Examining Climate Change Sentiment over Time through Twitter Data

Jason Sinn[1], **Chris T. Bauch**[1,2], **Madhur Anand**[3]
1 Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada
2 Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, Canada
3 School of Environmental Sciences, University of Guelph, Guelph, Ontario, Canada

***Abstract*— Anthropogenic climate change remains a debated topic among the general public, despite a strong consensus among climate scientists that it is a real phenomenon. Although the vast majority of scientific papers agree that humans play a substantial role in global warming, thoughts on social media are far more divided. In this paper, we conduct sentiment analysis on tweets using a support vector machine classifier. Afterwards, we present a new metric of measuring ambiguity in sentences using entropy, and use this metric to investigate possible causes of sentiment change over time. We then compare our results with a discrete approach to analyzing sentiment change over time and show that a continuous measure is more precise. Using this measure, we compare the entropy distributions of users, sentiment populations, and events. We also analyze the validity of using entropy to measure climate change sentiment over time. We finally conclude that denier tweets in the time period are less ambiguous than activist tweets, and that the Trump-Clinton debate had the most impact on climate change sentiment, among other findings.**

I. INTRODUCTION

Scientific papers in the field of anthropogenic climate change have shown that 97% of published research (N=2412) supports the consensus that humans are the cause of recent global warming [1,2]. However, a recent study by *Lieserowitz, et al.* has shown that climate change denial rose by 7 percent, up to 23 in November 2013 [3]. In the same study, it was concluded that 49 percent of Americans believed that global warming was caused by human activities. Furthermore, opinion polls on the popular site *Gallup.com* have shown that only 39 percent of American adults were concerned believers of anthropogenic climate change, whereas 25 percent were deniers as of March 2014 [4]. This rising denier population has been a cause of concern among activists and scientists.

In recent times, the amount of user sentiment data on social media websites has been growing exponentially. In particular, the micro-blogging site *Twitter* has become a popular location to share opinions on a variety of subjects. One of the subjects that is occasionally tweeted about is climate change. As a result, Twitter can be used to provide insights on the sentiments of the general population on topics such as climate change.

Meanwhile, machine learning has been a rapidly growing field, becoming popular in a large variety of areas over recent years. One of the areas of research in machine learning in the context of natural language processing is sentiment analysis. The goal of basic sentiment analysis is to learn text patterns in order to predict whether the sentiment of some text is positive, neutral, or negative. Common approaches to sentiment analysis include latent dirichlet allocation [5], naive-bayes classifiers [6], and support vector machines [6]. Other forms of sentiment analysis not related to machine learning include large semantic databases [7] and hedonometers [8]. One area of research that is still relatively unexplored however, is how user sentiment on global warming changes over time.

The explosion in machine learning research, easily accessible sentiment data from Twitter, and recent climate change events [10] have lead to an interest in analyzing the effects of various individuals and related events on a Twitter user's sentiment.

In the past, a number of techniques have been used to quantitatively measure sentiment on climate change. *Cody et al.* [8] discussed the use of a hedonometer, a technique that involves measuring the happiness of words using a database of survey-driven word-happiness ratings. They also analyzed specific dates and events, and found that climate change tweets were more politicized than the average tweet. It must be noted however, that the sentiment of a user and the sentiment of words used are not necessarily correlated.

In the field of machine learning, *Boussalis et al.* [5] discusses the use of latent dirichlet allocation, a form of unsupervised learning, to find underlying topic families. However, the data in this paper was in the form of large articles with explicit context. The topics that were found were also objective in nature, with topics such as "renewable energy", "fuel standards", and "oil production" dominating the trees.

Meanwhile *An et al.* [6] used both a naive-bayes implementation as well as a support vector machine in order to classify tweets on the dimensions of polarity and subjectivity. These two dimensions were separated, and the overall accuracy was low (combining both subjectivity and polarity, 60%). This was largely due to the large amount of uncertainty, diversity, and noise in combination with the highly implicit context that is present in Twitter data.

In *Soni et al.* [9], the concept of sentiment analysis over time was explored in the context of customer reviews. The paper discussed the use of a hidden markov model to find patterns in customer reviews. This naturally led to interest in sentiment analysis over time on Twitter.

This paper explores the relationships between user, sentiment, and activism over time on Twitter data. In order to combat the ambiguous nature of the data, a quantitative

measure of uncertainty is introduced. This measure is continuous in nature, and is defined with entropy, an information theory metric. The entropy changes over time are analyzed in accordance with several events with spikes in climate change tweets.

II. DATA COLLECTION

A Twitter scraping service was set up in order to process and record streaming data into text files. The data used in this paper spanned from April 7, 2015 to October 7, 2016. This is a total of 18 months, or 1.5 years. During this time, a total of 22,508,031 tweets containing ”climate change”, ”global warming”, or ”warming planet” were collected. Each tweet also contained associated metadata: the date tweeted, a unique user id, the username of the tweeter, location of the tweeter, number of retweets, number of friends and number of followers.

From this data, a set of 10000 tweets were sampled. This was distributed uniformly over the number of tweets (1 tweet per 2250). Each tweet was labelled one of four categories: activist, denier, other, and duplicate. Out of the 10000 tweets, 3317 were labelled as duplicate, leaving a final dataset of 6683 tweets. This dataset was partitioned into training data (N=6000), validation data (N=183) and test data (N=500).

