Submitted draft to Expert Systems With Applications (ESWA).
Published paper: http://www.sciencedirect.com/science/article/pii/S0957417412006100
Junqué de Fortuny, E., De Smedt, T., Martens, D., Daelemans, W. Media coverage in times of political crisis: A text mining approach. *Expert Systems with Applications* (2012).

Media coverage in times of political crisis: a text mining approach

Enric Junqué de Fortuny *
Faculty of Applied Economics
University of Antwerp, Belgium

Tom De Smedt
Faculty of Arts
University of Antwerp, Belgium

David Martens
Faculty of Applied Economics
University of Antwerp, Belgium

Walter Daelemans
Faculty of Arts
University of Antwerp, Belgium

February 1, 2012

Abstract

At the year end of 2011 Belgium formed a government, after a world record breaking period of 541 days of negotiations. We have gathered and analysed 68,000 related on-line news articles published in 2011 in Flemish newspapers. These articles were analysed by a custom-built expert system. The results of our text mining analyses show interesting differences in media coverage and votes for several political parties and politicians. With opinion mining, we are able to automatically detect the sentiment of each article, thereby allowing to visualize how the tone of reporting evolved throughout the year, on a party, politician and newspaper level. Our suggested framework introduces a generic text mining approach to analyse media coverage on political issues, including a set of methodological guidelines, evaluation metrics, as well as open source opinion mining tools. Since all analyses are based on automated text mining algorithms, an objective overview of the manner of reporting is provided. The analysis shows peaks of positive and negative sentiments during key moments in the negotiation process.

*Corresponding author, enric.junquedefortuny@ua.ac.be

1 Introduction

Belgium has recently recovered from the longest government formation period known to modern-day democratic systems (BBC Europe, 2011). It is well established that mass media and the internet in particular play an increasingly more important role in opinion formation (Savigny, 2002). On-line articles are easily accessible, providing us the unique opportunity to access, analyse and compare them over different newspapers. With huge amounts of articles available, it is no longer possible to analyse and interpret them manually. This challenge is overcome by using a text mining approach, which allows for automated analysis of all the articles. Using an automated technique strengthens the objectivity of the analysis: personal bias and opinion in scoring is reduced substantially due to the absence of manual human intervention. Please note that we refrained from interpreting the results on a political level as much as possible, yet we write to demonstrate how a text mining approach allows an efficient and objective analysis and to summarize news coverage in today's on-line media landscape.

1.1 The role of mass media in opinion formation

Nowadays, news papers and other news providers are updating their on-line news feeds in near real-time, allowing interested parties to follow the news in near real-time. The Internet has thereby become an important broadcast medium for politicians.

A study by Benewick et al. (1969) showed that high exposure to a party's broadcasts was positively related to a more favourable attitude towards that party for those with medium or weak motivation to follow the campaign. Knowing this, McCombs & Shaw (1972) raise the question whether the mass media sources actually reproduce the political world perfectly. In an imperfect setting, *biases* of media could propagate to the public opinion and therefore influence favouritism towards one or another party or politician, thus shaping political reality. Most of this bias is introduced by the editing and selection process. They conclude that the mass media may well set the agenda of political campaigns. In the digital era, Internet has taken up its own place as a mass medium next to TV (Fredricksen, 2010) and the aforementioned concerns are becoming increasingly more relevant for the Internet as well.

In a meta-analysis considering 59 studies D'Alessio & Allen (2000) found

three main bias metrics used to study partisan media bias. The first bias metric is derived of the fact that mere selection (and deselection) of news articles to be published by editors introduces a bias. This so called *gatekeeping bias* causes some topics to never surface in the media landscape and is therefore an interesting measure of a sampling bias, introduced by editors and journalists (Dexter & White, 1964). The problem when measuring the gatekeeping bias, however, is that it assumes knowledge of the whole universe of articles before actual selection. This turns out to be infeasible to determine for our purpose since information about rejection is generally undocumented and thus unknown.

A second bias metric considered, is the *coverage bias*, which measures the physical amount of coverage that each side of the issue receives. Traditionally, this is measured in column inches, amount of headlines or in broadcasts time devoted to sides of the issue. We measure the coverage as the amount of on-line articles. For political issues each party or politician can be seen as a 'side' or an entity. We argue that fair coverage is determined by an a priori distribution. This a priori distribution represents all entities in the population by their relative importance as measured by electoral votes in the last elections. Large deviations from the fair distribution tend to show some coverage bias towards one or another entity.

