Exploring Israeli News Website Bias using Simple Textual Analysis

Yuval Pinter[1], Oren Persico[2], and Shuki Tausig[2]

[1]Yahoo Labs
[2]The Seventh Eye
{yuvalpinter,orenpersico, shukitausig}@gmail.com

June 7, 2015

Abstract

Criticism of bias in Israeli media has been following a steep trend of escalation over the last few years, focusing on print journalism but by no means limited to it. We explore the hypothesis that various sources of bias are traceable in the main Internet news sites, using simple off-the-shelf Natural Language Processing tools to analyze the difference in style and language of the home page headlines among seven such sites over the course of a single month. Simple features prove powerful in a seven-class site identification task, reaching 49% accuracy using Random Forest. As part of this effort (still in progress) we created a dataset composed of a full year's headlines from these sites, which we plan to release for the benefit of the community.

1 Extended Description

The run-up to the recent Israeli general elections has seen the issue of media bias penetrate into the heart of the political discourse, including explicit accusations against specific news outlets by politicians as senior as the Prime Minister himself[1]. While continuous qualitative assessment of the daily news coverage by media experts and journalism critics (e.g. The Seventh Eye[2], Velvet Underground[3], Aviv Hurvitz[4]) provides valuable insights into the various flavors of media bias, to our knowledge a large-scale quantitative analysis has yet to be performed. We are also unaware of parallel undertakings elsewhere in the world.

Our research, still in progress, aims to assess the Israeli media ecosystem using standard Machine Learning and Natural Language Processing tools. As

[1]www.jewishpress.com/news/breaking-news/netanyahu-blasts-ynet-yediot-acharanot-owner-charges-smear-campaign/2015/02/09/
[2]www.the7eye.org.il
[3]velvetunderground.co.il
[4]mako.co.il/culture-weekend/media-fights

a first step, we created a corpus containing all homepage headlines[5] from seven Israeli news sites over the span of nearly a year, starting July 1, 2014: Ha'aretz, Israel Hayom, Ma'ariv[6], NRG, Ynet, Mako and Walla[7], totalling roughly 18K unique-per-site headlines with 150K tokens. Limiting ourselves to a subset containing the 1785 unique headlines extracted over the course of February 2015, we then extracted basic textual features from each headline and fed the features into Weka's [1] implementation of Random Forest, optimizing the features for a model which predicts with maximum accuracy from which news site a given headline was taken. We next created a model for the headlines from each pair of sites, analyzing the level of similarity (i.e. the model's difficulty in producing a correct prediction) between the various pairs.

The features we constructed were aimed at capturing several types of predicted bias, while striving to keep the model simple and not time-sensitive: **general editorial style** is represented by headline length features (by character and by word) and punctuation token count; **grammatical framing** is represented by features counting the letters which appear in the word form but not its lemma, as produced using HSpell [2] (for Hebrew, this feature captures both ordinary inflectional affixes and the adjoined particles); **topical bias** cannot be represented by explicit word features, but a proxy vocabulary feature set was used: minimum, maximum, median and average frequencies for words and lemmata from the headline based on publicly-available corpora counts[8].

As stated before, this work is still in progress, with respect to both computational approaches and subject-matter analysis. However we believe our approach is promising, based on the results so far: our 7-class predictor using no heavy word-form features achieves up to 49% accuracy on 10-fold cross validation on the February set; three of the 21 two-class predictors perform at over 90%, and all but two cross the 70% threshold. The three sites most difficult to differentiate are Mako, NRG and Walla, which are the three that stemmed from a digital-first ethos, suggesting that their similarity is a product of a common perception regarding the role of a website's headline.

Lastly, once our data and code stabilizes, we will release them in entirety for the benefit of the Hebrew-NLP and Israeli Journalism communities.

References

[1] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[2] Nadav Har'el and Dan Kenigsberg. Hspell-the free hebrew spell checker and morphological analyzer. In *Israeli Seminar on Computational Linguistics*, 2004.

[3] Tal Linzen. Corpus of blog postings collected from the israblog website. *Ms., Tel Aviv University*, 2009.

[5] More precisely, every headline that appeared at a round quarter-hour time, when each home page was scraped using HTTrack Website Copier: www.httrack.com.

[6] Starting August 27, when the site launched.

[7] respectively: www.haaretz.co.il, www.israelhayom.co.il, www.maariv.co.il, www.nrg.co.il, www.ynet.co.il, www.mako.co.il, www.walla.co.il

[8] Words: invokeit.wordpress.com/frequency-word-lists; lemmata: [3].