Identifying Political Bias in News Articles

Konstantina Lazaridou and Ralf Krestel

Hasso-Plattner-Institut, Potsdam, Germany
konstantina.lazaridou@hpi.de,
ralf.krestel@hpi.de

1 Motivation

The power of journalism has an undeniably crucial role in the formation of the general public opinion. Media outlet owners, journalists, editors, and also bloggers not only are in the position of informing the general public about the current events, but they also convey their personal accent while doing so. They are responsible for monitoring the current affairs, selecting which of them are worth reporting, deciding the amount of space (or time in the context of TV and radio stations) they should dedicate and finally presenting them to the readers (or audience).

Considering that each of the above-mentioned steps is conducted by humans, they could contain deliberate or accidental bias. Since this can potentially affect people's beliefs and by extension decisions, there are numerous political studies attempting to analyze bias of media outlets. They measure the scope of media bias [2], examine how it is generated [3] and detect patterns in newspapers' storytelling [1]. Furthermore, identifying media bias is not only useful to the readers, but also to the journalists, since they would be able to assess and reflect on their work.

Both traditional mainstream media and modern digital media decide what and how they will report on a daily basis. Since this process is perfomed by humans, it can incorporate levels of subjectivity and objectivity, negative or positive perspectives towards a person or an event, ideological comments etc. Consequently, user-generated text is usually a mixture of fact statements, but also opinions, sentiments, humor etc. This mixture is not easily tractable and even more difficult to be classified as fair or biased. The text can also refer to multiple entities and their attributes, which makes the problem of automatically identifying the overall expressed opinion a highly interesting yet challenging task.

Media slant can be divided into several types. The so-called *gatekeeping* or *selection* bias represents the choices of the media outlets to report an issue or not. Considering the vast amount of topics that could possibly be covered, the selection process media follow is an indication of what they find important and interesting enough in the world to discuss about. This process depends also on the profile of the news provider (e.g. mainstream newspaper or personal blog) and the total amount of available space. It can also be affected by the media outlet's owner, the target audience (in terms of age, interests, ideology, etc.), the space the outlet dedicates for advertisements and so on. *Gatekeeping* bias is not only topic related, but could also be geographical. When comparing local, national, or international newspapers, it becomes apparent that they tend to cover different events, since their readers' interests might differ respectively.

In addition, an alternative expression of media bias is called *issue and facts* or *coverage* bias. Assuming that an outlet decides to cover a certain event, this type of bias concerns how much space the event occupies, whether all aspects and information are included etc. A representative example could be the news coverage of the upcoming EU referendum in the UK. There have been statements by many politicians such as David Cameron and Bill Clinton — surveys and polls among economists have also been conducted. Hence, if a newspaper reports only a subset of the current events, readers would only relate to a few aspects of the EU referendum and thus their opinion and vote would be shaped accordingly.

Furthermore, another side of media slant is *framing* bias, which refers to the way facts are presented. Primarly, how a fact is described depends on the outlet that presents it, namely on the newpaper's guidelines and the consistent profile it wishes to maintain. On the other hand, the individual journalist that reports about it is responsible for conveying the truth and all aspects of it, regardless the provider's preferences or his/her own style. In spite of the latter, it often occurs that authors frame the content of an article with their beliefs, not allowing readers to capture all dimensions of an event. Moreover, the choice of language of the writer can also predispose one in a positive or negative manner to become a supporter or an opponent of an certain opinion.

Finally, *ideological stand* bias or *statement* bias refers to the perspective that the author comments from. Except for pointing out the facts, reporters usually make remarks according to their point of view. Similarly to *framing* bias, the author's remarks might be a mixture of one's ideological beliefs and the newspaper's ideology. When it comes to political issues, for instance the problem of unemployement and underemployment, the humanitarian rights and the refugee crisis, etc. then an author's comment might also indicate their political affiliation. The latter is not necessarily unfair for the readers, since one may prefer reading newspapers sharing their beliefs. That is, one is more likely to perceive bias the further the slant of the news is from their own position [2].

