Topics and their Salience in the 2015 Parliamentary Election in Croatia: A Topic Model based Analysis of the Media Agenda

Damir Korenčić[1] **Marijana Grbeša-Zenzerović**[2] **Jan Šnajder**[3]
[1]Deparment Electronics, Ruđer Bošković Institute, Croatia
[2]Faculty of Political Science, University of Zagreb, Croatia
[3]Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
damir.korencic@irb.hr grbesa@fpzg.hr jan.snajder@fer.hr

Abstract

There is a growing interest in automated content analysis for agenda-setting studies. While topic models have shown to be useful for this purpose, they are generally troubled with low topic quality and coverage. To alleviate this, Korenčić et al. (2015) proposed a semi-supervised topic modeling methodology. The aim of this work is to gain a better understanding of their methodology, by conducting a preliminary study of the media agenda during the 2015 parliamentary election in Croatia. Our goal is to analyze the topics and their salience during the official election campaign and the lively post-election negotiation period. We report on the methodological insights gained from this study and a preliminary analysis of the media agenda during the election period.

1 Introduction

Agenda-setting has been one of the most influential media effects theories for decades. Its underlying idea is that the media have the capacity to shape the public's perception of importance of particular issues (McCombs and Shaw, 1972; Scheufele, 2000). This effect is highly relevant during an election period, as it is widely acknowledged that the salience of media issues may influence voters' choices.

Agenda-setting studies often rely on quantitative content analyses of newspaper texts. In such studies, the media agenda is *measured* in terms of the salience of the issues: the newspaper documents are coded for issues, and the salience of each issue is taken to correspond to its frequency across the corpus. Recently, there has been a growing interest in the use of automated content analysis leveraging natural language processing techniques; in particular, *topic models* (Blei et al., 2003) have gained wide popularity. Existing studies from the domains of computer and political science demonstrate the usefulness of topic models for agenda analysis (Grimmer, 2010; Quinn et al., 2010; Kim et al., 2014). However, they also identify the problems related to topical coverage and quality (Chuang et al., 2013; Chuang et al., 2015), which may seriously hamper the validity of an agenda-setting study.

Recently, Korenčić et al. (2015) proposed a methodology based on topic modeling that mitigates the above deficiencies by using a semi-supervised, human-in-the-loop acquisition of topics, aiming for high-quality topics that better correspond to media issues. In a nutshell, the methodology consists of two steps: an agenda discovery step, in which topics are induced automatically and revised manually, and an agenda measuring step, in which the articles are tagged with topics. Korenčić et. al demonstrate that the approach can outperform a supervised classifier, while it additionally facilitates the discovery of topics. However, there is a number of non-trivial design choices associated with using their methodology, including technical (e.g., model parameters) and conceptual (e.g., the granularity and choice of topics) ones.

The aim of this paper is to put to practice the methodology of Korenčić et al., and gain an understanding of its advantages and potential caveats. To this end, we conduct a preliminary study on data collected during one of the most interesting periods in modern Croatian political history: the 2015 parliamentary election and the lively post-election period of negotiations on government formation. This paper makes two contributions: we report on (1) the methodological insights gained by applying this methodology and (2) a preliminary analysis of the media agenda during Croatian 2015 parliamentary election.

2 Corpus

We collected the data for our study from seven leading Croatian news sites: *Večernji list*, *Jutarnji* *list*, *Slobodna Dalmacija*, *Glas Slavonije*, *T-portal*, *Novi List*, and *RTL Televizija*. We first selected the news feeds that correspond to domestic and regional news, and then collected the articles published during the official election campaign (from October 21st to November 6th, 2015), as well as in the period between the election day and the constitution of the Parliament (from November 8th to December 28th, 2015). We next removed very short texts (those with less than 40 alphanumeric tokens) and non-texts (error messages, subscription previews, photo galleries, etc.). Finally, we performed deduplication by using word-level edit distance to form groups of almost identical texts and keeping from each group only one text per news outlet. After filtering and deduplication, the final corpus consists of 15,394 news articles.

