Controversy and Sentiment: An Exploratory Study

Kateryna Kaplun
Montclair State University
Montclair, NJ, USA
kaplunk1@montclair.edu

Christopher Leberknight
Montclair State University
Montclair, NJ, USA
leberknightc@montclair.edu

Anna Feldman
Montclair State University
Montclair, NJ, USA
feldmana@montclair.edu

ABSTRACT

Automatic keyword analysis is often performed around the world to limit individual access to online content. To enable citizens to freely and openly communicate on the Internet, research is required to study the predictive quality of single words to detect controversial content. This paper extends our previous work with a larger topic-diverse dataset of 1,068,621 words collected from 23 RSS feeds over a 2 month period. Reliability of prior results and the relationship between controversy and sentiment is examined by reproducing a crowd-sourced experiment. Results from the experiment suggest that controversial and not controversial words are classified by human annotators with a high degree of reliability, but unlike previous research we determine that single words are not useful for detecting controversy. In addition, while we cannot conclude that sentiment alone can be used to predict controversy we find that the variance of sentiment may be a useful metric for partitioning data into distinct clusters. Specifically, we find that higher sentiment variance provides greater discrimination quality compared to using positive and negative sentiment to classify controversial documents.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Natural language processing**; • **Social and professional topics** → *Censorship*;

KEYWORDS

classification, controversy, Internet censorship, sentiment analysis

ACM Reference Format:
Kateryna Kaplun, Christopher Leberknight, and Anna Feldman. 2018. Controversy and Sentiment: An Exploratory Study. In *SETN '18: 10th Hellenic Conference on Artificial Intelligence, July 9–15, 2018, Rio Patras, Greece.* ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.1145/3200947.3201016

SETN '18, July 9–15, 2018, Rio Patras, Greece

ACM ISBN 978-1-4503-6433-1/18/07…$15.00
https://doi.org/10.1145/3200947.3201016

1 INTRODUCTION

Automatic keyword analysis is often performed around the world to limit individual access to online content. Keywords are filtered and analyzed to identity "inappropriate content" that is often used as the criteria for blocking access to online resources or revoking rights to publish content. While we believe that censored content is topic independent, our efforts in this research focus on controversial content in online news media. We define controversial content as any contentious matter or argument that may spark public debate. Our approach is motivated by several factors. First, news is a significant part of our everyday lives. It shapes our beliefs and opinions on how we see the world and now more than ever people rely on a variety of online resources for their news. Consequently, online news sites are a prime candidate for censors to regulate as they have an enormous potential to disrupt the status quo. Several examples highlight the increasing threats to Internet Freedom and methods for restricting information access [9], [2], [3]. Second, the variety of news sources provide a diversity of topics necessary for analysis. Third, it is hypothesized that controversial content will garner more attention, and hence be more rapidly censored compared to content that is not controversial.

Therefore, we aim to identify the potential of using single keywords to detect controversial content or topics that will inform future work on Internet censorship. It is also important to know what kind of sentiment these topics emote for people as these ideas can be related. This can be used as a feature in determining if an article is controversial through the positive or negative words that occur in the article. By studying the relationship between sentiment and controversy in online news articles, we can better understand how news sources and people in general use language to share and foster discussion about certain ideas. Results provide insight into the accuracy and limitations with enforcing censorship using automatic keyword filtering.

1.1 Previous Work

There has been prior work on classifying controversial documents using probabilistic methods [10], logistic regression [12], support vector machine (SVM) [13], and nearest neighbor [7]. Term frequency - inverse document frequency abbreviated tf-idf gives a weight based on how often a lexicon appears in that specific document as a fraction of how often it appears in all documents. It is frequently used as a baseline for research in this field [6]. Sentiment and bias in language through news sources was examined using a crowd-funding technique

with human annotators that were tasked with classifying controversial articles [7].

They developed a list of strongly controversial, somewhat controversial, and not controversial terms and compared these terms against known sentiment and bias lists. Finally, they gave a score to each topic in the controversial and not controversial terms by using logistic regression, where the input features are the proportion of words from each lexicon and the training data is the manually-labeled data. Their results inspired the approach we take in our study and provide the basis for using sentiment lexicons in the future to understand type of content published by different news sources.

