Quantifying Content Polarization on Twitter

Muheng Yan*, Xidao Wen*, Yu-Ru Lin*, Lingjia Deng†
*School of Computing and Information, University of Pittsburgh
†Bloomberg L.P.
Email:{yanmuheng, xidao.wen, yurulin}@pitt.edu, ldeng43@bloomberg.net

***Abstract*—Social media like Facebook and Twitter have become major battlegrounds, with increasingly polarized content disseminated to people having different interests and ideologies. This work examines the extent of content polarization during the 2016 U.S. presidential election, from a unique, "content" perspective. We propose a new approach to quantify the polarization of content semantics by leveraging the word embedding representation and clustering metrics. We then propose an evaluation framework to verify the proposed quantitative measurement using a stance classification task. Based on the results, we further explore the extent of content polarization during the election period and how it changed across time, geography, and different types of users. This work contributes to understanding the online "echo chamber" phenomenon based on user-generated content.**

***Index Terms*—polarization, echo chamber, information bubbles, word embedding, presidential election, collective sensemaking, social media analysis**

I. Introduction

Social media like Facebook and Twitter have become major battlegrounds for political campaigns throughout the years. At their onset, social media were expected to be an open, democratic environment of information exchange. Over time, the tweets or posts about politics grew and became dominant traffic on social media platforms especially during election periods. It has been noted, however, that the "content" of these election-related messages have become increasingly homogeneous but only within individuals' social circles. The phenomenon resembles the "echo-chamber" effect that has been long observed in offline and mass media, as well as in online social networks [1], [2], [3], [4], [5]. It could be further reinforced by social media's algorithms to serve content that reinforces what we already know and like, and from like-minded people, while also to filter out those things we generally don't like – an online phenomenon that has been coined as "filter bubble" [6]. In a particularly intense political atmosphere, it makes people less informed or even blinded about different perspectives or consensus, and could lead to a more polarized society. In this work, we aim to analyze the content polarization on Twitter during the U.S. presidential election in 2016.

Studies have examined information polarization on social media since last decade. For example, Adamic and Glance found a polarized linking structure among political blogs [2]. Conover et al. discovered a polarized pattern in a "retweet" network among Twitter users [7]. Lin et al. compared the political mentioning on both social and news media and found that social media are more subject to polarization [8]. While existing studies have accumulated evidence of information polarization on social media, most them mainly focus on the structural patterns, such as those based on hyperlinks, retweets, or mentions, few have examined the extent to which the content created by users are themselves polarized.

In this work, we seek to examine the social media polarization from a unique, "content" perspective – that is, how similar or divergent linguistic elements (e.g., words) are distributed in the content generated by different groups of people. We seek to analyze the extent of polarization from user-generated content on Twitter during the four months of the 2016 U.S. presidential election period, focusing on tweets generated by the supporters of the two candidates, Hilary Clinton and Donald Trump. We propose a new approach to quantify the content polarization by leveraging the distributed representation *word2vec*, which learns the *semantic* similarities between words based on their distributional properties in a large sample of related tweets. We then verify our quantitative measurement using a *stance* classification task. Based on the results, we further explore the extent of content polarization during the election period and how it changed over time, geography, and different types of users.

The key contribution of this work is three folds:

- We propose a new method to quantify the content polarization from user-generated content by leveraging the word embedding (*word2vec*) method and clustering metrics. Our method not only measures the extent but also the statistical significance of the extent to which the semantics of content from two groups are polarized.
- We propose a novel evaluation framework to verify the proposed polarization measurement using a stance classification task. The results indicate the proposed polarization measurement are well aligned with the levels of difficulty in detecting users' stance.
- Using the polarization measurement, we analyze the content polarization across time, geography, and different types of users in our dataset. Our observation provides additional insight to understanding the "echo chamber" phenomenon in a fine-grained manner.

II. Related Work

A. Polarization on Social Media

Researchers have begun to examine the phenomenon of content polarization on social media in the early 2000s. Adamic and Glance examined the hyperlinks in the blog-sphere and

found that there were much denser links within two political groups (liberal and conservative) than those between the two groups, suggesting that these political blogs information flows were polarized [2]. Conover et al. used network analysis to analyze the communication networks on Twitter. Their results show a polarized pattern in "retweet" network, which expressing agreement or repeating others" words, and a non-polarized "mention" network, which representing interactions between individuals [7]. They also found that "hashtags" are relatively important in analyzing the social media polarization. Lin et al. compared the mentioning of political parties on both social and news media and found that social media are more subject to polarization as well as exogenous events [8]. Moreover, it is observed by Garimella et al. that using an 8-year tweet collection, the polarization on Twitter has grown gradually in all aspects including following network, retweet behavior, and hashtag content, and to some extent has reflected the offline polarization in political ideologies [9].

The polarization phenomenon in social media has been receiving much attention, and there are more similar works published. By analyzing a dataset of tweets that share news from 22 mainstream media, An et al. demonstrated the information exposure and sharing on Twitter follow the same polarized pattern [10]. Emotionally vigorous posts are more likely to be retweeted and consequently to be amplified more often and more quickly than those neutral posts on Twitter [11]. With their research on the Facebook dataset, Bakshy et al. argued that this pattern is solid across different social media platforms [4]. These researches are all focus on the network properties in social media – how people interact with each other; however, there are relatively few works that focus on the content itself. A case study by Yardi and Boyd in 2010 tracks the user content on Twitter after a shooting event happened for 24 hours, and the analysis of content suggests that people tend to hear agreeable voices at the beginning but would soon be actively engaging those who disagree [12]. Niculae et al. examined the characteristics of news medias by inspecting their quoting pattern and concluded that even news media would construct biased information even from the same truth [13]. Different from existing studies that mostly focused on the structural patterns, in this work, we seek to identify whether and to what extent the content created by users are themselves polarized.

B. Word Embedding

Word embedding [14] has recently become a popular approach for constructing features from text corpora. The idea of word embedding is to map words to multi-dimensional vectors based on their distributional properties in large samples of text data. In the recent years, a word embedding approach called *word2vec* have drawn much attention. This method was first proposed by Mikolov et al. in 2013, in which the words are embedded based on its surroundings [15], [14]. In a word2vec model, a vector that represents a word n-dimensional space is learned through a feed forward neural network, and the structure of the network varies with two different embedding architecture: *CBOW* (stands for the continuous bag of words) and *Skip-gram*. The former architecture predicts the target word by its context, while the latter one does the contrast: predicts surrounding words with one word given [14].

