Online Information Review

Utilizing overtly political texts for fully automatic evaluation of political leaning of online news websites

Maayan Zhitomirsky-Geffet Esther David Moshe Koppel Hodaya Uzan

Received 25 June 2015
Revised 8 December 2015
Accepted 18 December 2015

Utilizing overtly political texts for fully automatic evaluation of political leaning of online news websites

Maayan Zhitomirsky-Geffet
Department of Information Science, Bar-Ilan University, Ramat Gan, Israel
Esther David
Department of Computer Science, Ashkelon Academic College, Ashkelon, Israel, and
Moshe Koppel and Hodaya Uzan
Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel

Abstract

Purpose – Reliability and political bias of mass media has been a controversial topic in the literature. The purpose of this paper is to propose and implement a methodology for fully automatic evaluation of the political tendency of the written media on the web, which does not rely on subjective human judgments.
Design/methodology/approach – The underlying idea is to base the evaluation on fully automatic comparison of the texts of articles on different news websites to the overtly political texts with known political orientation. The authors also apply an alternative approach for evaluation of political tendency based on wisdom of the crowds.
Findings – The authors found that the learnt classifier can accurately distinguish between self-declared left and right news sites. Furthermore, news sites' political tendencies can be identified by automatic classifier learnt from manifestly political texts without recourse to any manually tagged data. The authors also show a high correlation between readers' perception (as a "wisdom of crowds" evaluation) of the bias and the classifier results for different news sites.
Social implications – The results are quite promising and can put an end to the never ending dispute on the reliability and bias of the press.
Originality/value – This paper proposes and implements a new approach for fully automatic (independent of human opinion/assessment) identification of political bias of news sites by their texts.

Keywords Automatic text categorization, Mass media, Objective evaluation, Online news sites, Political bias, Supervised machine learning

Paper type Research paper

Introduction

Digital newspapers and other mass media in democratic countries do not officially declare their political preferences. Nevertheless, most people do associate digital newspapers with a certain political tendency. Thus, numerous studies show that the public perceives the US media as imbalanced with a significant decline of trust in the past few decades (Schneider and Lewis, 1985; Ladd, 2010; Mak, 2011). Recent studies argue that media has the power to influence the public opinion by framing controversial political issues (Kuypers, 2002; Petrova, 2008). Furthermore, mass media might have a significant impact on voters' decisions during elections in democratic regimes (Ladd, 2010; Petrova, 2011). "Mass media provides a convenient means for manipulating public opinion, even when voters understand that the media can be biased" (Petrova, 2011).

Online Information Review
Vol. 40 No. 3, 2016
pp. 362-379

1468-4527
DOI 10.1108/OIR-06-2015-0211

Thus, identifying the political leaning of media has become an extensively explored topic for researchers, journalists and the wide public (D'Allesio and Allen, 2000; Gentzkow *et al.*, 2015; Groseclose and Milyo, 2005; Schneider and Lewis, 1985; Weatherly *et al.*, 2007). In these works political bias of media was measured by manual content analysis (coverage of topics specific to certain political wings, citations of politicians, selected politically charged term counting and sentiment/tone of coverage analysis of politicians' presentation). However, human-based evaluation of text by the above criteria is highly time-consuming and expensive and thus typically unavailable. In addition, human-based evaluation might be subjective and biased by the opinion of the assessors.

Therefore, in this paper we propose and implement a new methodology to objectively evaluate the political bias of the written media by automatic text categorization methods based solely on linguistic and statistical properties of the text. To this end, we apply a state-of-the-art supervised text categorization algorithm, which is trained on a sample of labeled data with known political tendency. Then, the learnt model is employed to automatically classify new unlabeled texts for which the political orientation is not known.

To ensure that the evaluation would be fully automatic, we decided to use overtly political texts with known political tendencies, such as parliamentary speeches of different politicians and Facebook pages of diverse political parties, as a training dataset for the algorithm. Thus, we experimented with three different text genres, all collected from Israeli websites and written primarily in Hebrew: parliamentary speeches and political parties' Facebook pages, each consisting of texts easily labeled for political orientation, as well as news sites, which make certain claims of political neutrality. In addition, we wish to check whether and to what extent the aggregated readers' perception of political bias of the news sites (as "wisdom of crowds") is similar to the automatic classification results. To this end, a survey has been arranged with a balanced group of subjects, readers of the news sites.

The significance of this research is that it introduces a method for objective evaluation of the written media on the web. This approach might lead to the resolution and end of the longitudinal dispute between readers, journalists, critiques, media watchdogs and researchers, who base their opinion on subjective human-coded evaluation. The proposed automatic method, based on standard text categorization techniques and freely available overtly political corpora, serves as a more objective and accurate method for political bias identification than human-coded evaluation. Its additional benefit is that it can be easily applied in cases when there is no available human-labeled data. From a multi-disciplinary perspective, the proposed methodology can contribute to methods in the field of media studies, mass communication, social and political sciences, in receiving a critical perspective of the different media sources and reduce the need in human-coded content analyses. Finally, if adopted by the mass media producers, it might help editors in becoming more balanced and regaining the broad public's respect.

The outline of this paper is as follows. In the next section, we review the related literature. Then, we present the applied methods. Our findings are presented in the results section. Finally, the discussion and conclusion section provides discussion and analysis of the results, their implications and main conclusions.

Literature review

In this section we mainly review works that propose various measures of political bias for mass media based on human judgments and also studies on automatic political profiling of overtly political texts.

As mentioned in the introduction, numerous studies have proposed measures for political bias of media. These measures are based on manual content analysis. For example, research examined whether media bias exists during presidential campaigns (D'Allesio and Allen, 2000). The following types of bias were defined: preference for selecting stories from one party or the other, the relative amount of coverage each party receives, favorability in coverage toward one party or the other. On the whole, no significant biases were found in the newspaper industry. Similar variables were examined in the research by Pew Research Center's Journalism Project Staff (2007). However, their findings for the elections of 2008 revealed that democrats generally received more coverage than republicans, (49 percent of the stories vs 31 percent). Overall, democrats also received more positive coverage than republicans (35 percent of the stories vs 26 percent), while republicans received more negative coverage than democrats (35 vs 26 percent). For both parties, a plurality of stories, 39 percent, were neutral or balanced.