However, this experiment has some limitations. Out of 21 million articles they used they selected the top 2000 most frequent words and due to issues with inter-rater agreement they only used 462 words or about 23%. The number of words is too small a sample and will be difficult to produce the same results they achieved in their study. To investigate the reliability of their results, we reproduce their experiment to evaluate predictive accuracy for potential use with other datasets.

Another method for detecting controversy is a nearest neighbor approach [7]. Their algorithm assumes that the controversy in a web document can be detected from controversy of related topics. It models topics that are related to Wikipedia articles and existing controversy labels on neighbors are used to decide for the original web document. This experiment has some limitations as well. The algorithm is dependent upon Wikipedia controversy indicators, produced from Wikipedia specific features. Searching for k nearest neighbors for each document is non-trivial and therefore, this could be practically inefficient [10]. Another limitation is that it is necessary for the topic to be covered by a Wikipedia article because if such an article does not exist, several parameters in the model cannot be calculated [10]. There are also generalization limitations with domain specific sources such as Wikipedia's edit history features because these cannot be generalized to any other sources [10].

A probabilistic method for detecting controversy based on the kNN-WC algorithm [7] uses binary classification such that for a document D P(C|D) is the probability that D is controversial, while P(NC|D) is the probability that D is not controversial. In order to perform binary classification, they test whether P(C|D) > P(NC|D). This is similar to a previous work [7], but now done on a probabilistic measure rather than a binary measure with logistic regression. They also extend the scoring function by removing the threshold and converting the aggregation function to a probability which normalizes over all of the nearest neighbor documents [10]. Their experiment [10] is also limited since it is based on domain specific Wikipedia features.

Some past work uses sentiment to detect controversy [5] but others argue that these two concepts do not overlap and that sentiment is a poor predictor of controversy [10]. Positive and negative sentiment words have been used to detect controversy using a mixture model of topic and sentiment [5] [10]. Another experiment that uses this method detects controversies surrounding celebrities on Twitter [13]. 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 [13]. Although this experiment uses Twitter, it does not use Twitter-specific features and can be more easily generalized to a wide variety of data. This paper is an extension of our previous work [11] that suggests (1) existing annotated datasets of controversial and not controversial terms provide poor predictive quality for detecting controversy in un-labeled documents, (2) while it appears that words that emote negative sentiment are found more in controversial documents and words that emote positive sentiment are found more in not controversial documents as in the previous study, [7] the results are not statistically significant. Therefore, sentiment does not help discriminate between controversial and non-controversial documents, unlike reports from a previous study [7]. A major limitation with our previous research [7] is the size of the dataset. Consequently, we further examine the aforementioned results with a significantly larger dataset (1,068,621 words vs. 317,361 words) to test if previous results were biased as a result of the sample size.

2 EXPERIMENTS

Three experiments inspired by previous research [7], explore the use of single words and positive and negative sentiment for detecting controversy in online news articles. Experiment I aims to test the reliability of previously annotated controversial words [7] for detecting controversy in unlabeled documents. Experiment II provides a descriptive analysis comparing the frequency of positive and negative words in our dataset compared to previously annotated sentiment datasets [5]. Experiment III statistically tests the claim that the proportion of negative sentiment in controversial text will be higher than the proportion of positive sentiment in non-controversial text.

3 METHODOLOGY

An application was developed to collect thousands of English-language news articles from 23 different RSS feeds. Next the application performs stemming, removes all stop words and generates a continuous bag of words (CBOW) for analysis. This forms the test data set that we compare with words that have been previously annotated as controversial terms, somewhat controversial terms, and not controversial terms [7]. In testing their terms against our datasets, we set up a baseline for our articles as seen in Table 1. Using the baseline datasets, we determine whether our dataset has sufficient terms that can be classified as controversial, somewhat controversial, and not controversial. We also compare our dataset with Wikipedia words that are in a list of controversial topics from Wikipedia from previous research [13]. The primary focus of our work is to determine the reliability and generalizability of results from previous work [7] or if they vary from the words in the Wikipedia lists [1]. In addition, we evaluate sentiment

