Prediction of Average and Perceived Polarity in Online Journalism

Albert Chu, Kensen Shi, Catherine Wong

***Abstract*— We predicted the average and perceived journalistic objectivity in online news articles based on the text of the article and data collected on the individual user reading the article, in order to better understand both how journalistic bias relates to the text of an article alone, and how reader demographics and political leanings influence the perceived objectivity of a given article. Using survey data on the perceived polarity of 284 articles ranked by multiple readers using the Mechanical Turk crowdsourcing platform, we applied supervised machine learning algorithms including L2-regularized L2-loss SVM, Naive Bayes, and Linear Regression to predict the mean polarity for each article and the perceived polarity of that article by any given reader. We were able to identify the words most strongly correlated with journalistic bias, predict a binary classification of mean and perceived bias for each article with relatively high accuracy, and determine the degree to which the article text alone and reader demographic information impacts perception of journalistic bias.**

I. Introduction

A. Background

Objectivity is a significant underlying principle in journalistic professionalism, and one that has become less standard in an Internet-oriented and often deeply partisan media environment, when anyone from a casual blogger to a major news outlet can disseminate information to the public. Understanding both the polarity of a given news article or media source, as well as the inherent biases influencing the perception of that media source by any given reader, presents especially interesting implications today. Recommendation engines and predictive algorithms filter the news presented to any given reader to cater to his or her own interests, often without the knowledge of the casual reader looking for what he may believe to be objective information.

Despite the importance of journalistic objectivity to the integrity and perception of a given article, previous machine learning research to automatically classify online news articles based on the presence of journalistic bias in a given article has been surprisingly sparse, especially in detection of bias within articles not specifically focused on political campaigns and figures. Sonal Gupta in 2009 [6] addressed a similar problem but focused only on articles concerning American politics, relying on the presence of previously identified political "memes" within the articles—phrases that had been manually identified as markers of a specific political inclination—to assign a limiting binary classification distinguishing between articles that favored the Democratic and Republican parties.

A. Chu, K. Shi, and C. Wong are undergraduate students at Stanford University, Stanford, CA. {achu8, kensens, catwong} at stanford.edu.

B. Goals

Using a combination of text analysis on online news articles alone and survey data collected on the demographics and subjective biases of the individual user reading a given article, we predict 1) the mean journalistic bias of a given online news article, based on labelings assigned by multiple readers of the same article, and 2) the perceived journalistic bias of a given article, which is the bias labeling assigned by a specific user to the article. Our goal was to better understand how journalistic bias relates to the text of an article alone, as well as the degree to which reader demographics and political leanings influence the perceived objectivity of the same article.

II. Data Sets and Feature Preprocessing

Since there is no prior dataset and little existing research on this topic, we created an entirely new dataset consisting of both articles and individual reader responses to those articles.

A. Article Text Collection

We collected URLs for 300 online news articles based on the top ten Google News search results for 30 high ranking keywords obtained from Google Trends [7], which provides data on high-volume search keywords across various categories in a given year.

Keywords were chosen to reflect a range of popular political and nonpolitical topics across categories, including the top trending keywords in 2013 overall, the top five business people of 2013, and the top male

and female politicians in 2013. The complete list of keywords can be found in the appendix.

We then used the Mechanical Turk crowdsourcing platform to isolate the text from each of the 300 articles. Due to nonstandard formats used to present article text online between various news sources, it is fairly challenging to automate text collection for online news articles, so the Mechanical Turk task was designed to allow crowdsourced workers to manually scrape article texts when presented with each article URL.

Article texts were preprocessed using the Python NLTK toolkit to remove punctuation, normalize character case, and stem tokens using the English Snowball Stemmer [4]. Certain common types of text features (such as URLs, dollar amounts, telephone numbers, etc.) were replaced with tokens representing each type. We then filtered the resulting article texts to remove broken links and articles with less than 50 words (which corresponded to articles that simply contained a video and article caption, and were considered to not contain enough text for a reader to consistently judge the bias of the article). After filtering, our final dataset consisted of preprocessed texts from 284 online news articles.

