Using topic models to reevaluate media effects

Frederik Hjorth[*]

VERY ROUGH EARLY DRAFT – PLEASE DO NOT QUOTE, CITE, OR DISTRIBUTE

Presented at the Amsterdam Text Analysis Conference
June 21-22, 2016

Abstract

The effects of media coverage on political attitudes is a perennial theme in the study of public opinion. However, it is difficult to obtain valid and reliable measures of the media content to which citizens are exposed. Existing observational studies typically measure media coverage of an issue using frequencies of pre-defined issue-relevant keywords. In this paper, I evaluate an alternative to this count-based method based on a topic model of the full text of news articles. By estimating trends in issue-relevant topic, the method can isolate the most issue-relevant information returned by keyword searches. Results suggest the topic model-based method produces larger point estimates of media effects, though the data used has insufficient power to detect a statistically significant difference.

[*] Corresponding author. Department of Political Science, University of Copenhagen, Øster Farimagsgade 5, DK-1353 Copenhagen K, (+45) 26 27 24 41, fh@ifs.ku.dk

1 Introduction

Since the inception of the modern study of public opinion, scholars have assigned a critical role to mass media in shaping attitudes, defining problems, and establishing evaluative criteria in the minds of citizens. Yet while the notion that mass media heavily influence public opinion is intuitively plausible, empirically demonstrating effects of mass media is challenging. Researchers have traditionally responded to this challenge by opting for experimental over observational designs, with ensuing threats to external and ecological validity. Researchers looking to identify effects of mass media using observational designs are faced with the challenge of accurately measuring the media environment faced by citizens.

In this paper, I describe and evaluate a method for improving the measurement of the mass media issue salience in observational studies. Simply put, the method uses results from a topic model of newspaper content to improve the traditional method using keyword counts, placing more weight on keyword counts that have a higher share of issue-relevant content. Results from the evaluation suggest that the method may in some cases reduce measurement error and in turn produce larger estimates.

The remainder of the paper consists of four sections. Section 2 presents the standard keyword count-based method and its limitations, and the logic of the proposed alternative. Section 3 presents the data used to evaluate the method. Section 4 presents results. Section 5 concludes.

2 Method

In order to focus the comparison, the examples and empirical evaluation in this paper revolve around the issue of immigration. Because native citizens' level of direct personal contact with immigrants is typically limited, studies of the politics of immigration typically assign a critical role to mass media in shaping attitudes about immigration. However, the lessons of this paper are likely to be transferable to other policy issues.

2.1 The keyword count method and its limitations

Observational studies examining the role of mass media coverage of immigration in shaping immigration attitudes typically employ a measure I will refer to here as the *keyword count method*. In short, the method runs as follows: start with a survey in which respondents are asked about immigration-related attitudes at several different points in time t. Then, define a search string consisting of a number of immigration-related keywords. For each time period t, count how many hits the search string returns in a select number of news outlets searchable in a major database. The number of hits, i.e. the keyword count, serves as a measure of immigration salience. In time periods when the search string returns a large number of hits, the measure thus indicates that immigration salience is high. In time periods when the search string returns few hits, the measure indicates that immigration salience is low.

The keyword count method is used in several prominent studies of effects of media coverage on immigration attitudes. Select examples include a study of priming among Danish citizens (Togeby, 2007), immigration media coverage in the Netherlands (Boomgaarden and Vliegenthart, 2007), and Germany (Boomgaarden and Vliegenthart, 2009), contrasting immigration coverage and real immigration flows (van Klingeren et al., 2014), voting for anti-immigrant parties (Burscher et al., 2015), and studies linking media effects with effects of ethnic diversity on local contexts in the United States (Hopkins, 2010) and the United Kingdom (Hopkins, 2011). In other words, the keyword count method can reasonably be considered a standard approach to measuring the media salience of immigration.

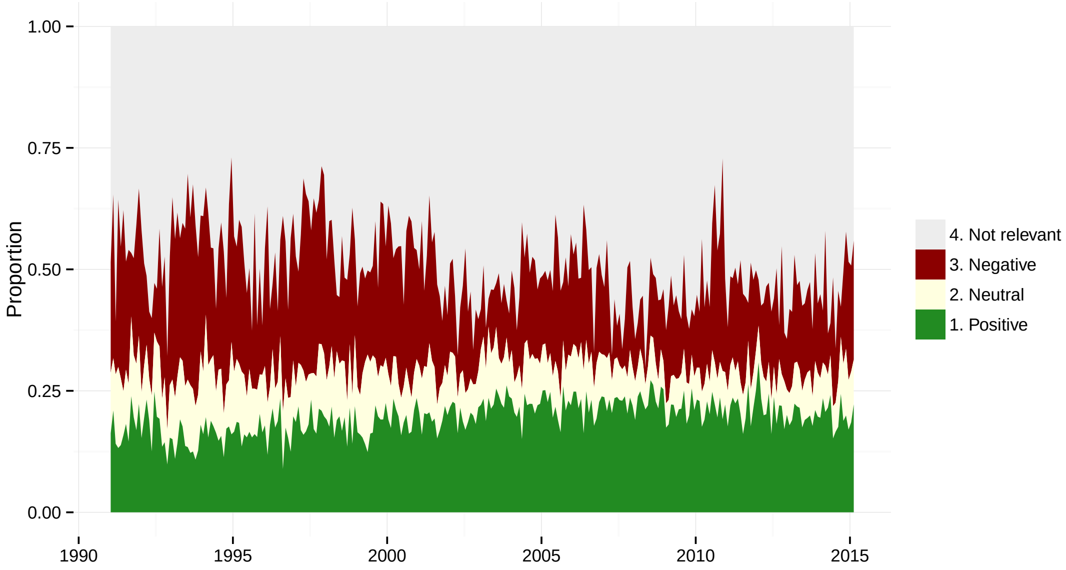

Figure 1: Estimated proportions of media coverage of immigration across categories. The estimates are based on an automated text analysis of ∼69,000 Danish newspaper articles. Full analysis presented in Hjorth (2016), Chapter 6.

