Automated profiling of the balance of optimism and pessimism in online news content

Tim Musgrove, Robin Walsh, Peter Ridge
Semantic Technology Group
Federated Media Publishing
San Jose, CA USA
{tmusgrove,rwalsh,pridge}@federatedmedia.net

***Abstract*—Using semantic techniques, we determined a probabilistic score indicating whether news stories were more optimistic (or solutions-oriented), versus their being more pessimistic (or threnodic). We observed over the length of our study that some news outlets, which were comparable in their topical coverage, quantity of output, and geographical focus, differed vastly in their level of optimistic or solutions-oriented news content. This did not seem to correlate with any perceived political bias (left vs. right) nor with the demographic of the target audience, and so raises questions of whether editorial culture or some other causal factor is at work, apart from the typical ideological or audience-driven biases. We found that it is indeed possible on a fully automated basis to profile media sources as falling more on the optimistic or pessimistic side of the spectrum.**

Keywords: semantic technology; text analytics; news filtering; media bias

I. INTRODUCTION

The "Slant Engine" is a semantic analysis system, conceived originally by TextDigger, Inc. and further developed by Federated Media Publishing, to automatically detect attitudinal or ideological leanings in a body of text. Currently, it is being applied at a new web site, OdeWire.com, to aggregate and publish optimistic news stories that are gathered from numerous sources around the Web. Herein we describe how this technology is utilized to develop profiles that illustrate how optimistic or pessimistic various news outlets are over the course of time.

II. TECHNIQUES

Previous work in analyzing and filtering news content has focused largely on one of three areas: determining topical classification of news [1][2][3], determining which news stories are reporting on the same event [4], or detecting political/cultural bias in new coverage and its effects on readers [5][6][7]. The relative negativity vs. optimism in news coverage has, for the most part, been left unaddressed.

Perhaps the most nearly-similar study of "slant" in the media is that of Groseclose & Milyo [8], in which they compared how often certain media sources quoted known conservative (or liberal) think tanks, compared with how many times the average member of Congress did. Over time, this gave them a good indication of left- vs. right-leaning publishers. However, our purposes were to determine something for which known think tanks are not established. Also, we needed not to determine a media source's overall slant, but the slant of particular news pieces, free of any other context. For this reason, we had to turn to analysis of the text content itself, irrespective of external references.

A. The Slant Engine Process

To estimate an article's overall optimism (or anti-optimism), a sentence function analyzer is used to perform a shallow parse and extract selected semantic elements, such as "solving a problem", "reaching a compromise", "alleviating suffering", etc. Note that these features are principally verb-object pairs (or in some cases, adjective-noun constructs). To prepare for extracting them, classes of appropriate objects, as well as classes of appropriate verbs, are predefined by using synsets from a word-sense indexed lexicon. These classes are expressed in a text file using a specialized syntax in order to be able to create flexible templates that the function analyzer uses for matching.

The semantic element definitions are organized into themes, with weight assignments that represent how much a match within an article will contribute to the theme. Themes are in turn assigned with positive or negative weights as contributing to optimism (or anti-optimism). Organizing these element definitions into themes is critical to making an effective working model of optimism. Both function themes (for verbs and adjectives) and object themes (for noun phrases) are created so that the attachment of certain function themes to certain object themes can count either for or against optimism. Examples of function themes include "progress in" [object theme], "positive signs of" [positive object theme], and "efforts against" [negative object theme]. Examples of object themes are "world problems", "social goods", "helpful actions", and "quality of life".

Human "knowledge editors" define these themes by examining a sample set of 70 news articles that are deemed "optimistic" and 30 articles that are deemed "anti-optimistic" and constructing a set of themes (with all their constituent elements) that represent all 100 model documents accordingly. This enables us to begin scoring new documents for overall balance of optimistic vs. pessimistic language based on the semantic model so created. For example, Table I contains the

978-0-7695-4492-2/11 $26.00 © 2011 IEEE
DOI 10.1109/ICSC.2011.85

snippets of text that instantiated "optimistic features" from one example URL.

Each article is then prepared for function analysis by retrieving the article from its source (e.g., downloading the HTML page from the Web), parsing the HTML elements in the page, locating the body content of the article and ignoring irrelevant elements such as navigation, template content and advertising, parsing the document into words (including collocations), and determining the part-of-speech for each word. The function analyzer then scans for the predefined semantic elements within the text.