3 Agenda Discovery

The first step in the methodology of Korenčić et al. (2015) is *agenda discovery*. This is an exploratory step, whose purpose is to chart a wide range of *topics* present in the media. We use the term "topic" instead of "issue" to avoid misunderstandings that may stem from a narrow understanding of the term "issue", which is commonly employed in agenda-setting studies to denote policy issues such as health, defense, economy etc., as opposed to less substantive campaign contents; cf. (de Vreese and Semetko, 2004; Zeh and Hopmann, 2013). In contrast, the term "topic" refers here to a broader range of different contents, varying from "issues" to more vague contents (such as intra-party conflicts and similar). Furthermore, we use the term *semantic topic*[1] to refer to topics as perceived by humans, including issues, processes, events, and entities. A semantic topic stands in contrast to a topic induced automatically by a topic model, to which we will refer as a *model topic*.

Ideally, model topics will correspond to semantic topics; in reality, however, model topics can contain noise or correspond to more than one semantic topic (Chuang et al., 2013). The objective of the agenda discovery step, then, is to detect the semantic topics and map them to model topics. To this end, we rely on human inspection of topics obtained by using several different models, each run on the same data. Namely, studies have shown that topics of a single model may not cover all semantic topics (Chuang et al., 2015). By analyzing the topics several models, we can compensate for the incomplete coverage of the individual models.

3.1 Topic models

To discover the semantic topics, we use the LDA topic model (Blei et al., 2003), available as part of the Gensim package (Řehůřek and Sojka, 2010). Text preprocessing consists of stop-word and nonword removal, and stemming using a Croatian stemmer of Ljubešić et al. (2007). Models are trained using a fast online learning algorithm (Hoffman et al., 2010). We set model hyperparameter $\alpha = 50/T$ (where T is the chosen number of topics), while we set $\beta = 0.01$ (Griffiths and Steyvers, 2004). Model learning parameters are set to $S = 1000, \tau_0 = 1.0, \kappa = 0.5$, as proposed by Korenčić et al. (2015). As input for the annotation process, we constructed three LDA models: two models with $T = 50$ topics (using different random seeds) and one model with $T = 100$ topics.

3.2 Semantic topic discovery

After obtaining the 200 topics from the three models, we presented the topics to seven human annotators: two authors and five master students of journalism. The annotators were instructed to perform a three-step annotation as follows. First, they were asked to deduce the meaning of the model topic by inspecting the list of words with high probability within the topic, and the list of news articles ranked by proportions of the topic within the article. After the first step, a number of semantic topics relating to the model topic were detected. In the second step, annotators consulted a shared list of extracted semantic topics to check whether the topics they detected have not already been detected by other annotators, and, if this was not the case, to add the topics to the list. Finally, the annotators used tags to link the model topics with the semantic topics, and vice versa, and also provided a short textual description for each model topic.

The actual annotation round was preceded by a training session, in which the annotation procedure was explained and demonstrated, followed by a test round and a discussion. Model topic inspection was performed with a GUI application deployed on a server and accessed via remote desktop clients. The annotators used the application to browse the topics and inspect the lists of words and news articles. Each annotator processed about 30 topics. The assignment balanced the topics across the three

[1] Korenčić et al. (2015) used the word "theme" for the same concept.

models. On average, an annotator spent 10 minutes on a single topic (min. 5.5 and max. 16.8). The total annotation effort was 33 person-hours.

3.3 Semantic topics revision

The procedure outlined above yielded 106 semantic topics. However, a closer inspection revealed errors in the annotations: some semantic topics were repeated, some were named ambiguously, while in some cases the link between the semantic topic and the underlying model topics was questionable. We speculate that the annotation quality could be ameliorated by investing more time in annotator training and by enforcing a more strict annotation procedure. We also observed that some topics, such as *weather reports* and *traffic disruptions*, while annotated correctly, are ultimately irrelevant for agenda analysis. For these reasons, we decided to carry out one additional revision round.

Another, less surprising finding was that the obtained topics are not mutually exclusive – rather, the topics are of different levels of abstractness and constitute a hierarchy. While inspecting the discovered topics and relations among them, we found it convenient to manually organize the topics into a taxonomy.[2] For instance, we put the semantic topic *election polls* under *election forecasts*, which, together with *election results*, we put under *electoral process*. We found that such a taxonomy was very useful for identifying and scoping the topics of interest. More concretely, we could use the taxonomy to chose a suitable level of topic granularity.