Third, there is the *statement bias* metric which is concerned with the distinction between favourable versus unfavourable or positive versus negative news. Given the large corpus used in this study, we choose an automated sentiment mining approach to measure the statement bias. A word of caution is in order about the interpretation of these 'sentiments' in news articles. When an article about an entity is classified as negative, this does not necessarily imply unfavouritism from a news source or journalist towards that entity. It might be that the entity is purposely interjecting negative criticism in the article. This measure should therefore be seen as an associative measure (i.e. a party is associated with a negative image when most of its coverage contains negative content). No conclusions can be made as to whether this image is build by the entity in question or by the news source, that is, we can only prove that it *exists*.

Belgium has seen a unique governmental crisis in 2007-2011 during which both political parties and politicians have had wide media coverage. In this study, we analyse the bias towards political entities during this period using a text mining approach. In order to do so, we must first elaborate on the unique setting in which this study took place.

Party	Ideology	Political figures
CD&V	christian democratic	M. Thyssen R. Torfs J.-L. Dehaene E. Schouppe S. de Bethune P. Van Rompuy
N-VA	Flemish nationalism, liberal conservatism	B. Wever H. Stevens
Open VLD	liberalism, social liberalism, market liberalism	A. de Croo D. Sterckx
SP.A	social democracy, third way	J. Vande Lanotte F. Vandenbroucke M. Temmerman
VB	Flemish nationalism, separatism, conservatism	F.Dewinter

Table 1: Flemish parties included in the corpus.

1.2 Belgium and its unique political crisis

At the basis of the Belgian crisis lies its dual federalist structure (the two major regions being the Dutch speaking Flemish Community and the French Speaking French Community). An overview of the political parties, including their ideology and their prominent figures is displayed in Table 1.

The mass media has played an important and sometimes controversial role in the courses taken by parties and - as is often the case in political conflicts - passionate assertions have been expressed towards the mass media and its favouritism (Niven, 2003). The question whether there truly exists a bias or not should however be answered by a systematic analysis. We accomplish this through the different bias metrics introduced in Section 1.1. Using text mining techniques we can calculate these metrics in an unbiased way, a rare feat in a political setting. Comparison with the actual election outcomes allows us to extrapolate information on the relative biases of mass media towards one or another party.

We use the 2010 Chamber election results as a golden standard against which we measure the mass media bias. The reasoning behind this choice is that the mass media know these results and therefore know the relative

weights of different political parties beforehand. A similar reasoning is used to compare politicians, but using the 2010 Senate election results since popular politicians are voted for directly in the Senate elections, whereas Chamber elections concern party votes.

1.3 Text mining and sentiment analysis

The information age offers an overwhelming number of text documents available on the Web and elsewhere. This poses a challenge since the amount of information is constantly increasing. Furthermore, text documents generally lack metadata such as language, topic, syntactic structure and semantic labels.

Text mining or knowledge discovery concerns the process of automatically extracting novel, non-trivial information from unstructured text documents (Fayyad & Piatetsky-Shapiro, 1996), by combining techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR) and knowledge management (Mihalcea, 2011). Common text mining tasks involve document classification, summarization, clustering of similar documents, concept extraction and sentiment analysis. Text mining has had a wide range of applications to date, prevalent applications include: forecasting petitions (Suh et al., 2010), guiding financial investments (Rada, 2008) and sentiment detection in reviews (Tang et al., 2009).

Textual information can be broadly categorized into two types: objective facts and subjective opinions (Liu, 2010). Opinions carry people's sentiments, appraisals and feelings toward the world. Sentiment analysis (or opinion mining) is a subfield of natural language processing that in its more mature work focuses on two main approaches. The first approach is based on subjectivity lexicons (Taboada et al., 2011), dictionaries of words associated with a positive or negative sentiment score ("polarity"). Such lexicons can be used to classify phrases, sentences or documents as subjective or objective, positive or negative. The second approach is by using machine learning text classification (see for example Pang et al. (2002)). Most work on sentiment analysis has been carried out on product reviews.