However, since this is not the general case, the media should present a complete view of the current affairs, without neglecting to report information or promoting a single aspect. In order to tackle the above-mentioned challenges and efficiently detect the potential media bias, we propose an automatic approach. Human annotations are not an entirely reliable source and may involve subjectivity. Although there are cases where they can be used to assist computer science models in the bias detection process, they are generally cumbersome, expensive to obtain and could contain a dose of personal bias (also known as *annotator* bias). Additionally, considering the extremely big amount of news documents online, an automatic method appears to be the most effective and promising choice to analyze these data and extract usefull information from it.

In this PhD thesis we aim at discovering a specific kind of media bias, that is the underlying political bias of a newspaper. Our goal is to detect the newspaper's choices that reveal its political affiliation. Our motivation derives initially from that the fact that news articles are supposed to convey all aspects of the truth regardless the newspaper's political position and secondly from the challenges that news media analysis entails. Newspapers do not explicitly express their preferences and thus make our task more difficult compared to analyzing political blog posts, tweets or political statements. In order to achieve this goal, all the above-mentioned types of media slant are of interest for us, yet not in terms of the newspaper's general preferences but its political ones. Interesting patterns to discover are whether an outlet covers an unusually large number of stories that involve a certain party or a certain range of parties (left-wing, right-wing etc.), whether the journalists consistently present a party's actions positively or negatively and finally whether they criticize or agree with a given political party in their remarks.

2 Related Work

2.1 Biased Language

Regarding the discovery of biased language in text documents, related research includes *framing* and *epistemological* bias detection in Wikipedia articles [4], where bias driven human edits are identified and a logistic regression model is trained that captures the biased words in sentences. Moreover, [5] performs an empirical study where users annotate sentences of political text as biased or not, and in the first case they are also asked to identify the biased word in the sentence. Another kind of text where bias is not expected and is considered misplaced are the academic papers. The authors in [6] use supervised models to predict the political leaning of an economist based on his/her research papers and they also perform an experimental comparison between supervised and unsupervised approaches for this purpose.

2.2 Prejudiced Perspectives and Disputed Views

Ideological bias detection is closely related to *perspective* or *view* identification of a writer. A news article could be written from the view of the conservatives, democrats, liberals, etc. In this direction, Lin et al. [7] analyze the political perspective in articles both on document level and sentence level. They use statistical models like SVM and naive Bayes to uncover the word choices of the authors that depict their ideological perspective.

A further related topic to political bias detection is to identify disputed issues that different people, such as reporters, present totally different opinions about. These opposite points of view could correspond to different political leaning as well. [8] identifies topics in UK and US newspapers by making use of DBpedia links and computes the sentiment that these topics contain in order to discover the conflicting ones. Park et al. [9] introduce a new version of HITS algorithm to identify the main disputants of a topic and they later use SVM to classify news articles into different viewpoints of a story.

2.3 Individuals' Political Leaning

An alternative kind of users that instead of being influenced by media slant participate actively in the information diffusion are social media users. Twitter users constitute a representative example, since they openly convey their opinions by tweeting or re-tweeting others posts. Fang et. al [13] address political classification on Twitter by using a naive Bayes classifier that utilizes topical information, while Cohen and Ruths [12] focus on the Twitter dataset selection criteria to ensure that the performance of political orientation classifiers is not influenced and overoptimistic.

2.4 Bias in Political Text

Another relevant research topic to political classification of tweets is the detection of the underlying political bias in political blogs. Jiang and Argamon [15] identify subjective sentences in blog posts and the involved opinions, which they further use to politically classify the text. Moreover, a well used corpus for this purpose is the Bitterlemons blog, where Israeli and Palestinian editors write their opinions about various issues. An example of work in the above-mentioned corpus is [17], which focuses on bias on a topical level.

Parliament speeches and politicians' public statements are opinionated as well. [18] uses US congress speeches and ideological books to identify biased sentences with recursive neural networks and Sim et al. [19] infer the ideological portions in the speeches of US presidential candidates, by using a hidden Markov model.