In another study a new measure of media political bias was proposed, which is a comparison of the number of citations of various think tanks or political groups in the newspapers to those of different politicians – congressmen labeled as democrat or republican (Groseclose and Milyo, 2005). All the examined news sources excluding Fox News' *Special Report* and *Washington Times* were found strongly liberally biased, more than the average congressmen. In a later research, a user study was conducted to explore whether a perception of bias would be found in the headlines of lead or major stories taken from the websites of two major American news organizations, CNN and Fox News, during the final two months of the 2004 presidential campaign (Weatherly *et al.*, 2007). Significant perceptions of bias were found. Overall, headlines taken from CNN were rated as significantly more liberal than those taken from Fox News. Another recent work checked parameters such as newspaper entry and exit, prices and circulation, readership, patronage such as the allocation of lucrative government jobs to newspaper editors, and the number of politicians from each of the wings mentioned in US newspapers from 1869 to 1928 (Gentzkow *et al.*, 2015). In general no bias was found toward the ruling party (except in reference to the issue of reconstruction of the south).

As can be observed the findings of these studies are quite controversial. This inconsistency might be attributed to the fact that they are based on different specified terms particularly picked by humans as politically charged (rather than considering all the terms in the corpus) or on subjective human-based evaluation of texts' sentiment. Thus, as opposed to the above studies, in this paper we propose and implement a new methodology to objectively evaluate the political bias of the media by automatic text categorization methods based solely on statistical properties of the text.

Numerous studies have been performed in the area of automatic recognition of an author's demographic profile. Text categorization methods have been used to identify an anonymous author's gender (Argamon *et al.*, 2003; Burger *et al.*, 2011; Filippova, 2012), age (Koppel *et al.*, 2006), native language (Koppel *et al.*, 2005) and personality (Pennebaker *et al.*, 2003). It has been shown that such demographic profiling can also be done on personal Facebook pages (Otterbacher, 2010; Popescu and Grefenstette, 2010; Gosling *et al.*, 2011). A survey of automated demographic profiling is presented in Argamon *et al.* (2009).

The application of these methods for the determination of political orientation is especially challenging. First of all, unlike demographic characteristics, an individual's political orientation may vary over time and is often complex and thus not easily captured by a single simplistic label such as left or right. Furthermore, conventions of

public expression often dictate that political views are stated in a subtle manner, if at all. A number of papers (Laver *et al.*, 2003; Grefenstette *et al.*, 2004; Efron, 2004; Lin *et al.*, 2006; Mullen and Malouf, 2006; Hassanali and Hatzivassiloglou, 2010) have considered the automatic identification of political tendency for overtly political documents, such as political blogs. A variety of studies have applied supervised learning for automatic perspective recognition of politically charged texts. For example, articles from the Bitter-Lemons website on the Palestinian-Israeli conflict were classified using lexical features (Lin *et al.*, 2006). The same corpus and three other politically polarized corpora were analyzed in Beigman-Klebanov *et al.* (2010) and it was shown that binary features are not less effective than frequency-based features. Similarly, articles taken from corpora concerning abortion and gun-rights (Hasan and Ng, 2012) and US congressional speeches (Yu *et al.*, 2008) were successfully classified according to party affiliation.

In general, these studies deal with texts in a manifestly political corpus in which labeled texts are relatively easy to find. Their methods are based on training with labeled data where the "ground truth" political orientation for each text is pre-supplied. As a result of the training process the optimal classifier is learned, which then can be applied to classify new texts. Such labeled data can be easily induced for overtly political texts, but typically there is no available politically labeled data for news sites most of which are self-declared as politically neutral. In addition, human-labeled data might be biased and rather subjective.

In this study, we endeavored to classify texts in genre of news websites without the use of labeled samples according to political tendency. The primary research question explored in this study is:

RQ1. Whether the classifiers learned on overtly political texts are useful for discerning political biases of news sources.

Our goal is to show that, using linguistic choices alone, it is possible to fully automatically identify the political orientation of news sources. To this end, we proposed to draw training data from other genres in which documents are easy to label. An initial step in this direction was proposed in Gentzkow and Shapiro (2010) and Groseclose and Milyo (2005) who identified a newspaper's political slant by measuring the similarity of its manually marked political lexicon to that of congressional republicans and democrats. Particularly, they compared the counts of the manually selected politically charged phrases in the congressional speeches to the frequencies in the media texts. We took this approach to the next level, where we used fully automated text categorization methods based on the entire corpus vocabulary rather than on manual word analysis and selection of manifestly political lexicon. As opposed to this approach, we do not consider some terms as more characteristic of political bias than others, but rather use the entire text and let the algorithm find the most distinguishing terms. Accordingly, our second research question is:

RQ2. Whether the most distinguishing terms for political bias in different corpora are in the semantic field of politics?

To answer this question we conducted a thorough analysis of such terms found by the algorithm.

Methods

In this section, we describe the proposed methodology for automatic identification of political bias in the digital news sites. First, we introduce the used corpora.

Corpora

The texts we considered are Hebrew texts written by Israelis. This presents a number of challenges and opportunities specific to this linguistic and political context. Since we used only lexical features, the morphological quirks of Hebrew did not present any special challenges. However, Israel's purely proportional single-region parliamentary election system provided an interesting opportunity. Unlike winner-take-all regional elections, which typically result in only two major parties, there are many medium-sized parties in Israel. While each of these parties can rather easily be identified as left, right or center, the parties differ widely in terms of the demographic group to which they appeal. In particular, because there are a number of self-declared centrist parties, we considered two-class (right/left) experiments, as well as three-class (right/center/left) experiments. We also explored which self-identified centrist voters are closer to the right and which are closer to the left.

