Telling Apart Tweets Associated with Controversial versus Non-Controversial Topics

Aseel Addawood
Illinois Informatics Institute
University of Illinois at Urbana-Champaign
aaddaw2@illinois.edu

Rezvaneh Rezapour
School of Information Sciences
University of Illinois at Urbana-Champaign
rezapou2@illinois.edu

Omid Abdar
Department of Linguistics
University of Illinois at Urbana-Champaign
abdar2@illinois.edu

Jana Diesner
School of Information Sciences
University of Illinois at Urbana-Champaign
jdiesner@illinois.edu

Abstract

In this paper, we evaluate the predictability of tweets associated with controversial versus non-controversial topics. As a first step, we crowd-sourced the scoring of a predefined set of topics on a Likert scale from non-controversial to controversial. Our feature set entails and goes beyond sentiment features, e.g., by leveraging empathic language and other features that have been previously used, but are new for this particular study. We find focusing on the structural characteristics of tweets to be beneficial for this task. Using a combination of emphatic, language-specific, and Twitter-specific features for supervised learning resulted in 87% accuracy (F1) for cross-validation of the training set and 63.4% accuracy when using the test set. Our analysis shows that features specific to Twitter or social media in general are more prevalent in tweets on controversial topics than in non-controversial ones. To test the premise of the paper, we conducted two additional sets of experiments, which led to mixed results. This finding will inform our future investigations into the relationship between language use on social media and the perceived controversiality of topics.

1 Introduction

The micro-blogging platform Twitter is a central venue for online discussions and argumentation. This service has also been widely used to disseminate information during emergencies and natural disasters, and to mobilize support for social and political movements (Lotan, Graeff, Ananny, Gaffney, & Pearce, 2011). As with many other outlets of public opinion, Twitter features the emergence of polarization around controversial issues (Addawood & Bashir, 2016; Garimella, De Francisci Morales, Gionis, & Mathioudakis, 2016), and provides a forum where people can express their opinions, which may be conflicting (Pennacchiotti & Popescu, 2010).

This paper focuses on the classification of tweets on topics that are perceived as controversial versus non-controversial. A distinction needs to be made between controversiality and controversy. "Controversy" can be understood as the dyadic or social act of discussing or arguing about an issue (Chen & Berger, 2013). This concept is not addressed in this paper. "Controversiality" means that multiple, potentially conflicting or opposing, viewpoints or opinions have been expressed on a given topic, and people may argue about them or not (Dori-Hacohen, Yom-Tov, & Allan, 2015). In this article, we focus on detecting tweets associated with controversial versus non-controversial topics. Our goal is to gain a better understanding of language-related and tweet-related features that people use in tweets on controversially versus non-controversially perceived topics.

The identification and characterization of controversial topics is difficult for several reasons. First, what is regarded as controversial depends on the senders and receivers of information as well as on the context of a topic in terms of space and time. Second, understanding or even resolving controversies on the individual level may require expertise that may not be part of everybody's general knowledge; making the construction of con-

Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, pages 32–41,
Vancouver, Canada, August 3, 2017.

sensus challenging in terms of creating a comprehensive and shared knowledge base in the first place. Third, the potentially continuously evolving nature of information and knowledge further adds to this challenge.

Previous research used Twitter for detecting both controversy and controversiality (Conover et al., 2011; Garimella et al., 2016; Pennacchiotti & Popescu, 2010). To date, much of the previous research on controversiality has used data from political debates (Adamic & Glance, 2005; Conover et al., 2011; Mejova, Zhang, Diakopoulos, & Castillo, 2014; Morales, Borondo, Losada, & Benito, 2015), news (Awadallah, Ramanath, & Weikum, 2012; Choi, Jung, & Myaeng, 2010; Mejova et al., 2014), and social media, such as blogs (Adamic & Glance, 2005), and Wikipedia (Dori-Hacohen & Allan, 2013; Kittur, Suh, Pendleton, & Chi, 2007; Rad & Barbosa, 2012).

To detect tweets about controversial versus non-controversial topics, we first built a questionnaire to identify such topics that are discussed in the U.S. by using social media and crowdsourcing. We then collected a total of 247,340 tweets from between January 1 to November 28 of 2016. Our research focuses on the underlying characteristics of tweets and demonstrates that the considered features are useful for distinguishing tweets on controversial versus non-controversial topics.

