Computationally Detecting and Quantifying the Degree of Bias in Sentence-Level Text of News Stories

C.J. Hutto Dennis Folds Scott Appling
Georgia Tech Research Institute (GTRI)
Georgia Institute of Technology
Atlanta, GA U.S.A.
email: {Clayton.Hutto, Dennis.Folds, Scott.Appling}@gtri.gatech.edu

Abstract—**Fair and impartial reporting is a prerequisite for objective journalism; the public holds faith in the idea that the journalists we look to for insights about the world around us are presenting nothing more than neutral, unprejudiced facts. Most news organizations strictly separate news and editorial staffs. Bias is, unfortunately, ubiquitous nevertheless. It is therefore at once both intellectually fundamental and pragmatically valuable to understand the nature of bias. To this end, we constructed a computational model to detect bias when it is expressed in news reports and to quantify the magnitude of the biased expression. As part of a larger overall effort, we conducted a survey of 91 people to investigate factors that influence the perception of bias in fictitious news stories. During this process, subjects provided ground-truth gold standard ratings for the degree of perceived bias (slightly, moderately, or extremely biased) for every sentence across five separate news articles. In this work-in-progress, we analyze the efficacy of a combination of linguistic and structural information for not only detecting the presence of biased text, but also to construct a model capable of estimating its scale. We compare and contrast 26 common linguistic and structural cues of biased language, incorporating sentiment analysis, subjectivity analysis, modality (expressed certainty), the use of factive verbs, hedge phrases, and many other features. These insights allow us to develop a model with greater than 97% accuracy, and accounts for 85.9% of the variance in human judgements of perceived bias in news-like text. Using 10-fold cross-validation, we verified that the model is able to consistently predict the average bias (mean of 91 human participant judgements) with remarkably good fit.**

Keywords-bias detection; bias quantification; linguistic model; text processing.

I. Dataset of Biased and Unbiased Text

A. *Perception of Bias in Unattributed News Stories*

In [1], people rated Presidents Bush and Obama on 25 adjectives and were then randomly assigned to read five fictitious news stories about one of them. Three of the stories described positive outcomes, and two described negative outcomes. In every story, one sentence was randomly manipulated to attribute the outcome to either an *internal* trait of the president or to *external* factors in an effort to observe the effects of moderating and mediating aspects of the attribution bias, whereby individuals typically assign greater attribution to internal/personal factors for positive outcomes when the person is someone they like, and to external/situational factors if the outcome is negative.

As part of the initial study, ninety-one people were surveyed. Participant demographics were skewed somewhat toward male (about 60%) and young adults under age 40 (over 50%). The political attitudes of the participants were of primary interest to [1], though, in particular, attitudes toward Presidents George W. Bush and Barack Obama. About two thirds of the sample had positive opinions about Obama and negative opinions about Bush, and one third exhibiting the opposite pattern. Participants were randomly assigned to provide ratings of one president first (Bush or Obama), followed by ratings of the second. Their responses were then used in a stratified sampling strategy to assign participants to read the five fictional news stories using either the name of the president they viewed most positively or most negatively (and 4 individuals who were neutral to both men were randomly assigned). Across the five stories, the story "target" remained the same once the participants were assigned to read about either Bush or Obama. We balanced the presentation order for the five stories to mitigate potential ordering effects. An example news story is presented below:

> *According to Forrester Research, an estimated 200,000 American jobs are lost annually due to offshore outsourcing. While in the past it was predominantly blue-collar jobs and low-level white-collar jobs that were relocated, the data show even mid- to high-level white-collar jobs are now being outsourced. During* ***{Bush/Obama}****'s presidential campaign, he maintained outsourcing is a part of globalization, which will be good for the American people in the long run. High unemployment rates led to growing public condemnation of outsourcing and demand for new regulations to stop or limit outsourcing. In response, corporations increased lobbying efforts to defend their ability to outsource jobs overseas, which they argued is necessary in order to remain competitive with international firms. Ultimately, President* ***{Bush/Obama}*** *rejected the proposal to implement trade protection policies that would discourage outsourcing. The President dismissed the proposal mainly because of…"*
>
> *"…* ***his unwillingness to stand up to corporate special interests****."(internal attribution)*
>
> *OR*
>
> *"…* ***intense pressure from corporations****." (external)*