A. Class Definitions

Sentiment analysis using traditional machine learning techniques requires data to be manually labeled by humans. Naturally, the classification of tweets is ambiguous due to the fact that sarcasm and the desire to be humorous is dominant in the domain. Furthermore, the topic of climate change is politically charged, and the lines between activists, deniers, and others can often become blurred. As a result, precise definitions of activism and denial are required for labeling purposes as well as for congruent interpretation.

Activism is an act of sentiment which raises awareness, generates resources for, or promotes activity acting against human caused climate change. This can also be an act of sentiment against the group known as deniers.

Denial is an act of sentiment that is against human caused climate change activism or one that raises awareness of other denier sentiments. This includes emphasizing the benefits of global warming and downplaying the effects of climate change. This also includes denial of the role of humans in climate change.

Neutral is a class that encompasses all tweets that do not fall into either of the two categories above. This includes items such as news reports, nonsense, tweets unrelated to climate change, or any tweet that does not contain a sentiment.

An *Ambiguous* tweet is one where the machine learning algorithm is unsure of its prediction. Ambiguity is defined in terms of entropy, and is discussed further in the methods section.

B. Event Identification

Throughout this paper, a number of major climate change events are identified to be analyzed. These include the COP21 conference in Paris (Nov 30-Dec 12, 2015), Leonardo DiCaprio’s Oscar speech (Feb 28, 2016), Earth Day 2016 (Apr 22, 2016), and the Trump-Clinton climate debate (Sep 26, 2016). A *climate change event* is a twenty day period - ten days before the event, and ten days after. A *control event*, a twenty day period with no significant spikes in the number of tweets about climate change, was also included. In the case of the COP21 conference, the final date to analyze was taken as the end date rather than the start date.

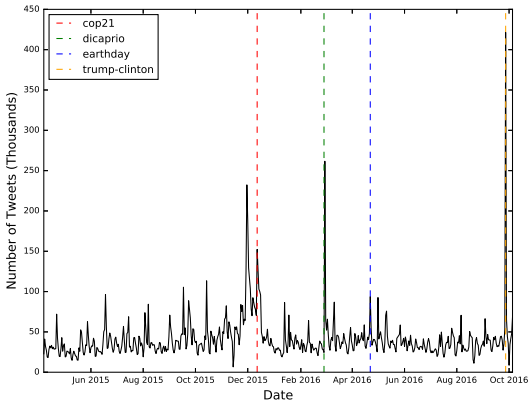


Fig. 1: A timeseries analysis of the total number of tweets grouped by day over the time period.

III. METHODS

Sentiment analysis is a specific type of classification problem in the context of natural language understanding (NLU). Traditionally, sentiment analysis is conducted by a feature extraction and feature selection model followed by a classifier trained to recognize those features. As a result, clever choice of both the feature model and the classification model are important in obtaining a well performing algorithm. Any further analysis conducted is also explained in detail.

A. Document Representation Model

A basic document representation model in natural language understanding is the *bag of words* model. This model takes each word (separated by space) as an individual feature, and as a result the number of features is equal to the number of unique words in the text. In traditional NLU work, the bag of words model is deemed as a primitive solution due to the models inability to understand temporally dependent information and deep context.

It must be noted that most of the data used in traditional natural language understanding contains information rich documents of significant length, providing explicit contextual details and deep relationships in terms of the entities

involved. However, a tweet on Twitter contains a maximum of 140 characters, leading to limited relationship depth. Furthermore, context in the data is often inferred rather than stated. As a result, complicated models that rely on large amounts of information to make decisions perform poorly in this particular situation, and a bag of words representation is significantly more effective.

Another important point is that language and speech linguistics on Twitter are informal, containing a significant amount of chatspeak as well as emojis. Additionally, data is noisy, and entities are often referred to by aliases. As a result, a text processor in order to standardize entities is required. This ensures that all features in the model are useful and not simply due to language discrepancies.

The following pseudocode details the algorithm of processing a tweet to be used in a bag of words model:

```
function PROCESS_TEXT(text)
    tokens ← split_by_space(text)
    tokens ← remove_retweet_prefix(tokens)
    tokens ← remove_hyperlinks(tokens)
    tokens ← remove_stop_words(tokens)
    for all tokens do
        token ← replace_unicode(token)
        token ← to_lower_case(token)
        token ← remove_prefix_suffix(token)
        token ← remove_unnecessary_punctuation(token)
    return join(tokens)
```

Afterwards, the processed training set was normalized and fed into a vectorizer, using a term frequency-inverse document frequency (tf-idf) indicator as well as a minimum frequency floor(min-df) to remove one-of features. Feature selection was done using a chi-squared statistic between each feature and the class it belonged to. The number of features was chosen using cross validation, with a final value of n = 740.

B. Classification Techniques

Classification by definition is a supervised learning problem. As a result, a number of different techniques can be used. In a paper by *An et al.* [6], it was found that support vector machines (SVM) were better than naive-bayes classifiers at determining the sentiment of a tweet.

