Treehugger or Petrolhead?

Identifying Bias by Comparing Online News Articles with Political Speeches

Ralf Krestel
L3S Research Center
Hannover, Germany
krestel@L3S.de

Alex Wall
Leibniz Universität
Hannover, Germany
alex@noligy.de

Wolfgang Nejdl
L3S Research Center
Hannover, Germany
nejdl@L3S.de

ABSTRACT

The Web is a very democratic medium of communication allowing everyone to express his or her opinion about any type of topic. This multitude of voices makes it more and more important to detect bias and help Internet users understand the background of information sources. Political bias of Web sites, articles, or blog posts is hard to identify straightaway. Manual content analysis conducted by experts is the standard way in political and social science to detect this bias. In this paper we present an automated approach relying on methods from information retrieval and corpus statistics to identify biased vocabulary use. As an example, we analyzed 15 years of parliamentary speeches of the German Bundestag and we investigated whether there is bias towards a political party in major national online newspapers and magazines. The results show that bias exists with respect to vocabulary use and it coincides with human judgement.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis

Keywords

Bias detection, news analysis, vector space model

1. INTRODUCTION

User generated content plays an important role in the Web 2.0 and the Social Web. Many Internet users express their experiences and opinions online. Product reviews, comments about videos or photos, or political blogs about current events constitute only some ways of user participation in online discussions. In addition, professional information providers like large news sites shape the way information is perceived in the Web. Identifying bias in individual texts or for particular sources is an important task to ensure that users can get a complete overview of a topic, product, etc.

The media, especially news providers, are responsible for how we perceive events, political decisions, discussions, or debates. A multitude of different news providers together with journalistic integrity helps ensure a balanced view of these news. Nevertheless, there is, for example in the United States, a common agreement that media outlets are somewhat biased. In Germany the individual newspapers or magazines are also considered to be biased towards different political camps [4]. The political spectrum is covered by various national and regional news providers allowing readers to pick a newspaper that represents their own political preferences. The bias in these newspapers is not explicit and difficult to grasp for the average reader. Especially if only one particular source of information is available and the confidence in objective coverage of the news is high.

Bias in news can have different shapes. [1] identifies three different types of media bias: gatekeeping bias, coverage bias, and statement bias. Gatekeeping bias is the effect of writers or editors selecting a story to be published or not from all possibly available news stories. Coverage bias is the amount of space a newspaper dedicates to a certain view, opinion, or event. And finally, statement bias takes place on the content level. Writers can incorporate their own opinions while reporting about an issue.

We focus on the third type of media bias termed statement bias. Identifying statement bias is usually done manually by experts using content analysis [3] techniques in political science or journalism research. We try to automate this process and find statistical measures to identify bias in documents. To circumvent possible gatekeeping bias and coverage bias we picked news stories that where covered by all online news sites under investigation. These were the four largest national daily newspapers in Germany (*Süddeutsche*, *Bild*, *Welt*, *FAZ*) and the major national weekly magazines: *Spiegel*, *Focus*, and *Stern*. We performed two kinds of experiments to find out: (1) What are the typical terms of the different parties in the German Parliament? (2) How is the vocabulary of the parties picked up by news providers? (3) Does similar vocabulary use correspond to the perceived bias of the newspaper? To answer these questions, we analyzed the speeches given in the German Parliament and compared the vocabulary with articles from different news providers.

Besides news articles, our method can also be used to identify vocabulary bias in comments, blog posts, or other documents. It can be used in the context of intelligent search or recommendations to find articles that represent a specific political position or allow for diversification of search results based on political view points.

2. BIAS DETECTION APPROACH

A very popular method in information retrieval (IR) for finding relevant documents to a query is using a vector space model. The most common term weights are based on tf*idf scores. To detect possible bias of online newspapers towards a certain political camp, we compare the content of the articles with the parliamentary speeches of all parties using a

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

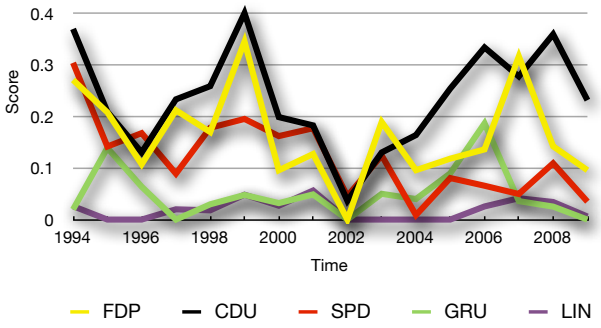

Figure 1: Analysis over time for term "Kernenergie" (nuclear energy)

vector space model. We consider the articles about a particular topic in one newspaper to be the query in an IR sense. All speeches of one party are then considered a document and we identify the most relevant document (party) for our query.

