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

Enric Junqué de Fortuny* Tom De Smedt David Martens Walter Daelemans  
Faculty of Applied Economics Faculty of Arts Faculty of Applied Economics Faculty of Arts  
University of Antwerp University of Antwerp University of Antwerp University of Antwerp  
December 23, 2011

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 analyzed all online news articles of Flemish newspapers. The results of our text mining analyses show interesting differences in media coverage and votes for some 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 even newspaper level. Since all analyses are based on text mining algorithms, a very objective overview of the manner of reporting is provided.

1 Introduction

Belgium has seen a unique governmental crisis during which both political parties and politician figures have had wide media coverage. The media has played an important role in the courses taken by parties (for better or for worse). In this study, we analyze the attention and sentiments of popular media sources with respect to political parties and politicians. We argue that news sources are sentiment-rich resources and extract the sentiments using a technique called sentiment analysis. This gives us an unbiased view of the general tone towards politicians and the political crisis.

2 Methodology

2.1 Data acquisition

Our corpus comprises of all articles published on websites of popular Flemish newspapers during a 10 month period (from January 1, 2011 to October 31, 2011). An overview of all covered newspapers is displayed in Table 1. Some newspapers (i.e. metro.be) were left out due to fact that they did not have an accessible on-line version.

All articles were gathered using a custom built web-crawler. The crawler extracted articles from the sources’ websites using their built-in search functionalities. The keywords of interest are all Flemish party names and leading figures of political parties (see Table 2). The criterion for being a party of interest is based on the votes for that party in the 2010 Chamber Elections, normalized over the Flemish parties. A leading figure is a politician with a top ten ranking amount of preference votes in the 2010 Senate Elections.

2.2 Data processing

A number of filtering steps are performed to clean up and whiten the raw data. First of all, the data is filtered so as to remove possible duplicate articles. Furthermore each article $a$ is converted to a bag of words $(w_1, w_2, ..., w_n)$ representation to allow computational processing. As a last preprocessing step, all stop-words are removed.

*E. Junqué de Fortuny and T. De Smedt contributed equally to this work.
2.3 Sentiment analysis

For sentiment analysis, we used the “Pattern” web mining module for Python\(^1\). The module contains a subjectivity lexicon of 3,000+ Dutch adjectives that occur frequently in product reviews, 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) = +0.9 and *belabberd* (lousy) = -0.6. The lexicon has been evaluated using 2,000 Dutch book reviews with a precision of .72 and a recall of .82 (De Smedt & Daelemans, submitted).

In each newspaper article, we look for occurrences of a Flemish political party. We then calculate the polarity of each adjective that occurs in a window of 2 sentences before and 2 sentences after. An article can mention several party names, or switch tone. The given interval ensures a more reliable correlation between the political party being mentioned and the adjective’s polarity score contrary to measuring all adjectives 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 name.

3 Results

3.1 Media coverage deviation

The coverage \( c(e, s) \) of an entity \( e \) by a newspaper \( s \) is defined as the amount of news articles published by the newspaper on that entity, normalized on the total amount of articles in the corpus \( A_s \). 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 \( E \)). The deviation of a media source is the difference between coverage and popularity. That is:

\[
\begin{align*}
c(e, s) &= \frac{\# \{ a | a \in A_s \land e \in a \}}{\# A_s} \quad (1) \\
p(e) &= \frac{\sum_{e' \in E} v(e')}{} \quad (2) \\
dev(e, s) &= c(e, s) - p(e) \quad (3) \\
dev(s) &= \sum_{e \in E} dev(e, s) \quad (4)
\end{align*}
\]

\(^1\text{http://www.clips.ua.ac.be/pages/pattern}\)

\(^2\text{Counted by the amount of readers of the printed version except for "De Redactie" which does not exist in a printed format. Instead, we used the number of unique visitors per day in 2009 as an estimation. Source: belga/odbs}\)

\(^3\text{Source: Federal Public Services Home Affairs (http://polling2010.belgium.be/)}\)
Figure 1: Discrepancy between media coverage and popularity for popular parties

where $a$ is a 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. For high values, the deviation can be an indicator for discrepancies between popularity in media and popularity by votes. Figure 1 shows that for some parties a significant deviation is found, with a maximal positive deviation towards CD&V and a maximal negative deviation towards Vlaams Belang (VB, Flemish Interest). This is in accordance with the fact that CD&V ran the interim government while the new government formations took place.

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 2, the deviation with respect to a politician varies irrespective of the party from which the politician comes. For instance, a positive deviation towards Bart De Wever is not reflected in the (negative) deviation of his party (N-VA). This implies either an underlying person cult or a negative coverage of all other party members, depending on causality.

It is also interesting to note the differences between different news source. To this extent we define a matrix, ranking all political parties by coverage per newspaper (Figure 3(a)). The major tendencies are similar to our

Figure 2: Discrepancy between media coverage and popularity for popular politicians
previous analysis, but some local differences do exist. We use the Hamming distance (Equation 5) to measure the amount of ranking difference for each newspaper, compared to the average ranking (see Table 1). As the Hamming distance increases, disagreement between the consensus ranking increases. A maximal hamming distance of 4 is found for regional newspaper GVA and Nieuwsblad. When we look at the total deviation of newspapers (Equation 4), we see the same pattern emerge (regional newspapers deviate more than global ones).

\[
d_H(v, w) = \sum_{i=1}^{\#E} \mu(v_i, w_i) \\
\mu(a, b) = \begin{cases} 
0 & a = b \\
1 & a \neq b
\end{cases}
\]

As we stated earlier, it follows from Figure 1 that the deviation between different parties is not uniformly distributed among all parties (i.e., equal to zero). A more fine grained analysis (Figure 3(b)) shows that general tendencies propagate to the local level (i.e., Vlaams Belang is under-represented in all newspapers). Interestingly though, significant local differences exist as well. For instance regional newspaper Het Belang Van Limburg has a large negative deviation towards N-VA.

### 3.2 Sentiments

#### 3.2.1 Sentiments by party

For a given political party, we take the distribution of positive vs. negative assessments (i.e., adjective polarity scores) as an indicator of the party’s overall sentiment in media. Figure 3.2.1 shows the distribution for each party. Overall, 30-40% of newspaper coverage is assessed as negative.

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. 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. This has sparked media controversy, a possible explanation for the score.
3.2.2 Evolution of sentiments in time

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

Figure 5(a) 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 negotiations resumed around August 15th, the peak drops.

Figure 5(b) 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. We have shown that there exists a deviation from popularity to media coverage for both political parties and their influential figures. The sentiment analysis results further provided a graphical overview of the tone of reporting throughout this period, where interesting changes are observed at key moments in the negotiations.
Figure 5: Sentiment of news items during 2011, for each party (a), and for each newspaper (b).