Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion

Stephanie Lukin, Pranav Anand, Marilyn Walker and Steve Whittaker
Computer Science, Linguistics and Psychology Depts.
University of California, Santa Cruz
Santa Cruz, Ca. 95064
`slukin,panand,mawalker,swhittak@ucsc.edu`

Abstract

Americans spend about a third of their time online, with many participating in online conversations on social and political issues. We hypothesize that social media arguments on such issues may be more engaging and persuasive than traditional media summaries, and that particular types of people may be more or less convinced by particular styles of argument, e.g. emotional arguments may resonate with some personalities while factual arguments resonate with others. We report a set of experiments testing at large scale how audience variables interact with argument style to affect the persuasiveness of an argument, an under-researched topic within natural language processing. We show that belief change is affected by personality factors, with conscientious, open and agreeable people being more convinced by emotional arguments.

1 Introduction

Americans spend a third of their online time on social media, with many participating in online conversations about education, public policy, or other social and political issues. Our hypothesis is that online dialogs have important properties that may make them a useful resource for educating the public about such issues. For example, user-generated content might be more engaging and persuasive than traditional media, due to the prevalence of emotional language, social affiliation, conversational argument structure and audience involvement. Moreover, particular types of people may be more or less convinced by particular styles of argument, e.g. emotional arguments

may resonate with some personalities while factual arguments resonate with others.

Factual: Death Penalty	
Q1:	I'm sure there have been more repeat murderers than innocent people put to death. As far as the cost goes, is that really an issue? Execution Room = $10,000. Stainless Steel Table = $2,000. Leather Straps = $200. Lethal Injection Chemicals = $5,000. Knowing this person will never possibly be able to kill again = PRICELESS
R1:	Actually the room, straps, and table are all multi-use. And the drugs only cost Texas $86.08 per execution as of 2002.
Emotional: Death Penalty	
Q2:	You mean, the perpetrator is convicted and the defender acquitted? Yes, that's the rule and not the exception. Notice here how no-one ended up dead, or even particularly seriously injured. Additionally the circumstances described are incredibly rare, that's why it makes the news.
R2:	The defender shouldn't even have been brought to trial in the first place. That doesn't make it any better. Somebody breaks into your home and threatens your family with rape and murder, they deserve serious injury at the very least.

Table 1: Factual vs. Emotional dialog exchanges `4forums.com`. **Q** = Quote, **R** = Response.

For example, contrast the two informal dialogic exchanges about the death penalty in Table 1 with the traditional media professional summary in Table 2. We might expect the argument in Table 2 to be more convincing, because it is carefully written to be balanced and exhaustive (Reed and Rowe, 2004). On the other hand, it seems possible that people find dialogic arguments such as those in Table 1 more engaging and learn more from them. And indeed, about 90% of the people in online forums are so-called lurkers (Whittaker, 1996; Nonnecke and Preece, 2000; Preece et al., 2004), and do not post, suggesting that they are in fact **reading** opinionated dialogs such as those in Table 1 for interest or entertainment.

Research in social psychology identifies three

Curated Summary: Death Penalty
PRO: Proponents of the death penalty say it is an important tool for preserving law and order, deters crime, and costs less than life imprisonment. They argue that retribution or "an eye for an eye" honors the victim, helps console grieving families, and ensures that the perpetrators of heinous crimes never have an opportunity to cause future tragedy.
CON: Opponents of capital punishment say it has no deterrent effect on crime, wrongly gives governments the power to take human life, and perpetuates social injustices by disproportionately targeting people of color (racist) and people who cannot afford good attorneys (classist). They say lifetime jail sentences are a more severe and less expensive punishment than death.

Table 2: Traditional balanced summary of the death penalty issue from `ProCon.org`.

factors that affect argument persuasiveness (Petty and Cacioppo, 1986; Petty and Cacioppo, 1988).

- the ARGUMENT itself
- the AUDIENCE
- the SOURCE of the argument

The ARGUMENT includes the content and its presentation, e.g. whether it is a monolog or a dialog, or whether it is factual or emotional as illustrated in Table 1 and Table 2. The AUDIENCE factor models people's prior beliefs and social affiliations as well as innate individual differences that affect their susceptibility to particular arguments or types of arguments (Anderson, 1971; Davies, 1998; Devine et al., 2000; Petty et al., 1981). Behavioral economics research shows that the cognitive style of the audience interacts with the argument's emotional appeal: emphasizing personal losses is more persuasive for neurotics, whereas gains are effective for extraverts (Carver et al., 2000; Mann et al., 2004). The SOURCE is the speaker, whose influence may depend on factors such as attractiveness, expertise, trustworthiness or group identification or homophily (Eagly and Chaiken, 1975; Kelman, 1961; Bender et al., 2011; Luchok and McCroskey, 1978; Ludford et al., 2004; McPherson et al., 2001).