A support vector machine is a classification technique which employs the use of *support vectors* to categorize data into one of several classes. It is characterized by its use of *kernel functions*, which allows for an implicit mapping onto an infinite dimensional feature space. This is useful for sentiment analysis as the input data is usually an encoded vocabulary set, and the feature space therefore contains many dimensions. We will denote the input data (the contents of a tweet) as X, and its predicted sentiment as Y from now on.

In terms of parameterization, cross-validation is once again used to choose a L2 regularization technique, and a probability distribution is calculated for each of the three classes. The baseline classifier chooses a corresponding class to output according to the formula:

$$Y = argmax([Pr(x) \, \forall x \in C])$$

In this equation, x refers to each individual probability distribution with the classes, C.

For the continuous case, we instead introduce a new parameter, α, which denotes a confidence threshold. This parameter is used in calculating the corresponding class according to the following rule:

$$Y = argmax([1 \; if \; Pr(x) > \alpha, else \; 0])$$
$$Y = Y \; if \; Y \; not \; all \; 0 \; or \; U$$

This formula first checks if the probability of a class is above the threshold α, and if so, then that class is selected to be the final class. However, if none of the classes have a sufficiently high probability, then the ambiguous class U is assigned instead. $\alpha = 0.7$ was chosen by cross-validation.

C. Measuring Ambiguity

Furthermore, for analysis purposes, the *entropy* of the classification is calculated. Traditionally, entropy is a term in information theory used to quantify the *disorder* of a system. In this case, the entropy, H, is re-defined as follows:

$$H = -\sum_{x \in C} Pr(x) \, log \, Pr(x)$$

It is interesting to investigate how this equation models uncertainty and ambiguity in the context of sentiment classification. Essentially, $h = -Pr(x) \, log \, Pr(x)$ is small when $Pr(x)$ is close to 0 or 1, and large when $Pr(x)$ approaches $\frac{1}{n}$. As a result, $H = \sum h$ is small when there is a single probability that approaches 1 and the others approach 0, and H is large when the probabilities approach a uniform distribution.

At inference time, both the probability distribution and the entropy were calculated for each tweet and stored along with the predicted sentiment for analysis purposes.

IV. CLASSIFIER COMPARISON

A confusion matrix (C_{NxN}) is one method of evaluating the accuracy of a classifier. In the context of climate change sentiment analysis, $C_{i,j}$ denotes the number of test examples with predicted sentiment i and labeled sentiment j. It can be used to calculate a number of conditional probabilities in order to give a quantitative measurement on the performance of a classifier. The final confusion matrices for the validation set after cross-validation for both the discrete and continuous classifiers are shown in Table 1 and Table 2.

		Predicted			Total
		A	D	N	
Actual	A	60	10	9	79
	D	2	20	1	23
	N	23	18	40	81
Total		85	48	50	183

Table 1: Confusion matrix for validation set (Discrete)

		Predicted			Total
		A	D	N	
Actual	A	19	1	0	20
	D	0	8	0	8
	N	1	0	12	13
Total		20	9	12	41

Table 2: Confusion matrix for validation set (Continuous)

The discrete classifier produces a final validation accuracy of 65.57%. Meanwhile, the continuous approach gives an accuracy of 95.12%. However, it must be noted that the size of the nonambiguous validation set shrinks from N=183 to N=41.

We are also interested in the test set accuracy. Table 3 and Table 4 show the confusion matrices for the test set in the discrete case and the continuous case, respectively.

		Predicted			Total
		A	D	N	
Actual	A	172	27	18	217
	D	21	50	6	77
	N	66	52	88	206
Total		259	129	112	500

Table 3: Confusion matrix for test set (Discrete)

		Predicted			Total
		A	D	N	
Actual	A	43	4	1	48
	D	1	18	3	22
	N	7	5	37	49
Total		51	27	41	119

Table 4: Confusion matrix for test set (Continous)

For the discrete case, we find an accuracy of 62%, while we find an overall accuracy of 82.35% for the continuous case. This is a significant improvement of 20.35% when using the continuous classification measure. Similarly to the validation set, the size of the dataset with a nonambiguous sentiment shrinks from N=500 to N=119 in the continuous case.

We are also interested in finding out the distribution of the classifiers. We first let

$$Z = i; \quad i \in [A, D, N]$$

denote that the tweet Z had actual sentiment i. These tweets can be concatenated, i.e. $Z = AD$ means a tweet chain containing an activist tweet, and then a skeptic tweet.

Similar, let

$$Y = j; \quad j \in [A, D, N]$$

denote that the tweet Y had predicted sentiment j. These tweets can be concatenated similarly as Z.

Then, we define

$$\hat{\theta}_{i,j} = \frac{C_{i,j}}{\sum_{x \in X} C_{x,i}}$$

$\hat{\theta}_{i,j}$ can be interpreted as the probability that a tweet with actual sentiment $Z = i$ will have predicted sentiment $Y = j$. The theta distributions calculated for the test confusion matrix can be found in Table 5 and Table 6.