Topics were revised and organized in a taxonomy jointly by all three authors, which took about three hours. After the second round, a list of 71 semantic topics remained, organized in a taxonomy with the following 21 top-level categories: *prosecutions of public figures*, *post-election negotiations*, *foreign policy*, *terrorism and refugee crisis*, *Catholic Church*, *institution of the president*, *armed forces / Croatian army*, *electoral process*, *ecology*, *energetics*, *education*, *tourism*, *decentralization and reform of local and regional government*, *health care*, *media and journalists*, *trade unions and workers' rights*, *economy*, *intra-party conflicts*, *agriculture*, *brain drain and demography*, and *independent events*. The last category mostly pertains to specific events that occurred during the election campaign, but which do not fit well in any other category.

4 Agenda Measuring

The detected semantic topics provide the analyst with a general overview of the media agenda. The next step is to measure the salience of the detected topics. For this preliminary study, we decided to focus on topics from two top-level categories: *electoral process* and *post-election negotiations*.

4.1 Defining custom semantic topics

We began our analysis with the inspection of the semantic topics in the two selected categories, using the same method as for the agenda discovery step. The inspection revealed that some topics overlap, while others seemed to be missing relevant content. We therefore decided to introduce new topics that better capture the issues of interest. We dub these topics *custom semantic topics*, as they are not the output of the topic discovery process, but were later constructed specifically for the purpose of agenda analysis. We defined six custom semantic topics of interest by combining existing semantic topics; an exception is the *party negotiations* topic, whose content we split into custom topics *negotiations* and *negotiations–substance*. The complete list of custom topics and semantic topics belonging to two selected categories is shown in Table 1.

What has become obvious at this point is the need for text exploration tools that would complement and improve exploration based on the inspection of topic models (browsing topic-related words and articles). We envisage that such tools would enable keyword-based text retrieval for a deeper exploration of semantic topics, text similarity-based search for tracing rare issues, ability to seed entirely new topics, as well as the interactive modification of topic models, perhaps along the lines of (Hu et al., 2014). We consider this an interesting challenge for future work.

4.2 Measuring topic salience

After defining the custom topics, we proceed to measure their salience by counting the news articles in our corpus that deal with this topic. We do this by tagging each document with custom semantic topics. This process is essentially used as a proxy for human coding of the articles with topics.

To perform the tagging, Korenčić et al. (2015) propose to build a new topic model specifically customized towards the topics of interest. Namely, when using a non-customized model, there is no guarantee that model topics produced by the

[2] We note that there exist models specifically designed for the extraction of topic hierarchies; e.g. (Griffiths et al., 2004).

Top-level categories	Custom semantic topics	Semantic topics	Description
Electoral process	**election mathematics**	election forecasts, election polls, election results	pre- and post-election polls, speculation and statistics, forecasts, turnout, results, parliament combinatorics
	election procedures and regulation	election procedures and regulation, voting outside the place of residence, election rules and DIP, irregularities	election calendar, candidacy, monitoring, Ivan Turudić, electoral commission, candidates' debates, ethical commission, voting rules, irregularities
	election programs and campaign related events	economic election program, media coverage of elections	election programs, communication and bickering of parties and politicians
Post-election negotiations	**negotiations**	party negotiations, split within Most	negotiations and position taking, accusations and bickering, split within Most
	negotiations–substance	party negotiations	reform of local government, exclusive economic zone, economic and fiscal measures
	appointment of the PM designate and constitution of the Parliament	appointment of the mandate, presidential consultations, constituting the parliament	legal procedure and political process of the PM candidate appointment and constituting the Parliament

Table 1: List of semantic topics and derived custom semantic topics for the selected top-level categories

stochastic inference procedure will match any of the semantic topics of interest. Even if they would match to a certain extent, one would still have to manually inspect them and map to semantic topics.

