A Comparison of Lexicons for Detecting Controversy

Kateryna Kaplun, Christopher Leberknight, Anna Feldman
Montclair State University
1 Normal Avenue, Montclair, New Jersey 07043
{kaplunk1, leberknightc, feldmana}@montclair.edu

Abstract

We collect a corpus of 1554 online news articles from 23 RSS feeds and analyze it in terms of controversy and sentiment. We use several existing sentiment lexicons and lists of controversial terms to perform a number of statistical analyses that explore how sentiment and controversy are related. We conclude that the negative sentiment and controversy are not necessarily positively correlated as has been claimed in the past. In addition, we apply an information theoretic approach and suggest that entropy might be a good predictor of controversy.

Keywords: controversy, online news, sentiment analysis

1. Introduction

In many countries around the world access to online information is strictly regulated. The news is a large part of our everyday lives. News media brings social, economic, political, and all other issues to the forefront to facilitate discussions about these topics. Some of these topics may be considered controversial in that they spark debate among those with firm opposing beliefs. It is also important to know what kind of sentiment these topics emote for people. This can help determine if an article is controversial through the positive or negative words that occur in it. By studying the sentiment and controversiality of articles, we can better understand how news is censored and how news sources and people in general use language to share and promote certain ideas. In this paper, we perform a statistical analysis of sentiment and controversiality.

1.1 Previous Work

Mejova et al. (2014), Pennacchiotti et al. (2010), Dori-Hacohen & Allan (2015) and Jung et al. (2010) use a logistic regression classifier, a support vector machine classifier, a nearest neighbor method, and a probabilistic approach, respectively, to detect controversial content. There is also work on censorship tracking that uses bag-of-words models (e.g., Crandall et al. 2007). Mejova et al. (2014) conduct an experiment in which they quantify emotion and bias in language through news sources. To establish a baseline for measuring controversy, they use a crowdfunding technique with human annotators whose task was to classify controversial articles. They develop a list of strongly controversial, somewhat controversial, and non-controversial terms. They use their new lexicon to analyze a large corpus of news articles collected from 15 US-based news sources. They compare controversial and non-controversial articles in terms of a series of bias and sentiment lexicons and discuss the differences in the strength with which annotators perceived a topic as controversial and how it was perceived in the media. Mejova et al. (2014) report that in controversial text, negative affect and biased language prevail. While the results of this experiments are definitely interesting, the researchers use a relatively small number of annotators.

The size of their new dataset is small, too. They classify 462 words in their experiments. Such a small sample size adversely impacts discrimination quality and classification accuracy. To investigate the reliability of their results, we reproduce their experiment to evaluate predictive accuracy for potential use with other datasets. However, since we could not gain access to their dataset, we could not reproduce the experiment used to classify documents using logistic regression.

Other approaches include Dori-Hacohen & Allan (2015) who use a nearest neighbor classifier to map webpages to the Wikipedia articulates related to them. The assumption is that if the Wikipedia articles are controversial, the webpage is controversial as well. Dori-Hacohen & Allan (2015)'s algorithm depends on Wikipedia controversy indicators, produced from Wikipedia specific features (Jang et al. 2015). Searching for k nearest neighbors for each document is non-trivial and therefore this could be practically inefficient (Jang et al., 2015). Another limitation is that it is necessary for the topic to be covered by a Wikipedia article (Jang et al., 2015). There are also generalization limitations with domain specific sources such as Wikipedia's edit history features and Twitter's social graph information (Jang et al., 2015).

Jang et al. (2015) extends Hacohen and Allan (2015)'s work by introducing a probabilistic method for detecting controversy based on the kNN-WC algorithm. Their approach uses binary classification and a probabilistic model that can be used for ranking (Jang et al., 2015). Their approach also uses Wikipedia, since it has domain specific features.

There has been work on controversy detection that explores sentiment. Jang et al. (2015), for example, demonstrate that utilizing sentiment for controversy detection is not useful. However, Choi, Jung, and Myaeng (2010) detect controversy using a mixture model of topic and sentiment and report good results. Pennacchiotti et al. (2010) detect controversies surrounding celebrities on Twitter. They use features such as the presence of sentiment-bearing words, swear words, and words in a list of controversial topics that come from Wikipedia. In our

work we investigate the relationship between controversial content and sentiment.