Table 1: Baseline datasets with number of words

Dataset	Type	Number of Words
MicroWNOP [5]	Positive	418
MicroWNOP [5]	Negative	457
General Inquirer [4]	Positive	1628
General Inquirer [4]	Negative	2000
Mejova [7]	Controversial	145
Mejova [7]	Not Controversial	272
Wikipedia [1]	Controversial	2133

by comparing our article dataset with two sentiment datasets, MicroWNOP [5] and General Inquirer [4]. Results from our experiments highlight the potential for using previously annotated controversial datasets [7] for classifying controversial documents.

4 DATASETS

As seen in Table 1, we used datasets of words that already exist as baselines and they were compared against our datasets. The controversial dataset comes from Wikipedia [1], where they have developed a list of controversial topics. Since our experiment focuses specifically on words, this dataset had to be filtered to contain single controversial words rather than topics. This dataset built by editors on Wikipedia was deemed controversial because they are constantly being re-edited in a cyclic way, have edit warring issues, or article sanction problems [1].

4.1 Baseline Datasets

4.1.1 Dataset I. Our dataset, referred in this paper as Dataset I, contains 1,068,621 word extracted from 4220 articles. This dataset is larger (200% increase) and contains a wider range of content compared to our previous work [11], which will help evaluate bias that may exist in the smaller dataset.

5 RESULTS

Each baseline dataset is analyzed with our dataset consisting of 4220 articles (1,068,621 words).

5.1 Crowd Source Experiment

An experiment was conducted with 33 annotators to classify previously labeled words [7] as controversial, somewhat controversial, and not controversial. In total, there were 462 words and only 20 out of the 33 annotators classified all words. Data from all 20 subjects was analyzed and classified to the category that received the maximum number of votes. In the event of a tie, the word was discarded. Sixteen words were discarded and based on the previous study [7] 13/16 were controversial and 3/16 were somewhat controversial. While this suggests there may be more difficulty in evaluating controversial words, this only accounts for 3.5% of the words. The remaining results were very consistent to previous reports [7]. Table 2 presents the classification results. It can be observed that the best classification performance is with not controversial data (93%) followed by controversial data (71%). This suggests that words may create some notion of controversy for individuals yet it has not been rigorously demonstrated that the words can be used to classify unlabeled documents.

Table 2: Classification results

Classified	Precision	Recall	F1
Controversial	0.913978495	0.586207	0.714286
Somewhat Controversial	0.225	0.2	0.211765
Not Controversial	0.871794872	1	0.931507

Figure 1 further illustrates the average performance of annotated results compared to annotated results in previous research [7]. F1 for controversial data is slightly less compared to controversial data. This is most likely due to the difference in results for the 13/16 controversial words that were not classified into any of the categories due to ambiguities with inter-rater agreement. Also, Figure 1 suggests most performance measures are above 85% with the exception of data classified as somewhat controversial. While this requires further examination, the more interesting cases lie at the extremes. In terms of censorship, content classified as somewhat controversial would likely go through more extensive hence manual reviews before making a final determination. However, in the event of manual inspection, where decision confidence is high, there is still a possibility for human error. Based on our results this error is measured by a high false negative rate for controversial terms. For example, a small recall value for controversial content suggests a high false negative rate which implies a lot of controversial content is not being flagged as controversial. This is most critical for the censors. Similarly, small precision value for non-controversial content suggests a high false positive rate. Hence, a large portion of censored content is being classified as not controversial. In both cases the censors are performing poorly and allowing access for censored content. However, since we see good precision and recall for non-controversial content, the main point for future research will be on understanding how to exploit the limitation with classifying controversial content to facilitate the free flow of communication.