This first story was about a financial situation where the outcome was negative. The other four stories reported about:

1. The president's decision to eliminate a federal grant program for teachers who would no longer receive incentive grants to work in inner-city school districts due to budget concerns (a negative outcome).
2. The president's promise to seek funding to support better emergency planning efforts, particularly those aimed at assisting with disaster preparedness for individuals with disabilities (a positive outcome).
3. The president's pledge to improve healthcare services to veterans (a positive outcome).
4. A successfully foiled bioterrorism attempt to smuggle aerosolized Ebola virus aboard an airplane in New York City (also a positive outcome).

B. *Degree of Bias in News Stories*

The current work-in-progress is primarily concerned with automatically characterizing the *intensity* or *degree* of bias perceived to be present in these news stories. In [1], subjects first read an entire story in paragraph form, and then were presented each sentence one a time and asked to rate how biased they believed each statement to be. Response options consisted of a 7-point balanced rating scale, with an option for a neutral rating ([–3] *Extremely* biased AGAINST Bush/Obama, [–2] *Moderately* biased AGAINST Bush/Obama, [–1] *Slightly* biased AGAINST Bush/Obama, [0] Fair and Impartial, [+1] *Slightly* biased IN FAVOR of Bush/Obama, [+2] *Moderately* biased IN FAVOR of Bush/Obama, or [+3] *Extremely* biased IN FAVOR of Bush/Obama). As we are currently interested in quantifying the degree of bias (rather than the polarity), we simplify by using the absolute value of the numerically coded responses.

II. RELATED WORK

There is a rich literature on stance recognition and argument subjectivity that focuses on identifying which side an article takes on a two-sided debate (c.f., [2]), casting the task as a two-way classification of the text as being either for/positive or against/negative (e.g., [3]–[5]) or as one of two opposing views (e.g., [6], [7]). In contrast, our work is primarily interested in estimating the *magnitude*, rather than direction or polarity, of the bias perceived to be present at the sentence level across all five news stories.

Additionally, previous datasets consisted of texts that typically take an overt stance (such as product reviews, debate transcripts, or editorial news); in contrast, we desire the capability to gauge bias even within the much more subtle domain of so-called "objective" news reports. Our work follows in the same vein as [8] who analyze biased language in reference articles using page edits tagged for violating Wikipedia's Neutral Point Of View (NPOV) policy. Again, whereas [8]'s focus is on identifying specific words or phrases that signal bias in reference articles, our work is distinct in that we are interested in characterizing the *degree* of such bias in the context of *news stories*, which – as with reference articles – similarly strive for impartiality.

III. DETECTING AND COMPUTING DEGREE OF BIAS

Using the 7-point balanced rating scale described above (coded as ranging from [–3] to [+3]) and human judgements of perceived bias from 91 participants for each of the 41 sentences from 5 separate news stories, we calculate the mean and distributions of the ratings using the absolute value of the numerically coded responses. As we see from the example text in Table I, some sentences of the news story are clearly perceived by human judges as being somewhat biased (as [1] intended to subtly induce either internal or external attribution biases by manipulating the final two sentence options). Expanding on the insightful work of [8] with additional sentence-level features and a dataset of news stories rather than reference articles, we develop a computational model that reads in a given sentence of text and then extracts and computes the strength of 26 structural and linguistics features present in the text. We next describe these 26 features.