2. Experiments

Three experiments investigate the potential for using an existing annotated corpus of controversial and non-controversial terms (Mejova et al, 2014) to detect controversy in online news articles. Experiments use data comprised of Montclair Controversy Corpus (see section 3.1.1) of 1,554 online news articles collected from 23 RSS feeds with four lexical resources (Table 1). If lexicons of controversial words exist, can they be used to detect controversy in online news articles? We explore this question with a series of experiments inspired by previous research (Mejova et al, 2014) that suggests a positive correlation between negative sentiment and controversy in several news articles.

1. *Experiment I* aims to test the reliability of previously annotated controversial words (Mejova et al, 2014) for detecting controversy in unlabeled documents. This was done using existing lexical resources and 19 subjects who annotated the set of words from previous research (Mejova et al, 2014). We claim that the frequency of controversial terms in the MCC can be used to partition the data into controversial and non-controversial sets. We do not believe the lexicon can be used to detect controversy for individual documents, but believe it can be used to describe an aggregate view of the data.
2. *Experiment II* provides a descriptive analysis comparing the frequency of positive and negative words in our dataset compared to previously annotated sentiment datasets (Choi et al, 2010 and Chimmalgi 2010). The claim is that negative sentiment will be correlated to controversial documents and positive sentiment will be correlated to non-controversial documents.
3. Experiment III statistically tests the proportion of negative sentiment in controversial text will be higher than the proportion of positive sentiment in non-controversial text.
    a. **H1:** The proportion for overlapping words between the negative sentiment and controversial datasets is greater than the proportion for overlapping words between the negative sentiment and noncontroversial datasets.
    b. **H2:** The proportion for overlapping words between the positive sentiment and noncontroversial datasets is greater than the proportion for overlapping words between the positive sentiment and controversial datasets

3. Lexicons

As seen in Table 1, we use several resources for our experiments. MicroWNOp and General Inquirer are sentiment dictionaries. The Mejova lexicon is a set of words labeled with controversial/non-controversial categories. Finally, we use a set of words extracted from a list of controversial Wikipedia topics (Pennacchiotti and Popescu, 2010).

Lexicon	Type of Lexicon	# of Words
MicroWNOp	Positive	418
MicroWNOp	Negative	457
General Inquirer	Positive	1628
General Inquirer	Negative	2000
Mejova	Controversial	145
Mejova	Not Controversial	272
Wikipedia	Controversial	2133

Table 1: Lexicons with the number of words

This Wikipedia terms are deemed controversial because they appear in articles that are constantly being re-edited in a cyclic way, have edit warring issues, or article sanction problems (Wikipedia: List of Controversial Issues, 2018).

3.1.1 Montclair Controversy Corpus

The Montclair Controversy Corpus (MCC) contains 1554 news articles collected from 23 RSS feeds and has 317,361 word tokens in total after the stopwords (function words and punctuation) have been removed.

To create the MCC, we start by generating a dataset of hundreds of English-language articles through 23 RSS feeds. We then remove the stopwords from the MCC. We used crowdsourcing to label the words from the collected corpus as controversial, somewhat controversial, and not controversial. The final category was determined using a majority vote rule. We use these categories to make our new resource comparable to that of Mejova et al. (2014). We use a set of controversial terms, somewhat controversial terms, and not controversial terms that were also used in Mejova et al (2014) to test against the MCC. In testing their terms against our dataset, we used a set of lexical resources as seen in Table 1. Using different lexicons, we determine whether our dataset has sufficient terms that can be classified as controversial, somewhat controversial, and non-controversial. We also compare our dataset with Wikipedia words extracted from a list of controversial topics. (Pennacchiotti and Popescu, 2010).

Our goal is to determine how well Mejova et al. (2014)'s results generalize. To determine the sentiment of words included in the MCC, we apply two sentiment lexicons to it: MicroWNOp (Choi et al, 2010) and General Inquirer (Chimmalgi 2010).