In this work, we considered the following three Hebrew corpora:

(1) Posts on the Facebook pages of nine major Israeli political parties. Each party is labeled as left/center/right, with three parties assigned to each category. In particular, for the right-wing we considered the pages of "Habayit Hayehudi," "Likud-Beiteinu," and "Otzma le-Israel." For the center wing we considered the pages of Yesh Atid, Kadima, and Hatnua. Finally, for the left-wing we considered pages of "Haavoda," "Meretz" and "Hadash." While the assignments to categories are uncontroversial, the parties in each category are diverse in terms of their demographic appeal. The corpus consists of 646 posts, including over 550,000 words (229 posts for the right-wing, 208 posts for the left-wing and 209 posts for the center wing). For our study, chronologically consecutive posts were concatenated until they exceeded 1,000 words in aggregate. Party names and names of party leaders were omitted.

(2) Transcripts of all Israeli parliament members' speeches during a six-month period in 2011. This corpus includes 119 documents (one for each of the 120 members of Israel's parliament (Knesset), except one member who did not give any speeches during that period of time). The documents include over 1.6 million words. Each speaker belongs to a political party and is assigned to a category accordingly. Further, the splitting process separated the speeches into numerous chunks in such a way that each chunk contained speech quotes from a single speaker (of some Knesset member), while the original text contained multiple speakers. In addition, irrelevant information that appeared in the texts was removed. The texts in this corpus were assigned their political label according to the political opinion of the speaker (assuming the three categories of right, left, and center). After the splitting process, the corpus contained about 100-labeled texts for each category with an average of 6,000 words per text. As a result, 47 percent of the articles were labeled right-wing, 26 percent left-wing, and 27 percent centrist.

(3) News stories from the five most popular Israeli news websites (*Haaretz*, Walla, Ynet, NRG, Arutz 7) during a four-month period in 2011. This corpus contains about 3,800 articles including over 860,000 words. Chronologically consecutive stories were concatenated until they contained 500 words in aggregate. While two of the five news sites have a known self-declared political leaning, i.e., *Haaretz* is associated with the left-wing of the Israeli political spectrum, and Arutz 7 is the voice of the right-wing, the political leaning of the other three sites, Walla, Ynet and NRG, is disputable, and no "ground truth" judgment exist for them.

Experiments

To begin with we introduce the basic concepts from text categorization that we used in the experiments. First, each text in a set of labeled example texts is represented as a numerical vector reflecting the frequencies in the text of each feature (term or phrase) in a specified feature set. A machine learning algorithm is then used to learn a classifier that best distinguishes between training examples in different classes. These classifiers can then be used to classify new texts. The effectiveness of this method can be evaluated by applying a learned classifier to labeled test texts for which the correct answer is given. An evaluation method is that of k-fold cross-validation. We divide the training set into k roughly equal parts, train on $k-1$ parts and test on the holdout set, repeating this k times with a different part held out each time.

In this context, we performed all of the following experiments:

- For each of the two overtly political corpora (the parties and parliamentary speeches), we performed ten-fold cross-validation experiments to determine the accuracy with which we can train a political preference (right, center, left) classifier for a given genre. In these experiments the classifiers were trained and tested on ten different subsets of the same corpus. The goal of classifying these corpora is to assess their reliability. Thus, if we are able to classify such manifestly political texts with high accuracy then they can be used as a good source ("truth ground") for evaluation of the news sites' bias.
- We trained classifiers on training data in political corpora (speeches and party pages used jointly) and checked their effectiveness for classifying news articles according to the political bias.
- We used learned two-class (left/right) political preference classifiers to determine if news websites with disputable political leaning, and also self-identified centrist parties and parliament members are closer to the left or to the right.
- We conducted a reader survey to assess the "wisdom of crowds" evaluation. The results were analyzed and compared to the automatic classifier's evaluation.

In all these experiments each text was represented as a numerical vector (histogram) of 12,000 features encoding the frequency in the text of each of the 10,000 most common unigrams (single words) and 2,000 most common word bigrams (consisting of two consecutive words) in the corpus. For each experiment we only use features that appeared in the relevant corpus at least three times. Except in the case of k-fold cross-validation experiments, we used only the features for which frequency differences between classes were significant (t-test at $p < 0.05$) in the training examples. We used sequential minimal optimization (SMO) (Platt, 1999), an efficient implementation of support vector machine (SVM) (Joachims, 2002), a state-of-the-art machine learning algorithm. We used Weka system (Hall *et al.*, 2009) implementation for John Platt's SMO algorithm with default coefficients' value set, for training a support vector classifier with PolyKernel, epsilon of 1.0E-12, one seed, the complexity constant of 1 and tolerance value of 0.001.

Results

Classifying the overtly political corpora

In our first experiment, for each of the two corpora individually we considered the accuracy with which we were able to classify out-of-sample examples as having left or right political orientation. Thus, in these experiments for each corpus separately the

classifier was learned on a subset of texts and then applied on different texts from the same corpus. This was repeated ten times for different subsets of texts (as part of ten-fold cross-validation). Ground truth in each case is as described in the methods section above.

As noted, our feature set consists of all word unigrams and bigrams that appear in the corpus at least three times and we used SMO as our learning method. Results of ten-fold cross-validation experiments of each of the overtly political corpora are shown in Figure 2. As can be seen, the result accuracy for each case exceeds 90 percent (Figure 1).

For the three-class experiment, our feature set consisted of all word unigrams and bigrams that appear in the corpus at least three times and we used SMO as our learning method. Results of ten-fold cross-validation experiments of each of the overtly political corpora also exceed 90 percent as shown in Figure 2. These results demonstrate the ability of automatic text categorization methods to accurately predict the fine-grained (three-class) political orientation of texts in the manifestly political corpora. Thus, the corpora selected for evaluation of the news sites' bias were proven to be an easy case for political classification and thus can be considered reliable. We note that a slightly lower performance for the Parliament protocols can be explained by the higher complexity of automatic processing of spoken language transcriptions than processing of written language in Facebook party pages.

