Are Words Commensurate with Actions? Quantifying Commitment to a Cause from Online Public Messaging

Zhao Wang
Department of Computer Science
Illinois Institute of Technology
Chicago, IL, 60616, USA
zwang185@hawk.iit.edu

Jennifer Cutler
Kellogg School of Management
Northwestern University
Evanston, IL, 60208, USA
jennifer.cutler@kellogg.northwestern.edu

Aron Culotta
Department of Computer Science
Illinois Institute of Technology
Chicago, IL, 60616, USA
aculotta@iit.edu

***Abstract*—Public entities such as companies and politicians increasingly use online social networks to communicate directly with their constituencies. Often, this public messaging is aimed at aligning the entity with a particular cause or issue, such as the environment or public health. However, as a consumer or voter, it can be difficult to assess an entity's true commitment to a cause based on public messaging. In this paper, we present a text classification approach to categorize a message according to its commitment level toward a cause. We then compare the volume of such messages with external ratings based on entities' actions (e.g., a politician's voting record with respect to the environment or a company's rating from environmental non-profits). We find that by distinguishing between low- and high- level commitment messages, we can more reliably identify truly committed entities. Furthermore, by measuring the discrepancy between classified messages and external ratings, we can identify entities whose public messaging does not align with their actions, thereby providing a methodology to identify potentially "inauthentic" messaging campaigns.**

I. Introduction

Online social networks are increasingly used by public entities such as companies and politicians to speak directly to their constituencies. In addition to typical marketing and campaigning activities, these entities often post messages to foster cause-related associations such as eco-friendliness or public health, which are becoming important components of brand equity [1], [2]. However, due to the low effort and informal nature of such communication, it can be difficult for consumers and voters to determine an entity's commitment to a cause based on their public messaging. In the extreme case, this can result in "greenwashing", a deceptive marketing practice in which firms market their products or policies as more environmentally friendly than they truly are [3]. For example, a recent study suggests that 95% of environmental claims about products contain missing or misleading information [4].

However, alignment with a cause does not always require an explicit claim about a product or practice. For example, a tweet like *"Happy #EarthDay – Let's celebrate our love for the planet"* allows the entity to signal support for a cause without making specific claims about their actions. Contrast this with tweets like *"Today I've introduced legislation to support our fisheries and habitat."* or *"Our products are 100% sustainably sourced."* which indicate stronger commitment to a cause.

In this paper, we choose Twitter as our public messaging platform. We first introduce 4-point annotation scales to label training tweets by their commitment levels to a cause. We then propose a supervised text classification approach to categorize tweets into support and non-support classes, and next continue to categorize support tweets into low- and high- commitment classes. We explore a number of features, such as word embeddings, polarity, social interactions and so on. Meanwhile, we experiment with several classifiers and do grid-search with cross-validations to select the best model and corresponding parameters. Overall, we find that word embedding features significantly increase classification accuracy for short text.

Additionally, after model training, we apply support-classifier and commitment-classifier to all the historical tweets of hundreds of entities, and quantify the volume of tweets assigned to each commitment level as a measure of how entities' words align with causes. To determine the relationship between how entities talk and how they act, we collect entities' action-ratings from third-party sources – from GoodGuide[1] we collect environmental and health ratings for hundreds of brands, and from the League of Conservation Voters[2] we collect the Environmental Scorecard to rate Congress members base on the voting records. We then conduct a regression analysis to quantify how the volume of entities' cause-related tweets correlate with their third-party ratings. We find that entities who post many cause-supportive tweets often have high ratings, and that distinguishing between low- and high-commitment tweets improves this correlation.

Finally, by measuring the discrepancy between entities' volume of cause-related tweets and their action-ratings, we identify several entities that appear to express stronger commitment to causes in public messaging than their action-ratings would suggest. These results suggest that this methodology may be used to quantify the "authenticity" of an entity's public

[1]http://goodguide.com
[2]http://scorecard.lcv.org/

messaging with respect to a cause. Given the importance of authenticity to both firms and consumers [5], [6], the resulting model provides a method for managers and consumers to investigate relationship between a company's words and actions.

In the remainder of the paper, we first summarize related work, then describe our methodology for cause commitment classification and "inauthentic" entity detection; next, we present experimental dataset and results; finally, we conclude with limitations and future work.

II. Related Work

In this section, we discuss several areas of research that are related to yet distinct from the present work: sentiment analysis, stance detection, hedging, and deception detection.

Sentiment analysis. There is long-line of research in categorizing texts by positive or negative opinion of the author [7], [8]. While many approaches assume binary classification, some instead consider a point scale of sentiment. This is distinct from the present work because sentiment intensity does not necessarily contribute to a text's commitment level to a cause (E.g., distinguishing between entities who *"like"* the environment and those who *"love"* the environment has little effect on determining their actual commitment toward environment). Conversely, a high commitment message may not carry any sentiment (E.g., *"We planted 1000 trees this month"* shows high-commitment to the environment with neutral sentiment). Further exploration of how sentiments relate with commitments is shown in Section VI.