B. Bias Labeling and Reader Demographics Survey

Using a Mechanical Turk survey, each article URL was provided to 3 different survey respondents, who read and labeled articles on a discretized 1-10 scale according to the perceived journalistic bias of the article, where a ranking of 1 corresponded to no bias and 10 corresponded to a completely biased article. The survey also collected information on 9 additional features corresponding to reader demographic data and perceptions of the articles: reader age, education, gender, income, political leaning (on a discretized scale of 1-10 corresponding to left and right-wing leanings, with an additional option for no political leaning), reader location of residence, the perceived type of bias and political leaning of the article, and perceived bias ranking of the article source. Complete information on the survey features, including the options provided to survey respondents, is available in the appendix.

Prior to ranking the articles, survey respondents were provided with standard definitions of the terms "bias," "left-wing politically," and "right-wing politically," to ensure that survey respondents understood the terms and were looking for standardized attributes in the articles [1]–[3].

We assumed that the mean of the perceived bias rankings for each article represent the "ground truth" bias rankings on which we trained our models to predict mean bias. The perceived bias rankings assigned by a specific user to a given article were used to train our models to predict individual perceived bias. This dataset was collected through Mechanical Turk, which resulted in 843 individual responses (across 284 articles) after removing incomplete surveys.

III. Article Bias Prediction Models & Methodology

A. Mean Bias Labeling Prediction

Our first goal was to predict an article's average bias as reported by 3 readers per article only using the article's text. We first used L2-regularized L2-loss SVM (Support Vector Machine) for binary classification (with the Java LibLinear package [5], which we modified slightly to control randomization across runs) to predict average bias from the article text in bag-of-words format.

Initially, we considered all 2,931 words that appeared at least 5 times total and in at least 3 different articles. We realized that the feature space was too large (with only 284 training examples), so we applied forward search on the SVM to select 500 words to use as features, used throughout the remainder of the study. The number 500 was chosen because at this point, the accuracy improvement during forward search was tapering off. This search uses a mix of 50-fold cross validation and leave-one-out cross validation (LOOCV) with additional heuristics, based on word frequencies in biased and unbiased articles. We found that carefully choosing the words during forward search was especially important—in fact, using 20-fold CV in the forward search produced a much lower peak classification accuracy. At each step of the forward search, we first used 50-fold CV for efficiency and then used LOOCV to narrow down the top words found. Since we only had 284 training examples, the accuracy was always one of 285 choices (namely multiples of 1/284 between 0 and 1), so many words would have the same maximum accuracy. Hence, the heuristic was used to pick the single word to add at each step, instead of choosing them arbitrarily. The top five words (picked first by forward search) were "senat," "isnt," "court," "subject," and "council" (note that these are stemmed).

Each training example's feature data was normalized before feature selection but not after, so feature i actually represents the frequency of word i relative to the other 2,931 common words.

We then experimented with different numbers of bias classes to see how accurate the predictions could be. Since bias was reported on a scale 1-10, we split this range into equal parts to obtain the different bias classes. We report the LOOCV percent accuracy for 2 through 10 bias classes, where a "correct" prediction must exactly match the expected class. Standard Nave Bayes with Laplace smoothing and linear regression were self-implemented and run for comparison on the binary classification problem.

B. Perceived Bias Prediction

In this phase, we used the same set of 500 words found previously. The goal here is to predict a specific reader's rating of an article's bias, using both the article text and reader-specific information (age, education, gender, income, political leaning, and state of residence). We ran SVM on this new dataset (with 843 examples), reporting LOOCV percent accuracy for 2 through 10 bias classes computed as before. Similarly, standard Naive Bayes and Linear Regression were also implemented and run for comparison.

We then repeated this procedure with two modified datasets, one with only reader data and one with only article text for comparison.

IV. Results

Using the final set of 500 text features obtained after forward search feature selection, we predicted the mean bias class labelings using the SVM, linear regression, and Naive Bayes models to obtain the results in Figure 1. As described earlier, we experimented with different numbers of bias classes to determine how precise our bias predictions could be, by splitting the original bias labeling range, which was on a scale of 1-10, into equal parts to obtain between 2 and 10 bias classes. The results in Figure 1, which show the mean LOOCV accuracy of each model using the given number of bias classes, are plotted along with a "chance" baseline, corresponding to the results that would have been obtained by randomly assigning labels, to show improvement in the accuracy obtained by each model in comparison to random labeling. Using SVM with 2 bias classes gives a 93.66% LOOCV accuracy, Naive Bayes gives 84.88%, and Linear Regression gives 79.07%.