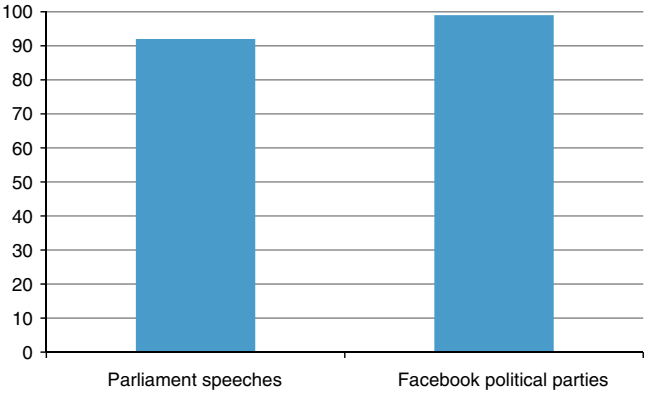

Figure 1.
Average accuracy in ten-fold cross-validation of each of the two corpora

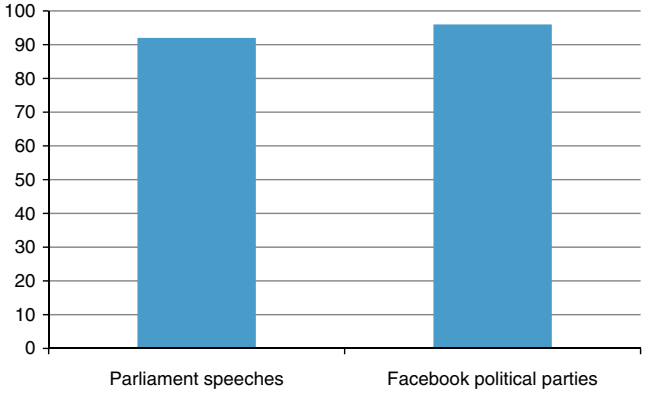

Figure 2.
Three-class accuracy results for the respective corpora

Distinguishing features

Consideration of the main distinguishing features (terms) for each experiment (as measured by student's *t*-test) yields insight into why successful classification is possible for each corpus. In general, we found that across all corpora, texts associated with the left are characterized by more frequent use of terms related to social protest, rights and minorities, as well as female pronouns and third-person pronouns. On the other hand, texts associated with the right are characterized by more frequent use of terms reflecting positive attitudes, references to religion, patriotism and use of first-person pronouns. The top features for each corpus are presented in Table I.

Speeches by members of right-wing parties are characterized by frequent mention of various stages in the legislation process (discussion, law, voting). This likely reflects that the governing coalition at the time consisted primarily of right-wing parties. In addition, speeches by members of left-wing parties include significantly more citation of particular political terms (rights, struggle, social, occupation, Arabs), as well as significantly more use of female third-person pronouns (she proposed, her claim). On the other hand, members of right-wing parties make significantly more frequent use of terms reflecting positive attitudes (dear, my friend, please, good) and religion (Jewish, Sabbath, God).

Right-wing party posts significantly more frequently note religious concepts (Rabbi, Torah, God, Sabbath) and positive attitudes (beloved, love you, be strong), while left-wing party posts significantly more frequently mention particular politically loaded terms (rights, struggle, social, two states, refugees, Palestinians) and female third-person pronouns.

Note that all mentions of "significant" differences in the above feature analysis are where $p > 0.05$.

In the news sites' articles a considerable number of ideologically-loaded terms are prominent (e.g. Judea and Samaria for the right, vs territories for the left). Similarly, references to religious concepts and patriotism (Zionist, to the Jews) are significantly more frequent in right-wing news stories, while certain politically loaded issues such as social protest, minorities, rights and peace process are significantly more frequent in left-wing news stories. Interestingly, overall prominent features on right-wing news sites there is a significantly more frequent use of first-person pronouns (our country, I feel), while left-wing news sites significantly more frequently contain third-person pronouns.

One of the most important finding of our feature analysis concerns pronoun use. The importance of pronouns as markers of attitude is generally well-attested (Chung and Pennebaker, 2007). In our study, we found an unusually strong correlation between pronoun use and political orientation. In Tables II-IV, for each of the three corpora we show the frequency (as a percent of total word occurrences) of each of first-person and third-person pronouns, including both singular and plural forms, in left-wing and right-wing documents.

Remarkably, we found that across all the corpora except the party Facebook pages, first-person pronouns are used more frequently in texts labeled as right-wing than in those labeled as left-wing, while third-person pronouns are used more frequently in texts labeled as left-wing than in those labeled as right-wing. All the differences are significant at $p < 0.05$ (apart from third-person pronouns in the news sites, with $p < 0.001$). In the case of party Facebook pages, the first-person vs third-person distinction is detracted by another difference: right-wing parties use significantly more plural pronouns of both types than left-wing parties and left-wing parties use

Right-wing features	Left-wing features
Party Facebook pages	
The Jewish Nation	She
The country of Israel	The only
Beloved	In her
Torah	Laws
Us	Alternative
Army	Struggle
Like you	Change
Be strong and brave	Agreement
God	Citizens
The religious	State
Love you	Rights
Zionism	Refugees
With G-d's help	Wages
Golan heights	Social
Me too	Arabs
We	Palestinians
Values	Annul
Amen	Democracy
Soldiers	Two states
Parliamentary speeches	
Bills	Her
With you	She
My friend	Arabs
We	Theirs
Knesset (Parliament) chairman	Palestinian
Hamas	Hurtful
To vote	State
I	In Israel
Law	Protest
The country of Israel	Citizenship
Committee chairman	Anti
Determines	Not only
Responds	But she
Opposition	Struggle
I can	Democracy
Please	Social
Legislation	Occupation
The committee	Israeli
Discuss	She cannot
Dear	Israeliness
News sites	
The state	Decided
Knesset member	Was not
We	Citizens
In Judea and Samaria	Politics
The left	Israeli Defense Forces (IDF)
Samaria	Social protest
Palestinians	To war

Table I.
The top distinguishing features for each political wing for different corpora

(*continued*)

Right-wing features	Left-wing features
Settlements	Security forces
State of Israel	Palestinian
Ariel	Occupation
Zionist	Prime minister
Honor	Investigation
Indeed	Demonstrations
The Rabbi's	Officer
Judea, Samaria and Gaza	Time
To the state	The events
To the Jews	Israel
To listen	Egyptians
Judea and Samaria	Peace with

Note: All of the expressions in the table are bigrams in Hebrew even when their English translation consists of more than two words

Table I.