Stance Detection. As defined in SemEval-2016 Task 6: Detecting Stance in Tweets means to automatically determine from text *"whether the author is in favor of the given target, against the given target, or whether neither inference is likely."* Related work includes using features such as word n-grams, character n-grams, sentiment lexicons, word vectors [9], [10], punctuation marks, syntactic dependencies and the dialogic structure of posts [11], [12] to do supervised stance classification. However, the present work not only focuses on an entity's stance towards a target cause, but more importantly on whether an entity fulfills its commitment to a cause. Stance detection is useful but not sufficient for the present task. Please refer to Table II for specific explanation.

Hedging. The term "hedging" was introduced by [13] to describe words *"whose meaning implicitly involves fuzziness,"* such as *"likely, potential, may"* etc. Many classification methods have been proposed to identify hedging sentences from news and bio-medical texts [14], [15], [16], [17], [18]. In this task, hedging may occur in low-commitment cause-related messages (e.g., *"We may need to address climate change"*), but is not a necessary feature of all low-commitment messages.

Deception detection. Another line of research investigates linguistic markers of deception – typically by analyzing data collected in laboratory settings in which one subject is instructed to deceive another [19], [20], [21], [22], [23]. Example linguistic markers include verbal immediacy, negative expressions, and emotion words. In our task, we expect outright deception to be rare; instead, we attempt to identify entities whose volume of high-commitment messages are elevated as compared to entities with similar third-party ratings.

Others. [24] gave a lexical analysis of brands' health, environment, and social justice communications on Twitter. However, in this paper, we combine lexical analysis and word embedding features, and provide a more fine-grained classification scheme to quantify volume of public messages' commitment levels.

Cause	Related Keywords
Eco	environment, ecosystem, biodiversity, habitats, climate, ecology, plantlife, pollution, rainforests
Health	healthy, nutritious, lowfat, wholesome, organic, natural, vegan

TABLE I
CAUSE RELATED KEYWORDS FOR RELEVANT TWEET RETRIEVAL

In the context of this prior work, the present paper offers the following **contributions**:

- This paper introduces the task of **cause commitment classification**, a new text classification task for public messaging data, and collect and annotate a new corpus.
- This paper introduces a new perspective to explore **whether entities' words align with their real actions** based on a combination of public messaging as the source of words and third-party ratings as a measure of actions.
- We investigate a number of features for this task, and perform an empirical comparison of several classifiers, indicating the feasibility of automating this task.
- We offer a method to detect potentially **"inauthentic" messages**, defined as high-commitment messages from low-commitment entities.
- We provide both quantitative and qualitative analysis of the results in 3 different domains, demonstrating the generalization ability of this framework, as well as conducting in-depth analysis.

III. Methods for Classifying Tweets by Commitment to a Cause

Our goals are: first, build text classifiers to categorize entities' historical tweets into different commitment levels; second, identify potentially "inauthentic" entities by comparing an entity's volume of high commitment tweets with third-party action-ratings. This section discusses the first task; the next section discusses the second.

Our first task is to build text classifiers that can categorize tweets by their support and commitment levels toward a cause. In this work, we consider three entity & cause pairs: consumer **brands** and **environment protection** ("eco" for short) cause; consumer **brands** and **health/nutrition** cause; **politicians** and **environment protection** cause. Entities along with their action-ratings are collected from third-party sources. Meanwhile, we collect Twitter timelines for each entity.

A. Identify cause relevant tweets

After an initial exploration of the data, we find that most tweets are not related to the target cause, so we use a high-recall method to first identify potentially relevant tweets.

Label	Description	Entity: brands; Cause: eco	Entity: brands; Cause: health
0	Not about the cause.	*Tourism is FL economy's lifeblood, providing 1.2 mil jobs.*	*Just saying hi, regards to hubby on this very special day!!*
1	About the cause, but does not indicate support.	*Check out the stunning landscape for our photoshoot: crisp river waters, mountains, fall foliage #NatureIsGreat*	*Our nutritional information is listed on each package.*
2	Indicates support of the cause in words but not actions. (**low commitment**)	*#CleanWaterAct protects drinking water, critical habitats, and waterways vital for the economy. #ProtectCleanWater*	*Here is a list of the top 10 foods to eat for healthy hair.*
3	Indicates that the entity has taken actions to support the cause. (**high commitment**)	*I've introduced legislation to help conserve our fisheries and habitat.*	*Bringing 23 new Certified Organic products to our fans in 2016.*

TABLE II
4-POINT ANNOTATION SCALES FOR CAUSE COMMITMENT CLASSIFICATION TASK AND EXAMPLES OF BRANDS' TWEETS WITH ECO AND HEALTH CAUSES

To identify topically relevant tweets, we first identify a list of keywords that we expect to be relevant for each cause, listed in Table I. These keywords are selected from each cause's most similar words returned by pre-trained GoogleNews Word2Vec model, which was fit on roughly 100 billion words from a Google News dataset, resulting in 300-dimensional real-valued vectors for 3 million words and phrases[3]. Vectors produced by Word2Vec encode semantic meaning and capture different degrees of similarity between words [25].