We present experiments evaluating how properties of social media arguments interact with audience factors to affect belief change. We compare the effects of two aspects of the ARGUMENT: whether it is monologic or dialogic, and whether it is factual or emotional. We also examine how these factors interact with properties of the AUDIENCE. We profile audience **prior beliefs** to test if

more neutral people are swayed by different types of arguments than people with entrenched beliefs. We also profile the audience for Big Five personality traits to see whether **different personality types** are more open to different types of arguments, e.g., we hypothesize that people who are highly agreeable (A) might be more affected by the combative style of emotional arguments. We provide a new corpus for the research community of audience personality profiles, arguments, and belief change measurements.[1]

Audience factors have been explored in social psychological work on persuasion, but have been neglected in computational work, which has largely drawn from sentiment, rhetorical, or argument structure models (Habernal and Gurevych, 2016b; Conrad et al., 2012; Boltuzic and Šnajder, 2014; Choi and Cardie, 2008). We demonstrate that, indeed, undecided people respond differently to arguments than entrenched people, and that the responses of undecided people correlate with personality. We show that this holds across an array of different arguments. Our research questions are:

- Can we mine social media to find arguments that change people's beliefs?
- Do different argument types have different effects on belief change?
- Do personality and prior beliefs affect belief change?
- Are different personality types differently affected by factual vs. emotional arguments?

Our results show a small but highly reliable effect that short arguments derived from online dialogs do lead people to change their minds about topics such as abortion, gun control, gay marriage, evolution, the death penalty and climate change. As expected, opinion change is greater for people who are initially more neutral about a topic, than those who are entrenched. However personality variables also have a clear effect on opinion change: neutral, balanced arguments are more successful with all personality types, but conscientious people are more convinced by dialogic emotional arguments, and agreeable people are more persuaded by dialogic factual arguments. We describe how we use plan these findings to select and repurpose social media arguments to adapt them to people's individual differences and thus maximize their educational impact.

---

[1] `nlds.soe.ucsc.edu/persuasion_persona.`

2 Related Work

Previous work on belief change has primarily focused on single, experimentally crafted, persuasive messages, rather than exploring whether user-generated dialogic arguments can be repurposed to persuade. Recently however several papers have begun to investigate two challenges in argument mining: (1) understanding the structure of an argument and extracting argument components (Lippi and Torroni, 2015; Nguyen and Litman, 2015; Stab and Gurevych, 2014; Lippi and Torroni, 2015; Biran and Rambow, 2011); and (2) understanding what predicts the persuasiveness of web-sourced argumentative content (Habernal and Gurevych, 2016b; Fang et al., 2016; Wachsmuth et al., 2016; Habernal and Gurevych, 2016a; Tan et al., 2016).

Tan et al. (2016) study belief change in the Reddit /r/ChangeMyView subreddit (CMV), in which an original poster (OP) challenges others to change his/her opinion. They build logistic regression models to predict argument success, identifying two conversational dynamic factors: a) early potential persuaders are more successful and b) after 4 exchanges, the chance of persuasion drops virtually to zero. Linguistic factors of persuasive posts include: a) dissimilar content words to the OP, b) similar stop words, c) being lengthy (in words, sentences, and paragraphs), d) italics and bullets. Finally, susceptibility to persuasion is correlated with singular vs. plural first person pronouns, which the authors relate to the personality trait of Openness to Experience. The CMV reddit offers a unique window into how persuasion of self-declared open-minded people occurs online. However, while Tan et al. find potential proxies for personality traits, they cannot examine traits directly because they do not have personality profiles as we do here. They also do not examine the effect of argument style as we do.

Recent work (Habernal and Gurevych, 2016b; Habernal and Gurevych, 2016a) also examines what makes an informal social media argument convincing. They have created a new dataset of pairs of arguments annotated for which argument is more convincing, along with the reasons given by annotators for its convincingness. They test several models for predicting convincingness comparing an SVM with engineered linguistic features to a BLSTM, with both models performing similarly. In contrast to our experiments, they do not explore factors of the audience or explicitly vary the style of the argument.

Previous work also tests the hypothesis that dialogic exchanges might be more engaging, in the context of expository or car sales dialog (André et al., 2000; Lee, 2010; Craig et al., 2006; Stoyanchev and Piwek, 2010). Work comparing monologic vs. dialogic modes of providing information suggest that dialogs: (1) are more memorable and engaging, (2) stimulate the audience to formulate their own questions, and (3) allow audiences to be more successful at following communication (Lee et al., 1998; Fox Tree, 1999; Suzuki and Yamada, 2004; Driscoll et al., 2003; Fox Tree and Mayer, 2008; Fox Tree, 1999; Liu and Fox Tree, 2011).