2.5 Bias in News Articles

Regarding bias detection specifically in news articles, [27] demonstrates the characteristics of 100 English-speaking news outlets in terms of *gatekeeping*, *coverage* and *statement* bias. In addition, the framework "QUOTUS" [20] extracts politicians' quotations from news articles and blog posts that are used to construct an outlet-to-quote matrix. The authors utilize this matrix to predict whether an outlet will report a quote or not and to also discover the latent dimensions of political bias. Park et. al [21] deploy user comments to predict the political position of Naver News in South Korea, inspired by the idea that a liberal supporter will leave a negative comment to a conservative article and a positive one to an article that favors the liberals. Similarly, [24] examines whether user reactions on news opinion articles are positive or negative in order to classify the articles as liberal or conservative. Furthermore, Dallmann et al. [22] and Krestel et al. [26] cope with the bias detection problem in German, with the latter comparing parliament speeches and online news sources for this purpose. Finally, [25] quantifies the media slant by building a reference network to examine the citations of the 111th US Congress in social and news media.

3 Identifying Political Bias in News Articles

In this PhD thesis we are interested in revealing the different patterns that newspapers follow in their choice of language and the news story selection that potentially favor a political party. These tasks are concurrently important for the readers and the writers, but are also challenging since newspapers are not a typical example of opinionated and subjective text, such as product reviews. Initially, we examine *selection* bias, that is in terms of partisan politics how much space a newspaper dedicates for each political party. We examine how often politicians are mentioned and how often politicians are quoted in news articles. We also consider all political articles from the examined outlets and don't focus only on a single topic, politician or term of office. On the contrary, related work in [20] is only interested in quotes from Barack Obama both in newspapers and blogs, while [17] and [7] analyze the Bitterlemons political blog, which only refers to Israeli-Palestinian conflicts. We currently analyze the UK politicians' mentions and quotations in two UK newspapers during the period of 2000–2015.

3.1 Datasets

Our first dataset is the online newspaper `Guardian`. `Guardian's` articles are available online from 1996 till today. We currently use the *politics* section, since in `Guardian's` case it contains various worldwide and UK political news. In this way, we are also able to exclude irrelevant topics to us, such as lifestyle or sports articles. The total number of articles in the period 1996–2015 is 197,668.

The next dataset is the online newspaper `Telegraph`, which is also a well known UK media outlet. The `Telegraph` provides an archive with its past articles from 2000 to 2015 without explicit sections. In order to obtain political articles about the world and the UK, we examined the url patterns of the articles and crawled the ones that refer to *politics*, *UK* and *world* news, and *financial* issues. The news articles are 281,706 from 2000 till 2015. This is the common period between the two news corpora as well, hence we focus our analysis on this timeframe.

Besides UK newspapers, the third dataset, which is not yet fully exploited, is the collection of the UK parliament speeches from 1934 until 2015. Up to now, we have used this corpus to extract the politicians' names and parties from it, so that we search for them in the news articles. All datasets are analyzed (tokenized, stemmed) and indexed in an ElasticSearch instance, which is a search engine that enables full text search. Stopwords are also removed.

3.2 Ongoing Analysis

Our intuition is that if a newspaper mentions a party more frequently than the others, then this party is considered more significant, and thus their political orientation may be similar. Likewise, if a newspaper quotes more politicians' statements from the conservatives than from the liberals, they might share the same beliefs with the first ones. To this end, we count the times each party is mentioned in the news articles, namely sum the politicians' references for each party over time. This experiment is two-fold, since we use both text search and named entity search. The text search is performed in Elastic search, while the named entity recognition is performed by IBM Alchemy API, which provides us also with the direct quotations assigned to the discovered entities. Direct quotations refer to the case when media repeat what someone said using exactly the words they used, usually included in quotation marks.

Furthermore, the fact that a politician X is mentioned in a news article doesn't necessarily mean that X is presented favorably. It could occur that a politician Y criticizes X in his/her speech and the newspaper quotes Y's statement. It could also hold that the newspaper criticizes X negatively for his/her actions.