The rest of the paper is organized as follows: The literature review discusses how this work fills a gap in prior work. The data section describes the topic and corpus selection. In the method section, we explain the feature selection and classification. We then report the results of our empirical evaluation of the classifier. We conclude with a discussion of possible improvements and directions for future work.

2 Literature Review

2.1 Controversiality Detection in Online News

To quantify controversiality in online news, Choi, Jung, and Myaeng (2010) leveraged positive and/or negative sentiment words to compute the degree of controversiality. Mejova and colleagues (2014) report a high correlation between a) controversial issues and b) the use of negative affect and biased language. Awadallah and colleagues (2012) describe a method where opinion holders and their opinions as extracted facets from Web result snippets were identified through an iterative process based on a seed set of patterns that describe expressions in either support or opposition to an idea.

2.2 Controversiality Detection Using Other Sources

Some prior work on detecting controversiality leveraged Wikipedia, where structured data and revision histories provide relevant data related to conflicting opinions (Kittur et al., 2007). Using Wikipedia data, Rad and Barbosa (2012) compared five methods for identifying and modeling controversy and controversiality. Das and colleagues (2013) used controversy detection as one step in studying content manipulation by Wikipedia administrators. Knowledge about controversial articles on Wikipedia has been utilized to evaluate the level of controversy of other documents (e.g., web pages) (Dori-Hacohen & Allan, 2013). Finally, Wikipedia has been leveraged for developing a lexicon or hierarchy for controversial words and topics (Awadallah et al., 2012; Pennacchiotti & Popescu, 2010).

Another line of work has focused on controversy detection in blogs. Mishne and Glance (2006) present a large-scale study of blog comments and their relation to corresponding articles. They addressed the task of finding comment threads indicating a controversy as a text classification problem.

Finally, Tsytsarau, Palpanas, and Denecke (2011) focused on finding sentiment-based contradictions at scale by using data sets as disparate as drug reviews, comments to YouTube videos, and comments on Slashdot posts. Even though sentiment analysis seems an intuitive component for detecting multiple viewpoints (Choi et al., 2010; Pennacchiotti & Popescu, 2010), some researchers have argued that this technique is not sufficient and may not be the right metric with which to measure controversiality (Awadallah et al., 2012; Dori-Hacohen & Allan, 2013; Mejova et al., 2014).

2.3 Controversiality Detection in Twitter

The work closest to ours is that by Pennacchiotti and Popescu (2010), where they sought to detect controversiality about selected celebrities and events associated with them based on Twitter data. Their study measures the presence of terms explicitly associated with controversiality in

celebrity-related tweets, resulting in an average precision of up to 66% in predicting controversiality. The authors operationalized this task as a regression problem to predict a controversiality score of each tweet that mentions a specific celebrity and terms based on a list of controversial topics from Wikipedia. By contrast, we conceptualize this task as a classification problem where we predict if a tweet is about a controversial or a non-controversial topic. We do not address or measure if a tweet or sequence of tweets is controversial, in fact, we do not assume a relationship between the controversiality of tweets and of topics, and vice versa. While the work by Pennacchiotti and Popescu focused on celebrities, we address a broader range of topics.

Our work also relates to that of Conover et al. (2011), who studied controversy in political communication about congressional midterm elections using Twitter data. They found a highly segregated partisan structure (present in the retweet graph, but not in the mention graph), and limited connectivity between left- and right-leaning users.

Overall, we build upon previous work by adding additional features for the given task. We do not solely rely on sentiment analysis, but also extract other features. We also develop a lexicon to identify emphatic language used in tweets on the considered topics based on prior literature, and supplemented that with an existing lexical resource for profanity.

3 Data

3.1 Topic Selection

To identify a set of controversial and non-controversial topics, we first searched controversy-related web sources (i.e. Procon.org), Wikipedia controversiality lists, news media websites, and blogs. The results of this initial search helped us to develop eight claim statements (one statement per topic) on topics (see Table 1).

After formulating these statements, two online surveys were conducted in which the participants rated the statements pertaining to different topics on a 5-point scale ranging from controversial to non-controversial. Participants were randomly assigned to evaluate four out of the eight statements. Table 1 shows the selected topics and associated statements used in the survey.

The first questionnaire was run on Amazon's Mechanical Turk service (MTurk), an online crowdsourcing system. MTurk participants were compensated with $0.10 USD per survey. The survey was available only to U.S. residents with at least 95% approval rating (a screening option that is provided by MTurk). A total of 197 surveys was received from MTurkers, and 172 of them were valid. A response is considered invalid if it did not contain complete answers or was not validated through a validation question.