For an illustration of the potential limitations of the keyword count method, consider Figure 1, drawn from Hjorth (2016). Based on the machine learning method *ReadMe* (Hopkins et al., 2010), the figure shows the estimated proportion of articles categorized either by tone (positive/neutral/negative) or to be unrelated to immigration. The articles analyzed ($N \approx 69,000$) are identified using the keyword count method. The key notable feature of Figure 1 in the context of this paper is that, although the articles are all drawn from a search using immigration-related keywords, *around half* of the articles are estimated to be unrelated to immigration, marked as light gray in Figure 1.

The proportion indicated in Figure 1 may of course be overstated. For example, the share of articles deemed to be related to immigration inevitably depends on how narrowly immigration as a political issue is conceptualized. Nevertheless, the results shown in Figure 1 do strongly suggest that a substantial share of articles returned in keyword searches are not pertinent to the research question of interest.

The inclusion of non-relevant coverage in search results used in the keyword count method is consequential even if the variation in that coverage is random. Since immigration media coverage typically operates as the independent variable, random measurement error will result in attenuation bias. The inclusion of non-relevant coverage will thus result in underestimation of the effects of immigration media coverage. Conversely, measuring relevant coverage more precisely should translate into estimates less downwardly biased.

2.2 Using topic models to parse out relevant media coverage

I propose using topic models to estimate the proportion of results returned in a keyword search are relevant to the issue of interest, using these estimates to more precisely measure issue salience. Very briefly summarized, topic models are a type of clustering method for textual data, using information about word co-occurrences to identify semantic clusters (or 'topics') that vary in prevalence across documents.

The importance of non-relevant topics depends on whether their share of the observed total varies over time. Hence, understanding their role requires a method that allows for estimating how topic prevalence varies over time. I turn to the structural topic model (STM) (Roberts et al., 2014), which allows for estimating how document-level covariates (such as publication date) are associated with topic prevalence. Specifically, I propose parsing out the relevant variation in media coverage using the following procedure:

1. As in the keyword count method, search for articles based on keyword search string.
2. Download the full text of all returned articles.
3. Estimate the topic distribution across articles using an STM with time as covariate.
4. Identify subset of 'relevant' topics.
5. For each time period, sum across estimated proportions of relevant topics.

6. Assign estimated proportions as weights to original keyword count estimates.

For convenience, I will refer to this method as *the topic-weighted method.* The topic-weighted method 'piggy-backs' on the keyword count method by using standard keyword count estimates of issue salience, but using the STM to weight those estimates up when the share of relevant content is higher and down when the share of relevant content is lower.

Formally, let K be the (researcher-defined) number of topics in the topic model. Based on interpretation of the estimated topics, the researcher defines a number of issue-relevant topics $R < K$. Since time is included as a covariate in the STM, we can estimate the prevalence p of topic i at time t as p_{it}. Finally, let X_t denote the number of articles found using the keyword count method. Then, for each time t, the topic-weighted method estimates issue salience as

$$X_t \times \sum_{i=1}^{R} p_{it} \tag{1}$$

In the next section, I introduce the data I use to evaluate the method.

3 Data and modeling strategy

To evaluate the topic-weighted method, I combine two types of data: data on *media coverage* (to produce measures of the salience of immigration) and data on *public opinion* (to evaluate how different measures correlate with political attitudes).

For the *public opinion data*, I utilize data from a set of surveys conducted in Denmark from 1999 to 2003 (see also Togeby, 2004, 2007). In each of the four years, a computer-assisted telephone interview (CATI) was conducted on a nationally representative sample of about 800 individuals. The survey had a panel component, such that some respondents are interviewed in two waves and some in three waves. All in all, the survey has N=10,332 responses from 5,146 individuals. I do not exploit the panel structure of the data in this analysis, though I do take account of the structure by clustering standard errors by respondent ID. Crucially for the purpose of this analysis, the surveys include repeated questions about immigration attitudes, the importance of immigration as a political issue, and government evaluation.

For the data on *media coverage*, I follow the sampling strategy used in Togeby (2007), who uses the keyword count method in a study of priming of political attitudes. For each month in the survey and the preceding month, I search the Danish newspaper database *Infomedia* for mentions of the immigration-related keywords *indvandr** (immigration or immigrants) or *flygtn** (refugees). The search covers one right-leaning broadsheet newspaper (*Jyllands-Posten*), one left-leaning broadsheet newspaper (*Politiken*) and one major news bureau (*Ritzaus Bureau*). All in all, the search returns 15,771 articles. I download the full text of all articles and apply standard preprocessing steps (removing numbers, special characters, and stopwords, and stemming). To speed up estimation, I remove very rare words, defined as words mentioned in less than .2 pct. of the documents. The resulting term-document matrix has information about 2,747,720 mentions of 9,036 words.