Each occurrence of an element that is detected is given a relevance score between 0.0 and 1.0, inclusive, which is based on the structural qualities of where it is found within the article. These structural qualities may include the fact that an element occurs in headings or subheadings, has distinguishing font characteristics (e.g., bold, italics, size, is contained in a link, etc.), is in proximity to the start of the article, and so on. The relevance score is further adjusted by taking into account which words matched in the actual text, and their prior-probabilistic relationship to the words or synsets used in the element definition.

The matching element instances are then aggregated into distinct elements, taking the maximum score as a starting point and then increasing this score using each additional occurrence score. This effectively reduces the impact of the same idea being repeated within the article (e.g., "reduction of green house gases", "cut carbon emissions", etc.). The aggregated elements are then further reduced back to the element definitions, with a similar score computation (e.g., [reduction] of [World Problems]).

Scores for the themes are computed based on the scores of the occurring semantic elements, and their weightings as assigned to the theme. Positive and negative themes are combined using predefined weights to produce the total positive and negative theme scores. The final optimism score is then computed by reducing the total positive theme score by the negative theme score, using the same predefined weightings.

After calculating the overall balance of optimistic language over antithetical language, we choose a score threshold that satisfies a panel of human experts (trained news journalists) that most articles above that threshold were "good candidates". We then ran the system for several months, sending all of the system's candidate "optimistic" articles through to a professional editorial staff that was asked to accept only those that they deemed, with their human judgment, truly "optimistic" in nature.

TABLE I. EXAMPLE OF SEMANTIC FEATURE EXTRACTION

http://mondediplo.com/2010/09/15avatar
a participatory approach to world activism
environmentalists embraced Avatar
epic piece of environmental advocacy
directing attention to the rights of indigenous people
healthy scepticism towards the production of popular mythologies
creation for their own communicative purposes
attempts to regain lands
an empowered image of their own struggles
call attention to the plight
participatory culture
draw emotional power from engagement with stories
solidarity with the Iranian opposition party
repurposing pop culture towards social justice
participatory culture
shared narratives provide the foundation
culture gets created
building a grassroots infrastructure
sharing their perspectives

B. Application to OdeWire.com

OdeWire.com, a daily newsfeed of "solutions-oriented" articles from around the world, is now running on the basis of this engine, with some human editorial oversight. One interesting development is that the editors at OdeWire were comfortable letting the engine auto-publish a subset of articles that were above a very high confidence threshold (0.90). At this threshold about 1 in 12 articles so published were later retracted by editors as "not optimistic." Recall was measured at this output threshold, during a two-week period, at only 24%, but precision was 92.5%.

For practical purposes, i.e. to generate a more robust throughput of optimistic articles on a reliable basis, a lower threshold had to be used (0.60), which yielded 84% recall and 71% precision. In the end, articles scoring above 0.90 were auto-published, while those scoring between 0.50 and 0.90 were queued up for validation by OdeWire's editors.

Fig. 1 is a screen capture of the automatic publishing and queuing functionality from the point-of-view of an editor who is logged into the publishing system as an administrator. The story, "Holiday shoppers came out to spend in November", did not meet the threshold to be auto-published and, as a result, was set to Pending status so that it can be reviewed by an editor. Conversely, the story, "China to donate $23 mln for east Sudan development", was deemed to be sufficiently optimistic and published right away with no editorial intervention.

Figure 1. Published and pending stories in OdeWire's editorial control interface

Fig. 2 illustrates the overall flow of news stories into the Slant Engine to be scored and either queued for editorial review or published automatically to the web site.

Overall, this process of leveraging the Slant Engine to automatically publish stories that are clearly optimistic and only requesting human intervention when there is sufficient uncertainty results in approximately a 95% labor reduction when compared with OdeWire's editors manually gathering and curating for optimistic news stories by manual means.

III. FINDINGS

We wanted to determine whether certain news outlets were more consistent producers of optimistic news than others. We selected 45 representative news sources from around the world, including both well-known and lesser-known sources. After six months of running the process on a daily basis (from November 17, 2010 through May 17, 2011), we compiled a table of the percentage of articles published, for each publisher, which were deemed optimistic by the engine with a confidence of 0.50 or higher.

Meanwhile, professional editors also graded the documents with a simple yes or no answer as to whether they were optimistic or not, respectively. Editors also could take articles not even recommended by the engine, and deem them optimistic anyway, and publish them to OdeWire. Thus we created an environment allowing us to compare the engine's recommendation profile for each publisher, with the resultant "track record" of each publisher according to editors.