2.1 Corpus Generation

We used two corpora which we crawled from the Internet: *German Parliament speeches* and *online news articles from major German newspapers*. Some non-trivial preprocessing was needed to get the full text of each parliamentary speech together with the party of the speaker and the full text of the news articles for a particular topic on the Web removing boilerplate text.

The German Bundestag maintains an archive of all plenary discussions that took place in the parliament [2]. They are available as PDF documents converted from stenographic notes. We extracted the speeches from legislative period 13 to 16 (1994–2009) resulting in over 900 plenary sessions. For each speech we extracted together with the text the party of the speaker and removed template content.

To get topically relevant, political news from Germany we crawled the German GoogleNews page (`http://news.google.de`) between January and February 2010. They index over 700 German speaking news sources and cluster the individual articles into categories like business, sport, etc. Since there is no distinct category for politics, we took the category "Deutschland". We randomly selected 10 topics where we had a couple of news articles from all the major national newspapers.

2.2 Results

To identify characteristic terms for each party in the parliamentary speeches we computed tf*idf scores for each term for each year and for each party. When we look at the top terms for each party for the different legislative periods based on aggregated tf*idf scores, we can identify the focus of the different parties. We can also see the evolution within the parties and what they consider important topics, e.g. between 2005 and 2009 the top terms for the conservative (CDU) party were "growth" and "challenges" whereas for the social democrats (SPD) "soldiers" and "employees" were most important.

A temporal analysis for the term "Kernenergie" (nuclear energy) can be seen in Figure 1. It shows that the Green Party does not use the term "nuclear energy" very often in contrast to CDU and FDP. An explanation for this is the fact that the Green Party prefers to use the term "Atomenergie" instead, as a way to indicate their bias against the use of nuclear energy. In addition, different peaks mirror particular events: In 1999 e.g., the German government voted for a nuclear phaseout.

Table 1: Deviation from Average cosine similarity over all topics for selected newspapers in ‰

Newspaper	FDP	CDU	SPD	Grüne	Linke
Süddeutsche.de	-0.7	-1.4	-1.4	+0.6	+2.0
Bild.de	-5.8	+5.5	+3.6	-3.6	-1.1
Welt Online	-8.3	+1.4	-0.8	-1.5	+8.0
FAZ.net	-4.2	+1.9	+4.0	-3.2	+1.0
Focus Online	-5.1	+2.9	+0.6	-1.7	+2.5
Stern.de	-4.7	+0.1	+5.4	-1.9	+0.2
Spiegel Online	-5.8	+0.7	+3.0	-1.1	+1.9

The overall similarity over all analyzed topics is shown in Table 1. The deviations of the averages of the relative cosine similarity gives an impression for the bias of each newspaper. The data used for this experiment is from the 16th legislation period (2005–2009) where the government was formed by a coalition of the two major parties CDU and SPD. What can be seen in this table is that the *Süddeutsche.de* is more biased towards the left opposition of GRUENE and LINKE. The *Bild* newspaper on the other side uses more terms from the governing parties CDU and SPD. These statistical findings are in accordance with human judgement about newspaper bias [4]. The rather high overlap of all newspapers with the LINKE seems to indicate that their representatives in parliament use a more catchy language which is closer to the genre of news articles. To verify this last interpretation of the data, more investigation is needed.

3. CONCLUSIONS

In this paper we did a quantitative analysis of major German online news providers and of the speeches given by members of different parties in the German Parliament during 15 years. We identified typical terms for each party during different legislative periods revealing the different focuses of each party. We also compared the political speeches with current news articles based on corpus statistics and used vocabulary. The results show that newspapers tend to have a slight bias towards a political camp with regard to the vocabulary use. An analysis of individual news topics showed the details of this effect.

4. REFERENCES

[1] D. D'Alessio and M. Allen. Media Bias in Presidential Elections: A Meta-Analysis. *Journal of Communication, Autumn*, pages 133–176, 2000.

[2] Deutscher Bundestag. Stenografische Berichte. Online, January 2011. `http://dip.bundestag.de/`.

[3] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Inc, 1980.

[4] Redaktion Eigentümlich-Frei. Übersicht: Politisch meinungsbildende Zeitungen und Zeitschriften in deutscher Sprache. Online, December 2009. `http://ef-magazin.de/2009/12/23/1761-uebersicht-politisch-meinungsbildende-zeitungen-und-zeitschriften-in-deutscher-sprache`.