I evaluate the topic-weighted method by comparing how strongly the keyword count method measure of immigration salience vis-à-vis the topic-weighted measure correlate with political attitudes in three standard models of media effects. I describe these three in the remainder of the section.

3.1 Comparison 1: Agenda-setting

The first model reflects what is in some sense the most basic way in which the media salience of immigration should affect attitudes: by making the issue more salient in the minds of citizens. This effect is typically referred to as *agenda-setting* (McCombs and Shaw, 1972; McCombs, 2004).

In order to measure agenda-setting effects, I utilize the fact that respondents in each wave are asked to name what is currently the most important political problem. Respondents often list immigration as the most important problem. In fact, across all responses, immigration is listed 43 pct. of the time. Hence, agenda-setting effect would imply that citizens are more likely to list immigration as the most important issue when the media salience of immigration is higher. I estimate models of the following form:

$$ImmImp_{it} = \beta_1 + \beta_2 ImmSal_t + \mathbf{X_{it}}\gamma + \mathbf{X_i}\delta + \varepsilon_{it} \tag{2}$$

Where $ImmImp_{it}$ is a binary indicator of whether respondent i lists immigration at time t and $ImmSal_t$ is the measure of the salience of immigration at time t. $\mathbf{X_{it}}$ represents time-varying controls (age) and $\mathbf{X_i}$ represents time-invariant controls (gender and education level). The coefficient of interest is β_2, which I expect to be positive. Though the dependent variable is dichotomous, I use estimate the relationship using OLS in the interest of simplicity.

3.2 Comparison 2: Attitude change

The second model captures whether citizens' attitudes toward immigration change when the media salience of immigration changes. Here, I refer to this effect with the generic label *attitude* *change*. This effect is often the focus of studies of the consequences of media coverage of immigration (e.g., Boomgaarden and Vliegenthart, 2009; van Klingeren et al., 2014). Due to the 'thermostatic' nature of public opinion, the effect is typically expected (and found) to be negative, i.e. public opinion turns more restrictive when the issue of immigration is more prominent Erikson et al. (2002).

To measure attitudes toward immigration, I construct an index based on three items: respondent's answers to whether *immigrants should be free to practice their religion*, whether *Denmark should accept more or fewer refugees*, and whether *immigrants should gain the right to vote sooner or later compared to today*. The index has acceptable reliability (Cronbach's $\alpha = .66$). I estimate models of the following form:

$$ImmAtt_{it} = \beta_1 + \beta_2 ImmSal_t + X_{it}\gamma + X_t\delta + \varepsilon_{it} \tag{3}$$

Where $ImmAtt_{it}$ is respondent i's score on the index of immigration attitudes at time t and the remaining model terms are as described above. Note that the immigration attitudes index is coded such that higher values correspond to more negative attitudes. Hence, I expect β_2 to be positive.

3.3 Comparison 3: Priming

Lastly, I estimate models of *priming*, whereby higher immigration media salience causes voters to rely more on their immigration attitudes when evaluating government performance. Priming is the focus of Togeby (2007), the study for which the data was originally produced. The key observable implication of priming is that in times of higher immigration media salience, immigration attitudes should be more strongly associated with evaluations of government performance. Hence, the dependent variable is a four-point scale of how well the respondent thinks the government is doing its job. I estimate models of the following form:

$$GovtPerf_{it} = \beta_1 + \beta_2 ImmSal_t + \beta_3 ImmAtt_{it} + \beta_4 ImmSal_t \times ImmAtt_{it} + X_{it}\gamma + X_t\delta + \varepsilon_{it} \tag{4}$$

Where $GovtPerf_{it}$ is respondent i's evaluation of the government's performance at time t. The remaining model terms are as described above. Since priming implies changes in the relationship between two variables, model 4 is an interaction model.

I expect respondents with strong anti-immigration attitudes to evaluate the government more negatively when the issue of immigration is primed. Hence, I expect β_4 to be negative. Lastly, note that the label of priming should be used with some caution. The correlation between issue attitudes and evaluations of government performance can change for other reasons than priming (Lenz, 2012). On the other hand, group-level predispositions such as immigration attitudes may be particularly easy to prime Tesler (2014).

In the next section, I present the results from estimating these models on the data described above. Before doing so, I present the result from the topic model fitted to the full set of newspaper articles.

4 Results

4.1 Topic model results

When setting up a topic model, a crucial decision on the part of the researcher is setting K, the number of topics to fit to the data. Here, I present results for $K = 50$, a number chosen in part based on the advice given in Roberts et al. (2014), in part based on the desired granularity of the analysis. Table 1 presents each of the 50 topics. For each topic, Table 1 shows a label (manually assigned) as well as the seven terms with the highest FREX (frequency-exclusivity) score, a measure combining how typical a phrase is for a topic and how exclusive it is relative to other topics (Bischof and Airoldi, 2012).

Since Table 1 presents the terms in their original Danish, some examples are in order. Some topics are very clearly relevant to the domestic immigration debate. For example, topic 14, *Multiethnic society*, is associated with the terms *foreign, society, discrimination, multiethnic, culture, xenofobia*, and *debate*. Similarly, topic 32, *Crime*, is associated with the terms *arrest, perpetrator, jail, knife, fire, copenhagen*, and *police*. Similarly, topic 40, *Christianity and Islam*, is associated with the terms *muslim, islam, christianity, muslim, christian, religion*, and *imam*.