Table II. Frequency of pronouns as distinguishing features per 1,000 words averaged by number of documents in Facebook party pages corpus

	Right	Left
First person	9.0	9.0
Third person	12.7	*13.3*

Table III. Frequency of pronouns as distinguishing features per 1,000 words averaged by number of documents in parliamentary speeches corpus

	Right	Left
First person	*20.1*	17.6
Third person	13.6	*16.7*

Table IV. Frequency of pronouns as distinguishing features per 1,000 words averaged by number of documents in news sites corpus

	Right	Left
First person	*4.1*	2.6
Third person	12.2	*13.0*

significantly more singular pronouns than right-wing parties. These findings can be explained by the fact that the government was led by the right-wing parties at the time when these corpora were created, thus the right-wing texts mostly contain first-person pronouns that reflect the rhetoric of the decision makers. On the other hand, the left-wing parties and voters tend more to criticize the right-wing's actions and decisions, thus utilizing third-person pronouns.

At the training phase SVM learning algorithm uses the feature vectors of the training set texts (the parliament speeches and political party pages) to learn an optimal function (classifier), which finds the boundary between left-wing, centrist and right-wing text samples. Then, the learnt classifier is applied to classify feature vectors of the new texts (the news site articles). The more similar the features of the new texts are to those of the trained sample texts, the easier it is for the classifier to classify them more accurately. Thus, these similar sets of prominent features found in different corpora (as shown in the above analysis) provide the basis for effective classification of news site texts with a classifier trained on the parliament speeches and political party pages.

Classifying the news sites stories: learning across corpora

As shown by the above feature analysis, similar features are representative of the same political tendencies in different corpora. Thus, we further used the automatic classifier trained on a combination of party pages and parliamentary speeches to classify individual newspaper stories as right-wing or left-wing.

Haaretz (a left-wing self-declared site) was identified as such by our classifier (63 percent left-biased texts), the Arutz 7 site (a right-wing self-identified site) was clearly classed as right-biased (61 percent). Thus, we conclude that our classifier trained on the overtly political corpus correctly distinguishes between the right and left-leaning news sites. The other sites (without self-declared tendency) were identified as somewhat biased to the left (58-73 percent) as shown in Figure 3.

Centrist texts

Furthermore, for each of the respective corpora we wanted to determine to which of the two political wings self-identified centrists are more similar. To this end, we reverted to

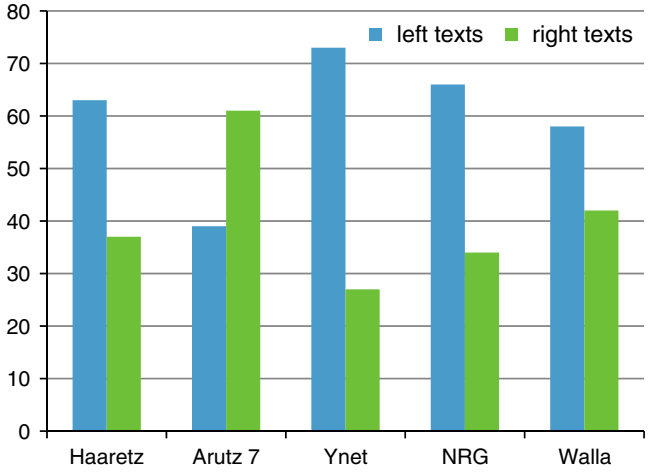

Figure 3.
The percentages of right and left texts as evaluated by the automatic classifier trained on the overtly political corpora

the two-class (left/right) classifiers for the respective corpora and used them to determine the percentage of centrists assigned to each class. The results are shown in Figure 4. It is apparent that for the Parliamentary speeches the centrist texts lean slightly more to the right, while slightly over 60 percent of the posts of centrist parties on Facebook were classified as right. This phenomenon might be interpreted as an attempt to attract more right-wing voters. However, the preponderance of articles in the news site with undeclared and thus deemed "centrist" political orientation are classified as closer to the left-wing texts than to the right-wing texts.

This result can be explained by the relatively high resemblance between the most characteristic features in all the corpora as shown next.

The "wisdom of crowds" evaluation of political leaning of the news sites

An alternative method for evaluation of political stance of news sites is based on the "wisdom of crowds." To this end, 101 random Israelis (ranging in age, place of residence, gender and political views) were asked to assess the percentage of the articles on each of the websites biased to the right and those biased to the left (as two separate questions). The readers were guided that the two percentages do not have to sum up to 100 percent, in case that they assess the rest of the articles on a website as neutral/unbiased. The average age of the participants was 32; 26 percent of them self-identified as left-wing voters, 31 percent were centrist, and 43 percent voted for the right-wing. These percentages are approximately similar to those of the corresponding political wings in the Israeli Parliament.

The average (over all the 101 readers) results of the survey are depicted in Figure 5.

