article they were coding was not on topic. Almost 5% of the coded articles were flagged by at least one coder as not being on topic, and were dropped from further analysis.

The variance in selected positions for an article is the quantity of interest in evaluating coder agreement on a scale-rating task. Table 3 presents the averaged variances for each of the 11 topics, along with the averaged positions of the coded articles for each of the topics, which allow for a more substantive interpretation. Coder deviations were relatively small for most of the topics, and the average article positions match our expectations based on previous knowledge of the 2015 context, in which there was an increased anti-European sentiment, and the Conservatives obtained surprisingly good results, partly driven by their stance on the economy, foreign policy and immigration.

3.7 Ideological scaling - Wordscores and Wordfish

We applied two of the most commonly used methods for scaling political text to the full dataset of topic-classified articles, and to the sub-sample of articles which were human-coded on ideology. The predictions from Wordfish and Wordscores are correlated, but the correlation is smaller than the one reported for other types of political data, such as party manifestos, and is around 0.4 for most topics. Figure 5 presents an example of predicted Wordscore document positions and Wordfish estimated thetas for one of the topics, the Budget, for the subset of articles which had a perfect agreement among CrowdFlower coders

Figure 3: Confusion matrix



Test size 25%.

based on topic, and which were subsequently coded on ideology. The ordering of news outlets is given by the averaged position of their articles. The correlation between the predicted document potions is 0.41 in this case. Wordfish seems to capture more within-outlet variation, but the ordering of news sources for Wordscores matches common beliefs about the ideological landscape of UK media more, with publications that are economically conservative, such as the Financial Times, Daily Mail and Telegraph on the right, and publications that are economically liberal, such as the Guardian, Scotsman and the Mirror on the left.

To evaluate if the positions predicted by Wordfish and Wordscores are capturing ideology, we compared them to the human coded positions. Table 4 presents the correlations for each topic. Overall, Wordfish estimates are more closely correlated with the human coding, although for each of the two measures the correlations are lower than we would like them to be. Wordscores requires a valid set of reference texts and highly confident estimates of the ideology positions of these texts, and the results are highly sensitive to the choice of reference texts. In this case, we used texts that were expert coded as being extreme on each topic, under the assumption that coder judgment relies exclusively on the definition of ideology

Table 3: Coder agreement on ideology coding

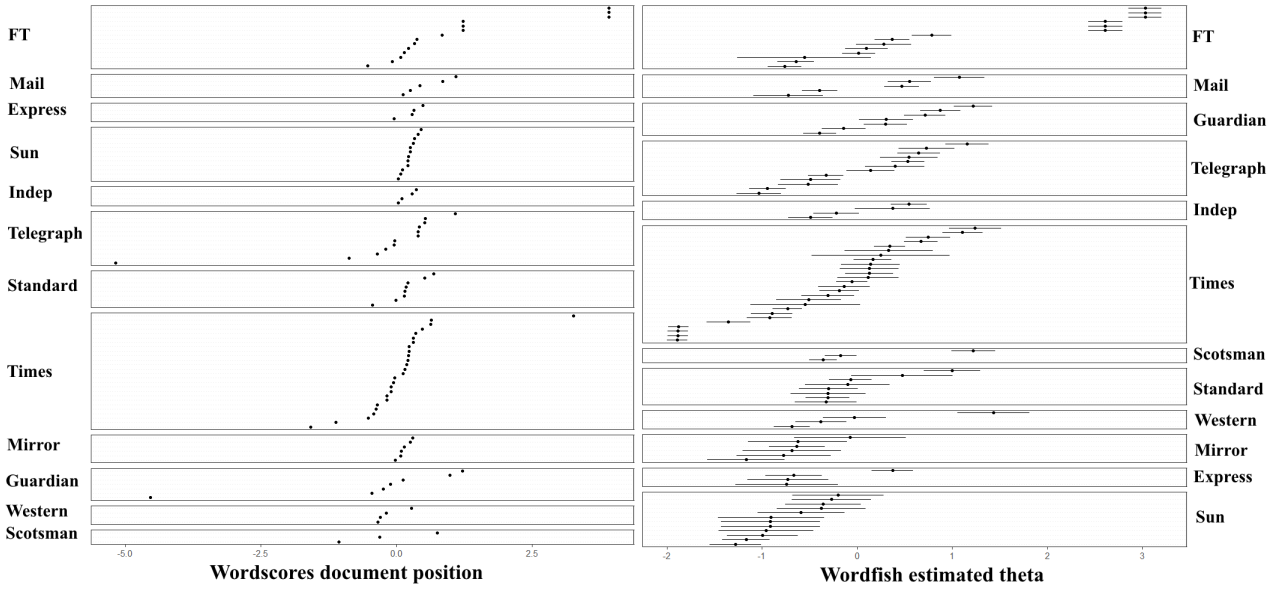
	Average coder position	Averaged coder deviation
Brexit	6.08	1.97
Budget	5.66	1.37
Crime	5.24	1.24
Economy	5.70	1.01
Education	5.08	0.27
Foreign policy	5.69	1.10
Housing	5.26	0.74
Immigration	5.32	0.84
Jobs	3.62	1.40
NHS	5.39	2.30
Welfare	4.83	1.83