A. *Structural Analysis at the Sentence Level*

In our sentence level analysis of the text, we observe characteristics of the text statement as a whole, considering syntactical, grammatical, and structural properties captured using the following five features:

1. **Sentiment score**: we use the freely available Python package VADER [9] to compute both the direction and intensity of the sentiment of each sentence (values range continuously from –1.0 [Extremely Negative] to +1.0 [Extremely Positive]). VADER is a highly accurate and well-validated sentiment analysis processing engine that implements numerous empirically derived sentiment processing rules related to textual syntax, grammar, punctuation, capitalization, negation, and other word-order sensitive elements of text [10].
2. **Subjectivity score**: we use Pattern.en [11] to compute the subjectivity of the sentence (values between 0.0 and 1.0). Pattern is a web mining module for Python, and the Pattern.en module is a natural language processing (NLP) toolkit that leverages WordNet to score subjectivity according to the English adjectives used in the text [12].
3. **Modality (certainty) score**: we use Pattern.en to compute the modality, or certainty, of the sentence (values range between –1.0 and +1.0, where values greater than +0.5 represent facts).
4. **Mood**: we use Pattern.en to compute the mood of the sentence. The mood of the sentence can be INDICATIVE (used to express facts, beliefs, e.g., "*It's raining*".), IMPERATIVE (used for commands or warnings, e.g., "*Make it rain!*"), CONDITIONAL (used for conjectures, e.g., "*It might rain today*") or SUBJUNCTIVE (used to express wishes or opinions, e.g., "*I hope it rains today*.").
5. **Readability**: we implement the Flesch-Kincaid Grade Level (FKGL) formula [13] to compute the readability of the sentence and associate it with a typical requisite grade level of reading comprehension. The higher the grade level, the more difficult the text.

TABLE I: MEAN (STANDARD DEVIATION) FOR 91 RATINGS OF PERCEIVED BIAS [SCALE: 0=UNBIASED TO 3=EXTREMELY BIASED]

	Sentence Level Text (for sentences from the first news story)	Mean (SD)
1	According to Forrester Research, an estimated 200,000 American jobs are lost annually due to offshore outsourcing.	0.10 (0.42)
2	While in the past it was predominantly blue-collar jobs and low-level white-collar jobs that were relocated, the data show even mid- to high-level white-collar jobs are now being outsourced.	0.11 (0.46)
3	During Bush/Obama's presidential campaign, he maintained outsourcing is a part of globalization, which will be good for the American people in the long run.	0.71 (1.00)
4	High unemployment rates led to growing public condemnation of outsourcing and demand for new regulations to stop or limit outsourcing.	0.20 (0.64)
5	In response, corporations increased lobbying efforts to defend their ability to outsource jobs overseas, which they argued is necessary in order to remain competitive with international firms.	0.12 (0.51)
6	Ultimately, President Bush/Obama rejected the proposal to implement trade protection policies that would discourage outsourcing.	0.70 (1.04)
7e	The President dismissed the proposal mainly because of intense pressure from corporations.	1.35 (1.22)
7i	The President dismissed the proposal mainly because of his unwillingness to stand up to corporate special interests.	1.90 (1.21)

B. *Linguistic Analysis at the Sentence Level*

Motivated by [8], we implement several linguistic features aimed at detecting either *epistemological* bias (features 6-9) or *framing* bias (features 10-12). To these, we add several additional linguistic features that we hypothesize may effect human perceptions of bias in text (features 13-26). For all of our sentence level linguistic features, we normalized the count of observations of the feature in the sentence by the total number of words in the sentence, producing values between 0.0 and 1.0 for each.