We also calculated and compared the average scores (percentages) assigned by the readers from different political wings. Interestingly, as can be observed from Figures 6-8, there were significant differences (by student t-test at $p < 0.01$) between the right-wing readers and the other two groups. The right-wing voters judged most of the news sites (except for Arutz 7) as more biased to the left than the centrist and left-wing readers. The centrist and left-wing readers' evaluation was quite close to each other for most of the sites. Despite the differences there was a consensus between all the groups regarding the sites *Haaretz* and Ynet (that are clearly leaning to the left) and Arutz 7 (which is clearly leaning to the right). Two other sites received controversial judgments by different reader groups: left-biased by the right-wing readers and right-biased by the left-wing readers. A comparison to the automatic classifier's results reveals that for the three consensus news sites, the classifier similarly determined the

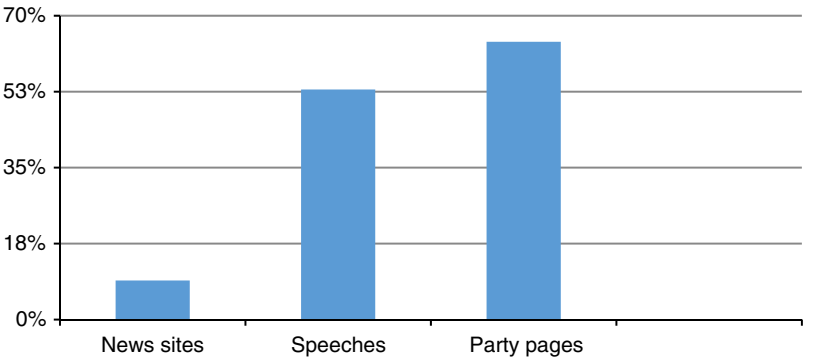

Figure 4.
The number of centrist text classified as right for the three individual corpora

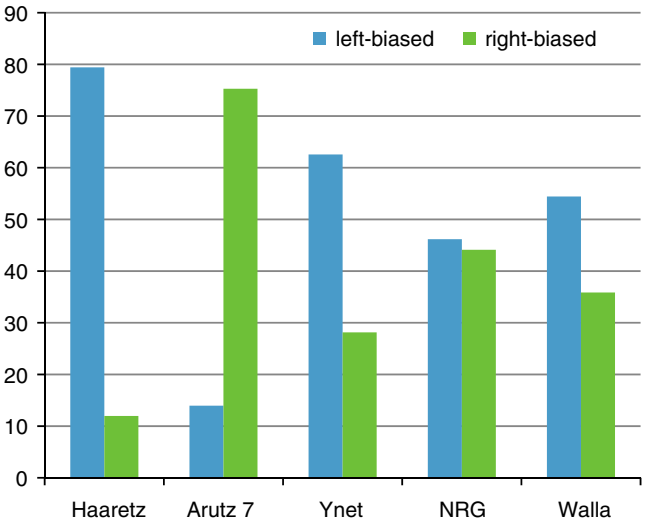

Figure 5.
The average percentage of articles on each of the news site biased to a certain political wing (left or right)

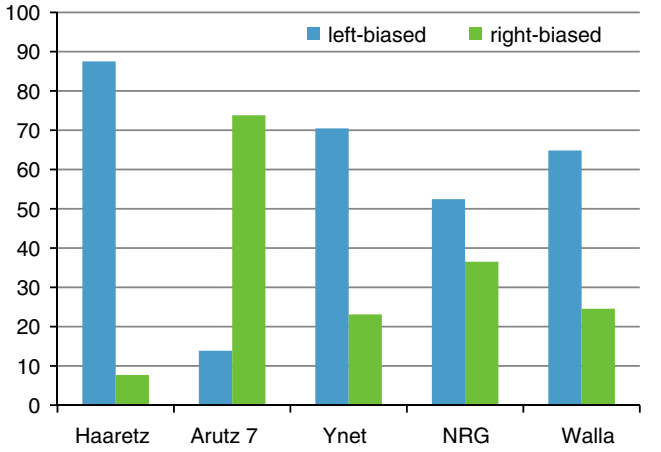

Figure 6.
The average percentage of articles on each of the news site biased to a certain political wing (left or right) according to the right-wing voters

same clear bias. For the two controversial sites (NRG and Walla) the classifier found the majority of their texts as leaning to the left. This tendency is generally consistent with the average reader decision (but with smaller differences between the percentages for the readers than the classifier).

Finally, there is a high correlation (with a Pearson coefficient of 0.75) between the aggregated readers' perception ("wisdom of crowds") and the classification of individual articles by the learned classifier trained on easily assembled inherently tagged data, like party pages and parliamentary speeches.

Discussion and conclusions

Evaluation of political bias of the media has been widely researched and is considered a controversial topic. In this paper we propose a new approach for fully automatic

evaluation of the political bias of the online news websites by their texts. We conducted a multi-disciplinary study that combines state-of-the-art approaches in political studies with advanced computer science tools. The proposed methodology was implemented and attested on the case of five popular news sites in Israel.

In particular, we revealed that the same automatic text categorization methods can be used effectively in each of the three different genres of varying degrees of political expressiveness. Specifically, we found that training and testing in the same corpus (genre) yields strong results (over 90 percent) even for three-class classification (right/center/left). Similar politically charged but also non-political textual features (e.g. pronouns and positive attitudes) are representative of each of the political wings in all the corpora. Therefore, automatic classifiers trained on the overtly political texts (e.g. party Facebook pages and parliamentary speeches) can be effectively applied to determine news sources' political orientation. In addition, the aggregated readers'

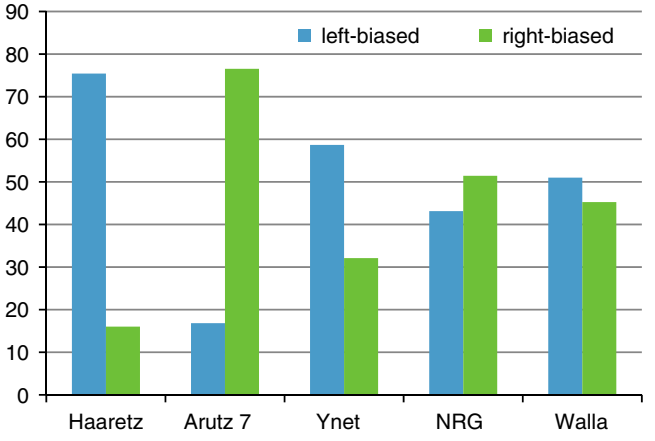

Figure 7. The average percentage of articles on each of the news site biased to a certain political wing (left or right) according to the centrist voters