$\hat{\theta}_{i,j}$		Y		
		A	D	N
Z	A	0.79	0.12	0.09
	D	0.27	0.65	0.08
	N	0.32	0.25	0.43

Table 5: Theta matrix for test set (Discrete)

$\hat{\theta}_{i,j}$		Y		
		A	D	N
Z	A	0.90	0.08	0.02
	D	0.04	0.82	0.14
	N	0.14	0.10	0.76

Table 6: Theta matrix for test set (Continuous)

We can see that the θs found using the continuous classifier are much more precise than the ones found using the discrete classifier. On a single tweet, the performance is 20.35% better. It must also be noted that sentiment analysis over time often requires a sequence of tweets. The uncertainty of a sequence of measurements increases as the number of measurements increases, and as a result, this 20.35% is much more significant in the context of this paper. This comes at the expense of sacrificing some aspects of validity. The effects of this continuous measurement are further discussed in the analysis and validity sections of the paper.

V. USER AND TWEET ANALYSIS

Fig. 2 gives the relative sentiment distribution of tweets over the time period. The overall average percentage of activist tweets was 41.65%, while the average percentage of deniers was 21.72%. These averages were calculated over the entire time period. This is similar to the distribution found in population surveys, with one study [3] citing a 41% activist population and 23% denier population and a second study [4] citing a 36% activist population and 25% denier population. This similarity in distributions may imply that the nonambiguous sentiments classified by the SVM are representative of the population. This relationship is discussed further in the validity section.

Similarly, the time series analysis of entropy grouped by sentiment is shown in Fig. 3. The entropy for a sentiment in a day was calculated by taking the average entropy of all nonambiguous tweets with that particular sentiment starting and ending at midnight. This was repeated for each day in the time period. The average entropy for ambiguous tweets throughout the time period was H=0.95, as compared to H=0.62, H=0.60, and H=0.69 for activists, deniers, and neutrals respectively.

It can be seen that ambiguous tweets ($Pr(x) < \alpha$) have significantly higher entropies throughout the time period, which may indicate that entropy is a representative measure of ambiguity. This is also discussed in the validity section.

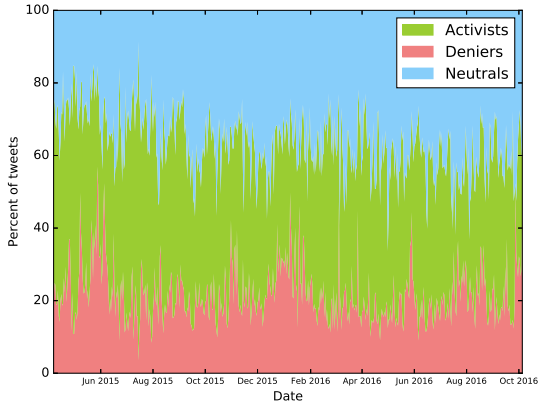


Fig. 2: A timeseries analysis of the tweet distribution over the time period.

It should also be noted that the average entropy for deniers is lower than the average entropy for activists, and that both of these entropies are lower than the average entropy for neutrals. This may be an indication that the classifier finds denier tweets less ambiguous, or that the classifier is slightly better at classifying denier tweets than activist tweets.

It is interesting that neutral tweets have a significantly higher entropy (10-15%). Most likely, this is due to the fact that we have a single layered SVM, which is required to infer both subjectivity as well as polarity at the same time. The classifier finds it more difficult to classify subjectivity, which is natural due to both the diverse language and nature of objective tweets. This is also congruent with findings from *An et al.*[6].

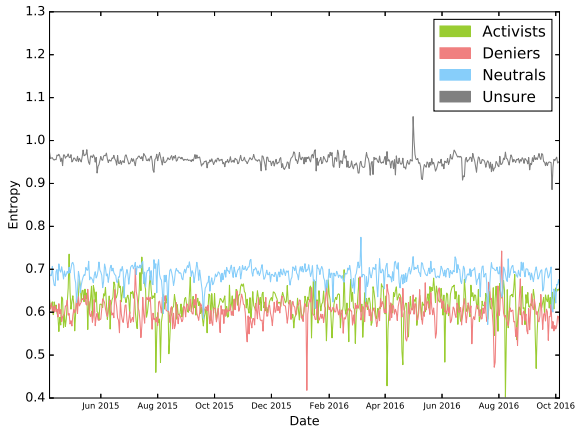


Fig. 3: A timeseries analysis of the average entropy over time, grouped by sentiment.

Another area of interest is in the *lifetime* of a tweet. This is useful as the lifetime will give some insight into how much influence a particular tweet has on the sentiment of the population. In order to measure the lifetimes, two different metrics were used. One measurement is the time between the original tweet and the last retweet, and another measurement is simply the raw number of retweets.

Fig. 4a shows the time elapsed, as grouped into buckets of 1 hour for histogram purposes. After initial raw frequencies were accumulated for each bucket, this data was normalized to show a relationship between the percentage of unique tweets and the lifetime of tweets. It can be seen that the data follows an exponential distribution. As a result, the lifetime of a tweet is asymptotic, and the average lifetime is not a sufficient statistic. Instead, we solve the following equation for the time t, given a parameter to characterize the percentage of data coverage, C:

$$C = \sum_{x=0}^{t} Perc(x)$$

For C=0.95, a lifetime of t=72 hours was found across all sentiments. It is interesting to note the small peaks around 24, 48, and 72 hours. Possibly, this is because of Twitter's activity algorithm, but this is not proven. The lifetime for activists was t=73 hours whereas for skeptics, t=70 hours. It is difficult to decipher whether or not this discrepancy is simply due to the difference in population sizes, however.