4. Results

4.1 Experiment I: Controversy

We compare Mejova et al 2014's controversial and non-controversial words, Wikipedia 2018 (only controversial words) and our new corpus. The results are summarized in the Table 2. The normalized proportion is calculated by taking the frequency of each word found the MCC and dividing it by the total number of words in the corpus in this case, 317361. Results in Table 2 suggest the Wikipedia list has a better coverage than the Mejova list of controversial terms. Overall, the results demonstrate that the controversial terms represent only a small fraction of words in the MCC, but it does appear that the MCC is

biased more toward controversial documents compared to non-controversial documents. This supports C1, but the small fraction of controversial words represented in the MCC suggests that controversial terms are not the only indicators of controversial documents.

Dataset	Type	Normalized Proportion
Mejova	Controversial	0.06130873
Mejova	Non-Controversial	0.054017349
Wikipedia	Controversial	0.193227901

Table 2: Normalized proportion of words vs lexicon

4.2 Experiment II: Sentiment

Experiment II evaluates previous claims that sentiment can help to identify controversial documents (Mejova et. al 2014). We matched the words in the MCC against the sentiment lexicons described above. The General Inquirer was approximately three or four times larger than the MicroWNOp dictionary and the frequencies were also approximately three or four times higher unlike in the controversy lexicons.

Figure 1 suggests that positive sentiment is found more in non-controversial documents across both lexicons compared to the fraction of words that emote negative sentiment in controversial documents. This appears consistent with previous results (Mejova et al, 2014). However, since results from experiment I suggest the proportion of MCC data contains controversial words we would expect that negative sentiment would also be more frequent in the MCC data. Results in Figure 1 indicate this is not the case. The statistical analysis in experiment III statistically evaluates this result.

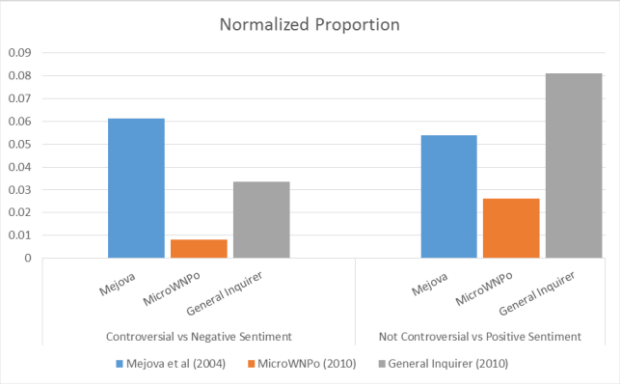

Figure 1: Normalized proportion of positive and negative sentiment

4.3 Experiment III: Controversy and Sentiment

Baseline lexicons are evaluated against each other in order to see how sentiment and controversy relate to each other. We ran four two proportion z tests to determine if words that indicate negative sentiment are more likely to appear in the lexicon of controversial terms. The following tests are summarized in Table 3.

1. Test 1 compares the controversial and non-controversial words from the Mejova lexicon in terms of the negative sentiment derived from the MicroWNOp dictionary.
2. Test 2 compares the controversial and non-controversial words in terms of the negative sentiment General Inquirer dictionary.
3. Test 3 analyzes the Wikipedia controversial words in terms of the negative sentiment obtained from the General Inquirer.
4. Test 4 compares Wikipedia list of controversial words against the Mejova non-controversial words in terms of the negative sentiment MicroWNOp dictionary.

Our alternate hypothesis (H1) is that the proportion for overlapping words between the negative sentiment and controversial datasets is greater than the proportion for overlapping words between the negative sentiment and noncontroversial datasets.

Test	z-statistic	p-value
Test 1	3.37927	0.000363
Test 2	2.00707	0.022371
Test 3	-1.45454	0.927101
Test 4	3.4985	0.000234

Table 3: Negative sentiment tests with z stats and p-values

Based on the frequencies and proportions, there is not a lot of overlap between the sentiment dictionaries and controversial lexicons.