6. **Factive verbs**: are verbs that presuppose the truth of their complement clause (c.f., [8] for use in detecting epistemological bias in reference articles).
7. **Implicative verbs**: implicative verbs imply the truth or untruth of their complement, depending on the polarity of the main predicate (c.f., [8]).
8. **Assertive verbs**: are verbs whose complement clauses assert a proposition. The truth of the proposition is not presupposed, but its level of certainty depends on the asserting verb (c.f., [8]).
9. **Hedges**: used to reduce one's commitment to the truth of a proposition, evading any bold predictions (c.f., [8]).
10. **Strong subjective intensifiers**: are adjectives or adverbs that add (subjective) force to the meaning of a phrase or proposition (c.f., [8] for detecting framing bias in text using [14]'s list of strong subjectives).
11. **Weak subjective intensifiers**: as in [8], we use [14]'s list of weak subjectives.
12. **Bias (one-sided) terms**: One-sided terms reflect only one of the sides of a contentious issue (e.g., *anti-abortion* versus *pro-life*). We use [8]'s lexicon.
13. **Opinion words**: signal the expression of positive or negative attitudes or opinions, which may be biased. We use [10]'s validated opinion lexicon.
14. **Degree Modifiers**: are contextual cues (often adverbs such as *extremely*, or *slightly*) that modify the intensity or degree of an action, an adjective or another adverb. We use [10]'s list of degree modifiers.
15. **Coherence Markers**: are words (*because*, *therefore*, *so*) or lexical phrases (*as a result*, *for that reason*) that may be used to bias a reader towards a particular conclusion. We use [15]'s list of coherence markers.

The Linguistic Inquiry and Word Count (LIWC) [16] is text analysis software designed for studying the various emotional, cognitive, structural, and process components present in text samples [17]. LIWC uses a proprietary dictionary of almost 4,500 words organized into one (or more) of 76 categories, of which we use several for our feature set:

16. **Causation words**: e.g., *create*, *founded*, *generate*
17. **Certainty words**: e.g., *absolutely*, *frankly*, *must*
18. **Tentative words**: e.g., *bets*, *dubious*, *hazy*, *guess*
19. **3rd Person Pronoun**: e.g., *he*, *him*, *she*, *hers*, *they*
20. **Achievement words**: e.g., *accomplished*, *master*, *prized*
21. **Work words**: e.g., *ambitious*, *resourceful*, *hard-work*
22. **Discrepancy words**: e.g., *inadequate*, *mistake*, *liability*
23. **Conjunctions**: e.g., *while*, *although*, *cuz*, *whereas*
24. **Prepositions**: e.g., *within*, *over*, *through*
25. **Adverbs**: e.g., *mostly*, *nearly*, *primarily*
26. **Auxiliary verbs**: e.g., *may*, *oughta*, *should*, *will*

IV. FINAL MODEL FEATURE SELECTION

We next processed the 26-item feature vectors for each sentence through an initial statistical linear regression model using both forward and backwards stepwise Akaike information criterion (AIC) to measure the relative quality of each feature for characterizing the degree of bias in text. Using step-AIC for feature selection in this way helped us restrict the feature space to the most useful and valuable features. For example, in the presence of [14]'s more detailed list of strong and weak subjective linguistic intensifiers, the sentence-level measure of *subjectivity* is less meaningful (we therefore removed it from the model). On the other hand, the sentence-level measure for *modality* *(certainty)* is a stronger indicator of bias than the linguistic cues associated with LIWC *certainty words*, so we removed the *certainty words* feature from the model. Unfortunately, there was not enough variation in the sample data to determine whether differences in sentence structure with regards to *mood* affected perceived bias. As one might expect in "objective" news stories, nearly all sentences (85.4%) were computed to be INDICATIVE; so, we removed *mood* as a feature from the model. We found Flesch-Kincaid Grade Level (FKGL) scores for sentence-level *readability* were unrelated to the degree of perceived bias. This might be due to grade-level reading scores being generally high across the sample. The majority of sentences in the news stories ranged from about an 11th grade reading level (high school junior) to an 18th grade reading level (graduate school) (Mean=14.57, Standard Deviation=3.22). We therefore removed *readability* as feature from the model. Finally, we found that measures for *implicative verbs*, *degree modifiers*, *coherence markers*, *causation words*, *conjunctions*, *prepositions*, *adverbs*, and *auxiliary verbs* were all relatively poor indicators of sentence level bias; we therefore removed those features from the final model.