Word2vec has been demonstrated to be effective. In the traditional usage of information retrieval, word2vec yields significantly better performance than previously invented language models [16]. Although there exists no consensus on how it achieved, word2vec to an extent can reconstruct the semantic meaning of the words embedded [15], and consequently succeeds in sustaining the linguistic relationships between words [17]. It is also implemented in Twitter sentiment analysis and has outperformed other feature construction methods [18], [19].

C. Sentiment and Stance Detection in Twitter

Existing works have used Twitter to track opinion shift or sentiment divergence around political events [20], [21]. In this work, we evaluate the proposed polarization measurement using a "stance" classification task [22]. Different from the sentiment classification and political alignment classification tasks [23], stance classification examines the coupling relationship of both aspects. Specifically, a *stance* depicts one's attitude toward a specific target, which reflects one's sentiment and political alignment when the targets are the political ideology, political parties, or party leaders. Detection of stance is an automatic classification process to determine if the text is in favor, or against the target which in political, social media posts includes people, events, and political acts, etc. [22]. Previous works that focused on analyzing political discourses [24], [25] have proved the effectiveness of stance classification in the political topics.

III. Research Questions

In this work, we are interested in the following research questions:

- How do we quantitatively characterize the polarization of user-generated content? How do we evaluate the quality of the quantitative measurement?
- How does the content polarization evolve over time during the election?
- How does the content polarization vary across geography (states)?
- How does the content polarization vary across different types of users (i.e., users with different level of activity and influence)?

IV. Dataset

This section describes the data used in this work. We first describe the selection of users and then describe the iterative data collection process.

Our data collection was based on those *exclusive followers* – i.e., Twitter users who followed only one candidate but not the other, whom we assume to be supporters of the candidates. To test this hypothesis, we examine the extent to which the exclusive followers expressed their support explicitly. Twitter

TABLE I
RESULTS OF ODDS RATIO TESTS

	Trump Followers		Clinton Followers		
Hashtag	Adopted	Not Adopted	Adopted	Not Adopted	OR
Trump2016	1628	38189	134	37686	11.989
MakeAmericaGreatAgain	2128	37703	107	37703	19.888
DropOutHillary	80	39533	5	37796	15.297
NeverHillary	1415	38386	57	37749	24.413
ImWithHer	1139	38676	4419	33635	0.224
Hillary2016	255	39413	338	37506	0.718
NeverTrump	610	39128	1282	36650	0.446
DumpTrump	44	39559	544	37308	0.076

users frequently use hashtags to declare their support or disapproval with respect to a political agenda or target. For instance, the hashtag "#makeamericagreatagain" was used to express the support to Trump. Thus, we assume exclusive followers are associated with a skewed distribution of these for/against hashtags. Specifically, we examine the odds ratios based on a set of most frequently observed support or disapproval hashtags. The *odds ratio* (OR) with respect to the use of a particular hashtag is defined by:

$$OR = \frac{A_{trump}/N_{trump}}{A_{clinton}/N_{clinton}}, \quad (1)$$

where A_X represents the number of tweets from exclusive followers of candidate X that adopted the particular for- or against-declaring hashtag and N_X stands for the number of tweets from followers of candidate X that *did not* adopt the specific hashtag. If the value of OR is above 1, the usage of that specific hashtag is likely associated with exclusively following Trump, where a higher value reflects a stronger association. If the value of OR is smaller than 1, it reveals that the hashtag is likely associated with exclusively following Clinton, where a lower value indicates a stronger association.

As shown in Table I, eight hashtags are selected in the representation of support or disapproval. The hashtags are either from the campaign slogans of the candidates or are simply stating the support. In the experiment part which will be explained in Section VI-A, this hashtag set is further expanded with top co-occurring hashtags. As the results show, the four hashtags of supporting Trump are highly associated with "exclusively following Trump", while the other four hashtags of supporting Clinton (especially "#ImWithHer" and "#DumpTrump") are associated with "exclusively following Clinton". These associations suggest that exclusive followers can be considered as proxies of supporters.

The data collection involves three steps. First, we acquired the followers of Trump and Clinton at two time points, respectively. These two are Aug. 30, 2016, and Nov. 15, 2016 (one week after the election). Then, we cross-identify the users who exclusively follow Trump at both time points (not following Clinton), and vice versa for Clinton. After getting the list of exclusive followers, we acquire the user profile of a sampled 600k users from each set, and further among which we obtain a stratified sample of 25% users based on the level of user activity defined by the tweet count. The users are segmented into ten quantiles, where each of them

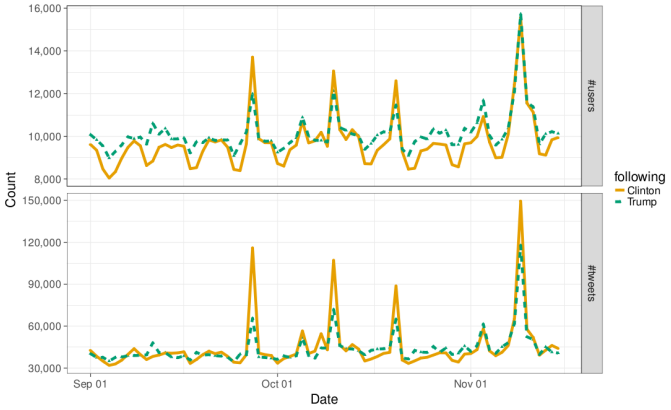

Fig. 1. **Number of users and number of tweets over time.**

contains 10% of all population based on the amount of tweets users published. Then, for each quantile, we randomly sample 25% of the users to construct the final set of users so that our users are representative of the Twitter population with respect to different levels of Twitter activities. The sampling ratio is limited as 25% due to the data collection capacity. In the next step, we acquired this set of 200k users timeline tweets. Finally, we select the users for whom we can trace back their tweets till Aug. 30, 2016, which resulted in 87k for Trump followers and 74k for Clinton followers. Figure 1 shows the volume of daily active users and volume of tweets over time. As expected, we observe both volumes peak on dates of debates and finally reach their highest on the election day.