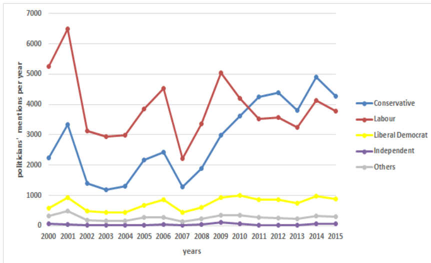
According to [1] most newspapers cover a wide variety of issues concerning all parties. However, the different parties are criticized disproportionately. Media generally point out the errors of all parties, but particularly the republican outlets portray the *Democratic* party in a more negative manner and vice versa. On the contrary, [2] and [3] discover in news articles strong political media bias and partisan language respectively. As indicated in [20], given the topic of president Obama and the timeframe of his tenure, their latent semantic analysis on numerous English-speaking outlets reveals that the media sources group by their ideology, whether they declared their political beliefs, and whether they are local or international media. Nonetheless, it shouldn't be disregarded that [20] considers both news media and social media (blogs, social networks etc.) in their analysis, with the latter presenting much more bias slant. In addition, the authors identify Barack Obama's quotations in the text by searching for specific parts of his speeches, instead of extracting reported speech from it.

4 Experiments

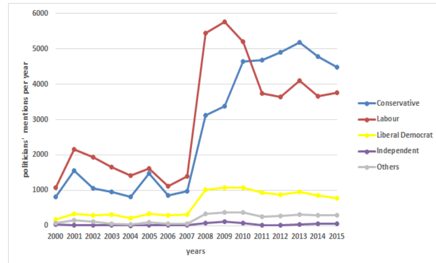
Our first experiment is to quantify the mentions of all UK politicians in `Guardian` and `Telegraph` from 2000 till 2015 using *ElasticSearch*. We perform phrase queries with the politicians' first and last names and search in the title, subtitle, body and image captions of the articles.

4.1 Politicians' mentions

The search results for the `Guardian` and the `Telegraph` are shown in Figures 1a and 1b, respectively. Regarding the most mentioned parties, the red curve corresponds to the *Labour*, whereas the blue one to the *Conservatives*. It is apparent that both newspapers discuss more about the current governments than the remaining parties and the two above-mentioned curves cross at the general election year of 2010. That is, both parties are mentioned more during their term — 1997-2010 in the case of *Labour* and 2010-2015 for the *Conservatives*.

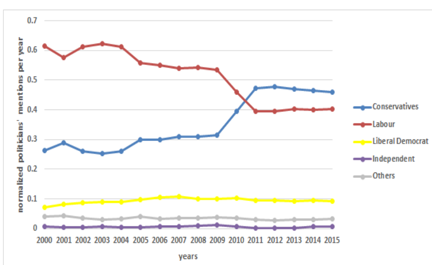


(a) Politicians' mentions in `Guardian`

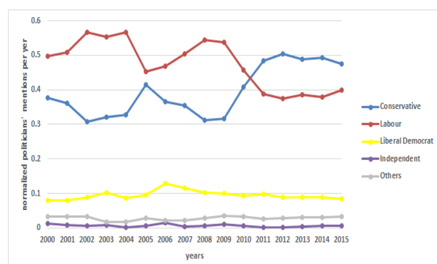


(b) Politicians' mentions in `Telegraph`

Although the above pattern holds for both outlets, the difference between the mentions of *Labour* and the *Conservative* party in `Guardian` is significantly higher compared to `Telegraph`. Namely, in 1a until 2010 *Labour* is discussed almost two times more than the *Conservatives*. The latter is in accordance with `Guardian`'s Wikipedia page stating that its politician alignment is centre-left. However, in 1b while *Labour* is not exceptionally discussed during Tony Blair's term (1997-2007), the references grow rapidly during Gordon Brown's tenure (2007-2010). As 1b also depicts, the *Conservatives*' mentions are increasing during Gordon Brown's term as well. This is an interesting discovery that motivates us to analyze further the sentiment around the mentions, as this will indicate whether the parties' references are positive, negative or neutral.