At the same time, the topic model also identifies topics that do not seem highly pertinent to the debate over immigration as a domestic political issue. For example, topic 6, *Mediterranean*, is associated with the terms *spain, immigrant, spanish, italian, illegal, italian*, and *morocco*, and appears to capture coverage of illegal immigration across the Mediterranean. Similarly, topic 49, *Development aid*, is associated with the terms *million, kroner, money, nielson, amount, aid, fund*, and appears to capture coverage of third world aid.

In sum, while the topic model does seem to reveal semantically coherent clusters in the corpus of articles, it also indicates that several of those topics are in a sense ‘false positives’, captured by the search string but not directly relevant to immigration as a domestic political issue.

Table 1: Summary of extracted topics

	Label	Highest FREX score terms
1	Los Angeles	ritzau, angel, los, sort, hvid, rac, farm
2	Refugee camps	asylansøg, asylsøg, udlændingestyr, asyl, afslag, asylcentr, sandholmlejr
3	Football players	jen, jens, fodbold, ole, sørens, aarhus, mads
4	Immigration statistics	pct, procent, statistik, undersøg, antal, stig, steg
5	Israel/Palestine	israelsk, palæstinensisk, arafat, vestbred, israel, gaza, palæstinens
6	Mediterranean	spani, immigrant, spansk, itali, illegal, italiensk, marokko
7	Political leaders	jespers, kar, radikal, khad, jelved, marian, nas
8	Labor market integration	ansæt, arbejdskraft, arbejdsgiv, job, arbejdsplads, virksom, arbejdsmarked
9	Municipal politics	kommun, kommunal, borgmest, byråd, farum, brixtoft, almennyt
10	Globalization	demokrati, globalisering, elit, forandring, værdi, fællesskab, pittelkow
11	Prime ministers	nyrup, rasmus, fogh, statsminist, lykketoft, poul, bjerregaard
12	East Timor	østtimor, australsk, indonesisk, australi, tampa, dili, indonesi
13	Holocaust	jød, nazist, verdenskr, jødisk, historik, holocaust, hitl
14	Multiethnic society	fremmed, samfund, diskrimination, multietnisk, kultur, fremmedhad, debat
15	Film	instr, film, skuespil, instruktør, teat, sko, publikum
16	EU politics	forbehold, eu-land, euro, medlemsland, eus, eu's, topmød
17	War On Terror	bin, taleban, osama, terrorangreb, afghanistan, bush, usa
18	Verbs	havd, gik, fik, ham, stod, måt, skul
19	Kosovo War	milosevic, nato, beograd, slobodan, serbi, jugoslavi, jugoslavisk
20	Word jumble 1	tænk, bar, tror, måsk, ting, selvfølg, vel
21	Sending refugees home	flygtning, modtag, bosni, nærområd, midlertid, hjem, vend
22	Sweden	svensk, sver, holg, socialdemokrat, persson, socialdemokrati, venstrefløj
23	Development aid	mio, kron, peng, nielson, beløb, nødhjælp, fond
24	France	chirac, pen, fransk, frankr, franskmænd, jospin, jacqu
25	Danish People's Party	kjærsgaard, pia, glistrup, folkepartis, folkeparti, enhedslist, fremskridtsparti
26	Family reunions	haard, bertel, familiesammenføring, konvention, lovforslag, integrationsminist, stramning
27	Exhibits	udstilling, museum, muse, dronning, kunstn, www, bibliotek
28	Immigrant communities	minoritet, indsam, nydansk, etnisk, forening, tvangsægteskab, indvandrerorganisation
29	Human rights violations	kina, kinesisk, krigsforbryd, rwanda, amnesty, massakr, beijing
30	Macedonia	makedoni, albani, makedonsk, albansk, tirana, kuk, alban
31	Netherlands	fortuyn, haid, pim, østr, frihedsparti, jörg, hollandsk
32	Crime	anhold, gerningsmænd, varetægtsfængsl, kniv, brand, københavn, politi
33	Public expenditures	skat, kontanthjælp, udgift, sektor, afgift, efterløn, ældr
34	Books	bog, forfat, roman, bøg, forlag, udkom, udgiv
35	Word jumble 2	smil, spis, grin, drik, klok, kaf, tænd
36	Families	famili, mor, pig, gift, dat, barn, kvind
37	Ghettos	vollsmos, betjent, nørrebro, vold, ung, overfald, ballad
38	Refugee crime	dom, advokat, dømt, retssag, fængsel, sag, flygtningenævn
39	Germany	schröder, fisch, berlin, gerhard, stoib, cdu, tyskland
40	Christianity and Islam	muslim, islam, kristendom, muslimsk, kristn, religion, imam
41	War-torn areas	hus, landsby, brænd, bjerg, bebo, ruin, met
42	Media	blad, artikl, medi, indlæg, annonc, artikel, avis
43	Health care	patient, sygdom, psykisk, sygeplejersk, syg, sygehus, læg
44	Chechnya	tjetjeni, tjetjensk, grosnij, putin, russisk, moskva, rusland
45	United Nations	udenrigsminist, annan, kofi, sikkerhedsråd, blair, torsdag, onsdag
46	School integration	elev, tosproged, undervisning, modersmålsundervisning, skol, folkeskol, lær
47	Iraq	irakisk, irak, saddam, tyrki, hussein, bagdad, kurdisk
48	Charity	kor, flygtningehjælp, rød, rand, frivil, indsaml, chemnitz
49	Africa	afrika, fattigdom, global, kroati, region, fat, økonomisk
50	Ethiopia	etiopi, eritrea, angola, sudan, etiopisk, landmin, uganda

In Table 1, I have highlighted the topics I have assigned as 'relevant'. As described above, the topic-weighted measure of immigration salience is then constructed by summing across the estimated prevalences of relevant topics for each time period and weighting the keyword count estimate by this sum. Figure 2 shows how the two measures compare.