V. METHOD

In this section, we aim at measuring the content polarization during and after the election period. The U.S. presidential campaign lasted through almost the entire year of 2016 and became increasingly intense in the last quarter of the year. In particular, we want to quantify the level of content polarization between groups, which defined by the candidate supporters, in the period starting ten weeks before the election day (Aug. 30th, 2016), and five weeks afterward (Dec. 13th, 2016). We seek to quantify the content of tweets by projecting the tweets and users to a multidimensional space and further analyze them to observe the polarization in the contents.

This work includes two major steps: (1) learning the semantic representation of content, and (2) quantitatively characterize the content polarization. In the first step, we adopt word2vec embedding [14] to project words in the tweets through corresponding vectors. We then verify the word-embedding vector representation of tweets through a stance detection task. The effectiveness of the word embedding was often evaluated by the classification results [26]: the quality of the content embedding should be reflected in the levels of difficulty in stance classification. Specifically, the more separable the content represented by word-embedding vectors, the clearer the stance detected in the content, and hence the easier the classification task. The following step is to use the word

vectors to quantify the content polarization. Based on the word vectors, mean distances are calculated to measure average distances between or within tweets aggregated by groups. We compute distances for tweets in different pre-defined groups (e.g., exclusive candidate follower group, geolocation group, etc.). The correlation between the distances and AUCs from classification tasks are calculated to exam the effectiveness of the quantification.

A. Learning Word Embedding

Word embedding maps words in a multi-dimensional vector space where words with similar semantic meanings are close to each other in the embedded space. Word embedding has a history in NLP (natural language processing), but all other methods besides word2vec have a limitation that they depend on a fundamental assumption that words appearing in the same document should somehow share similar meanings. However, the approach of word2vec has broken the limitation down. This relatively new approach is claimed to be a more efficient model for learning word embeddings from raw text and meanwhile does not depend on the naive assumption described above. There are two common architectures of word2vec model, continuous bag of word (CBOW) model, and skip-gram model. These two forms are similar, for instance, they all feed forward neural network models, except that skip-gram model predicts surrounding words from the input word, while CBOW model does the inverse, and predict the target word from its surroundings [14]. Mikolov argued that CBOW is smoothed over distribution information, which is useful for small documents, while skip-gram works better when the dataset is relatively large [14]. In our study, since tweets are short documents that contain less than 140 characters, CBOW model would be a better choice.

We trained the word2vec model based on our collected dataset, which contains the tweets posted between Aug. 30, 2016, and Dec. 25, 2016. 7,730,752 tweets in total were included in the wor2vec training. The tweets were preprocessed before the training: they were stemmed with stopwords removed. In the training step, two parameters are the most critical to the model – the window size, which is the maximum distance between the surroundings and the predicted word within a document, and the vector size, which is the dimension of the representing vectors. The window size was set to be 10 in tests with candidates from 5 to 12 (chosen around 15.2, which is half of the average length of tweets) in which 10 works the best. The vector size was set to be 100 because this set of parameters yields the best results in the classification.

B. Quantifying the Content Polarization

With tweet text embedded, every word in the vocabulary is represented as a vector in a n-dimensional space (where n is empirically chosen to be 100 in this work). Then, we can generate the representation of tweets by averaging all word vectors they contain and further quantify the relationships between tweets by numerical distance. After every tweet embedded in the high dimensional space, there are two typical distance metrics for quantifying the relationships of the embedded vectors: Euclidean distance and cosine similarity. The former one is better reflecting the real distance between vectors, while the later one is simulating the closeness between vectors by measuring the cosine value of the angle between them. Previous work by De Boom et al. has provided the evidence suggesting that Euclidean distance is a better choice over others when embedding short documents with the word2vec approach [27].

To quantify the polarization of contents, we consider two clustering metrics for group-level distance metrics. The within-group average distance, defined by the mean of all pairwise distances between tweets within a group, provides the variance caused by individual differences of tweets in a single group. Meanwhile, the between-group average distance, defined by the mean of all pairwise distances between tweets across two groups, reflects the variance resulted from the differences between tweets from the two distinct groups. Let there be two groups labeled *A* and *B*, and then the between-group distance is defined by:

$$D_{AB} = \frac{\sum_{p \in A, q \in B} \|\vec{V_p} - \vec{V_q}\|}{N}, \tag{2}$$

where V_p and V_q are the embedding vector representations of tweet p and q, respectively, and $\|\vec{V_p} - \vec{V_q}\|$ denotes the Euclidean distance between the two vectors.

While the within-group distances are defined by:

$$D_X = \frac{\sum_{p \in X, q \in X, p \neq q} \|\vec{V_p} - \vec{V_q}\|}{N}, \tag{3}$$

where $X \in \{A, B\}$ indicates the group label.

The group-level distances are designed to quantify and represent the relationships and differences of raw texts that are manually assigned into groups. Consequently, each predefined group could be represented as a *information bubble* by the metrics described above: within-group variance as bubble size and between-group variance as the distance between bubbles. Figure 2 illustrates the distance measures quantifying the information bubbles. Through observing the bubble sizes and distances, we can quantitatively characterize how user-generated content evolve through the election period, and how it differed across distinct types of users and over various geographic locations.

VI. Validation

A. Validation of the Polarization Mesurement through Stance Classification

Since word2vec is unsupervised, the effectiveness of embedding is usually measured on a supervised task preceding that is similar to document classification. In this work, we use stance classification to evaluate the quality of the word embedding representation of content.