We also predicted perceived bias labelings, the bias labelings assigned by a specific reader to an article, using the same range of 2-10 discrete bias classes corresponding to increasingly accurate discretizations of the original bias labeling range. Figure 2 shows the

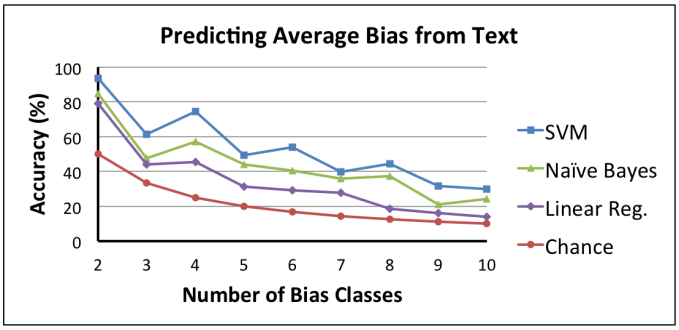

Fig. 1: LOOCV accuracy of the SVM, Naive Bayes, and Linear Regression models in predicting mean bias with varying numbers of bias labeling classes.

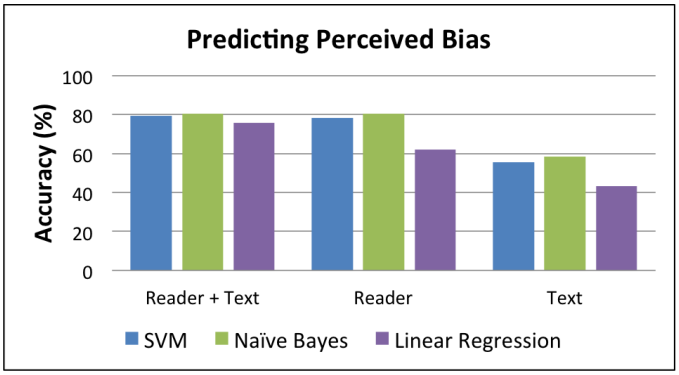

Fig. 2: LOOCV accuracy of the models in binary classification of perceived bias labeling with text features alone, survey reader demographic data alone, and a combination of the feature sets.

comparative LOOCV binary classification accuracy of the SVM, Naive Bayes, and Linear Regression models when trained with only the 500 text features, only the reader-specific demographic features obtained from the survey, and a combination of these two feature sets. For binary classification, SVM with both text and reader information achieves a 79.30% LOOCV accuracy. With reader information alone, SVM obtains a 78.29% accuracy, and with text alone, it gets a 55.63% accuracy.

For better comparison of trends and differences obtained by training the models on these different feature sets—that is, when we trained on text features alone, reader-specific features alone, and the combination of these two feature sets—Figure 3 shows an expanded graph of the resulting LOOCV accuracy of the SVM model on 2-10 discretized bias labeling classes.

V. Conclusion

After forward search feature reduction to reduce overfitting, the performance of the SVM and Naive Bayes models showed that we were able to predict

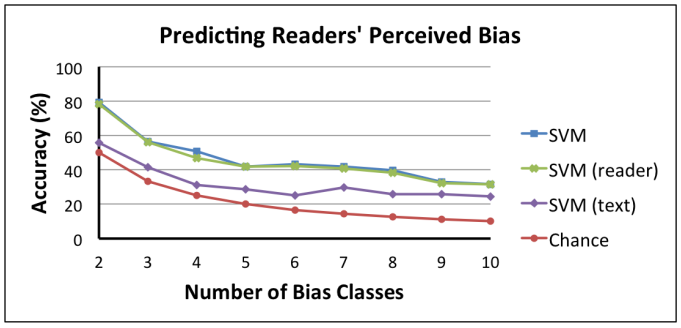

Fig. 3: LOOCV accuracy of the SVM model on varying numbers of discrete perceived bias labeling classes, with the text features alone, reader demographic features alone, and a combination of the feature sets.

mean bias labelings, especially a relatively rough binary classification of bias labelings, with surprisingly high accuracy using text features alone. This suggests that any article does in fact have a "ground truth" level of journalistic objectivity based on the text. Despite variation in the bias labelings assigned by each of the survey respondents to any given article, it does in fact appear that there exists an overall degree of bias associated with the article, independent of the specific reader. However, it is also interesting to note that perceived bias for any one individual reader seems to be far more strongly correlated with the specific demographic and political leanings of that reader than with the article text alone; even adding article text features to the reader-specific features does not strongly impact model performance for perceived bias prediction.