News sources are sentiment-rich resources for which we can extract qualitative information using similar techniques as described above. Using a subjectivity lexicon for Dutch adjectives we analyse the general tone associated with politicians and their parties during the political crisis. Similar research has been conducted by Schumaker & Chen (2009) for stock market

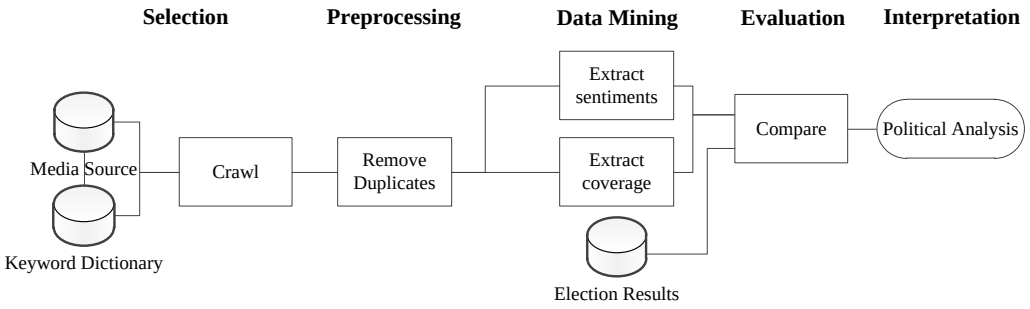


Figure 1: Processing steps used to build and analyse the corpus.

prediction, Godbole & Skiena (2007) for news and blogs and Balahur et al. (2010) for newspaper articles.

2 Material and methodology

In our analyses, we followed the KDD methodology (*knowledge discovery in databases*) by Fayyad & Piatetsky-Shapiro (1996) that describes the different steps in a KDD application. The resulting process as applied to our problem is displayed in Figure 1.

2.1 Data acquisition and selection

The corpus used in this study comprises of all articles published in on-line versions of all major Flemish newspapers in 2011 until the end of the political crisis. The corpus contains over 68,000 articles, spanning a ten month period (from January 1, 2011 to October 31, 2011). An overview of the eight covered newspapers is displayed in Table 2.

All articles were gathered using a custom built web-crawler. The crawler extracts articles from the sources' websites using their built-in search functionalities. The crawling process is the equivalent of a typical database *selection process* in which relevant data are selected using the given query criteria. The query keywords are all major Flemish party names and leading figures of political parties (see Table 1). The criterion for being a *party* of interest is based on the votes for that party in the 2010 Chamber Elections, we only included major parties who were allowed in the Chamber (i.e. parties with at least one chair). A leading *figure* is a politician with a top ten ranking

amount of preference votes in the 2010 Senate Elections (cf. Section 1.2). An overview of all political entities included in this study is displayed in Table 1.

2.2 Data Preprocessing

In a second *preprocessing phase* all data is filtered so as to remove possible duplicate articles. To see why this is necessary, consider the case in which an article concerns two parties at the same time. As a consequence of the crawling process, this article will be presented twice in the dataset (once for the first party name search, once for the second party name search). The second article present is a redundant entry and must be removed for the frequency information to be correct. This leaves us with a corpus of all unique articles containing the keywords presented in Table 1.

2.3 Data Mining: Sentiment analysis

For sentiment analysis, we used the previously created *Pattern* mining module for Python[1]. The module contains a subjectivity lexicon of over $3,000$ Dutch adjectives that occur frequently in product reviews, manually annotated with scores for polarity (positive or negative between $+1.0$ and -1.0) and subjectivity (objective or subjective between $+0.0$ and $+1.0$). For example: "boeiend" (fascinating) has a positive polarity of $+0.9$ and "belabberd" (lousy) has a negative polarity of -0.6. A similar approach with one axis for polarity and one for subjectivity is used by Esuli & Sebastiani (2006) for English words.

In previous research, the lexicon was tested with a set of $2,000$ Dutch book reviews. Each review also has a user-given star rating. The set was evenly distributed over negative opinion (star rating 1 and 2) and positive opinion (star rating 4 and 5). The average score of adjectives in each review was then compared to the original star rating, with a precision of 72% and a recall of 82% (De Smedt & Daelemans, 2011).

In our approach, we look for occurrences of Flemish political parties (see Table 1) in each newspaper article. We then calculate the polarity of each adjective that occurs in a window of two sentences before and two sentences after. An article can mention several party names, or switch tone. The given

[1] http://www.clips.ua.ac.be/pages/pattern

interval ensures a more reliable correlation between the political party being mentioned (the "target") and the adjective's polarity score, contrary to measuring all adjectives in the article. A similar approach for target identification with a 10-word window is used in Balahur et al. (2010). They report improved accuracy when compared to measuring all words in the article. We furthermore exclude adjectives that score between -0.1 and $+0.1$ to reduce noise. This results in a set of $366,613$ assessments, where one assessment corresponds to an adjective score linked to a party or politician.