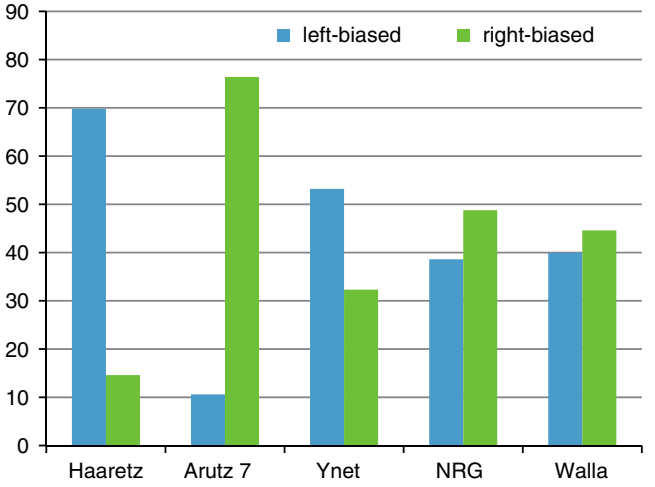

Figure 8. The average percentage of articles on each of the news site biased to a certain political wing (left or right) according to the left-wing voters

perception of the political bias of news sites is generally confirmed by the automatic classifier. Moreover, our findings show that while most politicians and parties in Israel use right-wing rhetoric, Israeli media clearly leans to the left. A possible interpretation of this behavior is that the press reflects an anti-power bias, which happens to be a left bias with a right-wing government in power.

Ultimately the validity of this work rests on corpora chosen as "ground truth" for training the classifier. We argue that texts produced by politicians and political parties seem to be the optimal choice for this purpose. This choice was also supported by our experimental results. However, it should be noted that the experiments of this study were limited to Israel and to the specific period of time with certain characteristics, power balance and players in the political arena. As political situation in the modern world might change dynamically, to generalize the proposed methods for additional political systems, periods and situations, the overtly political "ground truth" corpora should be constantly updated to contain texts of the current leaders and parties.

In summary, the proposed fully automatic approach for evaluation of mass media solely by its texts supported by the above findings is of great importance for researchers of political and information studies and mass communication, editors and journalists, politicians and for the broad public. It provides an objective critical perspective of the different media sources with no need in manual content analyses. As media has significant influence on the election results and many other democratic processes, the more balanced media can better serve the democracy of a state and regain the trust of the public.

The novel methodology proposed in this study is not limited to the digital news sites. It can be employed to analyze and evaluate political bias in different types of media, such as transcripts of radio and TV shows, social networks (e.g. tweets and posts of individual users, forums and organizations on Facebook and Twitter), blogs and wikis. Another interesting application could be identifying political bias of user comments in online news sites and comparing it to the bias of the sites' articles.

In future work, we intend to extend the study to additional time periods, in particular, when the left-wing was leading the governing power, and also to other countries and media sources.

References

Argamon, S., Koppel, M., Fine, J. and Shimoni, A.R. (2003), "Gender, genre, and writing style in formal written texts", *Text*, Vol. 23 No. 3, pp. 321-346.

Argamon, S., Koppel, M., Pennebaker, J.W. and Schler, J. (2009), "Automatically profiling the author of an anonymous text", *Communications of the ACM*, Vol. 52 No. 2, pp. 119-123.

Beigman-Klebanov, B., Beigman, E. and Diermeier, D. (2010), "Vocabulary choice as an indicator of perspective", in Hajic, J., Carberry, S. and Clark, S. (Eds), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computer Linguistics, Uppsala, pp. 253-257.

Burger, J.D., Henderson, J., Kim, G. and Zarrella, G. (2011), "Discriminating gender on twitter", *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, pp. 1301-1309.

Chung, C.K. and Pennebaker, J.W. (2007), "The psychological function of function words", in Fiedler, K. (Ed.), *Social Communication: Frontiers of Social Psychology*, Psychology Press, New York, NY, pp. 343-359.

D'Allesio, D. and Allen, M. (2000), "Media bias in presidential elections: a meta analysis", *Journal of Communication*, Vol. 50 No. 4, pp. 133-156.

Efron, A. (2004), "Cultural orientation: classifying subjective documents by co-citation [sic] analysis", *Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design*, pp. 41-48.

Filippova, K. (2012), "User demographics and language in an implicit social network", *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics*, pp. 1478-1488.

Gentzkow, M. and Shapiro, J.M. (2010), "What drives media slant? Evidence from US daily newspapers", *Econometrica*, Vol. 78 No. 1, pp. 35-71.

Gentzkow, M., Petek, N., Shapiro, J.M. and Sinkinson, M. (2015), "Do newspapers serve the state? Incumbent party influence on the US press 1869-1928", *Journal of the European Economic Association*, Vol. 13 No. 1, pp. 29-61.

Gosling, S.D., Augustine, A.A., Vazire, S., Holtzman, N. and Gaddis, S. (2011), "Manifestations of personality in online social networks: self-reported facebook-related behaviors and observable profile information", *Cyberpsychology, Behavior, and Social Networking*, Vol. 14 No. 9, pp. 483-488.

Grefenstette, G., Qu, Y., Shanahan, J.G. and Evans, D.A. (2004), "Coupling niche browsers and affect analysis for an opinion mining application", in Fluhr, C., Grefenstette, G. and Croft, W.B. (Eds), *Proceedings Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) – RIAO 2004, 7th International Conference*, CID, Avignon, pp. 186-194.

Groseclose, T. and Milyo, J. (2005), "A measure of media bias", *The Quarterly Journal of Economics*, Vol. CXX No. 4, pp. 1191-1237. doi:10.1162/003355305775097542.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009), "The WEKA data mining software: an update", *ACM SIGKDD Explorations Newsletter*, Vol. 11 No. 1, pp. 10-18.