We then create one vector per cause by averaging vectors of cause keywords and call it **cause vector**, where words' vectors are produced by pre-trained GoogleNews Word2Vec model. Similarly, we create one vector per tweet by averaging vectors of all words in a tweet and call it **tweet vector**.

Finally, we calculate cosine similarity between a tweet vector and a cause vector as a tweet's **relevance-score with a cause**. After some initial experiments, we set a threshold of 0.3 as the minimum cosine similarity allowable for a tweet to be considered potentially relevant to a cause, and thus serve as a candidate for further classification by the subsequent phases.

B. Select and annotate training data

We sort tweets by their cause relevance-scores, and then select each entity's high relevant tweets as our training dataset.

Additionally, after some initial text analysis, we define a four-point annotation scale for labeling, shown in Table II. We then label the training tweets with this annotation criteria.

C. Feature representation

To represent each tweet, we augment a traditional bag-of-words representation with a number of linguistic features for public messaging, as well as features derived from Word2Vec.

1) Linguistic Cues:

- **Polarity:** This focuses on capturing negative polarity terms (e.g., *not, don't*) that may reverse the meaning of a message (e.g., *"It's not organic"*).
- **Pronoun usage:** We mark first, second, and third persons in tweets to distinguish between tweets talking about the entity itself or others. For example: *"I've introduced legislation to help conserve our fisheries and habitat in South #Louisiana."* A tweet using first person to talk about a cause may be more likely to show high-commitment.
- **Keywords:** We identify 100 most similar words for each cause using GoogleNews Word2Vec model, and then search tweets containing these keywords and calculate the number of keywords in each tweet. This helps to figure out whether explicitly and frequently used cause related keywords will promote a tweet's commitment level.
- **Context:** We mark every keyword's left and right context words as separate features, to investigate if there are common phrases or language patterns relate to the cause (e.g. *"planting trees"* often show up in context of eco, and *"staying in hospital"* often appear in context of health). This is different from n-gram because we only extract keywords' neighbor words but not with keywords. And it is a complementary to the previous keywords features.
- **Social interactions:** Self-mention and re-tweet are common signs in social interactions. We check whether an entity mentions itself in a tweet (e.g. *"@our_company's products are all organic"*), and search for re-tweets that mention the entity itself. For example: a congress member named *RepMarkTakai* posted a tweet: *"RT @CivilBeat: Sen @RepMarkTakai introduce bill to support coral reef conservation."* In this tweet, *RepMarkTakai* is said to be supportive of the environment by *CivilBeat*, providing evidence of RepMarkTakai's efforts toward a cause. If an entity re-tweets a message that mentioned itself, then this message is likely to mention the entity's positive actions towards a cause, which means high-commitment.
- **Part-of-Speech tags:** We use NLTK toolkit[4] to do part-of-speech tagging for tweets, and expect to capture action verbs, which may correlate with high-commitment tweets.

2) Word embedding features: While word embedding features have been used in prior work, here we attempt to customize feature representations for the present task.

- **Tweet vector:** A tweet vector is calculated by averaging vectors of tweet's words. We compute tweet vector and its cause relevance-score as described in III-A. Vectors produced by Word2Vec encode semantic meaning and capture different degrees of similarities [25]. If a tweet vector has high cause relevance score, then this tweet tends to expresses high commitment towards that cause.

[3]https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

[4]http://www.nltk.org/

- **Keywords vector:** Keywords are words that have high cause relevance-scores. To get a keywords vector for each tweet, we first calculate words' cause relevance-scores, and then for each tweet, we sort and extract its top-n (*n=3,5*) cause relevant words as keywords in this tweet. Furthermore, we calculate keywords vector for a tweet by averaging vectors of extracted keywords. Keywords vector serves as a measure to determine whether the most cause-relevant keywords can represent the commitment level of a whole tweet.
- **Keywords' context vector:** This is the vector representation of keywords' context words. Context vector is a complementary vector of keywords vector, it helps to know whether certain context contribute to a tweet's cause commitment level.

We search for optimal combinations of linguistic features and word embedding features in the experiments below.

D. Classifying tweets by support and commitment

We train two separate binary classifiers using labeled training data: the first classifier (**support** classifier) to distinguish non-support tweets (*labels: 0, 1*) from support tweets (*labels: 2, 3*), and the second classifier (**commitment** classifier) to distinguish low-commitment tweets (*label 2*) from high-commitment tweets (*label 3*). We adopt these 2 binary classifiers instead of a single, multi-class classifier because we find that the optimal set of features for each classifier is different.

IV. Methods for detecting "inauthentic" entities

Our second task is to measure the discrepancy between public messages' commitment to a cause and external action ratings for that cause, and then identify "inauthentic" entities whose public messaging does not align with their actions.