For example:

*Bart De Wever (*N-VA*) verwijt de Franstalige krant Le Soir dat ze aanzet tot haat, omdat zij een opiniestuk publiceerde over de Vlaamse wooncode met daarbij een foto van een Nigeriaans massagraf. De hoofdredactrice legt uit waarom ze De Wever in zijn segregatiezucht hard zal blijven bestrijden. "Verontwaardigd? Ja, we zijn eerst en boven alles verontwaardigd."*

*Bart De Wever (*N-VA*) accuses the French newspaper Le Soir of inciting hatred after they published an opinion piece on the Flemish housing code together with a picture of a Nigerian mass grave. The editor explains why they will continue to fight De Wever's love of segregation hard. "Outraged? Yes, we are first and above all outraged."*

The adjective "hard" scores -0.03 and is excluded. The adjective "verontwaardigd" (outraged, indignant) scores -0.4. In overall, the passage about the political party N-VA is assessed as negative.

3 Results

We analyse the biases and sentiments throughout the whole corpus for each political entity. We will first be looking at the frequency of occurrence (i.e. the coverage) and afterwards at the tone of articles.

3.1 Media coverage bias

[2]Source: Federal Public Services Home Affairs (http://polling2010.belgium.be/. All percentages in this study are normalized over the Flemish parties and the selection, so as

Source	Regional	#Articles	#Readers[2]	$d_{H,c}$	$d_{H,v}$	Bias
De Redactie	No	5,767	146,250	0	2	16.16%
De Morgen	No	11,303	256,800	0	2	21.53%
GVA	Yes	9,154	395,700	**6**	4	**27.30**%
HBVL	Yes	3,511	423,700	3	**5**	**37.43%**
Nieuwsblad	No	7,320	1,002,200	2	4	20.24%
De Standaard	No	9,154	314,000	2	4	23.32%
De Tijd	No	10,061	123,300	3	4	22.71%
HLN	No	11,380	1,125,600	3	4	21.62%

Table 2: News sources used for the analysis including their respective number of articles and readers as well as bias metric values for the hamming distance from consensus ($d_{H,c}$), the hamming distance from votes ($d_{H,v}$) and the deviation from election outcomes (*bias*).

The *coverage* $c(e, s)$ of an entity e by a newspaper s is defined as the number of news articles published by the newspaper on that entity, normalized on the total amount of articles by that newspaper in the corpus $\mathcal{A}_s$:

$$c(e,s) = \frac{\#\{a | a \in \mathcal{A}_s \wedge e \in a\}}{\#\mathcal{A}_s} \quad (1)$$

The *popularity* $p(e)$ of a political party e is defined as the relative amount of preference votes $v(e)$ for that entity (as compared to other entities in the top ranking set $\mathcal{E}$):

$$p(e) = \frac{v(e)}{\sum_{e' \in \mathcal{E}} v(e')} \quad (2)$$

The popularity is used as the a priori fair distribution. The *coverage bias* (henceforth referred to as the bias) of a media source is the difference between the real distribution and the fair distribution. That is:

$$bias(e,s) = c(e,s) - p(e) \quad (3)$$

$$bias(s) = \sum_{e \in \mathcal{E}} bias(e,s) \quad (4)$$

where a is an article, represented as a bag of words $\{w_1, w_2, ..., w_{n_a}\}$ with n_a the amount of words in the article. Figure 2 shows that for some parties a

to be able to make statistically sound comparisons.

substantial bias is found, with a maximal positive bias towards *CD&V* and a maximal negative bias towards the far-right *Vlaams Belang* (*VB*). The first result can be justified by the fact that *CD&V* ran the interim government while the new government formations took place. The latter result gives supportive evidence to previous research outcomes that the party is being quarantined by the media (Yperman, 2004).