Early results indicated a particular problem of discourse analysis. We found that journalists often included the opposing slant, as it were, in their writings. This manifested as a sort of implicit diatribe, i.e., while writers did not explicitly engage in posing counterpoint questions, they nonetheless mentioned declaratively what was an opposing, pessimistic view. Most often this occurred near the beginning or end of an article. For example, "Much consternation has been felt over the failure of reduced class size to immediately produce higher test scores in California schools, however a new study shows promising results when small class sizes are maintained for three years or more." We found that the contrasting connectives, e.g. "however", "despite", "nonetheless", etc. needed to be observed with care. Antithetical elements attached with these connectors needed to be substantially decremented in their negative point value, in order to prevent the article's final optimism score from being improperly low.

Another early finding was really a human-factors limitation stemming from our working in partnership with professional news editors: a strict requirement for transparency in the underlying causes of an article being deemed optimistic or not. The opacity of machine learning methods, which generate thousands of features in large feature vectors, would have been impossible for us to explain or defend. This is why we elected to use hand-constructed ontologies and rule-based feature extraction; it was exactly what news editors needed in order to feel they understood what the Slant engine was doing. News editors view the automated system like a "cub reporter"—a less competent but still useful version of themselves who can take the first pass at editing the daily influx of news. From this view, the system needed to make decisions readily understandable to the human editors themselves. More to the point, we, as their technology partners, needed to be able to answer questions of the form, "Why did the system not consider this article optimistic?" in a way that the layman could readily understand. The ability to say that, "the occurrence in the opening paragraph of the words 'victims', 'horror' and 'prolonged agony' triggers the Human Tragedy theme, and there are meanwhile no positive phrases tempering that theme within the entire paragraph, so right away it registers a very high anti-optimism sub-score," sufficed to inform human editors that there is a logical reason for the decisions that the engine makes.

Table II shows how closely the engine's ranking of optimistic news producers resembled the human editors' ranking, gauged by the percentage of total stories that were deemed optimistic.

With a few notable exceptions, the human and engine rankings tracked fairly closely. Note that of the ten most optimistic sources according to editors, six were also in the top ten according to the Slant Engine. Conversely, of the bottom ten or least optimistic sources, also six were placed in the bottom ten as well by the Slant Engine. Comparing the two arrays of scores of all 45 sources by both editors and the engine yielded a rather large positive Pearson correlation of 0.605.

Anecdotally, the cases wherein the Slant Engine differs largely from the editors seem to be those wherein either (a) optimistic language is used for a trivial matter not deemed newsworthy by our editors (pet tricks, diet recipes), or (b) a strong mix of positive language is used in an article whose overarching message is nonetheless negative. It would be interesting in future work to seek remedies to these types of problem cases.

IV. SECONDARY FINDINGS

Further to our previous findings, we wanted to see if the difference in optimism would align with another factor, such as political leanings. While it is notoriously difficult to develop definitive lists of which news outlets are indeed left-leaning or right-leaning, we wanted some basis for comparison to see

whether optimism or pessimism correlated with the perceived political leanings of a publication.

TABLE II. RANKINGS OF NEWS SOURCES WITH MOST OPTIMISM

News Source	Optimism rank by editors	Optimism rank by Slant Engine
Le Monde Diplomatique	1	1
Treehugger	2	8
Huffington Post	3	24
IPSNews	4	3
Wall Street Journal	5	22
Mother Jones	6	5
The Guardian	7	6
CNN	8	10
Christian Science Monitor	9	4
AllAfrica	10	21
Times of India	11	34
Jerusalem Post	12	25
Denver Post	13	27
New York Times	14	14
Common Dreams	15	2
Mercury News	16	17
ABC News	17	19
Der Spiegel	18	12
CBS News	19	32
Arab News	20	15
European Voice	21	39
Japan Times	22	30
The Economist	23	20
Al Jazeera	24	42
Chicago Tribune	25	23
Los Angeles Times	26	18
Financial Times	27	40
BBC	28	36
Reuters	29	29
Business Insider	30	35
Fresno Bee	31	41
Bangkok Post	32	31
China Daily	33	16
Haaretz	34	9
Boston Globe	35	11
The Age	36	38
New Scientist	37	37
Maclean's	38	33
Belfast Telegraph	39	26
Irish Times	40	13
London Telegraph	41	28
Environmental News Network	42	43
National Post	43	7
Financial Post	44	44
New Zealand Herald	45	45

In an attempt to find a roughly neutral standpoint, we found that the Wikipedia article on "bias in the media" was a particularly well-vetted article, having been edited hundreds of times by hundreds of users over a five-year period. The reason this is significant to us, is that particular sections of the article have remained largely uncontested by editors through many revisions, which merely state which US news outlets have been often accused of having a left or right bias. It is significant that over five hundred Wikipedia users, who apparently are interested enough in media bias to edit the article, are largely all in agreement with these lists.