After training and evaluating support and commitment classifiers, we apply them to all historical tweets of hundreds of entities to classify their tweets into different cause commitment levels. We then use three different measures to aggregate high-commitment tweets (*label-3*).

- The **number** of high-commitment tweets. Entities who post large number of high-commitment tweets show strong word commitments to a cause.
- The **fraction** of high-commitment tweets. Despite of large number of high-commitment tweets, higher fraction of high-commitment tweets indicates stronger word commitment to a cause. This aims to distinguish between cases where entity A posts 10 high-commitment tweets out of totally 100 tweets, while entity B posts 10 high-commitment tweets out of totally 20 tweets. In this case, entity B shows stronger commitment than entity A.
- The average posterior **probability** assigned to high-commitment tweets. Prediction probability measures the confidence of predicted label, higher probability means more confident prediction. If entities A and B have same number of predicted high-commitment instances, but the average prediction probability of label-3 in A is 0.9 while it is 0.7 in B, then this indicates that A shows higher commitment than B.

We find the top 50 entities according to each metric and take the intersection to get entities that have great number and fraction of confident high-commitment tweets, which means these entities are likely to express high-commitment towards the cause. We then sort these entities in ascending order of third-party ratings and select those below the mean as potentially "inauthentic" entities whose public messaging diverge from third-party ratings. An entity on this list with a low third-party rating may be attempting to align themselves with a cause in words more than their action ratings would suggest.

V. Data

In this section, we describe our experimental datasets used in the proposed approach.

A. Third-party ratings

We collect entities along with their third-party ratings of environmental actions (for brands and politicians) and health actions (for brands) from the following sources.

1) GoodGuide: GoodGuide[5] is a website that provides ratings for products from health aspects, company-level environmental and social issues. Products are scored from 0 to 10. We collect scores for 966 brands across 10 sectors.

2) The League of Conservation Voters (LCV): LCV's National Environment Scorecard[6] has provided objective and factual information about environmental legislation (e.g., *global warming, wildlife conservation, and so on*) and has become the standard bearer to determine the environmental record of Congress members since 1970. We collect scores for 514 Congress members.

B. Twitter

We choose Twitter as the public messaging platform. For each of the entities collected above(*brands and congress members*), we identify the corresponding Twitter account, and download the most recent 3,200 tweets from each account. Table III shows details.

VI. Experimental Results

We use datasets in Table III to validate the proposed approach[7] (*Note: All manually labeling-processes reached 90% agreement among annotators*). We focus on 5 research questions:

- **Sentiment:** Do high commitment tweets express more positive sentiment towards a cause?
- **Support classification:** How well can we distinguish between non-support (*labels: 0,1*) and support (*labels: 2,3*) classes? This indicates which tweets express at least a weak support for a cause.

[5]http://www.goodguide.com/
[6]https://www.lcv.org/
[7]Code is available here: https://github.com/tapilab/icdm-2017-causes

				Ratio of 4 commitment levels			
Cause	**Entity**	**Public Message**	**Labeled Instances**	**0**	**1**	**2**	**3**
Health	142 Brands	429,009 tweets	426	0.023	0.241	0.381	0.355
Eco	966 Brands	2,624,800 tweets	966	0.447	0.234	0.165	0.154
Eco	514 Congress members	1,118,962 tweets	514	0.063	0.197	0.467	0.273

TABLE III
TWITTER DATA COLLECTED FOR 3 PAIRS OF CAUSE-ENTITY TYPES AND CORRESPONDING COMMITMENT DISTRIBUTIONS

Label	**Brand Health**			**Brand Eco**			**Congress Eco**		
	pos	**neg**	**neu**	**pos**	**neg**	**neu**	**pos**	**neg**	**neu**
0	0.60	0.00	0.40	0.61	0.09	0.30	0.56	0.10	0.34
1	0.70	0.05	0.25	0.55	0.12	0.33	0.55	0.16	0.29
2	0.82	0.05	0.13	0.54	0.09	0.37	0.54	0.09	0.37
3	0.79	0.06	0.15	0.45	0.10	0.45	0.45	0.09	0.46

TABLE IV
RATIO OF SENTIMENTS IN 4 COMMITMENT LABELS FOR 3 DATASETS

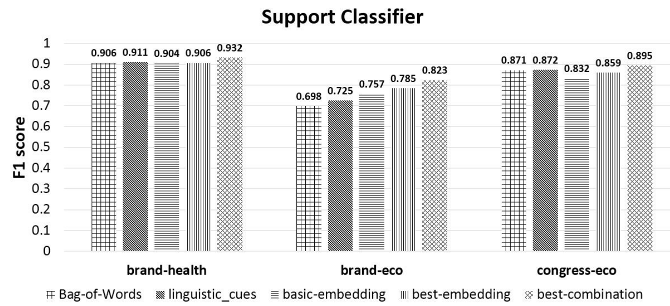

Fig. 1. Average 10-fold cross-validation F1 scores of support classifier with different sets of features. Bag-of-Words features serve as baseline, and linguistic-cues refer to features in III-C. Basic-embedding feature refers to tweet vector, and best-embedding refers to the set of embedding features that produce best F1 score. Best-combination is the combination of linguistic features and embedding features that produce best F1 score.