Stance is defined as *the attitude people have towards a specific target* [22]. Specifically, in our research, the targets concerned are the two candidates – Donald Trump and Hillary

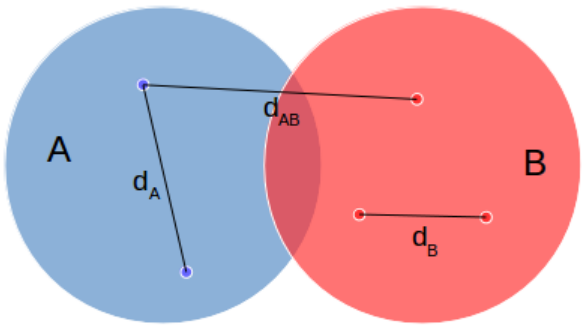

Fig. 2. **Illustration of Information Bubbles. The two circles refer the different tweet groups. d_A, d_B, and d_{AB} are examples of elements in the summation of D_A, D_B, and D_{AB}, respectively.**

Clinton, and meanwhile the *attitude* is defined as whether the tweet shows the appreciation or rejection to the target candidates. The attitude of neutrality is not considered in our experiment because our data were collected from exclusive followers of the candidates, and neutrality is less expected. Comparing to the sentiment detection or classification, we argue that stance classification can better test the coupling relationship of both attitudes (may correspond to sentiments) and the targets (corresponding to the supporters' for or against the candidates). The details of this classification task are described below.

Experiment Settings. The gold standard of stance can be obtained by screening hashtags in the tweets. Designated by the symbol '#', hashtags play a role of keywords in tweets. It often organizes a common topic or opinion around specific events or discussions, and the values of hashtags in political events have been illustrated in previous studies [28]. Throughout the presidential election of the United States in 2017, Twitter has become a major platform for shout-outs of both supporting and rejecting the candidates, which is naturally similar to our definition of stance. As shown in Table II, several hashtags have emerged as the primary media for users to share standings towards the candidates, and captured by us as our targets of stances in classification – some of them came out as the campaign slogans of the candidates, and others are statements of support. Comparing to the hashtag set in Table I, Section IV, which used in the odds ratio tests, we had extended the set by adding in the most co-occurring hashtags. This extended set is more comprehensive and representative.

As shown in Table II, each of the target hashtags is assigned to one of the four designated stances: Favor Hillary Clinton, Against Hillary Clinton, Favor Donald Trump, and Against Donald Trump. The stance of tweets is decided by the dominating category of hashtags. Those tweets with multiple dominating stances are discarded in the classification task. The number of tweets of each stance is listed in Table III

Features. The features used in the classifiers are the embedded vectors of each tweet. The tweet vectors are the average vector of all word vectors in the tweet. Thus the tweet vectors are in 100 dimensions, the same as word vectors. The embedded 100-dimensional vectors are treated as 100 features in the classification task, and no feature other than the embedded vectors is included.

TABLE II
SUMMARY OF HASHTAGS

Hashtag	Candidate	Attitude
trump2016	Donald Trump	Favor
makeamericagreatagain	Donald Trump	Favor
maga	Donald Trump	Favor
trumptrain	Donald Trump	Favor
imwithher	Hillary Clinton	Favor
hillary2016	Hillary Clinton	Favor
strongertogether	Hillary Clinton	Favor
lovetrumpshate	Hillary Clinton	Favor
nevertrump	Donald Trump	Against
dumptrump	Donald Trump	Against
trumptapes	Donald Trump	Against
notmypresident	Donald Trump	Against
dropouthillary	Hillary Clinton	Against
neverhillary	Hillary Clinton	Against
trumppence16	Hillary Clinton	Against
crookedhillary	Hillary Clinton	Against

TABLE III
NUMBER OF TWEETS IN EACH STANCE

Favor Trump	Favor Clinton	Against Trump	Against Clinton
22964.0	19340.0	10820.0	9275.0

Training Process. Correlating to the four classes, we trained four One-vs-All classifiers separately. In each of the four sub-classifiers, tweet with one of the four stances are treated as positive samples, and those with the two opposite stances are included as negative samples. The 'false negative' tweets with the different but logically similar stance are excluded in its corresponding sub-classifier (e.g. 'Against Trump' is a 'false negative' class when 'Favor Clinton' is a 'positive' class).

Due to the imbalance of the numbers of samples in the four classes, a sampling approach was adopted. For every sub-classifier, the number of samples is defined by: $2 \times \min(N_{positive}, N_{true-negative})$, where $N_{positive}$ stands for the corresponding number of positive samples in a sub-classifier, and $N_{true-negative}$ stands for the number of negative samples described above. Subject to the constraint, we randomly selected them so that the ratios of positive and negative samples became perfectly 1: 1 in the training sets.

Different classifiers including Naive Bayes, SVM (support vector machine), C4.5 Tree, and Neural Networks with one hidden layer, are tested through the training process. The best classification results are obtained by Support Vector Machine with these parameters: polynomial kernel, degree = 5, c = 0.5, gamma = 0.02.

Evaluation. The performance of classification was evaluated with a 10-fold cross-validation. Standard evaluation metrics including precision, recall, F-score, and AUC (area under the curve) are reported. The results are listed in Table IV. With the learned word embedding as features, all sub-classifiers have achieved relatively good performances. All of the AUCs are

TABLE IV
RESULTS OF CLASSIFICATIONS WITH SVM

Positive Group	Precision	Recall	F1-score	AUC
Favor Trump	0.86	0.82	0.84	0.84
Favor Clinton	0.86	0.85	0.86	0.86
Against Trump	0.77	0.86	0.82	0.80
Against Clinton	0.82	0.90	0.86	0.85

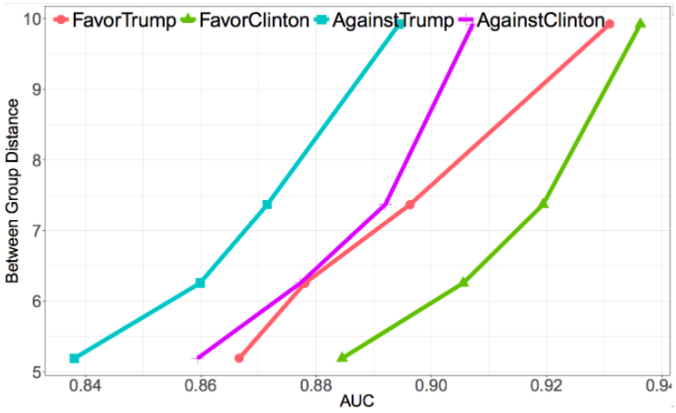

Fig. 3. **Correlation between AUC and Distance. Each line shows the mean values of sample correlation coefficients between AUC and Between Group Distance, with error bars representing standard errors (which are relatively small in this figure).**

above 0.8, and so are precisions and recalls. The precision of classifying 'Against Trump' tweets from others is low but still within an acceptable range. With these classification results, the validity of the learned word embedding is testified. In the further experiments, we will adopt the word embedding as an approach to quantifying the tweet texts.