Other work (Vydiswaran et al., 2012) explores how user-interface factors (e.g., number and order of argument presentation, whether and how arguments are rated) affect how readers process arguments. Several factors increased the number of passages read, including explicitly presenting contrasting viewpoints simultaneously. This exercise caused people with strong beliefs (about the healthiness of milk) to moderate their views after 20-30 minutes of concentrated study. We do not concentrate on interface factors, instead exploring how persuasiveness relates to audience factors and argumentative style. Also our experiments are run online with hundreds of users, rather than as a controlled study in the lab.

3 Experimental Method

Our experimental method consists of the following steps:

- Select user-generated dialogs with persuasive argument features from an online corpus of socio-political debates, exploring the role of **affect** (Sec. 3.1).
- Profile subjects for **personality traits** and **prior beliefs** about socio-political issues (Sec. 3.2).
- Expose subjects to user-generated, factual vs. emotional dialogic exchanges and compare the effects on belief change to balanced, curated arguments (Sec. 3.3).
- Conduct experiments to predict the degree of belief change as a function of prior belief, personality and type of argument.

The participants were pre-qualified using a reading comprehension task that checked their re-

sponses against a gold standard to ensure that they read the arguments carefully. Because we make many comparisons, and our experiments are conducted at large scale, all of our results incorporate Bonferroni corrections.

3.1 Dialog Selection: Identifying Socio-Emotional Arguments

Our work requires a new experimental corpus that is sensitive to readers' prior beliefs and personalities. We utilize online dialogs from `4forums.com` downloaded from The Internet Argument Corpus (IAC) (Walker et al., 2012c). The IAC contains quote/response pairs of targeted arguments between two people (Table 1) on topics such as: death penalty, gay marriage, climate change, abortion, evolution and gun control. Each argument is annotated to distinguish arguments making strong appeals to emotional factors versus straightforwardly factual arguments.

We selected a subset of extreme exemplars of factual (FACT) versus emotional (EMOT) arguments, defined as Q/R pairs reliably annotated to be at the extreme ends of the fact/emotion scale, i.e. responses with an average ≥ 4 annotation were considered factual, and those whose annotation averaged ≤ -4 were considered emotional on a scale of -5 to 5. Table 1 illustrates both factual (R1) and emotional (R2) arguments, with additional examples for other topics in Table 3.

In the IAC, 95% of the Q-R pairs are disagreements, the FACT and EMOT datasets were selected to contain a similar proportion. There was no correlation between agreement/disagreement and emotionality (r = 0.07, ns).

3.2 Personality

Personality is usually measured with a standardized survey that calculates a scalar value for the five OCEAN traits: opennness to experience O, conscientiousness C, extraversion E, agreeableness A, and neuroticism N (Goldberg, 1990; Norman, 1963) We first conducted an experiment to profile the Big Five personality traits of 637 Turkers using the Ten Item Personality Inventory (TIPI) (Gosling et al., 2003). The TIPI instrument defines each person on a scale from 1 to 7 with 0.5 precision. In order to guarantee reliablity of our results, we then verified that our pre-qualified Turkers are representative of the population as a whole, by comparing the means and standard deviations of our sample of 637 Turkers with the na-

	Factual: Abortion
Q3:	Not only that, to suggest that untold numbers of women would seek illegal abortions is a question-begging claim that has no grounding in history, logic, or reason. It is an unfounded, unproven claim. It is a betrayal to sound judgment to make decisions based upon unfounded predictions into the future.
R3:	But it is based on history. There is plenty of history showing that women had illegal abortions.

	Factual: Climate Change
Q4:	This is where the looney left gets lost. Their mantra is atmospheric CO2 levels are escalating and this is unquestionably causing earth's temperature rise. But ask yourself – if global temperatures are experiencing the biggest sustained drop in decades, while CO2 levels continue to rise – how can it be true?
R4:	Because internal variability from the likes of ENSO, which can cause short term swings of a full degree C, easily swamp the smaller increase we'd expect from CO2 forcing. Easy.

	Emotional: Abortion
Q5:	Undesired first pregnancy is an acute problem for many girls who choose to go under the surgical knife, even though that often ends up with infertility, broken life etc. Dry fasting is an alternative to first pregnancy abortion. If applied, up to 2-3 months old embryo gets dissolved after 15-16 days of the fast. Plus, there is no 'christian' sin.
R5:	No Christian sin??? Other then the intent to kill and then doing so:p

	Emotional: Gay Marriage
Q6:	Did anyone else expect anything less? These evil fundie christianists can have affairs, 2, 3, 4, or even 5 marriages yet gay people are a threat to marriage by wanting to get married.
R6:	You hear that cry...allowing gays to marry will cause the downfall of civilization...but you never hear 'how' or 'why'? More Chicken Little ####.