		support ($\{0,1\}$vs.$\{2,3\}$)			**commitment** (2vs.3)		
Entity	**Cause**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Brands	Health	0.935	0.929	0.932	0.782	0.765	0.773
Brands	Eco	0.860	0.789	0.823	0.800	0.714	0.755
Congress	Eco	0.890	0.902	0.895	0.708	0.721	0.712

TABLE V
PRECISION, RECALL AND F1 SCORES FOR THE BEST CROSS-VALIDATION RESULTS OF SUPPORT AND COMMITMENT CLASSIFIERS

- **Commitment classification:** How well can we distinguish between low-commitment (*label-2*) and high-commitment (*label-3*) classes?
- **Correlation with third-party ratings:** Do entities that tweet more about supporting a cause actually have high ratings with respect to that cause? Does distinguishing between low- and high- commitment tweets provide a stronger signal of third-party ratings?
- **Inauthenticity detection:** If an entity has a low third-party rating but shows high commitment in tweets, is this indicative of possibly inauthentic public messaging?

A. Relationship between sentiment and commitment

We use the TextBlob[8] Python library to classify labeled tweets. Table IV shows ratios of positive, negative and neutral tweets in 4 commitment labels of 3 datasets. Across all datasets, most instances have positive or neutral sentiment and only few have negative sentiment. However, there does not appear to be a strong relationship between positive sentiment and commitment level, which means positive sentiment alone is not sufficient to distinguish between different commitment levels, as suggested by our discussion in Section II.

B. Evaluation for support classification

In support classification, we distinguish between non-support class (*labels: 0,1*) and support class (*labels: 2,3*). For each of the 3 datasets, we experiment with a number of different classifiers (e.g., *LogisticRegression, SVM, MLP, DecisionTree*) and use GridSearchCV in scikit-learn[9] to select the best combination of features and model parameters that give highest F1 score for support class. In this task, LogisticRegression classifier outperforms others.

Fig. 1 shows F1 scores of support classifier for 3 datasets and score variations with different features. Meanwhile, Table V lists details of precision, recall and F1 score for the best classifier. Several conclusions can be drawn from Fig. 1: (1) Performance of Bag-of-Words feature is improved after adding linguistic cues (e.g., *re-tweet: "RT", mention: "@", hashtag: "#"*,); (2) Embedding features alone do not perform better than linguistic cues (embedding features only capture semantics for normal words but not specific symbols (e.g.,"@")); (3) Combination of linguistic features and embedding features gives the best F1 score. Linguistic features are effective only for seen words (words in the training set), and embedding features serve as a complementary to generalize to unseen words when they appear in similar context (words appearing in similar contexts have similar meanings and vector representations).

Table VI shows features that have high coefficients in support classifier. For support classifier, cause keywords play an important role to distinguish between support and non-support classes (e.g., tweets that support *health* tend to use *health keywords such as: natural, healthy, organic*, but tweets don't support health talks more about *flavor and non-healthy aspects: chocolate, cheese, animals' meat and so on*).

C. Evaluation for commitment classification

After classifying tweets into support and non-support classes, we continue to classify support tweets into low- and high- commitment classes with same tuning and evaluation process in support classifier, but the set of features and model

[8]https://textblob.readthedocs.io/en/dev/
[9]http://scikit-learn.org

Entity	Cause	**No Support** *(labels: 0,1)*	**Support** *(labels: 2,3)*	**Low commit** *(label-2)*	**High commit** *(label-3)*
Brands	Health	flavor, cheese, sleek, animals, chocolate	healthy, nutritious, organic, #vegan, natural	foods, eat, recipes, diet, veggies	our, natural, #organic, #nongmo, certified
Brands	Eco	skin, food, diet, fish, natural	sustainable, environment, planet, endangered, sustainability	planet, day, great, can, second_person	protect, _self_first__person, first_person, #sustainable, we
Congress	Eco	lives, rural, industry, jobs, economic	protect, habitats, conservation, epa, forests	plant, epa, historic, pollution, global	I, my, voted, must, _bill_

TABLE VI
FEATURES THAT HAVE HIGH COEFFICIENTS IN SUPPORT AND COMMITMENT CLASSIFIERS

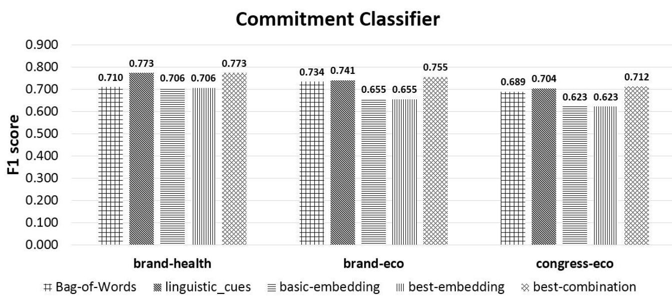

Fig. 2. Performance of commitment classifier with different sets of features, please refer to Fig 1 for explanation of 5 feature sets.