The second questionnaire was distributed on social media, specifically on Facebook and Reddit. Participants were not compensated for their contribution due to the need of preserving their anonymity. Empty responses and responses that did not contain complete answers were eliminated. A total of 120 responses was received and out of those, 71 were completed. In total (considering both surveys), a total of 243 valid responses was collected. The surveys were conducted over a period of three weeks in October 2016.

To measure the controversiality of a statement, participants were asked to rate how controversial they believed a statement was on a 5-point Likert scale (5 = "very controversial", 1= "not controversial at all"). Based on the participants' average rating of the presented topics (see Table 1), the three-top controversial and non-controversial topics were selected for further analysis: The controversial topics were (a) individual privacy versus national security, (b) the link between vaccination and autism, and (c) gun control. The non-controversial topics were (a) usage of seatbelts,

Categorization	Exemplary Topic	Statement	AVG
Controversial	**Privacy**	*"Citizen privacy takes precedence over national security"*	3.73
	Vaccine	*"MMR vaccine causes autism"*	3.63
	Gun control	*"Access to guns should be more restricted"*	4.10
Non-controversial	**Seatbelts**	*"Seat belt use can save lives in car accidents"*	1.30
	Child education	*"Every child should have access to education"*	1.49
	Sun exposure	*"Skin damage from excessive sun exposure"*	1.43

Table 1: Controversial and non-controversial topics considered in this study.

(b) access to education for children, and (c) detrimental effects of sun exposure.

3.2 Corpus Selection

We used Crimson Hexagon (Etlinger & Amand, 2012), a social media analytics tool, to collect public tweets posted in the time window from January 1, 2016 through November 28, 2016 on the given topics, based on queries we formulated. The sample only included tweets from accounts that set English as their language and that were geo-located in the U.S. The total number of collected and downloaded tweets is shown in Table 2. Out of the total 246,869 unique tweets that were collected, 148,677 were on controversial topics, and 98,208 were on non-controversial topics.

4 Method

4.1 Feature Selection

User-generated text can express various different thoughts in controversial and non-controversial tweets (Davidov, Tsur, & Rappoport, 2010). Our feature selection was motivated by the assumption that features that capture these thoughts would be effective for our classification task. Some of our features, e.g. sentiment (Pennacchiotti & Popescu, 2010), have been previously used for analyzing Twitter data, while others are novel for this task, and are motivated by pragmatic research into linguistic mechanisms related to engagement in controversial talk.

Emphatic Features

***Lexical Emphasis*:** In the pragmatics literature, it is believed that throughout conversation, speakers have a desire for their thoughts and beliefs to be accepted by their audience (Roberts, 1992). Since controversial topics can be expected to result in disagreement or dissent, we expect tweets on these topics to have a heavier reliance on emphatic language. Based on this intuition, we developed a lexicon to help detecting instances of lexical emphasis. We used a taxonomic grammar of English (Celce-Murcia, Larsen-Freeman, & Williams, 1999) to source a list of emphatic words, including emphatic adjectives (e.g., "awful," "horrible," "great," "fantastic," "superb," etc.) and intensifying adverbs (e.g., "perfectly," "extremely," "insanely," "ridiculously," etc.). We added these words to a lexicon of profanity in English (Ahn., n.d.), which was used since the use of swear words has shown to reflect the emotional state of the speaker (Jay & Janschewitz, 2008).

***Orthographic-Based Emphasis*:** Emphasis can also be achieved via orthographic stylistic expressions, including punctuation and upper casing (Davidov et al., 2010). We recorded instances of uppercase words. Social media users also occasionally use repeated exclamation marks to show sarcasm or emphasis. We recorded all instances of the use of one or more exclamation marks in tweets.

Language-Specific Features

Since a previous study showed that using lexical and syntactic features improve the accuracy of detecting controversy (Allen, Carenini, & Ng, 2014), we built upon this finding, but relied on a wider range of language specific features, namely grammatical and psychological features. We used the Python NLTK library (Bird, Klein, & Loper, 2009) and custom python scripts for grammatical features, and LIWC (Linguistic Inquiry and Word Count) (Pennebaker, Booth, & Francis, 2007) for psychological features.