As Figure 2 shows, although the two measures covary closely, the topic-weighted measure

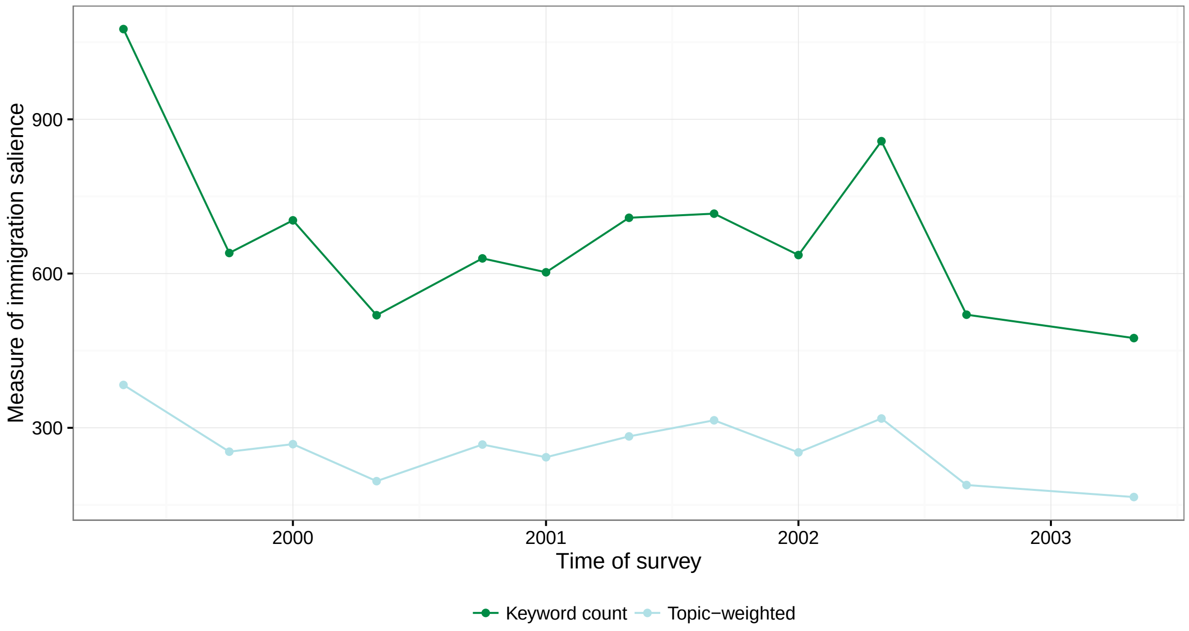

Figure 2: Measures of immigration salience for the keyword count method (dark green) and the topic-weighted method (light blue) for each survey wave.

assigns a smaller value to the survey waves in May 1999 and May 2002, since the estimated share of relevant topics at these times is relatively lower.

4.2 Comparison 1: Agenda-setting

Table 2 presents the results of estimating model (2) using keyword counts weighted by topic prevalence (column 1) and the standard keyword count method (column 2). In column 1, the coefficient for 'Rel. topic weights' represents the topic-weighted measure. The coefficient for 'Other topic weights' represents the same measure for the remaining, non-relevant topics.

To ensure that the coefficients have similar interpretations, the media salience measured have all been rescaled to range from 0 to 1. In all models, standard errors are clustered by respondent and survey wave.

As Table 2 shows, the topic-weighted estimate and the keyword count estimate have the same direction, but sharply different magnitudes. Whereas the keyword count estimate is statistically indistinguishable from zero, the topic-weighted estimate is nearly two orders of magnitude larger and statistically significant at the .05 level.

As a means of illustrating the differences in the estimated effects, consider Figure 3. The figure provides a simulation-based illustration of the coefficients of interest in Table 2. I show the

Table 2: Results for agenda-setting

	Imm. most important problem	
	(1)	(2)
Gender (m)	0.048***	0.046***
	(0.013)	(0.013)
Age	−0.003***	−0.003***
	(0.001)	(0.001)
Education level	0.002	0.0004
	(0.006)	(0.006)
Other topic weights	−0.907*	
	(0.474)	
Rel. topic weights	0.886**	
	(0.436)	
Counts		0.010
		(0.262)
Constant	0.409***	0.520***
	(0.090)	(0.085)
N	9,872	9,872
R^2	0.028	0.008
Adjusted R^2	0.028	0.007
Residual Std. Error	0.488 (df = 9866)	0.493 (df = 9867)
F Statistic	57.539*** (df = 5; 9866)	18.717*** (df = 4; 9867)

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

distribution of coefficients by plotting 100,000 draws from the multivariate normal distribution defined by the models' estimates and variance-covariance matrices, as described in King et al. (2000). The leftmost panel shows the distribution of the coefficient using the count-based method, closely centered around zero. The middle panel shows the distribution of the coefficient using the topic-weighted method, significantly above zero. The rightmost panel in Figure 3 shows the distribution of the difference between the two effects. The estimate using the topic-weighted method is thus clearly larger than the count-based estimate and in the hypothesized direction.