Therefore we deemed it prudent to take up these lists as our basis for comparison. We counted those news outlets accused both of being left- and right-biased, as being "neutral." We counted those subject to attack predominantly on one side as being either left or right, accordingly.

A subset of these appeared in our list of news sources, which allowed us to make at least a rough comparison of perceived political leanings and degrees of optimism. Examples of these news sources and their perceived leanings are shown in Table III.

TABLE III. NEWS OUTLETS PERCEIVED AS LEFT OR RIGHT LEANING

Perceived Political Leaning (gleaned from Wikipedia)	News Source	Optimism rank by editors	Optimism rank by Slant Engine
Left	Huffington Post	3	24
Right	Wall Street Journal	5	22
Left	CNN	8	10
Left	New York Times	14	14
Left	ABC News	17	19
Right	The Economist	23	20
Left	Los Angeles Times	26	18
Right	Financial Times	27	40

While any such compilation is imperfect at best, this does seem to suggest that the production of more optimistic stories does not correspond with the left/right leanings (real or perceived) of the news sources in question. Similar tables showing our estimated audience target demographics also failed to suggest any pattern.

We were left to speculate as to why some organizations exhibit a markedly different output of optimistic news (the lowest had zero, while the highest had 262 pieces of optimistic news in the same six-month period). Our admittedly loose conjecture is that some element of editorial "culture" might exist that does not correlate well with any numerical features that we have tested thus far.

V. CONCLUSIONS

A final result was clear as well. As many people have often suspected, the news is largely negative. The Slant Engine has processed well over 200,000 news articles at this point, with only around 3,000 passing the optimism test that allows them to show up on OdeWire.com. Even the most optimistic sources we found have only a small fraction of their total output come out as optimistic.

Table IV shows the total percentage of all stories that were finally deemed optimistic, for the top ten most optimistic news sources. Even the most optimistic of all of the news sources that were processed had less than 5% of all of its stories count as optimistic. Three of the 45 sources actually had 0% optimistic stories in a six-month period. The average

proportion of optimistic stories from all of the news sources was only 1.45%.

TABLE IV. PERCENT OF ALL STORIES DEEMED OPTIMISTIC

News Source	Percent Published
Le Monde Diplomatique	4.88%
Treehugger	4.60%
Huffington Post	3.48%
IPSNews	2.92%
Wall Street Journal	2.82%
Mother Jones	2.82%
The Guardian	2.40%
CNN	2.36%
Christian Science Monitor	2.24%
AllAfrica	2.11%

OdeWire.com will present later this year a series of awards for news agencies that produce the most "solutions-oriented" news pieces within their respective class of publishers. Their editorially-validated results from our system will be the basis of such awards. It is the hope of OdeWire's editors that these awards will encourage more news outlets to reexamine their balance of optimistic vs. pessimistic news content.

For further research, it would be worthwhile to attempt to improve the precision of the system, perhaps by bringing additional inputs such as social media commentary on the article from end users. Also, it would be interesting to investigate the sociolinguistic aspects of the phenomenon being studied, viz., possible causes of some organizations producing more optimistic content than others. Do certain topics, regions, times of year, or zones of coverage (urban vs. rural) draw more optimistic coverage (or less)? Does the size of the organization or its level of centralization have an influence? We hope to address some of these questions in our follow-up reports.

With regard to other applications of the Slant Engine, Federated Media will be incorporating this system into its "conversation targeting" of ads, a program designed for advertisers who want to show ads that can blend into the dialogue on blogs and social web content. For example, the marketer of an arthritis drug may want to capture online discussion threads where consumers are expressing misgivings and worries about the side effects of such drugs, in which case the advertiser has a message suited to that particular issue. In other words, many marketing messages are suited to the mood, attitude, or bias that consumers may have about a topic of concern, in which case the Slant Engine can serve the consumer's interest and the marketer's interest at the same time. Tests of such conversation targeting applications are in the early stages, and we hope to report further results within the coming year.