***Psychological Features*:** Controversial topics lead to disagreements in the audience (Dori-Hacohen et al., 2015), and controversial conversations can create misalignment effects that speakers might mitigate (Roberts, 1992). While the exact nature of how these effects occur in conversation can be hard to pinpoint, we included a set of psychological features as defined and provided by LIWC to help in capturing some of these effects from tweets. We extracted instances of the following selected categories available in LIWC (Pennebaker et al., 2007): (a) "Cognition Processes" such as words related to insight, cause, discrepancies, degree of certitude, and difference, (b) "Informal Language Markers" such as assents, fillers, and swear words, (c) "Personal Concerns" such as words related to work, leisure, home, money,

Topic	Number of Download	# After Removing duplicates
Privacy	99,549	73,593
Vaccine	63,137	41,005
Gun control	50,000	34,490
Seatbelts	89,912	73,271
Child education	46,931	10,808
Sun exposure	20,528	14,173

Table 2: Total number of tweets after removing duplicates.

Features	NB			DT			SVM		
	P	R	F1	P	R	F1	P	R	F1
Baseline (Twitter)	62.7	49.0	41.7	69.1	69.4	68.9	65.8	66.0	64.3
Twitter + Emphatic	63.2	50.3	44.2	69.8	70.1	69.7	65.8	66.0	64.3
Twitter + Language-Specific	77.6	77.7	77.4	87.6	87.6	87.6	86.3	86.4	86.3
Twitter + Emphatic+ Language-Specific	77.6	77.7	77.4	87.7	87.7	87.7	86.4	86.4	86.4

Table 3: Results of NB, DT, and SVM using 10-fold cross validation (values are %).

religion, and death, (d) "Social Words," such as words related to family and friends, (e) "Drives," which are words related to affiliation, achievement, power, risk and reward, (f) "Clout", (g) "Tone", (h) "Authenticity", and (i) "Analytical Thinking". LIWC is a dictionary-based tool which associates words with categories. As in the previous step, the presence of various words (in the respective category) is calculated per tweet and then normalized by tweet length.

Grammatical Features: We extracted or calculated the (a) presence of different parts of speech, (b) tweet length, (c) ratio of various pronouns, (d) time orientation of tweets as past, present, or future, calculated using different verb tenses and related adverbs, (e) ratio of comparisons, interrogatives, numbers and quantifiers, (f) sentiment of the tweets from Crimson Hexagon, and (g) the subjectivity or objectivity of tweets, using the MPQA subjectivity lexicon (Wilson, Wiebe, & Hoffmann, 2005). To capture the above-mentioned categories (c, d, e), we counted the number of related words in each tweet and normalized the counts by tweet length.

Twitter-Specific Features

Some text level attributes are specific to Twitter, such as mentions, URLs, and hashtags. Before preprocessing the data, we calculated the number of occurrences of each of these features in a tweet and added them to the set of attributes. We also incorporated the number of repetitions of each tweet in our data as a feature before removing the repeated tweets. In addition, we considered the gender, number of tweets, number of followers, and followings of accounts where available through Crimson Hexagon as Twitter-specific features. The gender of the authors was retrieved from Crimson Hexagon, where gender is calculated using "the distribution of the author names in census data and other public records" (Etlinger & Amand, 2012).

Overall, we considered a total of 90 features. We chose not to use some common features such as bag of word and top TF-IDF words to avoid overly strong domain dependence and topic specificity of the classifier.

4.2 Classification

After preprocessing and before building the classification models, we divided the data into training and testing data. Both sets included controversial and non-controversial topics. After dividing the data, the training set included the tweets from two controversial and two non-controversial topics: Privacy and Vaccines (controversial), and Seatbelts and Child education (non-controversial). The tweets from the other two topics, Gun control (controversial) and Sun exposure (non-controversial), were included in the test set.

As a first step, we compared classifiers that have frequently been used in related work: Naïve Bayes (NB) as used in Teufel and Moens (2002), Support Vector Machines (SVM) as used in Liakata and colleges (2012), and Decision Trees (DT, J48) as used in Castillo, Mendoza and Poblete (2011). We used Weka (Hall et al., 2009) and an R machine learning package (e1071) (Dimitriadou, Hornik, Leisch, Meyer, & Weingessel, 2011) as implementations of these classifiers.

To find the best features, we first built a baseline model using Twitter-specific features only. We then added the other two features to the baseline to find the impact of each set. Next, we conducted 10-fold cross-validation to find the best combination of features to train the model, and then used the best trained model on the test set to evaluate the predictability of tweets on controversial vs. non-controversial topics. In addition, before classifying the tweets, we chose the most efficient features using Information Gain (Eq.1).