We test the difference between the proportion of controversial words (first proportion) and non-controversial words (second proportion) in our dataset. The test evaluates if the first proportion is higher than the second proportion. For example, in Test 1, the two proportions that are being tested are 0.0414 and 0. Tests 1 and 4 are statistically significant as they have a p-value less than the 0.01 significance level. However, in these two cases the sample proportion that is being tested against is 0 because there was no overlap between the noncontroversial Mejova dataset (2014) and the negative sentiment MicroWNOp dataset (Choi et al, 2010). Therefore, the test is not particularly useful in determining if negative sentiment is more likely in a controversial dataset than a non-controversial dataset. Test 2 is significant at the 0.05 significance level but not strongly significant at the 0.01 significance level. Test 3 is not significant at all with a p-value of 0.927101. This test would actually have a better p-value if our assumption was that the non-controversial dataset had more overlap than the controversial dataset. Even still, this p-value would be 0.072899, which is still not significant.

This is an intriguing result. Unlike previous research (see e.g., Mejova (2014)) that has shown a positive correlation between controversy and negative sentiment, our statistical tests do not provide strong evidence to support this hypothesis and therefore *H1 is not supported*. In addition, we also ran four two proportion z tests to determine if words that indicate positive sentiment are more likely to appear in the noncontroversial dataset than the controversial dataset.

1. Test 5 analyzes the Mejova controversial words in terms of the positive sentiment derived from MicroWNOp dictionary.

2. Test 6 analyzes the Mejova controversial words in terms of the positive sentiment derived from the General Inquire dictionary.
3. Test 7 analyzes the Wikipedia controversial words in terms of the positive sentiment derived from the General Inquirer dictionary.
4. Test 8 analyzes the Wikipedia words in terms of the positive sentiment obtained from the MicroWNOp dictionary.

Test	z-statistic	p-value
Test 5	-1.09122	0.862412
Test 6	0.115636	0.453971
Test 7	-0.875432	0.809331
Test 8	-0.896901	0.815114

Table 4: Positive sentiment tests with z stats and p-values

Our alternate hypothesis is that the proportion for overlapping words between the positive sentiment and noncontroversial datasets is greater than the proportion for overlapping words between the positive sentiment and controversial datasets (H2). None of these four tests are significant indicating that there is no evidence that words that express positive sentiment occur more in noncontroversial data than in controversial data. Overall, results from our lexical resources suggest there is not enough conclusive evidence to determine that negative words are more likely in controversial words than noncontroversial words or that positive words are more likely in noncontroversial words than controversial words. *H2 is therefore not supported*. We hypothesize that it is the intensity of emotion rather than valence that correlates positively with controversy. We will address this issue in future work.

5. Entropy

Next, we ran an experiment in which we asked human subjects to rate all of the words in the Mejova lexicon (2014) in terms of controversy. The analysis is based on responses from 19 subjects. The subjects were presented with a single word and asked to label each individual word as controversial, somewhat controversial, or not controversial. The final category was determined based on which category had the majority of votes. Entropy is used to measure the amount of disorder or randomness in responses for each word categorized. Entropy is computed using the following formula:

$$Entropy = \sum_{i=1}^{n} \frac{(p(x_i, y_1))(\log(x_i, y_1))}{p(y_i)} \quad (1)$$

Word	Classification	Normalized Standard Deviation	Normalized Entropy
abuse	Controversial	0.00226879	0.00281437
aid	Not Controversial	0.00097444	0.00303359
america	Controversial	0.00168772	0.00289315
american	Controversial	0.00102444	0.00301502

Table 6: A small portion of the data with normalized standard deviation and normalized entropy

Table 6 presents a small portion of the data, which includes the word, its classification, its normalized standard deviation, and its normalized entropy

Entropy was normalized to account for words with a different number of responses but the same entropy value. Higher values of entropy suggest more randomness and hence more difficult to predict compared to low values of entropy. Entropy was computed for each word in the Mejova et al (2014) lexicon, the controversial, somewhat controversial, not controversial, and unknown categories, were plotted in Figure 2. The *unknown* category corresponds to words that did not have a majority vote. Figure 2 demonstrates that the entropy of words that are not controversial is lower than words that are controversial, somewhat controversial, and unknown. This indicates that entropy may be a useful feature in predicting controversial words. Additionally, of the 16 words that were classified as unknown within the survey, the classifications done by Mejova et al (2014) showed that 13 of the words were controversial and 3 of the words were somewhat controversial. Since they were controversial and somewhat controversial, they have a higher entropy just like the other controversial and somewhat controversial words. The main challenge this paper explores is how to detect controversial documents when the "ground truth" is unknown. We have shown existing controversial lexicons can be used to gain a global understanding of the degree of controversy and entropy can be used to cluster articles with the same label. For example, Figure 2 illustrates that articles with high entropy tend to fall within the controversial category. The differences between categories was shown to be statistically significant at the 0.05 significance level.