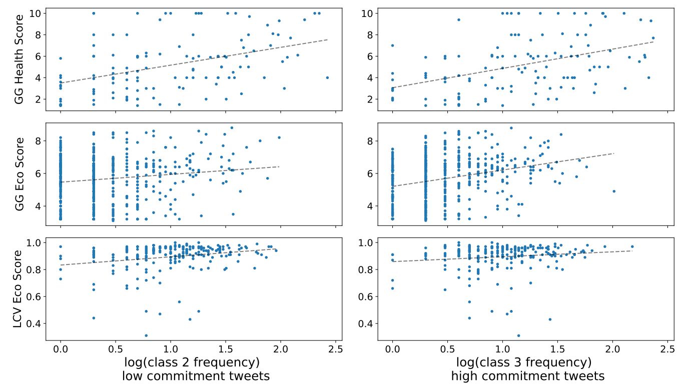

Fig. 3. Scatter plots for coefficients of low- and high- commitment tweets with third-party ratings

parameters that produce best F1 scores are different from support classification. Fig. 2 shows F1 scores of commitment classifier for 3 datasets and score variations with different features. Table V lists details of precision, recall and F1 scores for the best classifier.

According to Fig. 2, we find that: (1) F1 scores of commitment classifier are lower overall than support classifier, as expected given more nuanced distinction between low- and high- commitment levels; (2) Performance of Bag-of-Words feature is improved again after adding linguistic cues; (3) Embedding features alone perform worse than lexical features due to the lack of vector representations for special and distinguishable characters (e.g., *re-tweet: "RT", mention: "@", hashtag: "#"*) and thus not capturing subtle differences between low- and high- commitment classes; (4) Combination of linguistic features and embedding features also make improvements (from 0.710 to 0.773 for brand-health, 0.734 to 0.755 for brand-eco, and 0.689 to 0.712 for congress-eco).

Table VI shows features that have high coefficients in commitment classifier for each dataset. We find: the use of imperative words (e.g., *let, must, need*), pronouns (*I, me ,we*), @oneself together with action verbs (*vote, conserve*) provide the strongest signals (These features corresponds to definition of high-commitment in table II). Among these distinguishable features, most of them are discarded when generating embedding features (e.g., pre-trained GoogleNews Word2Vec model has no vector representation for @oneself), which explains the bad performance of embedding features in Fig. 2.

D. Correlation with third party ratings

We conduct a regression analysis to quantify how the volume of non-support, low- and high- commitment tweets correlate with third-party ratings. In this analysis, to reduce prediction noise, we only take instances with prediction probability greater than 0.7 as confident predictions (though results are similar without this threshold). We fit an Ordinary Least Squares regression model as follows:

	GG HealthScore		GG EcoScore		LCV EcoScore	
	coef	**p-val**	**coef**	**p-val**	**coef**	**p-val**
Non-sup	-1.501	0.033	-0.206	0.257	-0.034	0.202
Low-comt	1.168	0.015	0.101	0.614	0.027	0.149
High-comt	1.451	0.004	1.092	0.001	0.029	0.066

TABLE VII
COEFFICIENTS AND P-VALUES FOR NON-SUPPORT, LOW-/HIGH- COMMIT CLASSES' LOG FREQUENCIES WITH RESPECT TO THIRD-PARTY RATINGS

$$\mathtt{y} = \beta_0 + \beta_1 * \mathtt{x}_1 + \beta_2 * \mathtt{x}_2 + \beta_3 * \mathtt{x}_3$$

$\mathtt{y}$ refers to third party ratings, $\mathtt{x}_1$, $\mathtt{x}_2$, $\mathtt{x}_3$ represent log of non-support (*labels: 0, 1*), low-commitment (*label-2*), high-commitment (*label-3*) class frequencies respectively. β_1, β_2 and β_3 are corresponding coefficients.

Fig. 3 shows scatter plots of how the volume of low- and high- commitment classes independently correlate with third party ratings. And Table VII lists the coefficients and p-values for each class in 3 datasets (*GG refers to GoodGuide*).