B. Validation of the Polarization Distance

We verify our measures of polarization through the following rationale: as the content polarization among different groups become more separable, we expect higher accuracy in the task of detecting the stances based on the word embedding. One of the most natural characteristic that distance should represent is the separability of two groups of samples: samples from two close groups may not be necessarily inseparable. However, with other conditions fixed, samples from two groups with higher between group distance should always be easier to be separated. Considering the scenario of information bubbles, if the two bubbles move far apart from each other, it could be expected that the information becomes diverse, and thus it would be easier to classify the points in the two groups. To test the effectiveness of our quantification, the correlation between the quantified distance metrics and the classification results were examined.

In performing the comparison between the stance classification results and the quantifying results, four batches of tweet vectors (grouped by Favor/Against Trump/Clinton) with different levels of distance are sampled. We first calculate the overall center with all tweet vectors included. Then we sample tweet vectors into four batches (from close to faraway) based

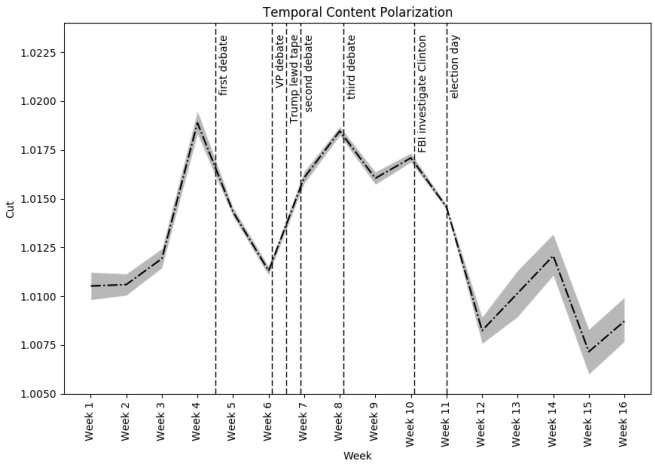

Fig. 4. **Content Polarization over Time. The line shows the NC (Normalized Cut) value of each week in the focused period, with a shaded region representing 95% bootstrap confidence intervals.**

on their distances to the overall center. From close to faraway, the distances between the sampled vectors in the batch and the center are within the range of 0%∼25%, 25%∼50%, 50%∼75%, and 75%∼100%. With the four batches of data sampled, we run the stance classification task on each of them.

Figure 3 shows the correlations between AUC from the classification tasks and the between-group distance with standard error (however the SEs are relatively small to be observed in the figure). The calculation of between-group distance follows the Equation 2, and the definition of *group* here in the term of *between-group distance* follows the definition of class in the classification task. For instance, if we are classifying "Favor Trump" tweets out from others, the two groups used in calculating the between-group distance are "Favor Trump" group and "Other" group. The calculation of SE follows the definition of $SE = \frac{SD}{\sqrt{n}}$ where SD is the sample standard deviation. As shown in the figure, the AUC of classification is highly correlated with the mean average pair-wise between group distance.

VII. CONTENT POLARIZATION ANALYSIS

Denotation of content polarization Before analyzing the content polarization, to make it clear, the real-life meaning of the metrics should be declared first. In the previous sections, we defined two distinct types of distance including between-group and within-group distance. The group here is a collection of tweets that meet specific conditions – a Clinton follower group contains all of the tweets in the election period from exclusive followers of Hillary Clinton and similarly a Trump follower group. Also, note that the representations of tweets are from word embedding. As a result, the distances reflect how different in language those tweets are – precisely the differences in choosing words and phrases. Thus a lower distance means the two tweets are similar in semantics, and vice versa. Extending in the two types of metrics described previously, the within-group

distance would explain the diversity of language usage in the group, while the between-group distance tells how different the two groups are in vocabulary. Ideally, the metrics would correctly reflect the differences in topic people are talking about on the social media; however, we were not expecting a large and significant distance between tweets because the different words and phrases only take a little portion in people's regular communications. The two measures, between-group, and within-group distances represent orthogonal aspects of the content polarization. To facilitate an efficient comparison of content polarization over time and space, we combine the two measures by employing the ***Normalized Cut (NC)***, which is defined as:

$$NC = \frac{\sum_{i=1}^{n} d_{AB}(i)}{\sum_{i=1}^{n} d_{AB}(i) + \sum_{i=1}^{n_A} d_A(i)} + \frac{\sum_{i=1}^{n} d_{AB}(i)}{\sum_{i=1}^{n} d_{AB}(i) + \sum_{i=1}^{n_B} d_B(i)} \tag{4}$$

where d_{AB} represents a pair-wise between-group distance between two vectors in group A and B respectively, with n indicating the number of such pairs; d_X represent a single pair-wise within-group distance between two vectors within group $X \in \{A, B\}$ with n_X indicating the numbers of such pairs. From this definition, the higher the cut is, the larger between-group distance is compared to both within-group distance, which consequently indicates that the information is more diverse, and vice versa.

To quantify the statistical significance of the measurement, a bootstrap re-sampling has been adopted. In calculation the cuts, we re-sampled the pair-wise distance with replacement to its original population in 1000 iterations, in which 1000 samples have been created through the bootstrap re-sampling, and consequently, we can acquire the 95% *confidence interval* for each cut calculated.

A. Analysis of Polarization over Time

Important Events. We expect novel signals when significant events occur. A controversial event would draw attentions from all communities and greatly enhance the possibility of people discussing the same topic whichever group they are in, or would rather make them focusing on whatever they would like to talk and share within the same community group. Thus reflected in the distances, it is assumed to be a fluctuation in content polarization whenever a drastic event occurs. To better analyze the content polarization, we have several predefined events highlighted, that is, the debates before the election day and the election day itself. Besides of the debates, two more shocking news that broke out in the election period should also be noted: the release of the lewd tape of Trump in early October, and the investigation on Clinton announced by FBI at the beginning of November. The Table V shows the list of important dates of the 2016 United States Presidential debate and the corresponding day in our following analysis.