V. Preliminary Results

Table 2 depicts preliminary results of the linear regression analysis for the improved 14-feaure model $F(14,26) = 11.3$, $p = 1.04\text{e-}07$, which accounts for over 85% of the variance in human judgements of bias ($R^2 = 0.859$). Figure 1 depicts the proportion of overall R^2 that each feature accounts for, using the mean of three regression techniques (feature added to model first, feature added to model last, and feature beta squared). We find that a linguistic model motivated by [8]'s list of features for detecting biased language in reference articles is a useful start for determining the intensity (degree) of bias in news stories.

TABLE II: COEFFICIENTS, ERROR, T-VALUES, AND *P*-VALUES FOR THE IMPROVED MODEL. *F*(14,26) = 11.3, *P* = 1.04E-07.

	b	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.56	0.19	-3.02	0.006
Strong subjective	5.10	1.07	4.74	0.000***
3rd Person Pronoun	8.36	1.95	4.30	0.000***
Weak subjective	4.87	1.19	4.08	0.000***
Modality (certainty)	0.52	0.15	3.42	0.002**
VADER Sentiment	0.35	0.11	3.13	0.004**
Tentative words	4.60	1.65	2.79	0.010**
Opinion words	-2.05	0.95	-2.16	0.040*
Achievement words	5.74	2.66	2.16	0.040*
Factive verbs	-16.64	8.39	-1.98	0.058`
Work words	9.81	5.20	1.89	0.070`
Hedges	3.06	1.75	1.75	0.092`
Assertive verbs	-3.58	2.16	-1.66	0.110
Discrepancy words	5.66	3.62	1.56	0.130
Bias (one-sided) terms	-0.95	0.74	-1.30	0.206

Signif. level codes: $p < 0.001$*** $p < 0.01$** $p < 0.05$* $p < 0.1$`

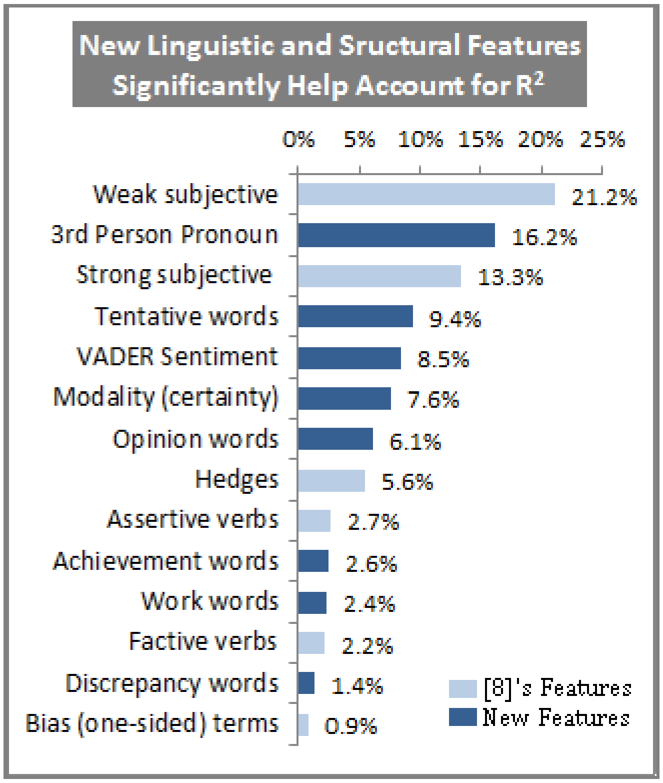

Figure 1. Proportion of variance accounted for by each feature in the improved model using the mean R^2 of three regression techniques (feature added to model first, feature added to model last, and feature beta squared).