Table 3: Factual vs. Emotional dialog exchanges. **Q** = Quote, **R** = Response.

tional standards given in Gosling et. al (2003). Table 4 shows that our survey means and standard deviations are very close to the national norms, suggesting our sample is representative of the public in general, and hence can be used to validate whether social media arguments could fruitfully be be used to educate the public.

	E	A	C	N	O
Our survey	4.30	5.19	5.50	4.82	5.53
Norms	4.44	5.23	5.4	4.83	5.38
Our survey σ	1.45	1.11	1.32	1.42	1.07
Norms σ	1.44	1.24	1.24	1.41	1.14

Table 4: TIPI σ and mean from our personality survey compared to the normal distribution

745

3.3 Prior Beliefs and Belief Change

Previous research suggests that people who are entrenched about an issue are unlikely to change their mind (Anderson, 1971; Davies, 1998; Devine et al., 2000), so we wanted to establish the baseline beliefs of our pre-qualified Turkers before they had been exposed to any arguments about a topic. We therefore collected each Turker's initial stance on a topic, by asking them to answer a simple **stance question** with no context, for example: *Should the death penalty be allowed?*. Likert responses were recorded on a -5 to 5 slider scale with 0.01 degrees of precision, with labels on the slider of "Yes", "No", or "Neutral".

	Curated Summary: Abortion
PRO:	Proponents, identifying themselves as pro-choice, contend that abortion is a right that should not be limited by governmental or religious authority, and which outweighs any right claimed for an embryo or fetus. They argue that pregnant women will resort to unsafe illegal abortions if there is no legal option.
CON:	Opponents, identifying themselves as pro-life, assert that personhood begins at conception, and therefore abortion is the immoral killing of an innocent human being. They say abortion inflicts suffering on the unborn child, and that it is unfair to allow abortion when couples who cannot biologically conceive are waiting to adopt.

Table 5: Traditional balanced summary of "Should abortion be legal?" from `ProCon.org`.

Our goal is to compare the belief change that results from social-media dialogs with the belief change from professionally-curated monologs. We selected the balanced, monologic, argument summaries from the website `ProCon.org` (in Table 2 with an additional example in Table 5). The arguments from `ProCon.org` are very high quality, and produced by domain experts.

After probing initial beliefs, we presented participants with one of the three different argument types to test their affect on belief change: a Curated Monolog (MONO) (Table 2), an emotional argument (EMOT) (R2 in Table 1), or a factual argument (FACT) (R1 in Table 1). After each person read one of these three types of arguments, we retested their reactions to the original stance question, while viewing the argument. Responses were again recorded on a -5 to 5 slider scale with 0.01 degrees of precision, with labels on the slider of "Yes", "No", or "Neutral". We computed belief change by measuring differences in stance be-

	N	Mean change	σ change
MONO entrenched	1826	0.50	1.09
MONO neutral	1359	0.62	0.71
FACT entrenched	258	0.27	0.79
FACT neutral	202	0.39	0.55
EMOT entrenched	213	0.35	0.87
EMOT neutral	187	0.37	0.54
ALL entrenched	2951	0.43	1.00
ALL neutral	2234	0.51	0.65

Table 6: Means and σ for belief change for neutral and entrenched participants presented with MONO, FACT, or EMOT argument types. Neutrals show more belief change, and all argument types significantly affect beliefs

fore and after reading each argument. We created 20 HITs on Mechanical Turk for this task, with 5 items per hit.

4 Experimental Corpus Results

4.1 Entrenchment and Belief Change

Our first question is whether our method changed participant's beliefs. Table 6 shows belief change as a function of argument type: monologs (MONO), factual (FACT) and emotional (EMOT). Belief change occurred for all argument types: and the change was statistically significant as measured by paired t-tests ($t_{(5184)}$ = 38.31, p <0.0001). This confirms our hypothesis that social media can be mined for persuasive materials. In addition, all three types of arguments independently led to significant changes in belief.[2]

One of the strongest theoretical predictions is that people with entrenched beliefs about an issue are less likely to change their mind when provided new information about that issue. Table 6 shows the relationship between initial beliefs and extent of belief change. We defined people as having more entrenched initial beliefs if their response to the initial stance question was within 0.5 points of the two ends of the scale, i.e. (1.0-1.5) or (4.5-5.0), indicating an extreme initial view.