The impacts of the events are supposed to emerge after the events, so it is likely to observe the consequences in the week before the event day, and also in the week after.

TABLE V
IMPORTANT EVENTS DURING ELECTION

Event	Date	Corresponding Day	Week of Impact
1st presidential debate	Monday, Sep. 26th	last day of week 4	week 4 & 5
Vice presidential debate	Tuesday, Oct. 4th	first day of week 6	week 5 & 6
Trump lewd tape released	Tuesday, Oct. 7th	third day of week 6	week 5 & 6
2nd presidential debate	Sunday, Oct. 9th	2nd last day of week 6	week 6 & 7
3rd presidential debate	Wednesday, Oct. 19th	2nd day of week 8	week 7 & 8
FBI investigate Clinton	Wednesday, Nov. 2nd	2nd day of week 10	week 9 & 10
Election day	Tuesday, Nov. 8th	1st day of week 11	week 10 & 11

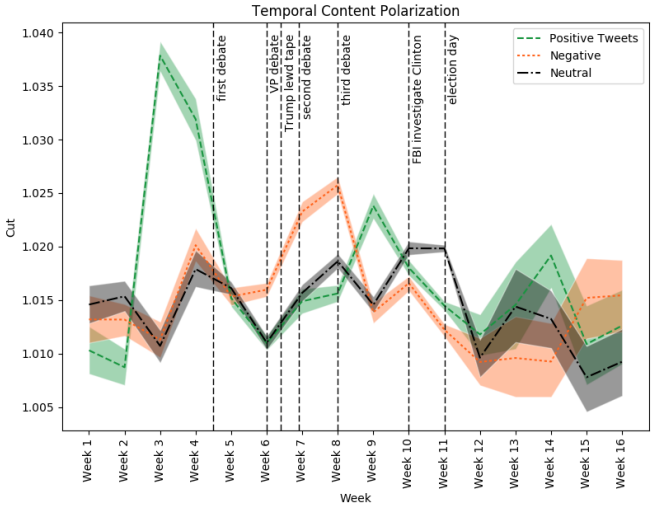

Fig. 5. **Content Polarization over Time by Sentiment. Each line shows the NC (Normalized Cut) value of a specific sentiment in the focused period, with a shaded region representing 95% bootstrap confidence intervals.**

Temporal Evolution of content polarization Figure 4 shows the temporal evolution of content polarization inferred based on the NC of all tweets. The x-axis of the figure is the number of weeks from week 1 (the week starting from Aug. 30th) to week 16, and the y-axis shows the value of normalized cut. Shades in the figure refer to the confidence interval of each data point. Overall, the metrics fluctuate over time. There are multiple peaks observed at week 4, 8, and 10, and eventually fades away drastically after the election day at week 12. The confidence interval went tight before the first presidential debate took place, and went back to a boarder range after the election ends. In general, the cut went high from the first debate to the election day, with a small range in confidence interval, indicating that discussions on Twitter are much more polarized while the election campaign ongoing. From the figure, we can observe there are usually peaks before or after major events and breaking news. The first peak appeared at week 4, the week when the first presidential debate happens. Although the polarization went small at the beginning of week 6, the three events have increased the intensity of polarization, and eventually, lead to a peak in week 8 when the third debate took place. Major events may evoke intense dispute between different follower groups, and thus enhance the intensity of polarization.

Because of the direct conflict between the two candidates, it is confusing to measure the content polarization across

sentiments. One may be talking good about the candidate which they are in favor of and talking bad about the other one at the same time. These two types of statements are converging in their nature but treated as diverging in the analysis above. To better understand the evolution of content polarization, we further tested the cuts on tweets with different sentiments separately. The sentiments are measured by a lexicon based method.

Figure 5 shows the result of the temporal content polarization breaking down into the three sentiments: positive, negative and neutral, separately. Each line in the figure only contains the tweets with the corresponding sentiment. Comparing to the general temporal content polarization, the major traits keep identical in this figure. The peaks of polarization appear around week 4, 8 and 10, which correspond to the first debate, the third debate, and the election day. So does the confidence interval, which kept tight during the debates, and went board after the election – in a more dramatic way. Also, there is an obvious increase in the value of normalized cut in all three sentiments with the influences of the VP debate, release of Trump lewd tape, and the second debate.

Separating the sentiments provides more insights in analyzing the content polarization. The peak of the polarization in the week four was mostly caused by those positive tweets, which indicates that during the first debate, people were more polarized when talking favorable things than talking in negative and neutral manners. After the first debate, the polarization in negative voices overwhelmed its counterparts in the other two types of views. This could be influenced by the increase of voices attacking Trump caused by that most people believed Clinton had outperformed Trump in the first debate. These increasing critics along with the views from Trump's defenders facilitated the increase in the negative polarization. Further enhanced by the scandal of the release of the lewd tape, the polarization in negative sentiment reached its peak and the leading position in week 8. Another interesting fact to be noted in the election period is that the polarization in positive tweets usually acts one week earlier than that in tweets with other two sentiments.

B. Analysis of Polarization by Types of Users

In the previous section, we provide the temporal analysis of the polarization based on the tweet-level examinations. In this section, we provide the analysis of polarization at the user level, where users are grouped based on several attributes, including the number of tweets, the number of followers, and the number of friends.