Classification Sets	Topics	NB			DT			SVM		
		P	R	F1	P	R	F1	P	R	F1
Train set (10-fold)	**Privacy, Child Education, Seatbelts, and Vaccine**	77.6	77.7	77.4	87.7	87.7	**87.7**	86.4	86.4	86.4
Test set	**Gun Control and Sun exposure**	60	61.3	60.5	66.5	60	61.7	66.5	62	**63.4**

Table 4: Results of the best NB, DT, and SVM models on the test set.

$$InfoGain\ (Class, Attibute) = H(Class) - H(Class|Attribute) \quad (1)$$

To assess the accuracy of the predictions, we used the standard metrics of precision (P), recall (R), and F-score (with $\beta = 1$) (F1). Table 3 lists the results of all features and classification algorithms.

5 Results

5.1 Classification

As shown in Table 3, the best performance of the baseline model (Twitter-specific features only) was achieved by the DT classification algorithm (69.9% F1-score). Adding the emphatic feature to the baseline increased the performance of DT and NB by around 1-2%, but did not change the result of the SVM classification. Adding language-specific features to the baseline only resulted in a jump in the performance of all three classifiers: The Precision, Recall, and F1-scores of all classifiers increased by 14-33%, which shows the effectiveness of this set of features (Table 3). Finally, combining all three features slightly increased the performance of DT and SVM by around 0.01%, but the performance of NB did not change. Overall, as the last row of Table 3 shows, we found the combination of all three features to provide the best performance.

After training, we tested the classifiers on the remaining two held out topics (test set) as a means of evaluating the best model (the combination of all three classes of features) in new controversial vs. non-controversial topics. As shown in Table 4, SVM outperformed the other models, and achieves a final average F1-score of 63.4%.

5.2 Feature Analysis

The Twitter-specific, emphatic, and language-specific features are the most helpful ones for the classification given task. To find the most effective attributes of each feature set, we ranked the attributes by their information gain score (Eq. 1). The attributes with the highest scores are listed in Table 5. The baseline model consists of nine attributes. From those, "Following" and "URL" are the highest ranked attributes. After combining Twitter-specific with emphatic features, "Following" and "URL" from the baseline model remained the top-ranked attributes, and "Uppercase words" benefitted the model more than other emphatic attributes. "Lexical emphasis" also ranked among the top ten attributes of this feature set. Also, we find that Twitter-specific features are more helpful for the detection tweets on controversial than non-controversial topics (Table 6).

The top ten attributes of the Twitter + Language-specific and the Twitter + Emphatic + Language-specific model were dominated by the language-specific features, both their grammatical and psychological attributes.

Regarding the emphatic features, the results show that the ratio of "Uppercase letters" is higher in tweets on controversial topics, while tweets on non-controversial topics have slightly more "Lex-

Feature Sets	Top-Ranked Attributes *(in order of internal ranking from left to right)*
Baseline (Twitter)	Following, URL, Followers, Hashtag, Mention, Gender, Posts, tweet count, Retweet
Twitter + Emphatic	Following, URL, Uppercase, Followers, Hashtag, Mention, Lexical emphasis, Gender, Posts, tweet count
Twitter + Language-Specific	Risk, Six letter, Personal pronoun, Adjective, Sentiment, I, Clout, Punctuation, Dictionary words, Authenticity
Twitter + Emphatic + Language-Specific	Risk, Six letter, Personal pronoun, Adjective, Sentiment, I, Clout, Punctuation, Dictionary words, Authenticity

Table 5: Top-ranked attributes of each feature set based on information gain score.