(a) Normalized politicians' mentions in `Guardian`



(b) Normalized politicians' mentions in `Telegraph`

Since the absolute number of mentions can vary extremely among different parties and different time periods, we further present the above results normalized. 2a and 2b depict the previous results, that is the annual politicians' mentions (summed for each party), but now normalized by the total number of mentions of all politicians per year. Initially, we still observe that the discussion dominating party changes from *Labour* to *Conservative* in 2010, as the *Conservatives* won in the general elections. Additionally, it is notable that the previous general elections have affected the media coverage as well. Particularly in the

`Telegraph` the two major parties' curves approach each other in 2005, while in the `Guardian` this phenomenon is more apparent in 2001.

Apart from analyzing the relative mentions of different parties, it is also beneficial to compare the media coverage with respect to the vote shares of the parties. Namely, examine whether all parties are represented by the news outlets according to their popularity. One can argue that an unbiased outlet is expected to report political events in accordance with the vote shares of the elected parties. As figure 3 presents, the *Liberal Democrats* increased their vote share from 18.3 in 2000 to 22.0 in 2005. This switch is not illustrated neither in 2a and nor in 2b, since the relative mentions of the *Liberal Democrats* during this period remain almost stable to 0.1. Moreover, the votes of the above-mentioned party dropped from 23.0 in 2010 to 7.9 in 2015 (during the coalition agreement with the *Conservatives*). The latter cannot be concluded in neither of the two examined news outlets.

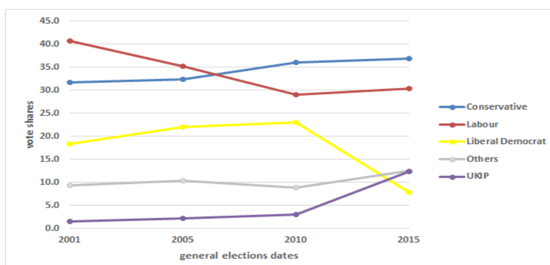


Fig. 3: UK general election vote share according to Wikipedia

Given that simple text search is not as reliable as discovering the named entities in the text, we further plot the UK politicians discovered as named entities in the news articles. For the rest of this section we will focus on the `Guardian`, as only these results are currently available and the reported results refer to absolute counts. As shown in 4 the above-mentioned pattern of citing mostly the current government's politicians is also apparent in the case of the named entity search. *Labour* is consistently more cited until the general elections in 2010, when the *Conservatives* begin to dominate the discussions. Figure 4 also presents the highest number of entity mentions for both parties in 2015, when the latest general election was held on 7th May. Furthermore, although the *Liberals*'s mentions rise in 2010, they continuously drop for the next five years and only increase during the latest general election year.

Since the named entity extraction task is challenging on its own, we will also use the Standford CoreNLP tool to ensure the completeness of the entity set and incorporate co-reference resolution into our approach.

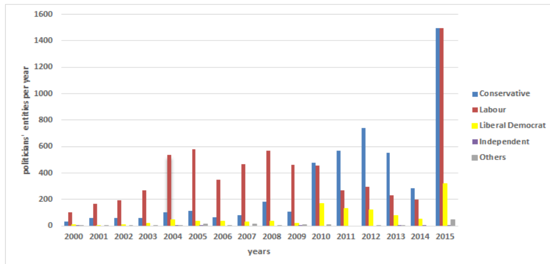


Fig. 4: Politicians' entities in `Guardian`

4.2 Politicians' quotations

Another possible indication of a newspaper's ideological perspective is its quoting pattern. Namely, the number of times media quote a party in general, or given a certain topic. To this end, we extract the politicians' quotes in `Guardian`, as shown in 5. *Labour*'s quotes in 2004 are three times more than the ones from the *Conservatives* and twelve times higher compared to the *Liberals*. Similarly to the previous experiments, 2010 is the first year that the *Conservatives* outperform *Labour* in terms of quotations and they maintain their prevalence until 2015.