We tested whether people who were more entrenched initially showed less change than those who were initially more neutral. We conducted a 2 Initial Belief (Entrenched/Neutral) X 3 Argument Type (MONO/EMOT/FACT) ANOVA, with Belief Change as the dependent variable, and Initial Belief and Argument Type as between subjects factors. Again, as expected, initially Entrenched

---

[2](For MONO, $t_{(3184)}$ = 32.65, p <0.0001, for FACT, $t_{(1019)}$ = 14.81, p <0.0001, For EMOT, $t_{(979)}$ = 14.35, p <0.0001).

people showed less change (M = 0.43) than those who began with Neutral views (M = 0.51), ANOVA ($F_{(1,5179)}$=5.97, p = 0.015).

4.2 Argument Type and Belief Change

We wanted to test whether the engaging, socially interesting, dialogic materials of EMOT and FACT might promote more belief change than balanced curated monologic summaries. We tested the differences between argument types, finding a main effect for argument type ($F_{(2,5179)}$=31.59, p <0.0001), with Tukey post-hoc tests showing MONO led to more belief change than both EMOT and FACT (both p <0.0001), but no differences between EMOT and FACT overall across all subjects (See Table 6). Finally there was no interaction between Initial Belief and Argument Type ($F_{(2,5179)}$=1.25, p >0.05): so although neutrals show more belief change overall, this susceptibility does not vary by argument type.

5 Predicting Belief Change

Our results so far show that our arguments changed people's beliefs as a function of their prior beliefs and argument type. However we aim to automatically **predict** belief change, and hypothesize that knowing a person's **personality** in combination with their **prior beliefs** will allow us to select social-media arguments that are more persuasive **for a particular individual**.

Thus, we vary whether providing a learner with features about a person's personality improves performance for predicting belief change, when compared with providing information about degree of entrenchment alone. We use different representations for personality and prior beliefs as features, the raw score from the Likert slider for belief change and the TIPI score, as well as normalizations of the raw scores according to the distributions per topic, and finally categorical binning of the transformed scores.

5.1 Feature Development and Selection

New features were created by computing the z-transformation score from the raw prior beliefs and personality traits scores. Applying Equation 1 to the raw data creates a normal distribution where the new mean is 0 and the standard deviation is 1. For prior beliefs, x_i is an individual prior belief for a particular topic, $\bar{x}_i$ is the mean, and σ_i is the standard deviation for the particular topic.

$$\frac{x_i - \bar{x}_i}{\sigma_{x_i}} \qquad (1)$$

Categorical bins are derived from the transformed scores to describe the direction of the belief change by comparing prior and final recorded beliefs. The belief change is positive or negative depending upon where the Turkers rate themselves on the belief scale, moving more towards one side (1) or the other (5). Next, to control for variance, we apply a z-transformation on change scores to create a normal distribution. We classify the resulting distribution into three bins: Low, Medium, and High. The interpretation of what stance the Low and High bins represent is strictly topic dependent. The Medium bin consists of z-transformation values between -1 and 1. These are the people whose belief change is less than one standard deviation from the transformed mean. The Low bin contains z-transform scores of less than -1 and translates to belief changes of a large magnitude (more than a standard deviation from the mean) in a negative direction, where again, the meaning of "negative" is dependent upon how the question was framed. The High bin contains z-transformation scores of greater than 1 and translates to belief changes of a large magnitude in a positive direction. For example, the stance question *Should the death penalty be allowed?'* has "no" at the -5 end and "yes" at the +5 end of the likert scale. A Low bin is indicative as moving in the direction of the "no" stance and High towards the "yes" stance.

Bins were also derived for the personality traits, e.g. for Openness, the High bin indicates someone who is very open, the Medium bin is average, and the Low bin is not open at all.

Finally, a binary feature was created to represent how entrenched an individual is in a particular topic. This feature is based on the raw prior belief score and is True if the prior belief score is within 0.5 points of either end point on the stance scale. This feature is different from the prior belief bins because this entrenchment feature groups together people who are in the extremes on both sides of the stance scale, while the prior belief bins distinguishes between the two ends.

We created a development set using data from a prior Mechanical Turk experiment which had 20 HITs, 5 questions per HIT, and 20 people who completed each HIT. In the same manner as the FACT and EMOT HITs, these Turkers (whose per-

sonality was already profiled) were asked about their prior beliefs about a topic, then presented with a factual or emotional argument. But in this case they were asked to rate the strength of the argument rather than to report their belief about the topic. We then identified the combination of features that best predicted argument strength in this development data, and then used this feature set for the belief change experiments below. Turkers who participated in this initial study did not participate in the belief change study and vice versa.

Results on the development set showed that the z-transformation scores for prior belief and personality performed better than the raw scores and bins. On the other hand, the belief change feature was most effective when represented as a categorical variable via binning and directionality. We also found that it is better to have **both** the z-transformed prior belief feature and the entrenchment feature. Thus our experiments below use these feature representations.