5.2 Experiment I: Controversy

Results from previous research [11] suggest that the controversial terms represented only a small fraction of words in dataset. This underscores larger lexicons are required and other features need to be considered as words alone do not provide enough context to discriminate between controversial and not controversial documents. This experiment measures how often words in our larger dataset of 1,068,621 words appear in the baseline datasets. The aim of the experiment is to test if existing annotated controversial and non-controversial

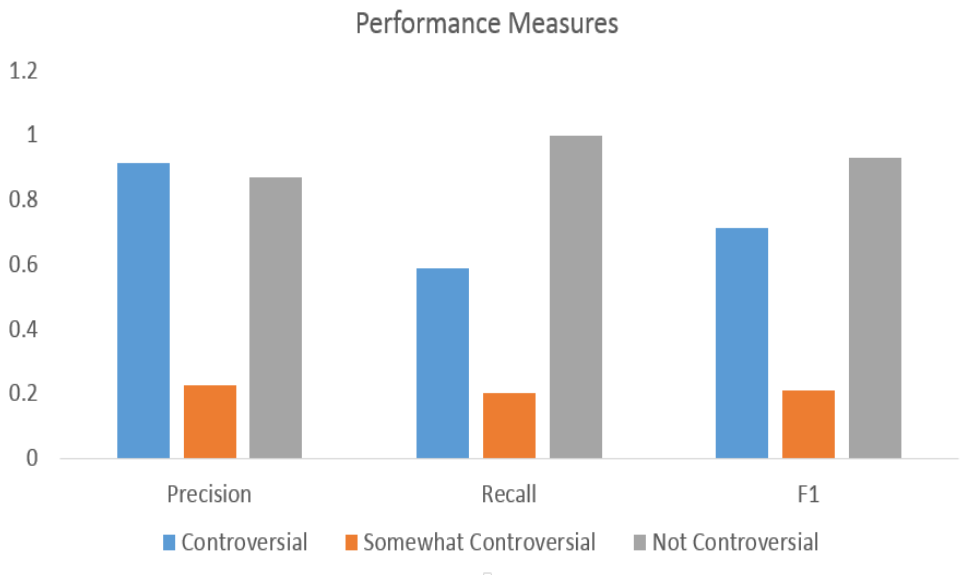

Figure 1: Relationships between classification measures

datasets (baseline datasets) comprised of single words can be used to classify unknown documents by measuring how frequently words in a document appear in each of the baseline datasets. Therefore, to evaluate if baseline controversial datasets represent a comprehensive list of words to classify unlabeled documents, the frequency of words in the baseline datasets is computed for dataset I. The normalized proportion is calculated by taking the frequency of the words found in our dataset and dividing it by the total number of words in the dataset, in this case, 1,068,621. The results indicate that while dataset I contains over 200% more words compared to data used in our previous work [11] there was only a 1.4% increase in the normalized proportion for the Mejova dataset [7].

The Wikipedia dataset [1] having almost twenty times the amount of words to search for in dataset I showed a 4% decrease in the normalized proportion of words found in our previous dataset. It can be concluded that even with over 200% more data in dataset I, the normalized values emphasize that existing controversial lexicons consisting of single words cannot accurately detect controversy. The words that matched each dataset were words that dealt with social and political controversy, which is entirely what the Mejova controversial dataset [7] is composed of whereas it is only a small portion of the controversial content that is in the Wikipedia dataset [1]. This can also be due to the fact that the Mejova dataset [7] was originally structured as specific words whereas the Wikipedia dataset was structured as topics and had to be sorted through to obtain single words from these topics, which could include adding words into the dataset that may not be very controversial. A lot of the most frequent controversial words in the Mejova dataset [7] are terms that are associated with politics and social issues. In the not controversial Mejova dataset [7], a lot of the words that appear most frequently refer to different representations of time such as lengths of time like *day*, *month*, *year*, various seasons, and various specific days of the week or months of the year.