We find similar results for the retweet graph. The data also follows an exponential distribution and as a result, we can solve for t with the above equation. We find a lifetime of t=22 retweets. This is split into t=21 retweets for activists and t=23 retweets for deniers.

VI. EVENT ANALYSIS

Previously, we looked at several trends and statistics throughout the time period. Next, four major climate change events that occurred during the time period will be analyzed. The effects of these events on the underlying sentiment of the population are naturally of interest as insight into prior events allows for better planning of the future. Two approaches will be used - one discrete in nature, and one which is continuous.

Tweet data is first collected in a ten day period prior to the event. After the event occurs, an additional ten days of tweet data is retrieved. The effect of an event is defined to be the difference in some quantitative measure between the pre-event data and the post-event data. It should be noted that the time period of data collection can be adjusted, and that t=10 days was chosen empirically with consideration to the lifetime analysis above. With more information on tweet impact and sentiment propagation, a more precise time period can be determined. In this specific case, all tweets collected are used in the analysis, but this can also be adjusted to only include certain event-specific words and phrases.

The first event to be analyzed is the COP21 climate change conference held in Paris between November 30, 2015 and December 12, 2015. During this time, the *Paris Agreement* was negotiated by 195 countries, including the United States [11].

The second event to be analyzed is the Leonardo DiCaprio Oscar speech on February 28, 2016. DiCaprio uses his speech

(a) Distribution of Lifetime of Tweets

(b) Distribution of Retweets of Tweets

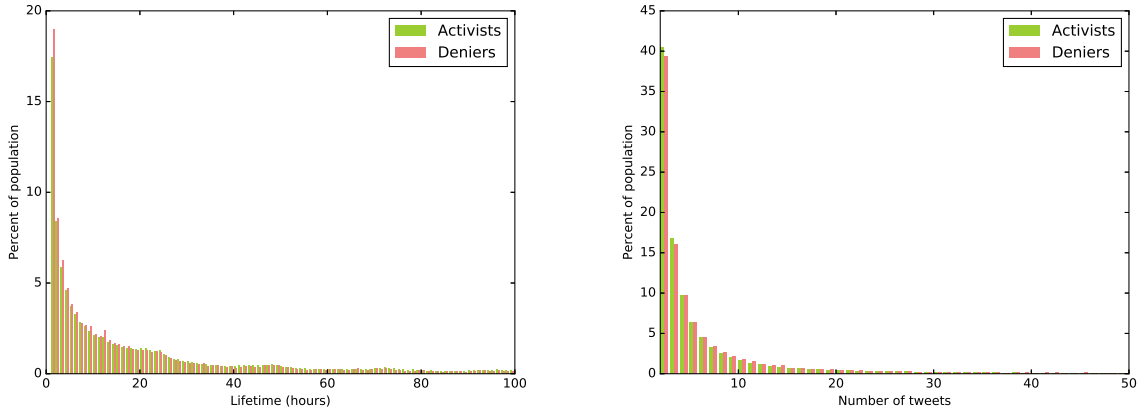


Fig. 4: The lifetime in 4a is the duration of time between the first and last occurrences of a tweet. The retweet graph depicts the distribution of tweets that were retweeted at least once. The x-axis is the total number of retweets (including the original tweet).

to address the seriousness of anthropogenic climate change and encourages all watchers to become activists [12].

The third event looked at in this paper is Earth Day 2016. On April 22, 2016, The Paris Agreement (from the COP21 conference) was opened for signing at a ceremony in New York on this day as well [13].

The final event that was investigated was the Trump-Clinton debate on September 27, 2016. On November 6, 2012, Donald Trump tweeted, *"The concept of global warming was created by and for the Chinese in order to make U.S. manufacturing non-competitive."* [14]. In the debate on September 27, this tweet was mentioned by Clinton after much deferral of the climate change debate in the presidential election [15].

A. Discrete Event Analysis

The discrete approach is, in essence, a frequency analysis of sentiments. The number of activists, deniers, and neutrals are counted prior to the event, and re-counted after the event. The number of users who switched sentiments according to either their first tweet or last tweet are then found. It must be noted that the accuracy of the classifier is directly called into question here.

In Table 7, we look at the expected number of sentiment switches assuming that the underlying sentiment of users does not change at all. In order to quantify these numbers, we look at two arbitrary tweets in a population n of n_A activists, n_D deniers, and n_N neutrals. Then we define

$$\begin{aligned} E[x_1, x_2] &= n_A \hat{\theta}_{A,x_1} \hat{\theta}_{A,x_2} + n_D \hat{\theta}_{D,x_1} \hat{\theta}_{D,x_2} + n_N \hat{\theta}_{N,x_1} \hat{\theta}_{N,x_2} \\ &= n \cdot (\hat{\theta}_{Z=x_1} \hat{\theta}_{Z=x_2}) \\ &\forall x_1, x_2 \in X \end{aligned}$$

Note that these calculations make the assumption that tweets are independent of each other. As a result, all calculations are symmetric, i.e. $E[x_1, x_2] = E[x_2, x_1]$. But then, any switches that occur should be cancelled out by the opposing one, and the net change should be 0.