As each tweet t is represented as the vector $\vec{V_t}$, we can further derive the user vector $\vec{V_u}$ for user u by averaging all the tweets posted by user u during the study period. With users represented as vectors, we then grouped users based on their traits on Twitter: the number of tweets, the number of followers, and the number of friends. Particularly, for each attribute, we divide the users into one of the four groups according to their quantile rankings, e.g., users with top 25% followers would fall into the group of *very-high* follower users

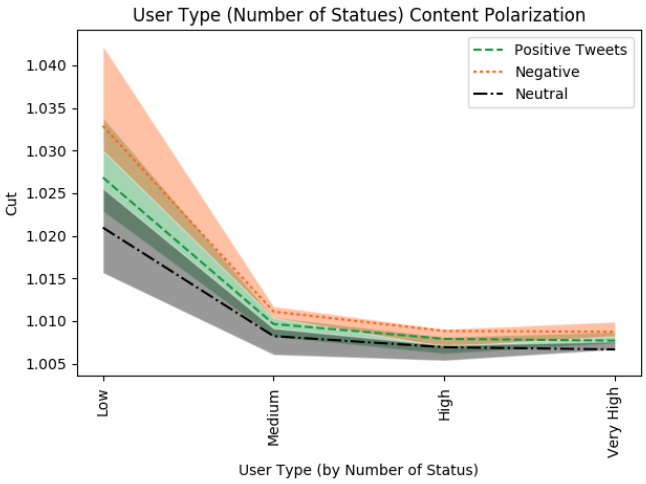

Fig. 6. **Content Polarization by User Category (categorized based on the number of tweets). Each line shows the NC (Normalized Cut) value of a specific sentiment by the corresponding type of users, with a shaded region representing 95% bootstrap confidence intervals.**

and users with the bottom 25% followers would belong to the group *low* follower users. Finally, we follow the Eq. 2 and Eq. 3 to compute the distances between different user groups.

Figure 6 to 8 show the content polarization when users are grouped by the number of tweets, the number of followers, and the number of friends, respectively. The x-axis shows different user groups, while the y-axis shows the normalized cut and different colors denote the sentiments. It is interesting to note that a similar trend can be observed across different settings: negative tweets seem to be the most polarized while those neutral ones are the least. This could be explained by the reason that, people would be more polarized when contesting on negative topics, and tend to be less polarized when the content they are debating is neutral.

Figure 6 shows the cut grouped by the number of tweets. For all the users we have, they are separated by quantiles into four groups: users that tweeted low, medium, high and very high amount of tweets. As expected, content generated from the group of least active users are more diverse, and as more tweets are included for analysis (e.g., for groups of more active groups), the polarization tends to converge for all three sentiments.

The analysis of content polarization over the number of friends and the number of followers reveals a similar pattern: content generated by users who have followers and friends in a middle range number are more polarized than other users. In Figure 7, the peaks of all three sentiments correspond to the "medium-size follower" group. This suggests that users with moderate influence (regarding their follower counts) exhibit greater polarization than other groups. Similar in Figure 8, the users with medium and high friends numbers show the highest polarization in all sentiments, indicating people that exposed to a medium level of information from others is more polarized than others when they are tweeting.

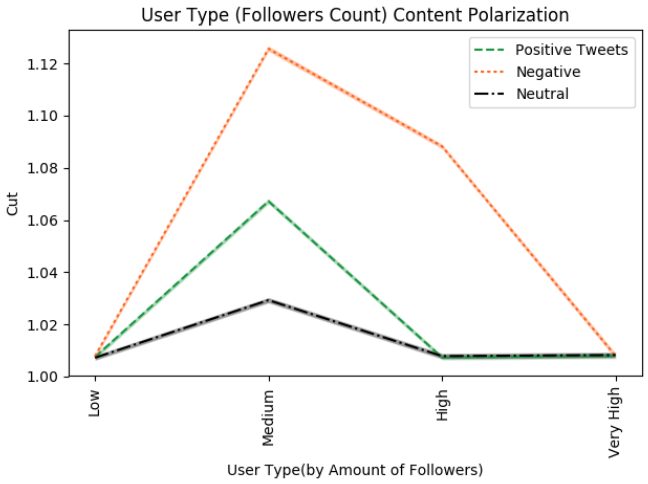

Fig. 7. **Content Polarization by User Category (categorized based on the number of followers). Each line shows the NC (Normalized Cut) value of a specific sentiment by the corresponding type of users, with a shaded region representing 95% bootstrap confidence intervals (which are relatively small in this figure).**

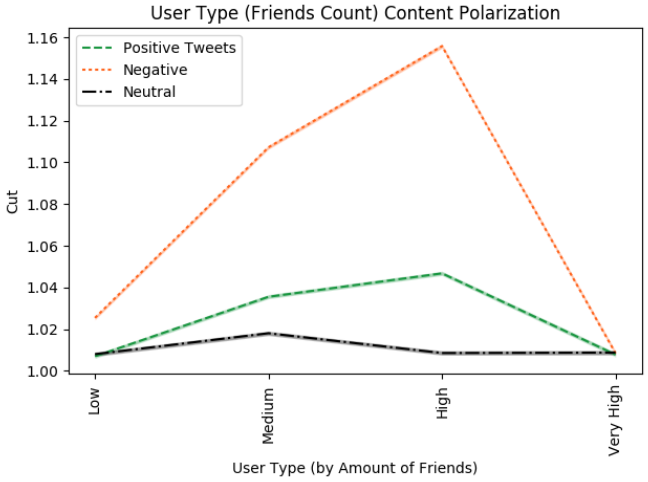

Fig. 8. **Content Polarization by User Category (categorized based on the number of friends). Each line shows the NC (Normalized Cut) value of a specific sentiment by the corresponding type of users, with a shaded region representing 95% bootstrap confidence intervals (which are relatively small in this figure).**

C. Analysis of Polarization by Geography

The three figures in Figure9 present the content polarization by states and by sentiment (positive-green, neutral-black, and negative-orange). We excluded states where an insufficient number of tweets collected. We observe that several states are consistently highlighted in all three sentiment conditions: Minnesota, Wisconsin, Colorado, Louisiana, Missouri and Pennsylvania. Most of these states, especially the state of Minnesota who achieved the highest cut in all three sentiments, are swing states where the election was more competitive than those safe states. In those swing states, it is naturally expected to be more polarization between the different candidate fol-