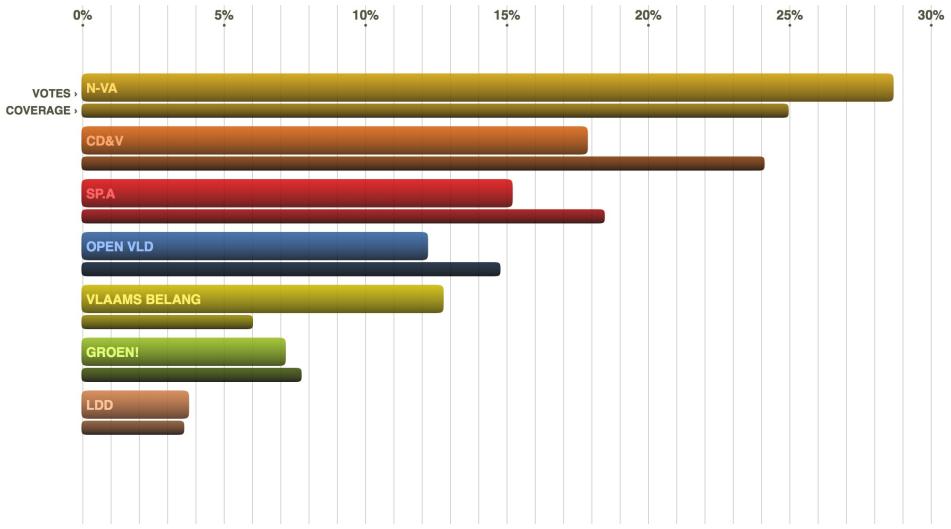


Figure 2: Discrepancy between media coverage and popularity for popular parties

We repeat the analysis for politicians, using the relative amount of preference votes for a party in 2010 as a comparison measure. As can be seen in Figure 3, the bias with respect to a politician varies irrespective of the party from which the politician comes. For instance, a positive bias towards *Bart De Wever* is not reflected in the (negative) bias of his party (*N-VA*).

It is also interesting to note the differences between different news sources. To this extent we define a matrix, ranking all political parties by coverage per newspaper (Figure 4(a)). The major tendencies are similar to our previous analysis, but some local differences do exist. We use the Hamming distance d_H (Equation 5) to measure the amount of ranking difference for each newspaper, compared to the average ranking (see Table 2, $d_{H,c}$). As the Hamming distance increases, disagreement between the consensus ranking increases. A maximal Hamming distance of 6 is found for regional newspaper *GVA*. When we look at the total bias of news papers (Equation 4), we