Additionally, the specific final text features chosen after forward search feature reduction give insight as to the words that appear most strongly correlated with prediction of journalistic objectivityinterestingly, although perhaps unsurprisingly, many of these were associated with politics, a topic common to many of the articles ranked as biased by readers.

Overall, our research shows that it is possible to assign mean bias labelings to articles based on their article text, and demonstrates a preliminary method to automatically score the journalistic bias present in a given article, results which could be used to better inform readers of the objectivity of news content presented online. The results also provide insight into the degree to which the previous biases and background of a reader influence his or her perception of the objectivity of a given article, showing that on an individual basis, reader perception of article bias is a subjective measure, influenced by a reader's own demographics and political leanings far more than the specific text of the article itself.

VI. Future Work

Future work could focus initially on expanding our dataset to include a larger scope of news articles across a greater range of sources and topics, with additional bias labelings and survey respondents for each article. A larger and more diverse dataset could reflect a more balanced demographic of readers, with more accurate "ground truth" labelings for mean bias overall.

Additionally, some features were collected in our bias labeling survey that were not used in this study, including the specific class of bias in a given article, the perceived political leaning of the article, or the bias ranking of the article source. Future work could attempt to predict these labelings and also take into account additional features into the prediction model, such as the perceived bias ranking of other works by a specific author or the general subject of the article, and feature reduction could be used to optimize text feature prediction to predict each of the additional possible labels.

Appendix

A. Search Keywords

Articles collected were selected based on top Google News results for these search terms, selected based on the following categories on Google Trends: Top 10 Trending 2013, Top 5 Business People, Top 1 Energy Company, Top 5 Female Politicians, Top 5 Male Politicians, Top 5 US Governors.

Search Keywords: 'Paul Walker', 'Boston Marathon', 'Nelson Mandela', 'Cory Monteith', 'iPhone', 'government shutdown', 'james gandolfini', 'harlem shake', 'royal baby', 'adrian peterson', 'oprah winfrey', 'willie robertson', 'charles r. schwab', 'bill gates', 'steve jobs', 'BP', 'Wendy Davis', 'Dianne Feinstein', 'Kathleen Sebelius', 'Janet Napolitano', 'Kay Hagan', 'Ted Cruz', 'Barack Obama', 'Hugo Chavez', 'Rand Paul', 'Arnold Schwarzenegger', 'Chris Christie', 'Jesse Ventura', 'Andrew Cuomo', 'Rick Perry'

B. Article Bias Labeling Survey Features

Mechanical Turk crowdsourced survey respondents were provided with online news article URLs, and asked to complete a survey after reading the article with their opinions on the article and demographic information. While the original Mechanical Turk survey cannot

be publicly accessed, a preview image of the survey, along with information on the survey data features, is available here: `http://bit.ly/cs229biassurvey`.

REFERENCES

[1] Bias. http://www.oxforddictionaries.com/us/definition/american_english/bias. Accessed: December 1, 2014.

[2] Left-wing politics. http://en.wikipedia.org/wiki/Left-wing_politics. Accessed: December 1, 2014.

[3] Right-wing politics. http://en.wikipedia.org/wiki/Right-wing_politics. Accessed: December 1, 2014.

[4] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. OReilly Media Inc., 2009.

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[6] Sonal Gupta. Finding bias in political news and blog websites. http://snap.stanford.edu/class/cs224w-2010/proj2009/report_Sonal_Gupta.pdf, 2009.

[7] Google Inc. Google trends 2013. http://www.google.com/trends/topcharts?date=2013. Accessed: November 12, 2014.