We test on three different datasets: MONO, FACT and EMOT, to elicit responses from reading the monologic summaries, and the factual and emotional dialogic arguments. The FACT and EMOT datasets have specific information in terms of scalar values about their degree of factuality or emotionality, on a scale of -5,+5 and a feature with this value is created for these datasets derived from the crowdsourced Turker judgments about the degree of Fact/Emotion in a Q/R pair, as described earlier. The monologic summaries (MONO) are assumed to be neutral and are not assigned a value for degree of factuality or emotionality.

5.2 Belief Change Experimental Results

Our dataset consists of 5185 items, with 3185 responses to the balanced MONO summaries, 1020 responses to FACT, and 980 responses to EMOT. We first applied 10-fold cross-validation with Naive Bayes, Nearest Neighbor, AdaBoost, and JRIP, from the Weka toolkit (Hall et al., 2005). Overall, Naive Bayes had the most consistent scores with our feature sets, thus we only report Naive Bayes experimental results below.

Seven feature sets were created for each of the three {MONO, FACT, EMOT} datasets. *None* feature sets are the no-personality baseline within each dataset. The baseline features contain **no information** about the personality of the unseen human subjects. We use the {MONO, FACT,

EMOT}+None feature sets for testing our hypothesis that personality affects belief change, and **our ability to predict belief change** using personality features. *All* feature sets have information about **all** of the human subjects' personality traits as 5 distinct features. The remaining five {O,C,E,A,N} feature sets examine the effect of providing information to the learner about personality using only **one personality trait at a time**, in order to determine if any personality trait is having a larger impact for belief change prediction.

Table 7 summarizes our key results, reporting accuracy, precision, recall, and F1 for predicting belief change as a discrete bin, Low, Medium, and High. We balanced each dataset to contain the same number of instances in bins, thus the accuracy for majority classification is 33% (Row 1).

After running Naive Bayes over all feature sets in the three datasets, we compared the experimental classifier performance of *All* and {O,C,E,A,N} against the None baselines using a Bonferroni corrected t-test for F1 measure. Using statistical ANOVA tests that control for pre and post test sample variance, we found small but highly reliable effects. We show all of our results, but focus our discussion below on statistically significance differences in F1. We boldface personality feature sets in Table 7 that are statistically significant when comparing {MONO, FACT, EMOT}+None with the other feature sets in the group.

The effect of argument alone (without personality information) can be seen by the no-personality baseline for each argument type, where we exclude personality information ({MONO, FACT, EMOT}+None). All these feature sets perform above the baseline of 33% (Row 1). This supports the results of our prior ANOVA testing over all subjects for belief change, and shows that the argument itself partially predicts belief change.

However, more interestingly, Table 7 also shows that providing the learner with information about personality consistently improves the ability of the learner to predict belief change. For all types of arguments, ie. the neutral, monologic summaries and the factual and emotional dialogs, the feature sets without any information about the personality traits of the unseen human subjects perform significantly worse than the feature sets that contain all five personality traits. MONO+None compared to MONO+All (rows 2 and 8 respectively) show a slight but significant increase in F1 from 0.51 to

row #	Dataset	TIPI	Accuracy	Precision	Recall	F1
1	**Baseline**		33%			
2	MONO	None	57%	0.50	0.58	0.51
3		**Open**	58%	0.52	0.59	**0.52**
4		Conscientious	58%	0.51	0.58	0.51
5		Extrovert	58%	0.49	0.57	0.49
6		Agreeable	58%	0.52	0.57	0.50
7		Neurotic	57%	0.49	0.55	0.47
8		**All**	58%	0.52	0.58	**0.52**
9	FACT	None	49%	0.47	0.47	0.46
10		Open	46%	0.45	0.47	0.46
11		Conscientious	48%	0.48	0.45	0.46
12		Extrovert	48%	0.46	0.45	0.44
13		**Agreeable**	51%	0.52	0.49	**0.49**
14		Neurotic	47%	0.45	0.44	0.43
15		**All**	50%	0.49	0.50	**0.49**
16	EMOT	None	53%	0.42	0.52	0.44
17		**Open**	56%	0.54	0.53	**0.51**
18		**Conscientious**	53%	0.49	0.51	**0.48**
19		Extrovert	49%	0.43	0.47	0.44
20		**Agreeable**	52%	0.48	0.50	**0.48**
21		Neurotic	53%	0.43	0.51	0.44
22		**All**	56%	0.55	0.57	**0.56**

Table 7: Predicting Belief Change with Naive Bayes for Three Data Sets: Statistics are based on 10-fold cross validation. Row numbers provided to reference particular results in the text.