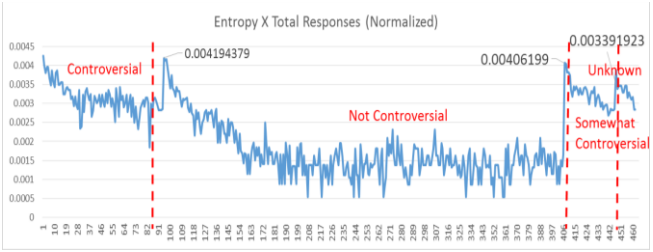

Figure 2: Normalized entropy by category

6. Limitations

There are some limitations with this study that are important to note. Our article dataset may not be fully representative of a variety of controversial topics. Most of the lexicons have a counterpart such as the Mejova noncontroversial words and the controversial words. However, there is no noncontroversial lexicon for Wikipedia, which hinders our ability to test that against the Wikipedia controversial set or compare to our sole noncontroversial dataset. Another limitation is that the negative sentiment MicroWNOp dictionary and the Mejova noncontroversial list have zero words that overlap, making it difficult to analyze these two together as well as run any tests. Also, some words are marked both as positive and negative by the sentiment dictionaries we experimented with affecting the results. This is because these words can be subject to interpretation and depending

on their context could be negative. For example, the word *help* can be positive when it is used in the sense that someone is assisting someone else with something whereas it can be seen as negative if someone is calling out for help because they are in trouble.

The final category that was assigned to each word using 19 annotators was determined with a majority vote rule. Out of 462 words, 180 had inter-rater agreement < 60%. We instead used a majority vote and results were very consistent with previous results (Mejova et al, 2014). Furthermore, entropy values need to be evaluated against "ground truth" to fully understand the benefit and reliability of this metric.

7. Future Work

Future research will investigate the potential of sentiment for controversy detection with larger news datasets and explore other methods and features for identifying controversial news. We will also build an automatic classifier to detect controversy using sentiment and entropy features. Determining the controversiality of news articles can assist future research by providing a predictor for censorship. If censorship can be predicted, a system can be designed to circumvent censorship allowing citizens to openly and freely communicate on the Internet.

8. Acknowledgements

This work has been supported by the National Science Foundation (NSF) under grant 1704113.

9. References

Choi Y., Jung Y., and S. H. Myaeng. (2010). Identifying Controversial Issues and Their Sub-topics in News Articles. Intelligence and Security Informatics, 140.

Crandall J. R., Zinn D., Byrd M., Barr E., and R. East. (2007). Concept Doppler: A Weather Tracker for Internet Censorship. *CCS '07: Proceedings of the 14th ACM Conference on Computer and Communications Security*, 352–365.

Dori-Hacohen S. and J. Allan. (2015). Automated Controversy Detection on the Web. Advances in Information Retrieval 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 – April 2, 2015. Proceedings, 423.

Garimella K., De Francisci Morales, G., Gionis A., and M. Mathioudakis. (2015). Quantifying Controversy in Social Media.

Jang M., Foley J., Dori-Hacohen S., and J. Allan. (2016). Probabilistic Approaches to Controversy Detection. Conference on Information & Knowledge Management, 2069.

Wikipedia: List of Controversial Issues. (2018).

Mejova Y., Zhang A. X., Diakopulos N., and C. Castillo. (2014). Controversy and Sentiment in Online News.

Pennacchiotti M. and A.M. Popescu. (2010). Detecting Controversies in Twitter: A First Study. Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, 1873-1876.

R. V. Chimmalgi. Controversy trend detection in social media. Master's thesis, Louisiana State University, May 2010.