Note that we do not expect a very strong correlation, given the expected presence of inauthentic messaging. Additionally, there are many outliers for small values of the x-axis, particularly for the GG eco score. This arises because there are many brands with high third-party eco ratings that nonetheless do not have many tweets identified as high-commitment, which suggests that some firms may not view environmental marketing as valuable, either because of the risk of being

Entity	Cause	Precision
Brand	Health	73.74%
Brand	Eco	94.80%
Congress	Eco	75.93%

TABLE VIII
PRECISION OF TWEETS (*from low-rated entities*) THAT ARE CLASSIFIED AS HIGH-COMMITMENT

Brand	Score	Inauthentic high-commitment tweets
littledebbie	1.5	*"RT @quintanarootri: Do you have a @LittleDebbie* ***nutrition*** *plan for #IMChattanooga? Simple* ***carbs*** *for quick* ***energy*** *on the* ***bike*** *#itspersonal"*
ampenergy	1.4	*"****Exercising*** *in the morning helps you stay* ***energized*** *throughout the day. Pack your bag the night before to make getting to the* ***gym*** *easier!"*
sprite	1.5	*@Randa_Rocks Randa, we have* ***vitaminwater zero*** *made with* ***stevia****. No other products yet but we're always coming up with* ***new ideas****!*

TABLE IX
EXAMPLES OF TWEETS CLASSIFIED AS HIGH-COMMITMENT FROM ENTITIES IDENTIFIED AS POTENTIALLY "INAUTHENTIC"

perceived as inauthentic, or because marketing managers do not believe such messaging fits into the overall personality of the brand. These results also explain why it is not advisable to fit a text classification model directly to the third party ratings – given the misalignment between words and actions, such a classifier would likely overfit to the outliers in these data.

Besides that, Table VII shows: (1) For all datasets, non-support classes have negative coefficients with third party ratings while low- and high- commitment classes have positive coefficients with third-party ratings, which means **entities that tweet more about supporting a cause actually have high third-party ratings**; (2) For all datasets, high-commitment classes show higher coefficients than low-commitment classes, suggesting that **high-commitment classes provide stronger signal of third-party ratings than low-commitment classes**.

In next section, we will investigate more closely outliers in scatter plots that may be indicative of inauthentic messaging.

E. Detecting potentially inauthentic entities

As described in Section IV, we first select a set of entities who show high-commitment in public messaging and then mark those who have low third-party ratings as potentially "inauthentic" entities. Below, we provide both quantitative and qualitative analysis to assess the feasibility of this approach. We emphasize that the identification of these entities does not necessarily indicate any wrong doing — there may be valid reasons for the misalignment, which we explore below.

1) Quantitative analysis: For a quantitative measure, for each of the 3 datasets, we select 10 detected "inauthentic" entities along with their tweets that are predicted as high-commitment with probability greater than 0.7. We then manually annotate these tweets to compute the precision of the high-commitment classifier. This provides both an additional validation measure for the classifier on unseen data, as well as a check to ensure that the tweets have been correctly identified as high-commitment.

Table VIII shows the percentage of correctly predicted high-commitment tweets for each domain. Precisions for brand-health and congress-eco datasets are consistent with the performance of commitment classifier. However, we find higher accuracy for the brand-eco dataset. This may in part be because some entities post many similar tweets and if one of them is correctly classified, then all of them are correctly classified.

2) Qualitative analysis: For a qualitative evaluation, we manually read the tweets of each entity to develop a better understanding of each domain.

In the congress-eco domain, the top three identified entities are moderate Democratic members from swing districts (Rep. Ann Kirkpatrick from Arizona, Rep. Kurt Schrader from Oregon and Rep. Joe Manchin from West Virginia). These members often tweet more narrowly about conservation and wildlife protection, as opposed to more broad appeals to prevent global climate change, which is common among more liberal Congress members. Rep. Kirkpatrick was a Democratic member of the House of Representatives until January 2017, representing Arizona's 1st congressional district, which is known to be a swing district — it voted for Republican presidential candidates in the last 5 elections, and of the most recent 6 representatives, 3 were Republicans, and 3 were Democrats. Thus, we would expect a politician in such a district to express more nuanced views toward the environment to cater to such a heterogeneous constituency [26]. Indeed, we find that Rep. Kirkpatrick has a very low lifetime environmental rating from the LCV (68%), which is the 7th lowest score given to a Democratic member of the House in the 2016. Despite such a low score, Rep. Kirkpatrick had 11 tweets classified as high commitment, of which 8 were manually verified as high commitment. For example, one message was a retweet from a non-profit in Arizona that focuses on energy and conservation ("*RT @SonoranArizona: Great meeting with RepKirkpatrick staff on I-11, renewable energy and conservation...*"). Many of the other messages involve her work on a bill to prevent forest fires in Arizona, which was often framed as conservation ("*Thx to CGDispatch for covering passage of my bipartisan bill to protect #AZ forests...*"). Thus, it appears that in the politics domain, entities may frame their public messaging to emphasize their actions with respect to a smaller, bipartisan subset of issues within the overall cause.

For brand entities, we find among the identified entities a number of brands whose public messaging attempt to align themselves with exercise and fitness, even though the product may not be so aligned. Table IX lists 3 examples from brands with low health ratings. Little Debbie produces many pre-packaged desserts, such as mini-cakes and brownies, that are high in saturated fat and sugar. Some of their messages reference how these snacks may be helpful for those training for an Iron Man race (#IM-Chattanooga). Amp Energy is an "energy drink" produced by PepsiCo. A variant of "Mountain Dew," the drink is mainly popular because of its high caffeine.