Event name	n_A	n_D	n_N	$E[A, D]$
COP 21 Conference	45872	5327	57972	1686.26
Leonardo Dicaprio Oscar	20589	1970	24742	720.95
Earth Day 2016	24938	2649	26151	856.89
Trump-Clinton Debate	25543	3557	45214	1082.93

Table 7: Expected switching per event for the discrete classifier based on accuracy alone. No actual switches occur, but the classifier will determine that a switch has occurred because of accuracy problems.

As a result, this means that if an uneven number of activists, deniers, or neutrals is found in the event period, these numbers are due to either temporal dependence, or sentiment changes in the population. However, it must be noted that all of these changes are directly affected by the accuracy of the classifier. This means that, for each sentiment switch that is found, for a classification error

$$Err_i = (1 - \theta_{i,i}) * 100$$

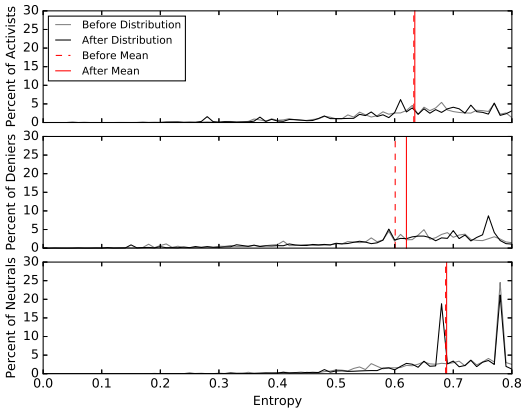
then the overall Error of the switch $Err_{i,j}$ is

$$Err_{i.j} = Err_i * Err_j$$

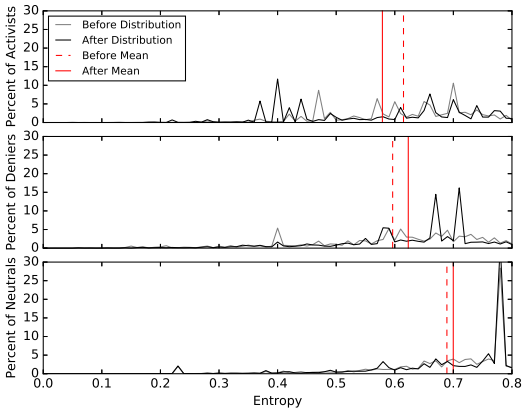
as the errors are independent of each other. As a result, the number of switches found using a discrete approach leans heavily in the favor of denier to activist simply because of the disproportionate population sizes. Throughout the entire time period, it was calculated that an overwhelming 63% of deniers who tweeted twice became activists. This shows that the discrete approach is heavily affected by classification error.

Table 8 shows the frequencies of each sentiment for each of the events discussed. Note that in several occasions, it is

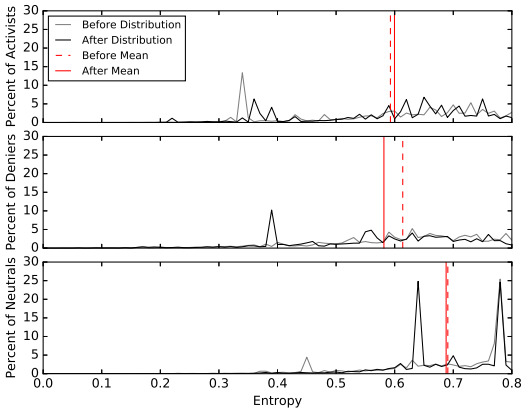
(a) COP21 Conference



(b) Leonardo DiCaprio Oscar



(c) Earth Day 2016



(d) Trump-Clinton Debate

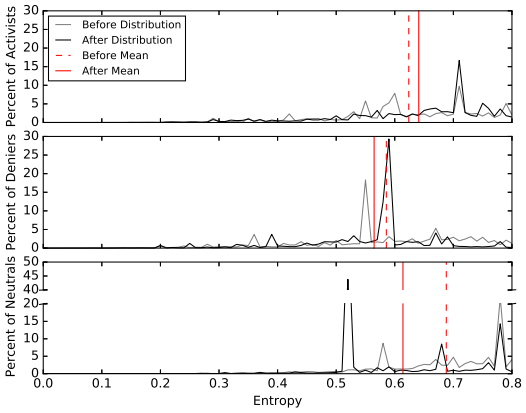


Fig. 5: Distributions of entropy prior to and after each event discussed. The (dotted) red line shows the (pre) post event entropy means.

difficult to measure the underlying sentiment changes and it is unclear what kind of effects were caused by the events.

Event name	A_{pre}	A_{post}	D_{pre}	D_{post}
COP 21 Conference	542707	369907	280225	355817
Leonardo Dicaprio Oscar	174578	387573	101684	212203
Earth Day 2016	239441	244540	137717	114197
Trump-Clinton Debate	203388	226286	125628	426658

Table 8: Actual tweet totals both before and after the event.