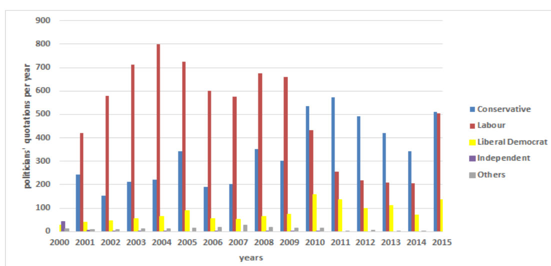


Fig. 5: Quotes from politicians in `Guardian`

Finally, since the Alchemy API discovers only direct quotations, our next step is to use an existing approach from the literature, such as [28] and [33], in order to consider indirect speech as well.

5 Conclusion and Future work

In this work, we are interested in identifying political bias in news media. Towards this direction, we consider different indicators for bias, that is the number of times media mention and quote politicians from different parties. We argue that if a newspaper discuss more a certain political party or group of parties belonging to a certain political direction, this behavior might indicate the newspaper's political beliefs. Our motivation stems from the fact that existing approaches are mainly manual. In the case of automatic approaches, the problem is tackled mostly for opinionated blog posts and the solutions focus on single topics or politicians. Quotation extraction from news articles is also part of our overall goal. That is why, whether media quote opposing sides of arguments or not is an indicator of how well informed the general public is.

Preliminary results on UK newspapers show that media tend to cover more stories from the current governing party and generally from the most popular parties. Coverage patterns vary in different election periods and among different newspapers. The latter grows our interest of analyzing these patterns in terms of the topics and sentiments involved. The comparison between the party mentions and the party vote shares shows that coverage is not always compliant with the votes that a party receives.

Regarding our future goals, an interesting initial task would be to use a topic model and discover the topics that are discussed both in the media and the parliament. This would feature the issues they all concentrate on and point out the common ones. Especially for the media outlets, we will use Mallet's LDA implementation and then we will examine the quoting patterns in each detected topic. To this end, the quote-to-outlet graph [20] is a concept of interest to us, especially in a temporal manner for observing how this graph evolves over time. Measuring the probability that the `Guardian` will quote in the future for instance the *Liberal* party for education issues and the *Labour* party for the health care issues would be an interesting achievement. Subsequently, the readers will have an insight in the media's political leaning and we will be able to recommend a set of news outlets that together provide a complete overview for a certain topic.

Moreover, political media bias detection consists of many aspects apart from *selection* bias, such as sentiment analysis and opinion mining in the text. We aim at detecting the expressed sentiment both around a politician's mentions [30] and in a complete news article [29]. Given the sentiment expressed by the media for the politicians, and inspired by the *signed networks* [31], we plan to capture these sentiment values in a signed graph. Namely, a bipartite graph G where the nodes V are outlets and politicians, while the edges E denote the sentiment that outlets show when they discuss a politician. A positive edge sign means positive language towards a person, while a negative sign refers to the case that an outlet criticizes negatively a person. Using the principles of *status* and *balance* theory, one could attempt to predict the sign of an unobserved edge, that is the sentiment an outlet will show about a politician.

Finally, our general goal is to analyze online media in order to observe how information and opinions travel and change among individuals. Different entities, for instance politicians, journalists, readers, commenters, etc. play different roles in the information diffusion and contribute to the people's influence. Since media's choice to promote certain pieces of information or persons shape general public's opinion, selection bias detection in political news is our first step in our analysis. Graph-based approaches such as the above-mentioned, seem appropriate to capture the dynamics of the involved entities. Namely, which politicians' quotes media report, how negatively or positively are political events described by media and by extension perceived by readers.