0.52 (using a paired t-test on 10 fold cross validation scores, (p = .001)). Similarly, FACT+None versus FACT+All (rows 9 and 15) shows a significantly greater increase in F1: from 0.46 to 0.49 (p = .0002), as does EMOT+None versus EMOT+All (rows 16 and 22) with F1 increasing from 0.44 to 0.56 (p = .00001). This confirms that the personality traits improves a model's ability to predict belief change in unseen human subjects.

Next we compared the effects of providing the learner with information about each individual personality feature **in isolation** by comparing {MONO, FACT, EMOT}+None with individual personality factors. For MONO, we found that adding personality information about Openness to Experience (MONO+O, row 3) improved F1 from 0.51 to 0.52 compared with a no-personality baseline (p = .0006). This suggests that open people are more persuaded by balanced monologic arguments.

A more interesting result is that Openness to Experience (EMOT+O, row 17) was also important for Emotional arguments, increasing F1 from 0.44 to 0.51 (p = .00001). In contrast, Openness had no effect for Factual arguments (Row 10) (p > 0.05). Models for predicting belief change for Emotional arguments also benefit from information about Conscientiousness and Agreeableness. Row 17 (EMOT+O), Row 18 (EMOT+C) and Row 20 (EMOT+A) all show significant differences in

F1, with EMOT+O better than EMOT+None (p = .00001), EMOT+C better than EMOT+None (p = .00001) and EMOT+A better than EMOT+None (p = .0001).

Information about Agreeableness also improves the quality of the belief change models for the factual dialogs (FACT+A, row 5) with an increase in F1 from 0.46 baseline to 0.49 (p = .004), suggesting that people who are more Agreeable are more influenced by factual arguments. This confirms one of our initial hypotheses that Agreeable people would be more sensitive to the fact/emotional dimension of arguments because of their desire to either avoid conflict (highly Agreeable people) or to seek conflict (Disagreeable people).

6 Conclusions

To the best of our knowledge we are the first to examine the interaction of social media argument types with audience factors. Our contributions are:

- A new corpus of personality information and belief change in socio-political arguments;
- A new method for identifying and deploying social media content to inform and engage the public about important social and political topics;
- Results showing at scale (hundreds of users) that we can mine arguments from online discussions to change people's beliefs;

- Results showing that different types of arguments have different effects: while balanced monologic summaries led to the greatest belief change, socio-emotional online exchanges also caused changes in belief.

Although our short question/response pairs did not induce as much belief change as the curated balanced monologs, we believe that these are striking results given that the materials we extracted from online discussions are not balanced or professionally produced, but instead are simple fragments extracted from online discussions.

Further, confirming prior work on persuasion (Eagly and Chaiken, 1975; Kelman, 1961; Petty et al., 1981), we found that these effects depend on audience characteristics. As expected, belief depended on the strength of prior beliefs so that initially neutral people were more likely to be persuaded than entrenched individuals, regardless of the type of argument. Again supporting our predictions, argument effectiveness depended on personality type. People who are Open to Experience were influenced by balanced and emotional materials. In contrast, Agreeable people are most affected by factual materials. Emotional arguments had very different effects from factual and balanced monologs: Openness is important but so too are Conscientiousness and Agreeableness.

How can we explain this? People who are more Open are typically receptive to new ideas. But our results for emotional arguments also show that Conscientious people change their views when presented with emotional arguments, possibly because they are careful to process the arguments however expressed. And Agreeable people may also be motivated to change belief by emotional arguments because they are less likely to be influenced by personal feelings.

Our results have numerous implications that suggest further technical experimentation. The fact that we can induce belief change by extracting simple discussion fragments suggests that belief change can be induced without the application of sophisticated text processing tools. While our results for balanced monologs suggest that summaries increase belief change, summary tools for such arguments are still under development (Misra et al., 2015). However, perhaps high quality summaries may not be needed if compelling argument fragments can be automatically extracted (Misra et al., 2016b; Subba and Di Eugenio, 2007; Nguyen and Litman, 2015; Swanson et al., 2015).

Our work also suggests the importance of personalization for persuasion: with different personality types being open to different styles of argument. Future work might be based on methods for profiling participant personality from simple online behaviors (Di Eugenio et al., 2013; Liu et al., 2016; Pan and Zhou, 2014; Yee et al., 2011), or from user-generated content such as first-person narratives or conversations (Mairesse and Walker, 2006a; Mairesse and Walker, 2006b; Rahimtoroghi et al., 2016; Rahimtoroghi et al., 2014). We could then select personalized arguments to meet a participant's processing style.