4.3 Comparison 2: Attitude change

Table 3 shows the results for the attitude change model, estimating model (3). As above, Figure 4, shows the distribution of the coefficients for the keyword count method, the topic-weighted method, and the difference between the two.

In this case, the results are slightly more ambiguous than for the agenda-setting model. The coefficient for the count-based method is negative, i.e. the opposite of the hypothesized direction. In contrast, the coefficient for the topic-weighted method is positive, but not statistically

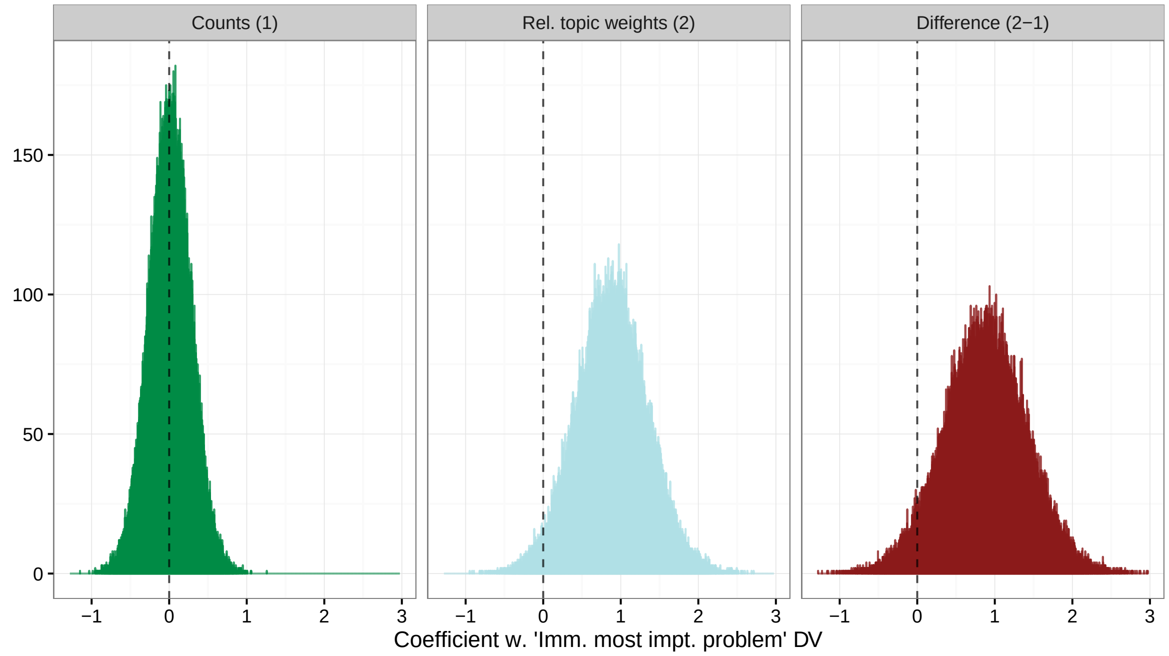

Figure 3: Estimated effects of immigration media salience on perceptions of immigration as the most important problem (agenda-setting) using the keyword count method (leftmost panel, dark green) and the topic model weighted method (middle panel, light blue) and the difference between the two coefficients (rightmost panel, dark red).

significant. The difference between the coefficients, shown in the rightmost panel of Figure 4, is thus in the hypothesized direction and significant at the .1 level.

4.4 Comparison 3: Priming

Table 4 shows the results for the attitude change model, estimating model (4). As above, Figure 5, shows the distribution of the coefficients for the keyword count method, the topic-weighted method, and the difference between the two.

Here, the coefficient of interest is the coefficient on the interaction term between the immigration salience measure and the measure of anti-immigration attitudes. The interaction for the count method is negative, i.e. in the hypothesized direction, but not statistically significant. The interaction term for the topic-weighted measure is around four times larger and statistically significant at the .1 level. As a consequence, the difference between the two coefficients, shown in the rightmost panel of Figure 5, is thus also in the hypothesized direction, though not statistically significant.

Table 3: Results for attitude change

	Anti-immigration attitude index	
	(1)	(2)
Gender (m)	-0.014^{**}	-0.015^{**}
	(0.006)	(0.006)
Age	0.003^{***}	0.003^{***}
	(0.0002)	(0.0002)
Education level	-0.048^{***}	-0.049^{***}
	(0.002)	(0.002)
Other topic weights	-0.112^{*}	
	(0.065)	
Rel. topic weights	0.063	
	(0.055)	
Counts		-0.048^{*}
		(0.029)
Constant	0.623^{***}	0.632^{***}
	(0.019)	(0.020)
N	10,295	10,295
R^2	0.123	0.122
Adjusted R^2	0.122	0.122
Residual Std. Error	0.243 (df = 10289)	0.243 (df = 10290)
F Statistic	287.311^{***} (df = 5; 10289)	357.833^{***} (df = 4; 10290)