Figure 2 shows the match between observed (measured) bias and the degree of bias predicted by the model; the fit is remarkably good. Many of our additional linguistic and structural features help to improve its predictive power:

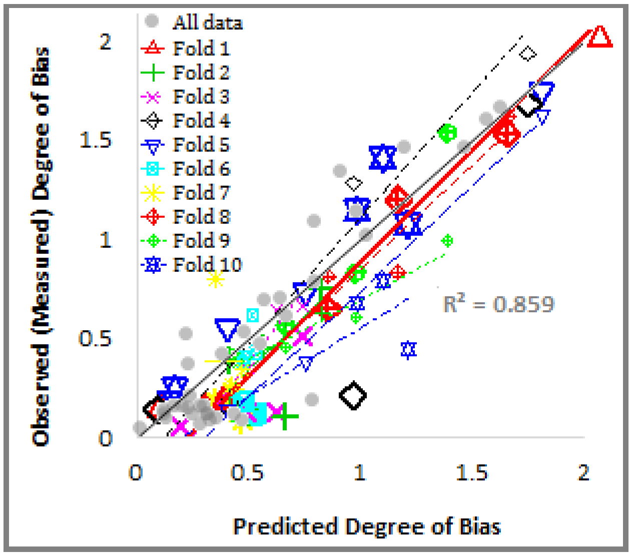

Figure 2. Results of 10-fold cross-validation analysis for fit between observed and predicted values of degree of bias in text.

REFERENCES

[1] D. J. Folds, “Perception of bias in unattributed news stories,” in *Procedures of the Annual Meeting of the Association for Psychological Science*, New York, NY, 2015.

[2] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann, “Which side are you on?: identifying perspectives at the document and sentence levels,” in *Proceedings of the Tenth Conference on Computational Natural Language Learning*, New York City, New York, 2006, pp. 109–116.

[3] P. Anand, M. Walker, R. Abbott, J. E. F. Tree, R. Bowmani, and M. Minor, “Cats rule and dogs drool!: classifying stance in online debate,” in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Portland, Oregon, 2011, pp. 1–9.

[4] A. Conrad, J. Wiebe, and and R. Hwa, “Recognizing arguing subjectivity and argument tags,” in *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, Jeju, Republic of Korea, 2012, pp. 80–88.

[5] S. Somasundaran and J. Wiebe, “Recognizing stances in ideological on-line debates,” in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California, 2010, pp. 116–124.

[6] T. Yano, P. Resnik, and N. A. Smith, “Shedding (a thousand points of) light on biased language,” in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, Los Angeles, California, 2010, pp. 152–158.

[7] S. Park, K. Lee, and J. Song, “Contrasting opposing views of news articles on contentious issues,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Portland, Oregon, 2011, pp. 340–349.

[8] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, “Linguistic Models for Analyzing and Detecting Biased Language,” in *Proceedings of the 51st Meeting of the Association for Computational Linguistics*, 2013, pp. 1650–1659.

[9] C. J. Hutto, “VADER Sentiment Analysis Software.” [Online]. Available: https://github.com/cjhutto/vaderSentiment. [Accessed: 28-Jul-2015].

[10] C. J. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,” in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014, pp. 216–255.

[11] CLiPS Research Center, “Pattern.en Software.” [Online]. Available: http://www.clips.ua.ac.be/pages/pattern-en. [Accessed: 28-Jul-2015].

[12] T. De Smedt and W. Daelemans, “Pattern for Python,” *J. Mach. Learn. Res.*, vol. 13, pp. 2063–2067, 2012.

[13] P. Kincaid, R. Fishburne, R. Rogers, and B. Chissom, “Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.,” National Technical Information Service, Springfield, Virginia 22151 (AD-A006 655/5GA, MF $2.25, PC $3.75), Feb. 1975.

[14] E. Riloff and J. Wiebe, “Learning extraction patterns for subjective expressions,” in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 105–112.

[15] A. Knott, “A Data-Driven Methodology for Motivating a Set of Coherence Relations,” PhD Thesis, Department of Artificial Intelligence, University of Edinburgh, 1996.

[16] Pennebaker Conglomerates, Inc., “LIWC Text Analysis Software.” [Online]. Available: http://www.liwc.net/. [Accessed: 28-Jul-2015].

[17] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC.net, 2007.