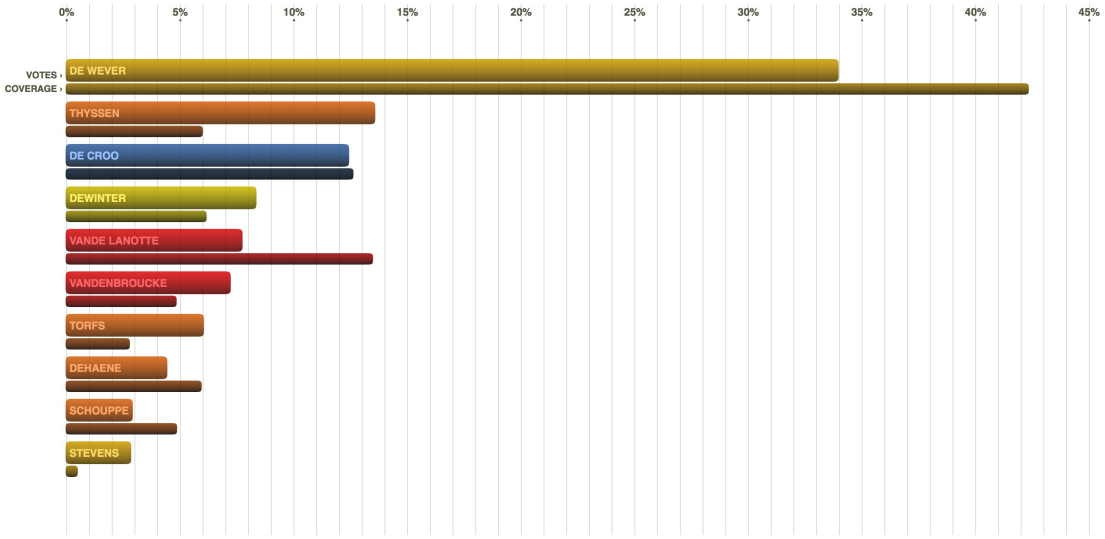


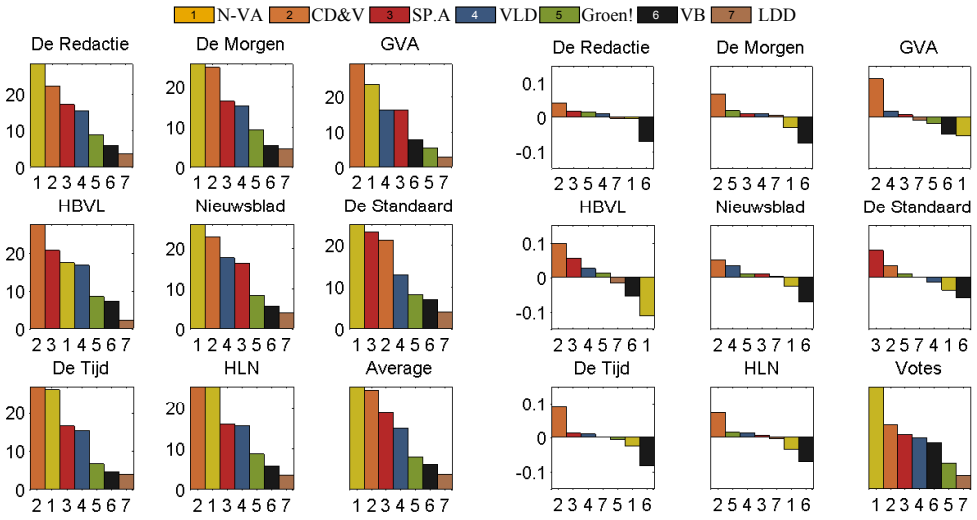
Figure 3: Discrepancy between media coverage and popularity for popular politicians

see that again regional newspapers (*GVA*, *HBVL*) deviate more than global ones. This is also visible in the Hamming distance between coverage and votes (see Table 2, $d_{H,c}$).

$$d_H(v, w) = \sum_{i=1}^{\#\mathcal{E}} \mu(v_i, w_i) \tag{5}$$

$$\mu(a, b) = \begin{cases} 0 & a = b \\ 1 & a \neq b \end{cases} \tag{6}$$

The bias between different parties is not uniformly distributed among all parties (i.e. equal to zero). A more fine grained analysis (Figure 4(b)) shows that general tendencies propagate to the local level (i.e. *Vlaams Belang* is under-represented in all newspapers). Interestingly though, some substantial local differences exist as well. For instance regional newspaper *Het Belang Van Limburg (HBVL)* has a large negative bias towards *N-VA*.



(a) Ranking (b) Bias

Figure 4: Comparison of media coverage of different parties by different newspapers: (a) coverage ranking (b) bias.

3.2 Sentiments

3.2.1 Sentiments by party

For a given political party, we take the distribution of positive versus negative assessments (i.e., adjective polarity scores) as an indicator of the party's overall sentiment in media. Figure 5 shows the distribution for each party. Overall, 30-40% of newspaper coverage is assessed as negative. Some statement bias is present under the assumption of uniformity of the sentiment distribution.

Highest negative scores are measured for the far-right *Vlaams Belang* (which is quarantined by the other parties): -30.4%, and for the *N-VA*: -28.8%. In 2010, the Dutch-speaking, right-wing *N-VA* emerged both as newcomer and largest party of the Belgian federal elections. The second largest party was the French-speaking, left-wing *PS* (founded in 1978). While the *N-VA* ultimately seeks secession of Flanders from Belgium, the PS is inclined towards state interventionism. During the following year they were unable to form a government coalition, which has sparked media controversy. This clash is a possible explanation for the negative score.

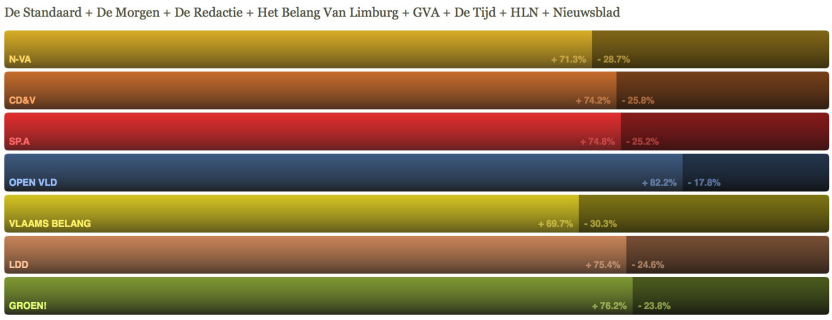


Figure 5: Sentiment for each political party, with the percentage of positive news items on the left and negatives ones on the right.

3.2.2 Evolution of sentiments over time

For a given political party, we group assessments in subsets of one week. We then calculate a simple moving average (SMA) across all weeks to smoothen fluctuation in individual parties and emphasize differences across parties.

Figure 6 shows the SMA of each political party across all newspapers. It is interesting to note the peak with all parties (except *Vlaams Belang*) around July-August. During this time, the negotiating parties (negotiating for a government coalition since 2010) were on a three-week leave. Once negations resumed around August 15th, the peak drops. In the Belgian political crisis, it seems, no news on the political front equals good news.