ACKNOWLEDGMENT

We wish to thank Jurriaan Kamp and the staff of OdeWire.com for their editorial help in the establishment of this project.

REFERENCES

[1] W.-L. Hsu, et al, "Classification algorithms for NETNEWS articles," Proc. of CIKM, 1999.

[2] Fabrizio Sebastiani, "Machine learning in automated text categorization," ACM Computing, Survey: 1-47, 2002.

[3] M. Steinbach, et al, "A comparison of document clustering techniques," KDD Workshop on Text Mining, 2000.

[4] Zhang Kuo, et al, "New event detection based on indexting-tree and named entity," Proc. of SIGIR, 2007.

[5] Dan Bernhardt, et al, "Political polarization and the electoral effects of media bias," CESifo Working Paper Series No. 1798.

[6] Samuel L. Becker, "Media bias: context, redundancy, and critical threshold as cultural factor," Annual Meeting of the International Communication Association, 1977.

[7] Stefano Della Vigna and Ethan Kaplan, "The Fox News effect: media bias and voting," University of California, Berkeley Mimeograph, 2005.

[8] Tim Groseclose and Jeffrey Milyo, "A Measure of Media Bias," The Quarterly Journal of Economics, Vol. 120, No. 4, pp. 1191-1237, November 2005.

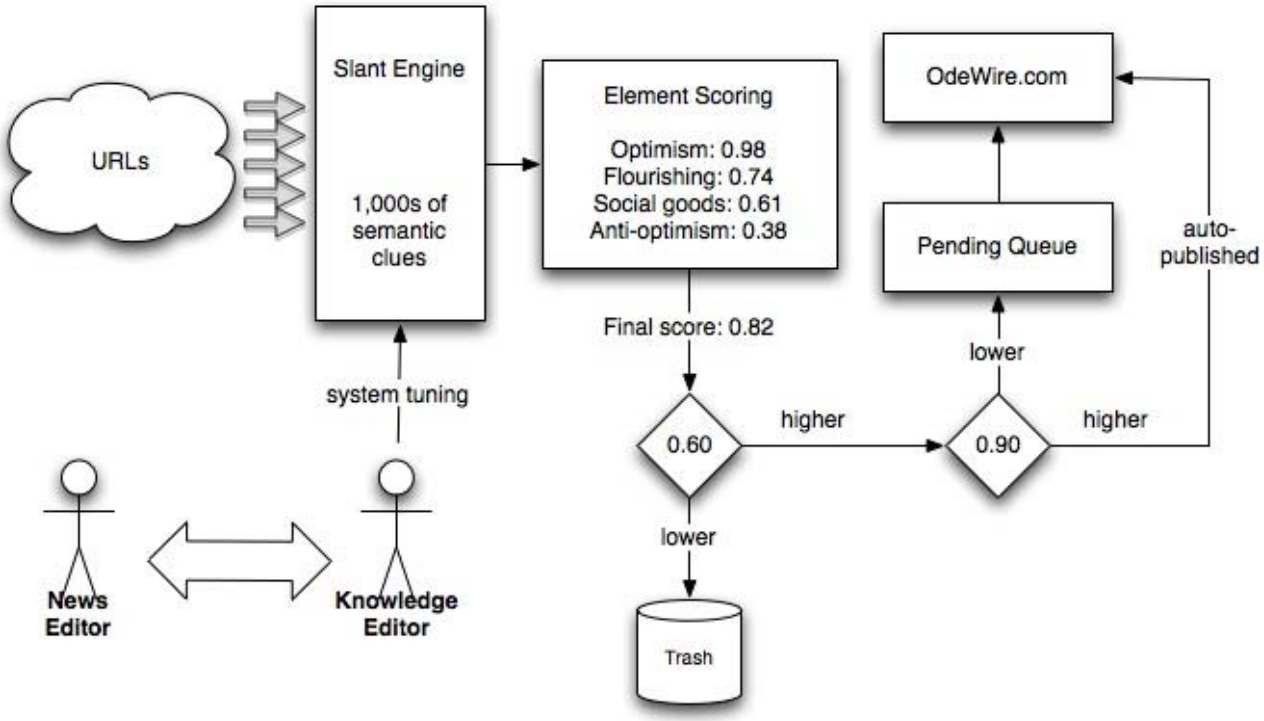

Figure 2. Slant Engine general flow diagram for OdeWire.com