Hasan, K.S. and Ng, V.I. (2012), "Predicting stance in ideological debate with rich linguistic knowledge", *Proceedings of the 24th International Conference on Computational Linguistics. Indian Institute of Technology Bombay, Mumbai, December*, pp. 451-460.

Hassanali, K.N. and Hatzivassiloglou, V. (2010), "Automatic detection of tags for political blogs", *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media. Association for Computational Linguistics*, pp. 21-22.

Joachims, T. (2002), "SVM-light", available at: www.svmlight.joachims.org (accessed May 9, 2016).

Koppel, M., Schler, J. and Zigdon, K.R. (2005), "Determining an author's native language by mining a text for errors", *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, Chicago, IL*, pp. 624-628.

Koppel, M., Schler, J., Argamon, S. and Pennebaker, J.W. (2006), "Effects of age and gender on blogging", *AAAI 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, Stanford, CA, March 27-29*.

Kuypers, J.A. (2002), *Press Bias and Politics: How the Media Frame Controversial Issues* (Praeger series in political communications), Praeger, Westport, CT and London.

Ladd, J.M. (2010), "The role of media distrust in partisan voting", *Political Behaviour*, Vol. 32 No. 2, pp. 567-585.

Laver, M., Benoit, K. and Garry, J. (2003), "Extracting policy positions from political texts using words as data", *American Political Science Review*, Vol. 97 No. 2, pp. 311-331.

Lin, W.H., Wilson, T., Wiebe, J. and Hauptmann, A. (2006), "Which side are you on?: identifying perspectives at the document and sentence levels", *Proceedings of the Tenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, NY, June*, pp. 109-116.

Mak, T. (2011), "Pew: public opinion of media never worse", *Politico*, September (accessed November 12, 2013).

Mullen, T. and Malouf, R. (2006), "A preliminary investigation into sentiment analysis of informal political dis-course", *Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs*, pp. 159-162.

Otterbacher, J. (2010), "Inferring gender of movie reviewers: exploiting writing style, content and metadata", *Proceedings of the 19th ACM international conference on Information and Knowledge Management. Association for Computational Linguistics*, pp. 369-378.

Pennebaker, J., Mehl, W. and Niederhoffer, K. (2003), "Effects of age and gender on blogging", *Annual Review of Psychology*, Vol. 54 No. 1, pp. 547-577.

Petrova, M. (2008), "Inequality and media capture", *Journal of Public Economics*, Vol. 92 Nos 1-2, pp. 183-212.

Petrova, M. (2011), "Newspapers and parties: how advertising revenues created an independent press", *American Political Science Review*, Vol. 105 No. 4, pp. 790-808.

Pew Research Center's Journalism Project Staff (2007), The invisible primary no longer: a first look at coverage of the 2008 presidential campaign", Project for Excellence in Journalism funded by the Pew Charitable Trusts and the Joan Shorenstein Center on the Press, Politics and Public Policy, Kennedy School of Government at Harvard University, Washington, DC, October 29, available at: www.journalism.org/2007/10/29/the-invisible-primaryinvisible-no-longer/ (accessed February 27, 2015).

Platt, J. (1999), "Sequential minimal optimization: a fast algorithm for training support vector machines", in Scholkopf, B., Burges, C. and Smola, A. (Eds), *Advances in Kernel Methods – Support Vector Learning*, MIT Press, Cambridge, MA, pp. 185-208.

Popescu, A. and Grefenstette, G. (2010), "Mining user home location and gender from Flickr tags", *4th Int'l AAAI Conference on Weblogs and Social Media, Washington DC, May*.

Schneider, W. and Lewis, I.A. (1985), "Views on the news", *Public Opinion*, Vol. 8 No. 4, pp. 6-11.

Weatherly, J.N., Petros, T.V., Christopherson, K. and Haugen, E. (2007), *The International Journal of Press/Politics*, Vol. 12 No. 2, pp. 91-104.

Yu, B., Kaufmann, S. and Diermeier, D. (2008), "Classifying party affiliation from political speech", *Journal of Information Technology and Politics*, Vol. 5 No. 1, pp. 33-48.

Further reading

Gillick, D. (2010), "Can conversational word usage be used to predict speaker demographics?", in Takao Kobayashi, K.H. and Nakamura, S. (Eds), *11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, ISCA, Makuhari, Chiba, pp. 1381-1384.

About the authors

Dr Maayan Zhitomirsky-Geffet received her PhD in Computer Science from the Hebrew University in Jerusalem, Israel. In her PhD thesis she explored automatic methods for ontological relationship recognition from large corpora and from the web. Currently, Dr Zhitomirsky-Geffet is an Assistant Professor in the Department of Information Science in the Bar-Ilan University and her main research fields include the Semantic web, internet research, social networks, and web-based information retrieval.

Dr Esther David received her PhD Degree in Computer Science from the Bar-Ilan University, Israel in 2003. Dr David has worked for three years (2003-2006) as a Senior Researcher at the Southampton University under the supervision of Professor Nicholas Jennings at the UK. Since 2006 she is a Senior Lecturer of the computer science department at Ashkelon academic college. Her research is primarily rooted in electronic commerce, mechanism design, game theory and auction theory. Her recent research includes also machine learning applications as building intelligent tutoring system for enhancing abilities in the domain of reading comprehension and author profiling for political tendency. For the last six years she has been one of the Organizers of the Agents Mediated Electronic Commerce conference (AMEC), which is jointly held with the AAMAS conference (one of the top AI and agent conferences).

Professor Moshe Koppel conducts research on a variety of machine learning applications including text categorization, image processing, speaker recognition and automated game playing. He is best known for his contributions to the branch of text categorization concerned with authorship attribution. More recently, he has begun researching fundamental problems in social choice theory.

Hodaya Uzan is an MA Student at the Department of Computer Science of the Bar-Ilan University in Israel. Her main areas of interest include: automatic text categorization, internet research and social networks.