Figure 7 shows the SMA of each newspaper across political parties. The curves with the highest fluctuation are those for *Het Belang Van Limburg* and *De Redactie*. With these newspapers we measure a standard deviation on the SMA of 0.08 and 0.07 respectively, where other newspapers are in the $[0.03, 0.05]$ range. *Het Belang Van Limburg* also has the highest average sentiment: $+0.15$ against $[0.13, 0.14]$ for all other newspapers. *De Standaard* newspaper appears to deliver the most neutral political articles.

4 Conclusions

We have analysed Flemish newspapers quantitatively during a period of political crisis using a custom built expert system. We have shown that there exists a coverage bias and provide support for the claim that some statement bias exists throughout the mass media. Methodologically we have shown that expert systems are a viable strategy for research in political text corpora.

This proposed text-mining approach based framework can be used as a general political barometer, to compare newspapers with regards to their (relative) reporting style, and can even serve political parties who could follow the news reporting on specific topics and adjust their stands if deemed appropriate. An on-line dashboard that updates the given graphs in real-time will facilitate such uses. In our future research, we envision the application of this framework for American political news, and more specific for the American presidential race. From a methodological point of view, we aim to compare the text mining approach using subjectivity lexicons with a machine learning classification approach, using a labelled dataset. Furthermore, the sentiment analysis could be refined by distinguishing negative content from negative opinion and by adding more classes of sentiment.

To conclude, we hope that further political analysis can benefit from this systematic and unbiased approach.

References

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., & Belyaeva, J. (2010). Sentiment Analysis in the News, .

BBC Europe (2011). *Belgium swears in new government headed by Elio Di Rupo*. http://www.bbc.co.uk/news/world-europe-16042750.

Benewick, R. J., Birch, A. H., Blumler, J. G., & Ewbank, A. (1969). The Floating Voter and The Liberal View of Representation. *Political Studies*, *17*, 177–195.

D'Alessio, D., & Allen, M. (2000). Media bias in presidential elections: a meta-analysis. *Journal of Communication*, *50*, 133–156.