5.3 Experiment II: Sentiment

Previous results [11] suggest that positive sentiment is found more in not controversial documents across both baseline datasets compared to the fraction of words that emote negative sentiment in controversial documents. This appears consistent with previous results [7], but a more thorough statistical analysis is required to confirm these results. Four baseline datasets for positive and negative sentiment were compared with dataset I. With the positive and negative sentiment datasets, the number of words in the datasets matched up more with the frequencies found in dataset I compared to previous work [11]. Therefore, there does appear to be a correlation between the frequency of positive sentiment and not controversial terms and the frequency of negative terms and controversial terms. The General Inquirer datasets [4] were approximately three or four times larger than their MicroWNOp [5] counterparts and the frequencies were also approximately three or four times larger unlike in the controversy datasets [7] , [1]. This demonstrates possible similarities in the datasets. In both cases, the frequencies and proportions for the positive datasets are almost double their negative counterparts.

5.3.1 Language examples. The words that are the most frequent include words that are generic words such as *good*, *well*, *full*, and *right* as well as verbs and nouns that reflect growth such as *promote*, *culture*, and *project*. There are also

Table 3: Top 10 most frequent Mejova controversial words [7] in our dataset

Rank	Word	Freq. in our dataset
1	China	12542
2	Chinese	5341
3	World	2644
4	Government	2373
5	Country	2089
6	President	1466
7	Law	1062
8	Economy	799
9	Media	763
10	Security	731

some words that demonstrate closeness such as *together*, *energy*, and *play*. Some of the words that appear most frequently involve general negative terms such as *take*, *need*, *get*, *demand*, *despite*, *return*, *waste*, and *sell*. Some of the words demonstrate a stronger negative connotation such as *hit*, *storm*, and *abuse*. The other words in this list such as *insurance*, *cause*, *story*, *case*, and *cover*, which can be negative but are more open to interpretation depending on the context that surrounds them.

The top twenty most frequent words of this dataset contain a lot of generic positive words such as *well*, *like*, *even*, *good*, *back*, and *great*. The remainder of the words promote growth and progress such as *help*, *cooperation*, *open*, *better*, *education*, and *support*. The words that overlap in the top twenty most frequent in the positive sentiment datasets from MicroWNOp [5] and the General Inquirer [4] are *good*, *well*, and *back*. There are similarities in the words in the sense that a lot of them promote growth and improvement but none of the actual words are similar besides the three that overlap.

There are mostly generic negative words and a few words that are extremely negative such as *poverty* and *war*. The types of words included in this list have similar connotations to the words in the MicroWNOp negative sentiment dataset [5] . The only words that actually overlap in the twenty most frequent words in the negative sentiment datasets are *need*, *get*, *hit*, and *opposition*. The negative and positive sentiment datasets for the General Inquirer [4] both include the words *make*, *even*, and *help*, which is subject to interpretation depending on context. The words from the Mejova dataset [7] that are found most frequently in our dataset concern topics such as countries, language, education, and powerful people or groups of people. The top twenty words found account for 62.5% of all of the Mejova words [7] found in our dataset. This indicates that these most frequent words contribute to a large portion of the controversial words in our dataset. The top 10 words are provide in Table 3:

5.4 Experiment III: Controversy and Sentiment

Results from prior work [11] suggest there is not enough conclusive statistical evidence to determine that negative words are more likely in controversial words than not controversial words or that positive words are more likely in not controversial words than controversial words. Significant results in a previous study [7] were obtained due to the fact that they tested their sources against the sentiment lexicons [7]. We examine the baseline datasets against each other in order to study the relationship between sentiment and controversy. The frequencies and proportions when these datasets are compared give the exact same frequencies and proportions as our previous study [11] . This implies that although dataset I is larger and contains a wider range of topics, since we are testing them against the same baseline datasets with the same words, there are no new words that appear in them. Subsequently, since the frequencies and proportions are the same, our two proportion z tests will use the same proportions generating the same z statistics and p-values. Therefore, using dataset I in conjunction with our baseline datasets still demonstrates that unlike previous work [7] using specific words that emote positive and negative sentiment provide poor predictive quality for detecting controversy. Consequently, further research is required to identify new features and methods for discriminating between controversial and not controversial text.