Feature	Contro. AVG±STD	Non-Contro. AVG±STD
Emphatic Features		
Lexical emphasis	0.66±0.88	**0.82±0.97**
Uppercase	**0.75±2.023**	0.48±1.83
# Exclamation	0.12±0.425	**0.17±0.47**
Language-Specific Features		
Personal pronoun	3.81±5.07	**9.50±8.28**
Preposition	8.69±5.84	**9.80±6.92**
Auxiliary verb	5.26±5.49	**5.97±6.05**
Adverb	2.78±4.19	**3.69±5.007**
Conjunction	2.85±3.94	**3.79±4.65**
Analytic	**74.65±28.33**	63.45±33.43
Authentic	21.59±28.65	**39.81±38.97**
Sentiment	**-0.23±0.61**	-0.08±0.69
Power	**4.55±5.14**	3.02±5.31
Risk	**3.92±4.40**	0.86±2.37
Focus past	1.58±3.20	**2.19±4.24**
Focus present	7.08±6.43	**9.07±7.77**
Focus future	0.67±1.96	**0.94±2.51**
Money	**0.60±1.94**	0.48±2.06
Religion	**0.19±1.11**	0.18±1.26
Death	**0.29±1.30**	0.23±1.26
Twitter-Specific Features		
Retweet	**0.0004±0.02**	0.00015±0.012
Mention	**0.42±0.49**	0.29±0.455
Hashtag	**0.315±0.46**	0.21±0.41
URL	**0.54±0.497**	0.37±0.48

Table 6: Data-driven feature analysis.

ical emphasis" and "Exclamation marks" (Table 6). This result might seem counterintuitive since we expected this set of features to be more significant for controversial topics. Furthermore, the results show that controversial topics have a higher ratio of negative sentiment (Table 6). Our findings support the insight from prior work that sentiment is a helpful feature for controversiality detection, but needs to be supplemented with other features (Awadallah et al., 2012; Dori-Hacohen & Allan, 2013; Mejova et al., 2014). Looking into some of the tweets on one of the non-controversial topic, i.e., "Seatbelts", we saw that these statements reflected an awareness of the dangers, risks, and negative outcomes that could result from ignoring seatbelts. In other words, deviations from socially agreed upon consensus or norms might spur attention and dissent. Alternatively, when tweeting about non-controversial issues, people might focus on controversial sub-aspects, for example, because they are lingering or emerging. Further research is needed to explain our observations and the engagement with non-controversial themes on social media.

		CT 1-vs-all	NCT 1-vs-all
CT	**Privacy**	86.5	80.6
	Vaccine	78.0	74.4
	Gun control	83.4	77.2
NCT	**Education**	84.2	86.1
	Sun exposure	77.4	80.2
	Seatbelts	72.7	76.3

Table 7: Classification results 1-vs-all (F1-measure values are %).

5.3 Testing the Premise of the Project

One potential critique of our study could be that we predict sets of topics rather than overarching, unifying characteristics (controversiality versus non-controversiality) of these set of topics. If that was true, then predicting tweets on controversial topics *CT* based on tweets from other controversial topics *OCTs* should result in higher accuracy than predicting tweets on *CT* based on tweets from non-controversial topics *NCT*. Analogously, predicting tweets on *NCT* based on tweets on other non-controversial topics *ONCTs* should result in higher accuracy than predicting tweets on *NCT* based on tweets on *CT*. We tested the premise of this paper by applying this logic in two ways.

First, we used a "one-versus-all" approach. Using all features, we built binary classifiers (using Naïve Bayes) for each type (CT, NCT) using the tweets on the other CTs or NCTs (the two remaining other topics from the same type, and three from the opposing type), and conducted 10-fold cross validation. Table 7 shows the resulting F-measure values. Using this test, we find that indeed, NCTs are predicted with higher accuracy when learning from tweets from other NCTs than CTs and vice versa in all tested cases, which support the general premise of this paper. This methodology is aligned with the learning methodology used in this paper (Table 3 and 4) where we perform binary classification to predict CT vs. NCT, the difference is that in this additional test, we predict only CT or only NCT.

Topic (CT)	Controversial topics	Non-Controversial topics (NCT)		
		(education, sun)	*(education, seatbelt)*	*(seatbelt, sun)*
Privacy	79.8	77.6	81.2	76.3
Vaccine	69.6	69.5	71.5	65. 2
Gun Control	70.9	69.7	72.0	74.7

Table 8: Prediction results for each CT from the other two CT as well as from all NCT (F1-measure values are %).

Topic (NCT)	Non-Controversial topics	Controversial topics (CT)		
		(privacy, vaccine)	*(privacy, gun)*	*(vaccine, gun)*
Education	88	83.8	81.1	80.1
Sun exposure	75.4	73.4	76.3	74.2
Seatbelt	64.9	68.7	77.3	69

Table 9: Prediction results for each NCT from the other two NCT as well as from all CT (F1-measure values are %).