Table 4: Ideology correlation with human coding

	Correlation with human coding	
	Wordscores	Wordfish
Brexit	0.17	0.25
Budget	0.16	0.19
Crime	-0.09	0.17
Economy	0.17	0.23
Education	-0.04	0.18
Foreign policy	0.22	0.17
Housing	0.13	0.24
Immigration	0.20	0.18
Jobs	0.15	0.21
NHS	0.10	0.15
Welfare	0.16	0.23

and left-right policy positions that we provided. On the other hand, Wordfish retrieves the dimension that explains the most variance in the word sample that is used as input. If CrowdFlower coders were using other cues in the text, apart from the stated ideological positions, such as the names, job titles or other elements related to political actors, then the correlation between coder positions and Wordfish estimates may be due to the fact that they are both capturing a dimension which may not correspond exactly to our understanding of ideology based on stated preferences for policies (which was what the expert coder selected reference texts on).

Figure 5: Wordscores, Wordfish on high confidence budget texts



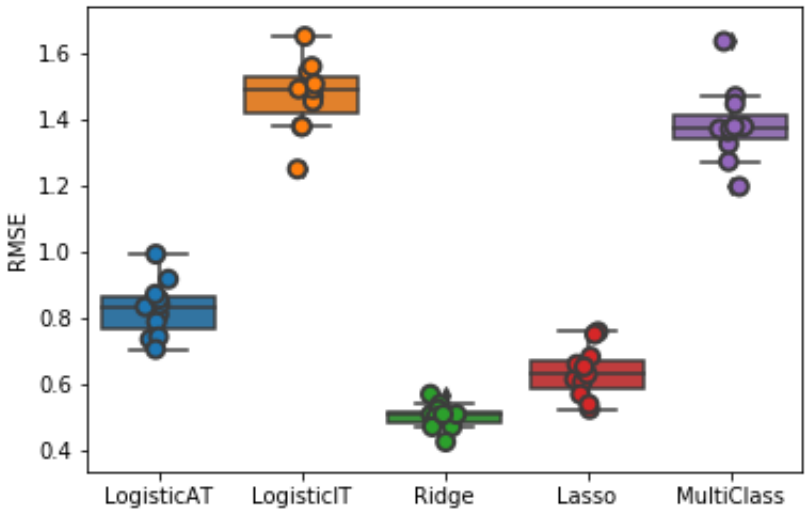
3.8 Ideology machine learning methods - regression and classification

Given the modest performance of the Wordscores and Wordfish algorithms, we turn to other machine learning methods, which we compare in Figure 6.

We can think of the ideology scores as discrete, un-ordered labels. In which case, we could use a multi-class classifier, such as multinomial logistic regression (labeled MultiClass in Fig. 6). If however we think of ideology as continuous values, which is what is usually assumed in the literature, then the problem becomes a typical regression task. Here we use a linear model with a squared *l2* regularization with a linear least squares loss function - a specification which is called Ridge regression (Fig. 6 Ridge), as well as a Lasso model (Nightshirt, 1996) which combines *l1* regularization with a least squares loss function. Finally, we could think of the ideological labels as discrete but ordered, which would be in line with the way ideology is displayed in the application, as well as the way it was coded by crowd-coders. Rennie and Srebro (2005) compare several of these methods. Their substantive focus is on predicting user rankings of products and services, and they use the example of movie review scores. They propose two types of threshold-based ordinal regression models. Both models introduce thresholds between the categories and combine penalties for threshold violations, but differ in whether they take only the immediate thresholds into account (LogisticIT in Fig. 6), or impose grater penalties for crossing multiple thresholds by summing all threshold violation penalties (LogisticAT in Fig. 6).

We plot root mean squared errors in 5-fold cross-validation, averaged across each of the topic areas for the following models: the all-threshold penalty version of the ordinal logistic regression model (LogisticAT), the immediate-threshold only penalty version of the ordinal logistic regression model (LogisticIT), the ridge regression model (Ridge), the Lasso model (Lasso), and the multi-class logistic regression classifier (MultiClass). The models

Figure 6: Supervised classification - ideology



were implemented in Python using the packages *scikit-learn* and *mord*. The best performing models are the Ridge and Lasso regression. Both of these models are able to handle sparse features well, but the Ridge regression proved to be less computationally efficient than the Lasso (Nightshirt, 1996), which it yields sparse solutions and has been found to have a number of desirable properties when applied to text (Genkin et al., 2007). Among the ordinal regression models, the specification with the all-threshold penalty performs better, while the immediate threshold version is overall comparable or worse than a multi-class classifier.