6 TESTING ON MEJOVA'S DATA [7]

Mejova et al [7] received statistically significant results in finding that negative terms occurred more frequently in controversial topics compared to not controversial topics and positive terms occurred more frequently in not controversial topics compared to controversial topics. Mejova et al. [7] performed statistical tests on a news source compared to a proportion of words in a lexicon and obtained significant results. Our results, however, did not take the news source into account and we simply tested the proportions of controversial and non-controversial words. Our results could have differed due to the fact that previous study [7] tested combinations of sources and sentiment datasets, whereas we compared the controversial and non-controversial datasets with the sentiment datasets. Since our results differed from previous work [7], we next investigate sentiment variance which has shown to have some promising results for detecting controversial documents [8].

6.1 Total Sentiment Variance

Total sentiment variance in the words separated by each category is examined using SentiStrength [8]. The sentiment scores are calculated on a 1 to 5 scale for positive sentiment, where 5 is the strongest. Negative sentiment is computed on a -1 to -5 scale where -5 is the strongest. The strongest positive and negative sentiment is computed to examine the difference in variance between sentiments. The results are summarized in Table 4.

The program finds the word that emotes the strongest positive sentiment and outputs the corresponding number for a positive sentiment score. It then finds the word that emotes the strongest negative sentiment and outputs the corresponding

Table 4: Positive and Negative Sentiment Scores

Classification	Positive	Negative	Difference
Controversial	2	-5	7
Somewhat Controversial	1	-2	3
Not Controversial	2	-2	4

number for a negative sentiment score. These words and their corresponding numbers can be found in EmotionLookUpTable, which is the name of the dictionary that SentiStrength uses. If there are no positive words inputted that are in the dictionary, SentiStrength outputs a 1 and if there are no negative words that are in the dictionary, SentiStrength outputs a -1. For example, in the sentence "I hate Paul but I encourage him," the word *hate* has a score of -4 and the word *encourage* has a score of 2. These are the highest and therefore, the positive sentiment score is 2 and the negative sentiment score is -4, making their difference 6. The possible differences range from 2 (a positive score of 1 and a negative score of -1) to 10 (a positive score of 5 and a negative score of -5).

Table 4 indicates that the words that are classified as controversial in [7] have a larger difference among their positive and negative scores than the words that are classified as somewhat controversial and not controversial. The larger the difference, the more positive and negative sentiment variation exists in the data. It is interesting to note that the words classified as somewhat controversial have a smaller variation than the words that are classified as not controversial. However, since controversial words have a larger variation, this feature is more useful to use than the positive and negative sentiment words themselves.

7 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 datasets have a counterpart such as the Mejova not controversial dataset and the controversial dataset [7]. However, there is no not controversial dataset for Wikipedia, which hinders our ability to test that against the Wikipedia controversial set [1] or compare to our sole not controversial dataset. Another limitation is that the negative sentiment MicroWNOp dataset [5] and the Mejova not controversial data [7] have zero words that overlap, making it difficult to analyze these two together as well as run any tests. Also, some words appear on both the positive and the negative sentiment datasets, which can be 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.

8 DISCUSSION

This work differs from previous work [11] because we explore the words themselves more closely and hypothesize that many of the overlaps in our datasets are words that have multiple connotations. This makes it difficult for the words to be used as predictors of controversy because in certain context a word can be controversial but in other context it would not be considered controversial. Therefore, we determine that the positive or negative sentiment words themselves are not enough to use in controversy detection. Due to this idea, in this paper, we aim to locate other features. Now, we quantify the extreme positive sentiment and extreme negative sentiment in order to explore if sentiment variance is a more useful predictor in detecting controversy. In this way, it is evident that the controversial data in the Mejova experiment has a larger sentiment variance than the somewhat controversial and not controversial data. This demonstrates to us that sentiment variance can be a beneficial feature, whereas sentiment lexicons themselves were not.

9 CONCLUSION

While we have observed that there is a good degree of reliability for classifying controversial terms in our human subject experiment, the frequency of these terms and the use of sentiment analysis is not sufficient for classifying unlabeled documents. It is evident that the most frequent words occur in large proportion out of all of the Mejova words. This indicates that words that are deemed controversial appear very frequently but the words alone do not occur frequently over the entirety of the dataset, making it inefficient to use them as a predictor for controversy. Additionally, although some research has shown that sentiment and controversy have a relationship, we come to the conclusion that sentiment lexicons are not enough to detect controversy but exploring sentiment variance across the dataset is a more useful indicator of a corpus being classified as controversial. Unlike previous research, our research does not show that negative sentiment occurs more in controversial data and positive sentiment occurs more in not controversial data but rather that words with stronger connotations that are positive and negative appear in controversial data and words with weaker connotations appear in not controversial data.