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

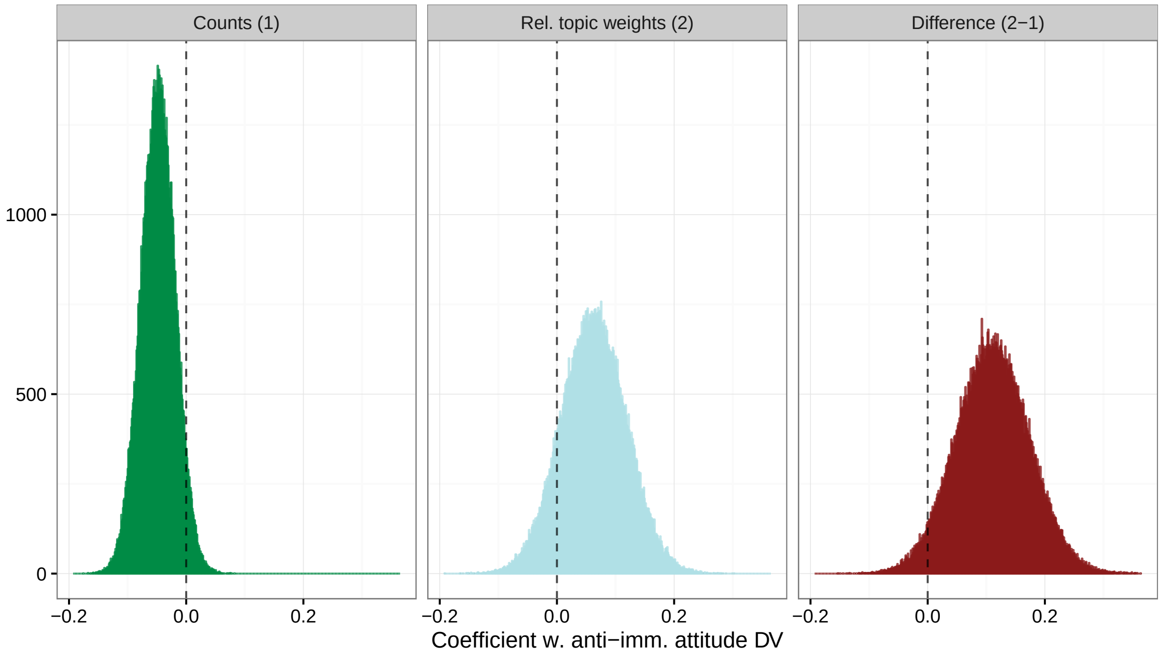

Figure 4: Estimated effects of immigration media salience on anti-immigration attitudes (attitude change) using the keyword count method (leftmost panel, dark green) and the topic model weighted method (middle panel, light blue) and the difference between the two coefficients (rightmost panel, dark red).

Table 4: Results for priming

	Evaluation of govt. performance	
	(1)	(2)
Gender (m)	0.001	0.003
	(0.009)	(0.009)
Age	0.0001	0.0001
	(0.0003)	(0.0003)
Education level	0.003	0.004
	(0.003)	(0.003)
Other topic weights	0.269	
	(0.727)	
Counts		-0.269
		(0.246)
Anti-imm.	0.041	-0.031
	(0.085)	(0.088)
Rel. topic weights	-0.574	
	(0.913)	
Other topic weights × Anti-imm.	0.785	
	(0.577)	
Rel. topic weights × Anti-imm.	-0.829^{*}	
	(0.429)	
Counts × Anti-imm.		-0.171
		(0.264)
Constant	0.728^{***}	0.643^{***}
	(0.223)	(0.115)
N	10,044	10,044
R^2	0.160	0.088
Adjusted R^2	0.159	0.088
Residual Std. Error	0.275 (df = 10035)	0.287 (df = 10037)
F Statistic	239.142^{***} (df = 8; 10035)	161.609^{***} (df = 6; 10037)