4 Discussion and future work

The analysis above suggests that the supervised classification methods for political topics work very well, with the understanding that the model will have to be constantly recalibrated, since the language used to discuss a certain topic may change over time, even if the overall topic itself stays the same. The topic classification process can be improved by eliminating the human input element used to prune the lists of keywords from the initial topic model. On the other hand, two classical ideological scaling methods, Wordfish and Wordscores perform worse that we would like them to. Ridge and Lasso regression show promising results, which could potentially be improved by adding a number of article and journal level covariates in the model.

The application is designed as a Chrome browser extension built upon IBM Watson services, and it is currently in test phase. We plan to recruit a panel of users representative of the UK voting-age population and have them install the browser extension. Upon installing the extension and providing informed consent, users are given the option to take a short online survey covering basic demographic information as well as information about their ideological positions, political preferences and patterns of news consumption. Parts of the survey will be repeated at later dates to evaluate changes in attitudes. We are monitoring

user web browsing activities and storing the links visited if they are located on domains which are from our pre-defined list of news domains, and we analyse media exposure in relation to the individual characteristics recorded through the user surveys. We also record user choice to click on suggested articles, as well as the topic and ideological position of those articles, and analyse these choices in relation to their web browsing histories and individual characteristics. We plan to manipulate the articles to be displayed as suggestions to users, in order to evaluate how different sets of choices with various ideological positions affect individual decisions to read an article.

References

[1] AllSides. AllSides | Balanced news via media bias ratings for an unbiased news perspective.

[2] S. A. Banducci, I. Cioroianu, T. Coan, G. Katz, E. Kolpinskaya, and D. Stevens. Content Analysis of Media Coverage of the 2015 British General Election. Technical report, UK Data Service, 2017.

[3] Pablo Barbera and Gaurav Sood. Follow your ideology: Measuring media ideology on social networks. In *Annual Meeting of the European Political Science Association, Vienna, Austria. Retrieved from http://www. gsood. com/research/papers/mediabias. pdf Google Scholar*, 2015.

[4] Bruce Bimber and Richard Davis. *Campaigning Online: The Internet in U.S. Elections*. Oxford University Press, September 2003.

[5] Tawanna R. Dillahunt, Christopher A. Brooks, and Samarth Gulati. Detecting and Visualizing Filter Bubbles in Google and Bing. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1851–1856, New York, NY, USA, 2015. ACM.

[6] Discover. Discover - Content Analysis of Media Coverage of the 2015 British General Election.

[7] Anthony Downs. An Economic Theory of Political Action in a Democracy. *Journal of Political Economy*, 65(2):135–150, 1957.

[8] Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1):298–320, January 2016.

[9] Matthew Gentzkow, Bryan T Kelly, and Matt Taddy. Text as data. Technical report, National Bureau of Economic Research, 2017.

[10] Matthew Gentzkow and Jesse M. Shapiro. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1):35–71.

[11] Matthew Gentzkow and Jesse M. Shapiro. Ideological Segregation Online and Offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, November 2011.

[12] Gentzkow Matthew and Shapiro Jesse M. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1):35–71, February 2010.

[13] Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297, 2013.

[14] Tim Groseclose and Jeffrey Milyo. A Measure of Media Bias. *The Quarterly Journal of Economics*, 120(4):1191–1237, November 2005.

[15] Richard Hanna, Andrew Rohm, and Victoria L. Crittenden. We're all connected: The power of the social media ecosystem. *Business Horizons*, 54(3):265–273, May 2011.

[16] Robert Klemmensen, Sara Binzer Hobolt, and Martin Ejnar Hansen. Estimating policy positions using political texts: An evaluation of the Wordscores approach. *Electoral Studies*, 26(4):746–755, December 2007.

[17] Michael Laver, Kenneth Benoit, and John Garry. Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, 97(2):311–331, May 2003.

[18] Michael Laver, Kenneth Benoit, and John Garry. Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, 97(2):311–331, May 2003.

[19] MediaBias. Media Bias/Fact Check - Search and Learn the Bias of News Media.

[20] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[21] Diana C. Mutz. Facilitating Communication across Lines of Political Difference: The Role of Mass Media. *American Political Science Review*, 95(1):97–114, March 2001.

[22] Fabian Pedregosa-Izquierdo. *Feature extraction and supervised learning on fMRI : from practice to theory.* phdthesis, UniversitÃ© Pierre et Marie Curie - Paris VI, February 2015.

[23] Patrick O. Perry and Kenneth Benoit. Scaling Text with the Class Affinity Model. *arXiv:1710.08963 [cs, stat]*, October 2017. arXiv: 1710.08963.

[24] Markus Prior. Media and Political Polarization. SSRN Scholarly Paper ID 2265184, Social Science Research Network, Rochester, NY, May 2013.

[25] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

[26] Jason D. M. Rennie. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, pages 180–186, 2005.

[27] Slapin Jonathan B. and Proksch Sven Oliver. A Scaling Model for Estimating Time Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722, July 2008.

[28] Cass R. Sunstein. *Going to Extremes: How Like Minds Unite and Divide*. Oxford University Press, May 2009. Google-Books-ID: jEWplxVkEEEC.

[29] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[30] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, January 2016.