B. Continuous Event Analysis

Meanwhile, the analysis of sentiment over time with the continuous approach looks at the changes in the distribution of entropy before and after the event. The mean and standard deviation are calculated for both distributions, and the differences in entropy for each class are found. Visual representations of the probability distributions prior to each event as well as after each event are also included.

Note that in this case, the continuous measurements that are being analyzed are probability distributions, and as a result, uncertainty and error of the measurements are naturally characterized through the standard deviation of the data distribution. Table 9 shows the mean entropies for activists, deniers, and neutrals (respectively) before and after the events (using the same data as in the discrete case). Meanwhile, Table 10 shows the standard deviations for each class corresponding to the means.

We also take a look at the difference in probability distributions between the before and after entropy distributions. Using the Wilcoxon signed-rank test, we find the W test statistic, which allows us to find p-values for non-normal distributions. All of the p-values can be found in Table 11.

Firstly it must be noted that the pre-event means are similar to values found in the overall tweet entropy means. This indicates that the periods of time prior to the event are relatively calm (average).

In terms of the COP21 conference, a comparison between the pre and post entropy means tells us that the activist and neutral entropies are almost unchanged, but the denier

Event name	μ_{pre}	μ_{post}
COP 21 Conference	$[0.63, 0.60, 0.69]$	$[0.63, 0.62, 0.69]$
Leonardo Dicaprio Oscar	$[0.61, 0.60, 0.69]$	$[0.58, 0.62, 0.70]$
Earth Day 2016	$[0.59, 0.61, 0.69]$	$[0.60, 0.58, 0.69]$
Trump-Clinton Debate	$[0.62, 0.59, 0.69]$	$[0.64, 0.56, 0.61]$

Table 9: Entropies of tweets grouped by population before and after the events. Entropies are in order of activists, deniers, and neutrals, respectively.

entropy mean went up (and the denier entropy uncertainty decreased). This could be attributed to the fact that the conference presented many facts with credible references, which made denier tweets less certain. Also note that, for the activist and neutral populations, $p > 0.05$, indicating that the activist and neutral entropy distribution did not change significantly. Meanwhile, we found that $p < 0.001$ for deniers, supporting our findings.

Event name	σ_{pre}	σ_{post}
COP 21 Conference	$[0.12, 0.15, 0.11]$	$[0.13, 0.15, 0.09]$
Leonardo Dicaprio Oscar	$[0.12, 0.15, 0.12]$	$[0.15, 0.12, 0.12]$
Earth Day 2016	$[0.15, 0.14, 0.12]$	$[0.14, 0.14, 0.09]$
Trump-Clinton Debate	$[0.12, 0.14, 0.11]$	$[0.13, 0.11, 0.11]$

Table 10: Standard deviation of entropies of tweets grouped by population before and after the events. Similarly in order of activists, deniers, and neutrals, respectively.

We can find similar results for Earth Day 2016. We can see that the activist and neutral tweets did not change in ambiguity, but the denier tweets became less ambiguous in this case. Also interesting to note is that the standard deviations decreased for activists and neutrals, but increased for deniers. What is particularly interesting about Earth Day is that, unlike the COP21 conference, the entropy distribution of the neutral sentiment changed (according to the p-values). However, the overall mean entropy was unchanged, leading to the possibility that Earth Day raised awareness but was unable to affect the ambiguity of tweets.

We can contrast these numbers with the ones found in the Leonardo DiCaprio event and the Trump-Clinton debate. In Dicaprio's case, activist tweets became less ambiguous while both denier and neutral tweets became more ambiguous overall. Meanwhile, we see the opposite effect from the Trump-Clinton debate. These effects are supported by the p-values calculated, and are shown visually in Fig. 5.

There is a clear difference between the activism events and the social media events analyzed in terms of effect on entropy. In both the COP21 conference as well as on Earth Day 2016, only the entropy of the denier tweets was affected. This leads to the natural conclusion that these events did not noticeably affect the two populations - as in, it did not make the activist tweets any less ambiguous, nor did it cause the neutral entropy to change (perhaps indicating that it did not attract any attention to users who were already neutral).

What is particularly interesting about the Trump-Clinton debate in comparison to the other events is that there was a significant impact on the neutral class. This shows that users who were originally neutral became more involved due to the event.

Event name	$p_{activist}$	p_{denier}	$p_{neutral}$
COP 21 Conference	0.0536	< 0.001	0.3082
Leonardo Dicaprio Oscar	< 0.001	< 0.001	< 0.001
Earth Day 2016	0.2613	< 0.001	< 0.001
Trump-Clinton Debate	< 0.001	< 0.001	< 0.001

Table 11: p-values for each sentiment in an event. We use the standard rejection value of p=0.05, and all p-values less than p=0.001 are rewritten.

We can also determine the impact I of an event simply by taking a weighted sum of the entropy deltas with the sentiment distributions as weights. This is expressed in the following formula:

$$I = \frac{1}{n}(n_A \Delta H_A + n_D \Delta H_D + n_N \Delta H_N) * 100$$

From this equation, we calculate the impact of each event and find that the Trump-Clinton debate has the highest impact on the overall population, with an impact of I=4.44 as compared to I=0.44, I=2.04, and I=1.07 for the COP21 conference, the DiCaprio speech, and Earth Day 2016, respectively.