References

1. Budak C., Goel Sh. and Rao M.: Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. Public Opinion Quaterly, Vol. 80, Special Issue, p. 250–271 (2016)
2. Groseclose T. and Milyo J.: A measure of media bias. The Quartely Journal of Economics 120, Issue 4, p. 1191–1237 (2005)
3. Gentzkow M.: What drives media slant? Evidence from us daily newspapers. Econometrica, Volume 78, Issue 1, p. 35–71 (2010)
4. Recasens M., Danescu-Niculescu-Mizil C. and Jurafsky D.: Linguistic Models for Analyzing and Detecting Biased Language. Proceedings of ACL (2013)
5. Yano T., Resnik Ph. and Smith N.: Shedding (a Thousand Points of) Light on Biased Language. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (2010)
6. Jelveh Z., Kogut B. and Naidu S.: Detecting Latent Ideology in Expert Text: Evidence From Academic Papers in Economics. Proceedings of EMNLP (2014)
7. Lin W., Wilson Th., Wiebe J. and Hauptmann A.: Which Side are You on? Identifying Perspectives at the Document and Sentence. Proceedings of CoNLL (2006)
8. Clercq O., Hertling S., Hoste V., Ponzetto S. and Paulheim H.: Identifying Disputed Topics in the News. Proceedings of the LD4KD Workshop at ECML/PKDD (2014)
9. Park S., KyungSoon L. and Song J.: Contrasting Opposing Views of News Articles on Contentious Issues Souneil. Proceedings of ACL : HCT (2011)
12. Cohen R. and Ruths D.: Classifying Political Orientation on Twitter: It's Not Easy! Proceedings of ICWSM (2013)
13. Fang A., Ounis I., Habel Ph., Macdonald C. and Limsopatham N.: Topic-centric Classification of Twitter User's Political Orientation. Proceedings of ACL (2015)
15. Jiang M. and Argamon Sh.: Exploiting subjectivity analysis in blogs to improve political leaning categorization. Proceedings of SIGIR (2008)
17. Xing E. and Ahmed E.: Staying Informed: Supervised and Semi-Supervised Multi-view Topical Analysis of Ideological Perspective. Proceedings of EMNLP (2010)
18. Iyyer M., Enns P., Boyd-Graber J. and Resnik Ph.: Political Ideology Detection Using Recursive Neural Networks. Proceedings of ACL (2014)
19. Sim Y., Acree B., Gross J. and Smith N.: Measuring Ideological Proportions in Political Speeches. Proceedings of EMNLP (2013)
20. Niculae V., Suen C., Zhang J., Danescu-Niculescu-Mizil C., Leskovec J.: QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns. Proceedings of WWW (2015)
21. Park S., Ko M., Kim J., Liu Y. and Song J.: The Politics of Comments: Predicting Political Orientation of News Stories with Commenters Sentiment Patterns. Proceedings of CSCW (2011)
22. Dallmann A., Lemmerich F., Zoller D. and Hotho A.: Media Bias in German Online Newspapers. Proceedings of ACM HyperText (2015)
24. Zhou D., Resnik P. and Mei Q.: Classifying the Political Leaning of News Articles and Users from User Votes. Proceedings of ICWSM (2011)

25. Lin Y., Bagrow J. and Lazer D.: Quantifying Bias in Social and Mainstream Media. ACM SIGWEB Newsletter (2012)
26. Krestel R., Wall A. and Wolfgang N.: Treehugger or Petrolhead? Proceedings of WWW (2012)
27. Saez-Trumper D., Castillo C. and Lalmas M.: Social media news communities: gatekeeping, coverage, and statement bias. Proceedings of CIKM (2013)
28. Pareti S., O 'Keefe T., Konstas I., Curran J. and Koprinska I.: Automatically Detecting and Attributing Indirect Quotations. Proceedings of ACL (2013)
29. Socher R., Perelygin A., Wu Y., Chuang J., Manning C., Ng A. and Potts C.: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Proceedings of EMNLP (2013)
30. Engonopoulos N., Lazaridou A., Paliouras G. and Chandrinos K.: ELS: a word-level method for entity-level sentiment analysis. Proceedings of WIMS (2011)
31. West R., Paskov H., Leskovec J. and Potts C.: Exploiting Social Network Structure for Person-to-Person Sentiment Analysis. Proceedings of TACL (2014)
33. Scheible Ch., Klinger R. and Pad S.: Model Architectures for Quotation Detection. Proceedings of ACL (2016)