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

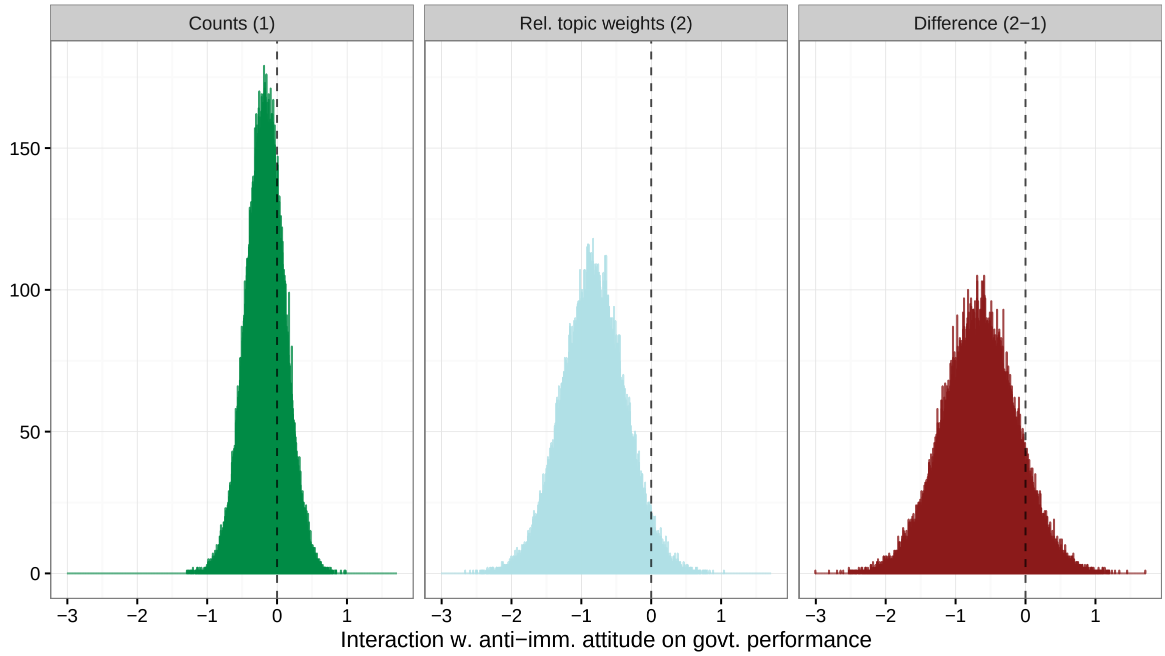

Figure 5: Estimated interactions of immigration media salience and anti-immigration attitudes on evaluations of government performance (priming) using the keyword count method (leftmost panel, dark green) and the topic model weighted method (middle panel, light blue) and the difference between the two coefficients (rightmost panel, dark red).

5 Conclusion

In this paper, I set out to evaluate whether using topic models to more accurately measure the content of media coverage can improve estimates of issue salience used in studies of media effects. In short, the proposed method, here labelled the topic-weighted method, uses a topic model to estimate the share of issue-relevant content in keyword count estimates and then weights the keyword count estimate by that share. I evaluated the method using data from a 12-wave running survey of immigration attitudes combined with topic data from 15,771 newspaper articles.

At this early stage, the results provide grounds for a very cautiously positive assessment. Across three models examined, the topic-weighted estimate differed from the keyword count estimate in the hypothesized direction, suggesting that the topic-weighted estimate does indeed reduce measurement error. However, the differences between the two coefficients are not in all cases statistically significant. Since the two are by design strongly correlated, considerable statistical power is needed to statistically distinguish between them. In this context, the 12-wave survey used in this evaluation is likely not sufficient. Hence, increasing statistical power should be a key focus in additional efforts to evaluate the topic-weighted method.

This paper has presented merely one possible way to combine the emerging field of text-as-data with traditional studies of the role of media in public opinion. It remains to be established whether the additional effort required to retrieve and process textual information is justified by added measurement precision. It may turn out not to be the case, or alternatively that the most promising method is a different one altogether. In any case, given the sophisticated tools available in the text-as-data approach, it is at the very least worthwhile to reassess received wisdom about the relationship between media coverage and public opinion.

References

Bischof, J., Airoldi, E.M., 2012. Summarizing topical content with word frequency and exclusivity, in: Proceedings of the 29th International Conference on Machine Learning.

Boomgaarden, H.G., Vliegenthart, R., 2007. Explaining the rise of anti-immigrant parties: The role of news media content. Electoral Studies 26, 404–417.

Boomgaarden, H.G., Vliegenthart, R., 2009. How news content influences anti-immigration attitudes: Germany, 1993-2005. European Journal of Political Research 48, 516–542.

Burscher, B., van Spanje, J., de Vreese, C.H., 2015. Owning the issues of crime and immigration: The relation between immigration and crime news and anti-immigrant voting in 11 countries. Electoral Studies 38, 59–69.

Erikson, R.S., Mackuen, M.B., Stimson, J.A., 2002. The Macro Polity. Cambridge University Press.

Hjorth, F., 2016. Ethnicization in Welfare State Politics. Ph.d. dissertation. University of Copenhagen.

Hopkins, D., King, G., Knowles, M., Melendez, S., 2010. ReadMe: Software for automated content analysis. Institute for Quantitative Social Science .

Hopkins, D.J., 2010. Politicized Places: Explaining Where and When Immigrants Provoke Local Opposition. American Political Science Review 104, 40.

Hopkins, D.J., 2011. National Debates, Local Responses: The Origins of Local Concern about Immigration in Britain and the United States. British Journal of Political Science 41, 499–524.

King, G., Tomz, M., Wittenberg, J., 2000. Making the Most of Statistical Analyses: Improving Interpretation and Presentation. American Journal of Political Science 44, 341–355.

van Klingeren, M., Boomgaarden, H.G., Vliegenthart, R., de Vreese, C.H., 2014. Real World is Not Enough: The Media as an Additional Source of Negative Attitudes Toward Immigration, Comparing Denmark and the Netherlands. European Sociological Review 31, 268–283.

Lenz, G.S., 2012. Follow the Leader?: How Voters Respond to Politicians' Policies and Performance. University of Chicago Press, Chicago.

McCombs, M., 2004. Setting the Agenda: The Mass Media and Public Opinion. John Wiley & Sons, Inc.

McCombs, M., Shaw, D., 1972. The agenda-setting function of mass media. Public Opinion Quarterly 36, 176–187.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G., 2014. Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science 58, 1062–1084.

Tesler, M., 2014. Priming Predispositions and Changing Policy Positions: An Account of When Mass Opinion Is Primed or Changed. American Journal of Political Science , n/a–n/a.

Togeby, L., 2004. Man har et standpunkt... - om stabilitet og forandring i befolkningens holdninger. Magtudredningen, Aarhus University Press.

Togeby, L., 2007. The Context of Priming. Scandinavian Political Studies 30, 345–376.