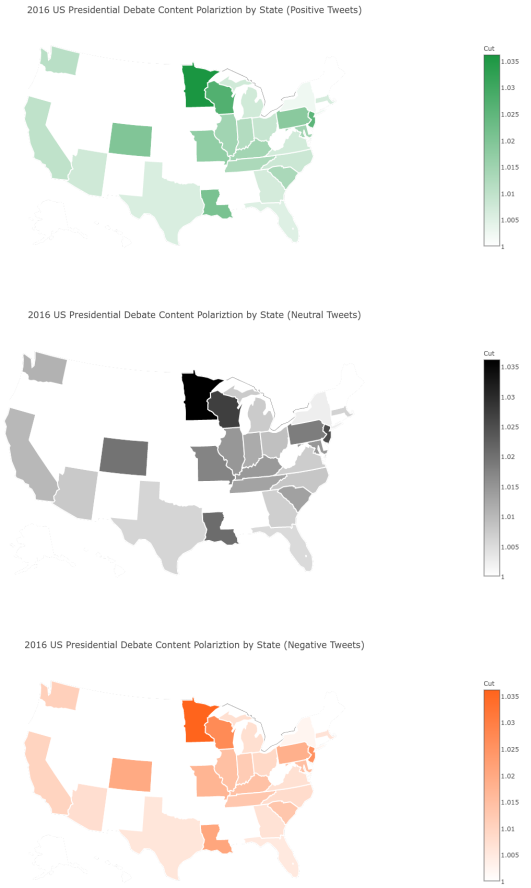

Fig. 9. **Content Polarization by State. Each sub-figure shows a heat map of content polarization by state. A darker color represents a more intense polarization.**

lower groups and converging contents within groups, which results in higher cut values.

VIII. Conclusion

Our research provides insights of content polarization on social media from a new aspect. We have shown that word2vec embedding performs well in quantifying the Twitter content polarization. The embedding has been validated by the stance classification results.

With the tweets quantified, we further explored the content polarization on Twitter in the 2016 U.S. presidential election period from Aug. 30th to Dec. 25th, and concluded as follows:

- It was discovered that the polarization is more intense during the election period, and was ignited and enhanced by major events including debates and breaking news.
- When looking into user level vectors, it was found that in general people are more polarized on negative topics and less polarized on neutral topics.
- By analyzing the polarization among user characteristics, it was found that individuals with moderate influence (regarding numbers of followers) and people exposed to

a medium amount of information (regarding numbers of friends) tend to be more polarized than others.
- Polarizations across the different states is also analyzed, and the results indicate that those states with a higher level of polarization are more likely to be swing states – states could be reasonably won by any candidate from either party.

Unlike other studies that focused on structural properties, this paper focused on the semantic similarities of contents among two ideological groups. We have shown a promising quantification approach for analyzing the polarization of contents of micro-blog style materials. However, there are still some limitations in our work. Firstly, in validating the word embedding, we used predefined hashtag sets as the gold standard for classifying tweets. Future works can use human annotations to leverage similar tasks. Secondly, we only included original tweets in the exploration of content polarization, however, adding re-tweets in along with network analysis may depict a more comprehensive picture.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support from the NSF Grants #1634944 and #1637067. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the funding sources.

REFERENCES

[1] K. H. Jamieson and J. N. Cappella, *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, 2008.

[2] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005, pp. 36–43.

[3] E. Gilbert, T. Bergstrom, and K. Karahalios, "Blogs are echo chambers: Blogs are echo chambers," in *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*. IEEE, 2009, pp. 1–10.

[4] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[5] E. Colleoni, A. Rozza, and A. Arvidsson, "Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data," *Journal of Communication*, vol. 64, no. 2, pp. 317–332, 2014.

[6] E. Pariser, *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[7] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter." *ICWSM*, vol. 133, pp. 89–96, 2011.

[8] Y.-R. Lin, J. P. Bagrow, and D. Lazer, "More voices than ever? quantifying media bias in networks." *ICWSM*, pp. 193–200, 2011.

[9] V. R. K. Garimella and I. Weber, "A long-term analysis of polarization on twitter," *arXiv preprint arXiv:1703.02769*, 2017.

[10] J. An, D. Quercia, M. Cha, K. Gummadi, and J. Crowcroft, "Sharing political news: the balancing act of intimacy and socialization in selective exposure," *EPJ Data Science*, vol. 3, no. 1, p. 12, 2014.

[11] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social mediasentiment of microblogs and sharing behavior," *Journal of Management Information Systems*, vol. 29, no. 4, pp. 217–248, 2013.

[12] S. Yardi and D. Boyd, "Dynamic debates: An analysis of group polarization over time on twitter," *Bulletin of Science, Technology & Society*, vol. 30, no. 5, pp. 316–327, 2010.

[13] V. Niculae, C. Suen, J. Zhang, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Quotus: The structure of political media coverage as revealed by quoting patterns," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 798–808.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[15] Y. Goldberg and O. Levy, "word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[16] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones, "Word embedding based generalized language model for information retrieval," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 795–798.

[17] T. Dao, S. Keller, and A. Bejnood, "Alternate equivalent substitutes: Recognition of synonyms using word vectors," 2013.

[18] D. Stojanovski, G. Strezoski, G. Madjarov, and I. Dimitrovski, "Twitter sentiment analysis using deep convolutional neural network," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2015, pp. 726–737.

[19] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification." in *ACL (1)*, 2014, pp. 1555–1565.

[20] Y.-R. Lin, D. Margolin, B. Keegan, and D. Lazer, "Voices of victory: A computational focus group framework for tracking opinion shift in real time," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 737–748.

[21] N. Cao, L. Lu, Y.-R. Lin, F. Wang, and Z. Wen, "Socialhelix: visual analysis of sentiment divergence in social media," *Journal of Visualization*, vol. 18, no. 2, pp. 221–235, 2015.

[22] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," *Proceedings of SemEval*, vol. 16, 2016.

[23] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 2011, pp. 192–199.

[24] K. S. Hasan and V. Ng, "Stance classification of ideological debates: Data, models, features, and constraints." in *IJCNLP*, 2013, pp. 1348–1356.

[25] M. A. Walker, P. Anand, R. Abbott, and R. Grant, "Stance classification using dialogic properties of persuasion," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 592–596.

[26] X. Rong, "word2vec parameter learning explained," *arXiv preprint arXiv:1411.2738*, 2014.

[27] C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, and B. Dhoedt, "Learning semantic similarity for very short texts," in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1229–1234.

[28] T. A. Small, "What the hashtag? a content analysis of canadian politics on twitter," *Information, Communication & Society*, vol. 14, no. 6, pp. 872–895, 2011.