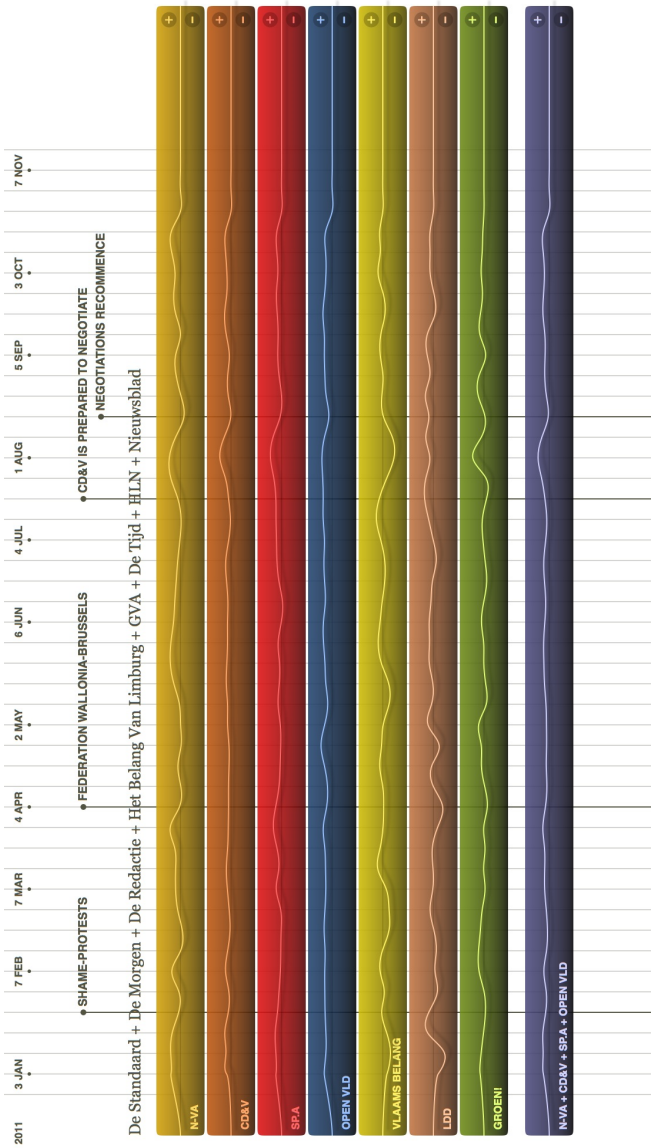


Figure 6: Sentiment of news items in 2011 for each party

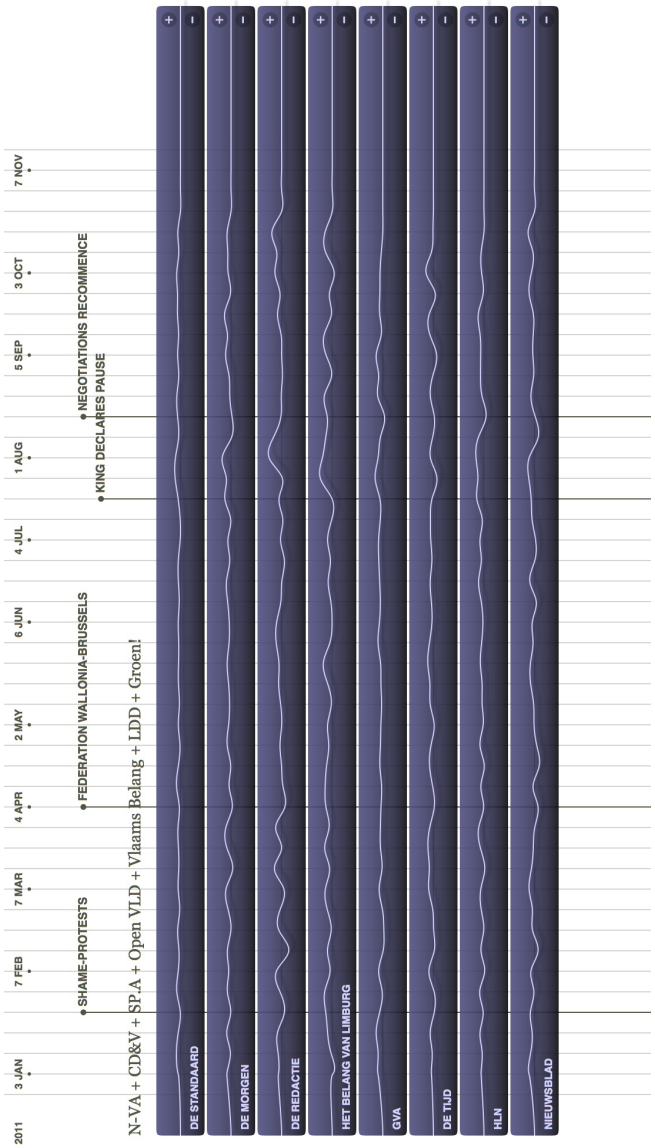


Figure 7: Sentiment of news items in 2011 for each newspaper

Dexter, L., & White, D. (1964). *People, society, and mass communications*. The Free Press of Glencoe, Collier-MacMillan LTD., London.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC* (pp. 417–422). Citeseer volume 6.

Fayyad, U., & Piatetsky-Shapiro, G. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*, 37–54.

Fredricksen, C. (2010). *Time Spent Watching TV Still Tops Internet*. http://www.emarketer.com/blog/index.php/time-spent-watching-tv-tops-internet/.

Godbole, N., & Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs (System Demonstration). In *ICWSM* Icwsm 07 (pp. 1–2).

Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*, (pp. 1–38).

McCombs, M. E., & Shaw, D. L. (1972). The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly*, *36*, 176.

Mihalcea, R. (2011). The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data Ronen Feldman and James Sanger (Bar-Ilan University and ABS Ventures) Cambridge, England: Cambridge University Press, 2007, xii+410 pp; hardbound, ISBN 0-521-83657-3, 70.00. *Computational Linguistics*, *34*, 125–127.

Niven, D. (2003). Objective evidence on media bias: Newspaper coverage of congressional party switchers. *Journalism and Mass Communication Quarterly*, *80*, 311–326.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques.

Rada, R. (2008). Expert systems and evolutionary computing for financial investing: A review. *Expert Systems with Applications*, *34*, 2232–2240.

Savigny, H. (2002). Public Opinion , Political Communication and the Internet. *Political Studies*, *22*, 1–8.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems*, *27*, 1–19.

Suh, J. H., Park, C. H., & Jeon, S. H. (2010). Applying text and data mining techniques to forecasting the trend of petitions filed to e-People. *Expert Systems with Applications*, *37*, 7255–7268.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *1*, 1–41.

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, *36*, 10760–10773.

Yperman, T. (2004). *Van Vlaams Blok naar Vlaams Belang*. Ph.D. thesis University of Ghent.