Model customization is achieved by constructing, for each semantic topic, a list of *seed words* – words highly indicative for that topic. Once we have such lists, the model is built with probabilistic priors set to enforce topics that assign high probability to seed words. The underlying idea is that models built in this way will produce topics that correspond to custom semantic topics we are interested in. We follow the procedure of Korenčić et al. (2015) for obtaining the seed words: for each custom semantic topic, we inspect the list of highly probable words for that topics, and for each such word, we inspect a list of news articles estimated as related to it by a word-article association measure. If the majority of articles indeed deal with the considered topic, we add the word to the seed words list for that topic. Table 2 shows the seed words for the considered custom topics.

Custom semantic topics	Seed words
election mathematics	mandate, result, poll, win, vote, voter, exit, preferential, turnout, advantage, constituency
election procedures and regulation	committee, DIP, donation, report, spend, donate, promotion, GONG, financing, law, electoral silence, violation, campaign, debate, complaint, observer
election programs and campaign related events	economic, program, VAT, promise, electoral, termination, Prnjavor, demographic, irrigation, debt
negotiations	Petrov, negotiation, Božo, Prgomet, meeting, non-party, independent, Petrina, Drago, key, Grmoja, tripartite, reply, support, forming, pressure
negotiations–substance	reform, local, self-governance, belt, devalvation, inflation, rationalization, model, termination, Lovrinović
appointment of the PM designate and constitution of the Parliament	PM-designate, signature, consultations, forming, Pantovčak, round, session, Reiner, constitutive, convocation, elected

Table 2: Seed words for the chosen topics

For document tagging, we follow the procedure outlined by Korenčić et al. (2015): for each news article, using the customized model, we first infer the topic probability distribution, and then tag the article with the semantic topic corresponding to its most probable model topic.

An important insight we gained in this step is that some semantic topics are difficult to detect using topic models. For rare issues, custom topic model seeded with issue-specific words will produce a topic that includes other similar topical content, ultimately decreasing the tagging precision. We believe that a better alternative to detecting such topics is to describe them with a set of discriminative keyphrases, similarly to traditional dictionary approach to coding (Krippendorff, 2012). A case in point are the *Ljubljana Bank* and *voting outside the place of residence* topics, for which tagging based on a boolean keyword-based query fared much better than tagging using a customized topic model.

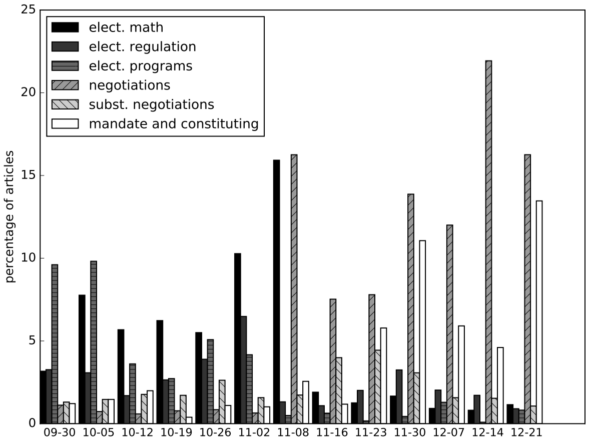

Figure 1: Electoral process and negotiations

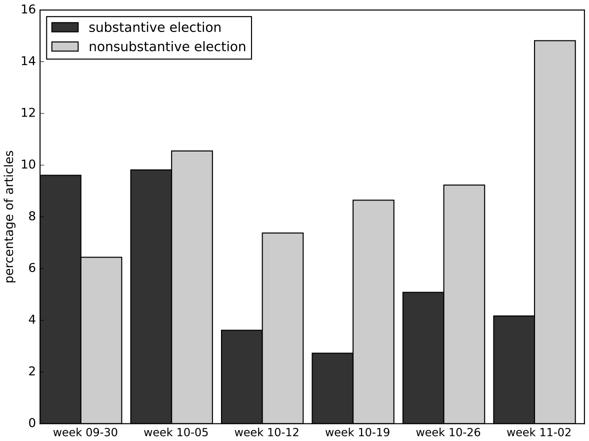

Figure 2: Substantive vs. non-substantive political topics in the pre-election period