Second, we predicted each CT from the other two CTs as well as from all NCTs (Table 8). Analogously, we predicted each NCT from the other two NCTs as well as from all CTs (Table 9). This methodology deviates from the learning methodology used in this paper (Table 3 and 4) in that it uses a more detailed approach to predict a single class. Therefore, this test challenges the premise of the paper more strongly or from a different methodological viewpoint than the main method, while the first premise validates our test. The results (Tables 8, 9) show that for each set of experiments, 5 of 9 test cases support the premise of this paper, and 4 out of 9 do not. Table 9 further shows that there might be topic related effects: Seatbelt, a NCT, is easier to be predicted from tweets associated with CT than tweets from NCT. These outcomes call for further research, including pragmatic analysis, into tweet characteristics that indicate tweet association with the controversiality of topics.

6 Discussion

Since noticing controversiality can be a hard task for individuals, we developed a supervised model that detects tweets associated with controversial versus non-controversial topics on Twitter. As a prerequisite for this study, we conducted an online survey where participants rated the controversiality level of sentences related to a selected set of topics. We then selected the topics that the crowd considered as most and least controversial. We trained and evaluated a classifier using three feature sets (Twitter-specific, emphatic, and language-specific features). We considered features new for this particular task, and the linguistic robustness of these features is backed by pragmatic research into the nature of disagreement between speakers during controversial talk (Roberts, 1992).

The considered features proved to be informative for the classification task, albeit with varying degrees of contribution: Twitter-specific attributes such as mentions, URLs, and hashtags helped to build a baseline that performed at 69.9% (F1 score) using the DT algorithm. This finding might be accounted for by the sociolinguistic insight that linguistic communication is socially distributed (Cox, 2005). In other words, Twitter users conform to social stylistic norms of using social media (enabled) features. Moreover, these features were more indicative of controversial than non-controversial topics, which may indicate that social media provides features that people use when making statements related to controversial themes (Table 6).

Emphatic features provide a small contribution to this task (about 1-2% increase in F1 when using DT and NB models). Such features have been previously used for the detection of sarcasm from social media text data (Davidov et al., 2010). Our results suggest that this feature can also improve the detection of controversiality (Table 6), which may be due to social stylistics or an element of sarcasm in the tweets, among other possible reasons. Finally, incorporating grammatical and psychological language-specific attributes resulted in a sizeable increase in the performance of all classifier models. These attributes are not equally distributed across the two types of labels.

7 Conclusion and Future Work

Our results show that focusing on the structural characteristics of tweets offers a means of detect-

ing tweets associated with controversial versus non-controversial topics. This work is limited in several ways. Linguistic and stylistic attributes of language use are subject to temporal and regional variations. Also, some of the features that we considered are not only affected by whether a tweet is related to a controversial topic or not, but also by the context and subject of the tweet. Even given these limitations, we believe this study expands prior work by a) distinguishing between controversy (a communication act or a social interaction, not addresses herein) and controversiality (an aggregate effect of potentially unrelated personal utterances, the object of study in this paper), and b) analyzing the contribution of features that can be assumed—based on prior work and theory—to help distinguish tweets on controversial versus non-controversial topics.

This work raises questions to be addressed in future research. First, we plan to test this approach on other social media platforms in order to study the utility and validity of these features across various outlets. Second, we intend to combine our data mining approach with close reading and qualitative text analysis techniques to explain the counterintuitive effects we have been observing, and to identify the relationship between a) expressions of consensus and dissent on the tweet level, and b) controversiality versus non-controversiality of topics.

Finally, yet importantly, the tests for validating the premise of the paper have provided mixed results: One strategy (one versus all) confirmed our basic idea and goal for all tested cases. This congruence might be due to the fact that the underlying strategy for partitioning the data and predicting classes was similar to the learning methodology. The second strategy (predicting NCT based on other NCTs versus all CTs, and vice versa) partially challenged our premise (confirmed it for 56% of the test cases, rejected for the other 44%). This test used a different logic than the learning experiments. We plan to further investigate the reasons for these discrepancies to inform our future work on identifying controversiality on social media.

Acknowledgment

We thank Professor Corina Roxana Girju from the Linguistics department at UIUC for her helpful insight and direction.

References

Adamic, L. A., & Glance, N. (2005). *The political blogosphere and the 2004 US election: Divided they blog*. In Proceedings of the 3rd International Workshop on Link Discovery, ACM.

Addawood, A. A., & Bashir, M. N. (2016). *"What is your evidence?" A study of controversial topics on social media*. In Proceedings of the 3rd Workshop on Argument Mining, ACL.