VII. Conclusions

In the context of climate change sentiment over time through twitter data, it was found that a continuous classification measurement using entropy performed 20.35% better than a discrete classification system. The continuous classifier was then used to infer 22 million tweets throughout the time period, producing a population sentiment distribution of 41.65% activist and 21.72% denier, which is consistent with population surveys. The entropy of the populations was analyzed throughout the time periods and denier tweets were found to be less ambiguous than activist tweets. Afterwards, the lifetime of a tweet was found to be 72 hours. This lifetime was used to create a time period from which to analyze climate change sentiment shifts due to different types of events. Both a discrete measurement of population frequency as well as a continuous measurement of entropy deltas were used, and the Trump-Clinton debate was found to have the highest impact with I=4.44, and was the only event able to influence the neutral population. It was also found that the climate change conferences and notable scientific publications were unable to make a significant difference in activist movements, but caused disruption in the denier population.

VIII. Validity

Due to the fact that a different class decision process is being used, not all tweets are given a nonambiguous sentiment at inference time. As a result, some of the data in the dataset is omitted. This naturally brings up several issues of validity. Specifically, in this section we will address the issues of soundness and completeness in the dataset representation, and discuss implications and limitations of this representation.

It should be noted that the sentiment distribution for individual time periods as well as the overall data collection period was found to be consistent. This implies that the data that is chosen is sound in terms of representativeness. However, the data is not complete, as it does not cover the full extent of tweets. As a result, we can use the data to find patterns in sentiment, but we cannot use these patterns to necessarily predict the data in the future or make conclusions about the population itself.

It could be argued that these sentiment patterns over time are still useful to to find which events produce large changes in sentiment. The new continuous metric described is also useful to find precise, quantitative patterns in sentiment over time when data is highly ambiguous.

Another concern for this paper is to determine whether or not entropy is correlated with sentiment ambiguity in users, or whether the entropy simply reflects the vocabulary learned by the classifier. Further work must be done to find a quantitative measure for the use of entropy in sentiment analysis, but a paper by *Park et al.* analyzes an entropy measure in multi-label classification, showing that classification accuracy is, in fact, correlated with entropy [16]. This leads to a strong hypothesis that entropy can be used as a measure of ambiguity in sentiment analysis. It then follows that, if machine learning techniques are able to correctly understand natural language, then entropy is also usable as a continuous and quantitative measure of ambiguity in population sentiment on Twitter.

ACKNOWLEDGMENT

The authors thank Justin Schonfeld and Zach Dockstader for guidance and inspiration throughout the project. We would also like to give special acknowledgment to Vincent Talbot and Demetri Pananos for their contributions in the collection of Twitter data and in machine learning techniques.

REFERENCES

[1] J. Cook, et al. "Consensus on consensus: a synthesis of consensus estimates on human-caused global warming", Environmental Research Letters Vol. 11 No. 4.

[2] P. T. Doran, M. K. Zimmerman. "Examining the Scientific Consensus on Climate Change", Eos Transactions American Geophysical Union Vol. 90 Issue 3.

[3] A. Leiserowitz, et al. "Climate Change in the American Mind: Americans Global Warming Beliefs and Attitudes in April 2013", Yale Project on Climate Change Communication.

[4] Gallup News Service, Gallup Global Warming Opinion Polls (March 6-9, 2014). Accessed: 2016-11-10.

[5] C. Boussalis, T.G. Coan. "Signals of Doubt: Text-Mining Climate Skepticism. Workshop, London School of Economics.

[6] X. An, R. Ganguly, Y. Fang, B. Scyphers, M. Hunter, G. Dy. "Tracking climate change opinions from twitter data", In Proceedings of the Workshop on Data Science for Social Good Held in Conjunction with KDD 2014.

[7] D. Maynard, K. Bontcheva. "Understanding climate change tweets: an open source toolkit for social media analysis", 29th International Conference on Informatics for Environmental Protection.

[8] E. Cody, A. Reagan, L. Mitchell, P. Dodds, C. Danforth. "Climate change sentiment on Twitter: An unsolicited public opinion poll", PloS One 2015.

[9] S. Soni, Aakanksha Sharaff. "Sentiment Analysis of Customer Reviews Based on Hidden Markov Model", 2015 International Conference on Advanced Research in Computer Science Engineering & Technology.

[10] The New York Times, September 27, 2016. Clinton and Trump climate change debate.

[11] United Nations. "Conference of the Parties, twenty-first session, Paris", FCCC CP 2015.

[12] Leonardo DiCaprio, transcribed by Molly Pier. "Watch Leonardo DiCaprio's 2016 Oscar Acceptance Speech for Best Actor", The Oscars website, Feb 29, 2016.

[13] Earth Day Network. "Earth Day - April 22", retrieved from earthday.org Campaigns, Dec 10, 2016.

[14] Donald Trump, Nov 6, 2012. Twitter social network database, retrieved Dec 10, 2016.

[15] Hillary Clinton. Trump-Clinton Presidential Election Debate, Sep 27, 2016.

[16] L. Park, S. Simoff. ”Using entropy as a measure of acceptance for multi-label classification”, Springer International Publishing, 2015.