As Table VIII suggests, precisions are not 100%, which means some entities may possibly be identified potentially "inauthentic" due to mis-classification. For example, the Sprite tweet in Table IX emphasizes that a sugar substitute used in their vitamin water. This may be interpreted as only a weak indicator of support. Overall, however, we find that all identified entities have at least one tweet manually verified as high-commitment.

Taken together, these results suggest that the proposed approach can identify potentially "inauthentic" entities, but we recommend manual verification to understand more precisely how the language relates to the cause being considered.

VII. Limitations and Future Work

This paper proposes a framework to investigate how entities talk and how they act. We use Twitter public messaging as source of entities' words, and third-party ratings as a measure of how they act. Entities who post high-commitment tweets but have low third-party ratings are detected as "inauthentic".

This framework can be generalized to any domain that has a collection of short texts with corresponding ratings (For example, Amazon's/ebay's product descriptions and ratings, restaurants' menu descriptions and customers' votings, and so on). The proposed commitment labeling criteria (Table II) is invariant to different domains. The linguistic features may be a little different for distinct domains but the idea of exploring and combining linguistic cues and embedding vectors to enrich feature representation for short texts can be applied to various domains.

However, there are several limitations need to address in our future work: (1) Our training datasets are manually labeled by experts, and we will automate this task in future work. (2) Once the labeling process is automated, we can get larger size of training data, and then apply more complex techniques such as deep learning. (3) We use pre-trained GoogleNews Word2Vec model to get word vectors, however, this model has no vector representation for some specific but important features in public messaging (e.g., *re-tweet, self-mention: @one's-name, hash-tags: #event*), and we will train Word2Vec models for public message to deal with this problem. (4) Also, we will search for more sources of third-party ratings to check for robustness of this framework.

Acknowledgments

Anonymous reviewers helped improve this paper. This research was funded in part by the National Science Foundation under grants #IIS-1526674 and #IIS-1618244.

References

[1] S. Sen and C.B. Bhattacharya. Does doing good always lead to doing better? Consumer reactions to corporate social responsibility. *Journal of Marketing Research*, 38(2):225–43, 2001.

[2] Khosro S Jahdi and Gaye Acikdilli. Marketing communications and corporate social responsibility (CSR): Marriage of convenience or shotgun wedding? *Journal of Business Ethics*, 88(1):103–113, 2009.

[3] William S Laufer. Social accountability and corporate greenwashing. *Journal of Business Ethics*, 43(3):253–261, 2003.

[4] TerraChoice. The sins of greenwashing: home and family edition, 2010.

[5] M. Schallehn, C. Burmann, and N. Riley. Brand authenticity: model development and empirical testing. *Journal of Product & Brand Management*, 23(3):192–199, 2014.

[6] F. Morhart, L. Malar, A. Gu'evremont, F. Girardin, and B. Grohmann. Brand authenticity: An integrative framework and measurement scale. *Journal of Consumer Psychology*, 2013.

[7] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[8] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[9] Parinaz, Sobhani, Saif M. Mohammad, and Svetlana Kiritchenko. Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semanticss (*SEM 2016),*, 2016.

[10] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2009.

[11] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 2011.

[12] Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. Stance classification using dialogic properties of persuasion. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012.

[13] G. Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2:458–508, 1973.

[14] G. Szarvas. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. Columbus, Ohio: Association for Computational Linguistics*, page 38–45, 2008.

[15] B Medlock and T Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 992–999, 2007.

[16] Melock, Briscoe, and Szarvas G. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proc 46th Meeting of the Association for Computational Linguistics; Columbus,Ohio: Association for Computational Linguistics*, pages 281–289, 2008.

[17] J Lafferty, A McCallum, and F Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001); Williamstown, MA, USA*, pages 282–289, 2001.

[18] C Friedman, PO Alderson, JHM Austin, JJ Cimino, and SB Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1:161–174, 1994.

[19] Bachenko, Joan, Eileen Fitzpatrick, and Michael Schonwetter. Verification and implementation of languagebased deception indicators in civil and criminal narratives. In *In Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK*, pages 41–48, 2008.

[20] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. Lying words: predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29:665–675, 2003.

[21] S. Adams. Communication under stress: indicators of veracity and deception in written narratives. In *Ph.D. dissertation, Virginia Polytechnic Institute and State University*, 2002.

[22] DePaulo, B. M., J.J. Lindsay, B.E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological Bulletin*, 129:74–118, 2003.

[23] Miller, G. R., and J. B. Stiff. Deceptive communication. In *Sage Publications*, 1993.

[24] Aron Culotta, Jennifer Cutler, and Junzhe Zheng. Finding truth in cause-related advertising: A lexical analysis of brands' health, environment, and social justice communications on twitter. *The Journal of Values-Based Leadership*, 8(2):7, 2015.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR Workshop*, 2013.

[26] Gary C Jacobson and Jamie L Carson. *The politics of congressional elections*. Rowman & Littlefield, 2015.