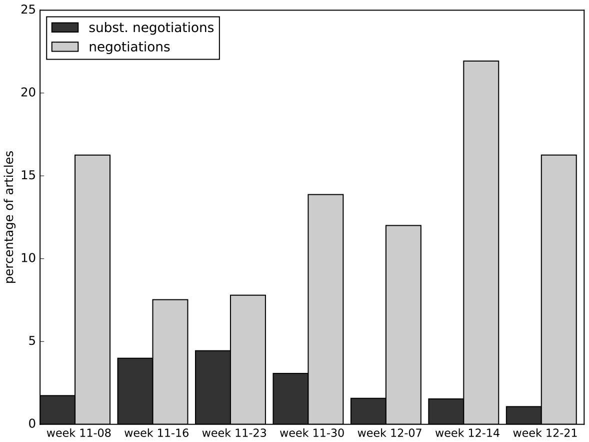

Figure 3: Substantive vs. non-substantive negotiation topics in the post-election period

5 Preliminary Analysis

In this section we present the results of using our model to conduct a preliminary analysis of the media agenda during Croatian 2015 parliamentary election.

5.1 Topic-event correlation

We note that validity is an important consideration when applying automated content analysis for political science research (Grimmer and Stewart, 2013; Lacy et al., 2015; Zamith and Lewis, 2015). While a thorough investigation of validity of our approach does not fit the scope of this paper, we ran a sanity check by analyzing how the inferred salience of topics correlates with real-life events.

In Fig. 1 we show the frequency of the articles across the six topics we considered. We find that the salience of semantic topics (defined as the number of articles tagged with the semantic topic) is very well correlated with real-life events. This correlation confirms the predictive validity (Grimmer and Stewart, 2013) of the model. Concretely, the *election mathematics* topic (including contents such as poll results, prediction of winners and losers, election results, etc.) was very salient in the week preceding the election day, and rocketed on the election day (November 8th).

Further evidence in support of the validity can be found by considering the events that took place after the election day. As none of the parties won the majority necessary to form the Government, both major parties – Social Democratic Party (SDP) and Croatian Democratic Union (HDZ) – tried to win over the newly established party of Most (The Bridge), which won a significant number of seats. Negotiations between the parties got excessive media coverage, which is successfully registered by our topic model. Furthermore, the week-to-week salience of the *negotiations* topic, also captured by our model, corresponds to the real-life events that triggered the visibility of this topic in the public discourse (bargains and disputes between the parties, search for the PM designate who would prompt Most to support one of the major parties, etc.). The same goes for the topic *appointment of the PM designate and constitution of the Parliament*, which was the most prominent in the days preceding the finally arranged constitution of the Parliament and formation of the Government.

5.2 “Game-schema” coverage

After the sanity check, we turned to a more insightful analysis from a political communication perspective. Building on the acknowledged distinction between substantive and less substantive election coverage (cf., for instance, Zeh and Hopmann (2013)), we divided the semantic topics into "substantive" and "non-substantive" ones. For the pre-election period, we categorize *election mathematics* and *election procedures and regulation* as non-substantive topics, and *election programs and campaign related events* as a substantive topic. For the post-election period, we differentiated between no-substance negotiation topics (such as conflicts between parties, bargains, etc.) and substantive negotiations that evolved around certain policy issues (cf. Table 1). Figures 2 and 3 show the week-to-week salience of these topics in the pre-election and post-election period, respectively.

The interesting finding is the clear dominance of the "non-substantive" content over the "substantive" content during the pre-election period. This primarily refers to the dominance of articles that focused on "horse-race" issues (e.g., opinion polling, who's ahead and who's behind, prediction of results) and the campaign hoopla, as opposed to articles that covered election programs and similar. Expectedly, the gap between substantive and non-substantive content was widening as the election was approaching. Interestingly, Figure 3 shows that this discrepancy was even stronger in the post-election period, suggesting that during the negotiation process media were more interested in parties' political bargain than in substantial content of negotiations. Whether this is due to media's interest in hoopla or due to the fact that politicians did not put substantial issues on the table is of course not the focus of this study. Overall, the analysis reveals the dominance of the "game schema" over more issue-centered information in the media coverage of elections, already witnessed in a number of countries (Patterson, 1993; Strömbäck and Dimitrova, 2006; Zeh and Hopmann, 2013).