10 FUTURE WORK

Future work will further explore sentiment analysis and other features as well as analyze information gain across varying amounts of content within a document. Determining the controversiality of article data can assist future research by providing a predictor for censorship. If censorship can be predicted, a system can be built to circumvent censorship. However, different countries have different censorship systems. Likewise, since their systems are different, the items they want to censor vary depending on the country. For example, a topic such as abortion is very controversial in the United States that sparks a lot of debate whereas in other countries this topic may not be as controversial.

We will develop a Bayesian model for classifying controversial documents inspired by previous research [10] that has shown some promising results. We will investigate the use of word embeddings to improve our classification accuracy. Word embeddings are a form of language modelling that uses a specific function, such as a neural network, to map words to a set of numbers in high-dimensional space [6]. Similar words are close to each other in the number space and dissimilar words are far apart [6]. This method differs from the kNN-WC algorithm because the kNN-WC algorithm tests if a document related to a controversial document is controversial whereas the word embeddings test if words similar to controversial words are controversial. This will create a lexicon of controversial words that extends the list in Mejova et al [7]. There are many lists of sentiment lexicons but there are no extensive lists of lexicons for controversy. We hope to create a list that can be used to more accurately classify controversial documents regardless of their genre.

ACKNOWLEDGMENTS

The work is supported by the National Science Foundation under Grant No.: 1704113, Division of Computer and Networked Systems, Secure & Trustworthy Cyberspace (SaTC).

REFERENCES

[1] 2018. Wikipedia: List of Controversial Issues. (2018).
[2] Ilhem Allagui and Johanne Kuebler. 2011. The arab spring and the role of icts introduction. In *International Journal of Communication*, Vol. 5.
[3] Thomas Chen. 2011. Governments and the executive internet kill switch. *Network, IEEE* 25, 2 (2011), 2–3.
[4] R. V. Chimmalgi. 2010. Controversy trend detection in social media. *Master's Thesis, Louisiana State University* (2010).
[5] Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. 2010. Identifying Controversial Issues and Their Sub-topics in News Articles. *Intelligence and Security Informatics* (2010), 140–153.
[6] Jedidiah R. Crandall, Daniel Zinn, Michael Byrd, Earl Barr, and Rich East. 2007. Concept Doppler: A Weather Tracker for Internet Censorship. *CCS âĂŹ07: Proceedings of the 14th ACM Conference on Computer and Communications Security* (2007), 352–365.
[7] Shiri Dori-Hacohen and James Allan. 2015. Automated Controversy Detection on the Web. *Advances in Information Retrieval 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 âĂŞ April 2, 2015.* 423 (2015).
[8] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media. *WSDM '16 Proceedings of the Ninth ACM International Conference on Web Search and Data Mining San Francisco, California, USA February 22 - 25, 2016* (2016).
[9] Mustafa E. Gurbuz. 2014. The Long Winter: Turkish Politics after the Corruption Scandal. 15 (2014).
[10] Myungha Jang, John Foley, Shiri Dori-Hacohen, and James Allan. 2016. Probabilistic Approaches to Controversy Detection. *Conference on Information and Knowledge Management* (2016).
[11] Kateryna Kaplun, Christopher Leberknight, and Anna Feldman. 2018. Measuring Controversy in Online News. In *Proceedings In International Conference on Language Resources and Evaluation (LREC), Workshop.*
[12] Yelena Mejova, Amy X. Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and Sentiment in Online News. (2014).
[13] Marco Pennacchiotti and Ana-Maria Popescu. 2010. Detecting Controversies in Twitter: A First Study. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management ACM New York, NY.*

Received February 2018; revised March 2018; accepted March 2018