Ahn., L. v. (n.d.). Offensive/Profane Word List [Lexicon]. Retrieved from https://www.cs.cmu.edu/~biglou/resources/bad-words.txt

Allen, K., Carenini, G., & Ng, R. T. (2014). *Detecting Disagreement in Conversations using Pseudo-Monologic Rhetorical Structure*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL.

Awadallah, R., Ramanath, M., & Weikum, G. (2012). *Harmony and dissonance: Organizing the people's voices on political controversies*. In Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM), ACM.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*: O'Reilly Media, Inc.

Castillo, C., Mendoza, M., & Poblete, B. (2011). *Information credibility on Twitter*. In Proceedings of the 20th International Conference on World Wide Web (WWW), ACM.

Celce-Murcia, M., Larsen-Freeman, D., & Williams, H. A. (1999). *The Grammar Book: An ESL/EFL Teacher's Course*. Boston, MA: Heinle & Heinle.

Chen, Z., & Berger, J. (2013). When, why, and how controversy causes conversation. *Journal of Consumer Research, 40*(3), 580-593.

Choi, Y., Jung, Y., & Myaeng, S.-H. (2010). *Identifying controversial issues and their sub-topics in news articles*. In Proceedings of the Pacific-Asia Workshop on Intelligence and Security Informatics, Springer.

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). *Political Polarization on Twitter*. In Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM), AAAI.

Cox, A. (2005). What are communities of practice? A comparative review of four seminal works. *Journal of Information Science, 31*(6), 527-540.

Das, S., Lavoie, A., & Magdon-Ismail, M. (2013). *Manipulation among the arbiters of collective*

intelligence: How Wikipedia administrators mold public opinion. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM), ACM.

Davidov, D., Tsur, O., & Rappoport, A. (2010). *Semi-supervised recognition of sarcastic sentences in Twitter and amazon*. In Proceedings of the 14th Conference on Computational Natural Language Learning, ACL.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2011). Misc Functions of the Department of Statistics (e1071), TU Wien. *R package version*, 1-6.

Dori-Hacohen, S., & Allan, J. (2013). *Detecting controversy on the web*. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM), ACM.

Dori-Hacohen, S., Yom-Tov, E., & Allan, J. (2015). *Navigating controversy as a complex search task*. In Proceedings of the ECIR Supporting Complex Search Task Workshop, Elsevier.

Etlinger, S., & Amand, W. (2012). Crimson Hexagon [Program documentation]. Retrieved from http://www.crimsonhexagon.com/wp-cotent/uploads/2012/02/CrimsonHexagon_Altimeter_Webinar_111611.pdf

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2016). *Quantifying controversy in social media*. In Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM), ACM.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18.

Jay, T., & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture, 4*(2), 267-288.

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). *He says, she says: Conflict and coordination in Wikipedia*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics, 28*(7), 991-1000.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., & Pearce, I. (2011). The Arab Spring the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication, 5*, 31.

Mejova, Y., Zhang, A. X., Diakopoulos, N., & Castillo, C. (2014). Controversy and sentiment in online news. *Computation+Journalism Symposium*.

Mishne, G., & Glance, N. (2006). *Leave a reply: An analysis of weblog comments*. In Proceedings of the 3rd Annual Workshop on The Weblogging Ecosystem, (WWW), ACM.

Morales, A., Borondo, J., Losada, J. C., & Benito, R. M. (2015). Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 25*(3), 033114.

Pennacchiotti, M., & Popescu, A.-M. (2010). *Detecting controversies in Twitter: A first study*. In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, ACL.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: LIWC [Computer software]. *Austin, TX: liwc. net*.

Rad, H. S., & Barbosa, D. (2012). *Identifying controversial articles in Wikipedia: A comparative study*. In Proceedings of the 8th Annual International Symposium on Wikis and Open Collaboration, ACM.

Roberts, J. (1992). Face-threatening acts and politeness theory: Contrasting speeches from supervisory conferences. *Journal of Curriculum and Supervison, 7*(No 3), 287-301.

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics, 28*(4), 409-445.

Tsytsarau, M., Palpanas, T., & Denecke, K. (2011). *Scalable detection of sentiment-based contradictions*. In Proceedings of the DiversiWeb 2011: 1st International Workshop on Knowledge Diversity on the Web, in conjunction with WWW, ACM.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis*. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, ACL.