6 Conclusion

We used a semi-supervised topic modeling methodology of Korenčić et al. (2015) to carry out a preliminary study of the media agenda during the 2015 parliamentary election in Croatia. The methodology consists of agenda discovery, in which model topics are manually mapped to semantic topics, and agenda measuring, in which news articles are tagged with topics. The primary purpose of our study was to gain a better understanding of the entire modeling process. In the agenda discovery step, the main methodological insights we gained is the need for a stricter annotation procedure and the importance of constructing a taxonomy of topics. In the agenda measuring step, we found the need for exploratory tools that would complement and improve the inspection of topic models and learned that some topics might be detected more precisely using keyphrases rather than topic model coding.

In a preliminary analysis of the media agenda, we were able to confirm the predictive validity of our model. Furthermore, we demonstrated the applicability of topic modeling by investigating the presence of substantive vs. non-substantive contents in the media coverage of the election. The results corroborate the common established assumption about the rise of game-oriented coverage at the expense of issue-related contents. It should be noted, however, that the findings presented here are just a few of the results that were obtained using topic modeling analysis, as a more detailed report would exceed the scope of this paper.

For future work, we plan to devise a stricter annotation procedure based on cross-checking, and test it using topic quality and coverage as the criteria. We also intend to experiment with text exploration tools complementary to topic models. Future validation should include more rigid quantitative validation measures and comparisons with the findings obtained by human coding. Finally, future research of media election coverage should be more focused in scope and include only articles specifically pertaining to election. This would weed out redundant content and may yield even more insightful results.

Acknowledgments

The first author has been supported by the Croatian Science Foundation project number 9623 and the The Croatian Policy Agendas Project[3] funded by the European Social Fund. We thank the students for their help in annotating the data: Frane Basioli, Doris Berečić, Janja Jagić, Mila Kovačević, and Karla Milevoj. We also thank the TakeLab[4] team for their technical assistance.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Jason Chuang, Sonal Gupta, Christopher Manning, and

[3] `www.cepis.hr`

[4] `takelab.fer.hr`

Jeffrey Heer. 2013. Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of ICML*, pages 612–620. JMLR Workshop and Conference Proceedings.

Jason Chuang, Margaret E. Roberts, Brandon M. Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. 2015. TopicCheck: Interactive alignment for assessing topic model stability. In *Proceedings of NAACL-HLT*, pages 175–184. ACL.

Claes H de Vreese and Holli A Semetko. 2004. *Political campaigning in referendums: Framing the referendum issue*. Routledge.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Thomas L. Griffiths, Michael I. Jordan, Joshua B. Tenenbaum, and David M. Blei. 2004. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems 16*, pages 17–24. MIT Press.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, pages 1–31.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning*, 95(3):423–469.

Yeooul Kim, Suin Kim, Alejandro Jaimes, and Alice Oh. 2014. A computational analysis of agenda setting. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pages 323–324. ACM.

Damir Korenčić, Strahil Ristov, and Jan Šnajder. 2015. Getting the agenda right: Measuring media agenda using topic models. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, pages 61–66. ACM.

Klaus Krippendorff. 2012. *Content Analysis: An Introduction to its Methodology*. Sage.

Stephen Lacy, Brendan R Watson, Daniel Riffe, and Jennette Lovejoy. 2015. Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*.

Nikola Ljubešić, Damir Boras, and Ozren Kubelka. 2007. Retrieving information in Croatian: Building a simple and efficient rule-based stemmer. *Digital Information and Heritage*, pages 313–320.

Maxwell E McCombs and Donald L Shaw. 1972. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187.

Thomas E Patterson. 1993. Out of order: How the decline of the political parties and the growing power of the news media undermine the american way of electing presidents. *New York: Alfred Knopf*.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. University of Malta.

Dietram A. Scheufele. 2000. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass Communication and Society*, 3(2–3):297–316.

Jesper Strömbäck and Daniela V Dimitrova. 2006. Political and media systems matter: A comparison of election news coverage in Sweden and the United States. *The Harvard International Journal of Press/Politics*, 11(4):131–147.

Rodrigo Zamith and Seth C Lewis. 2015. Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The ANNALS of the American Academy of Political and Social Science*, 659(1):307–318.

Reimar Zeh and David Nicolas Hopmann. 2013. Indicating mediatization? Two decades of election campaign television coverage. *European Journal of Communication*, 28(3):225–240.