While here we used crowdsourced judgments to select arguments of particular types. Elsewhere, we present algorithms for automatically identifying and bootstrapping arguments with different properties. We have methods to extract arguments that represent different stances on an issue (Misra et al., 2016a; Anand et al., 2011; Sridhar et al., 2015; Walker et al., 2012a; Walker et al., 2012b), as well as argument exchanges that are agreements vs. disagreements (Misra and Walker, 2015), factual vs. emotional arguments (Oraby et al., 2015), sarcastic and not-sarcastic arguments, and nasty vs. nice arguments (Oraby et al., 2016; Lukin and Walker, 2013; Justo et al., 2014).

An open question is to whether these effects are long term. Our approach limits us to examining belief change during a single session for practical reasons; long-term cross-session comparisons lead to significant participant retention issues.

Our results also suggest new empirical and theoretical methods for studying persuasion at scale. Only recently have studies of persuasion moved beyond small scale lab studies involving simple single arguments (Habernal and Gurevych, 2016b; Habernal and Gurevych, 2016a; Tan et al., 2016). Our research also suggests new methods and tools for larger scale studies of persuasion. While care must be taken in deploying these results, studies of juries and other decision making bodies suggest that exposure to a diversity of opinions and minority views are very important to countering extremism and understanding the issues at stake (Devine et al., 2000; Ludford et al., 2004). The ability to repurpose the huge number of varied opinions available in social media sites for educational purposes could provide a novel way to expose people to a diversity of views.

References

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats Rule and Dogs Drool: Classifying Stance in Online Debate. In *Proc. of the ACL Workshop on Sentiment and Subjectivity*.

N. H. Anderson. 1971. Integration theory and attitude change. *Psychological Review*, 78(3):171.

Elisabeth André, Thomas Rist, Susanne van Mulken, Martin Klesen, and Stephan Baldes. 2000. The automated design of believable dialogues for animated presentation teams. *Embodied conversational agents*, pages 220–255.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media*, pages 48–57. Association for Computational Linguistics.

Oram Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *2011 Fifth IEEE International Conference on Semantic Computing (ICSC)*, pages 162–168.

Filip Boltuzic and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proc. of the First Workshop on Argumentation Mining*, pages 49–58.

Charles S. Carver, Bjorn Meyer, and Michael H. Antoni. 2000. Responsiveness to threats and incentives, expectancy of recurrence, and distress and disengagement: Moderator effects in women with early stage breast cancer. *Journal of Consulting and Clinical Psychology*, 68(6):965.

Y. Choi and C. Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801. Association for Computational Linguistics.

Alexander Conrad, and Janyce Wiebe. 2012. Recognizing arguing subjectivity and argument tags. In *Proc. of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88.

S. D. Craig, J. Sullins, A. Witherspoon, and B. Gholson. 2006. The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. *Cognition and Instruction*, 24(4):565–591.

M. F. Davies. 1998. Dogmatism and belief formation: Output interference in the processing of supporting and contradictory cognitions. *Journal of personality and social psychology*, 75(2):456.

D. J. Devine, L. D. Clayton, B. B. Dunford, R. Seying, and J. Pryce. 2000. Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy, and Law*, 7(3):622–727.

Barbara Di Eugenio, Nick Green, and Rajen Subba. 2013. Detecting life events in feeds from twitter. In *ICSC*, pages 274–277.

D. M. Driscoll, S. D. Craig, B. Gholson, M. Ventura, X. Hu, and A. C. Graesser. 2003. Vicarious learning: Effects of overhearing dialog and monologue-like discourse in a virtual tutoring session. *Journal of Educational Computing Research*, 29(4):431–450.

A. H. Eagly and S. Chaiken. 1975. An attribution analysis of the effect of communicator characteristics on opinion change: The case of communicator attractiveness. *Journal of Personality and Social Psychology*, 32(1):136.

Hao Fang, Hao Cheng, and Mari Ostendorf. 2016. Learning latent local conversation modes for predicting community endorsement in online discussions. In *Conference on Empirical Methods in Natural Language Processing*, page 55.

J. E. Fox Tree and S. A. Mayer. —2008—. Overhearing single and multiple perspectives. *Discourse Processes*, 45(160-179).

J. E. Fox Tree. —1999—. Listening in on monologues and dialogues. *Discourse Processes*, 27:35–53.

Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59:1216–1229.

S. D. Gosling, P. J. Rentfrow, and W. B. Swann. 2003. A very brief measure of the big five personality domains. *Journal of Research in Personality*, 37:504–528.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. page 12141223.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

M. Hall, F. Eibe, G. Holms, B. Pfahringer, P. Reutemann, and I. Witten. 2005. The weka data mining software: An update. *SIGKDD Explorations*, 11(1).

Raquel Justo, Thomas Corcoran, Stephanie M .Lukin, Marilyn Walker, and M. Inés Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*.

H. C. Kelman. 1961. Processes of opinion change. *Public Opinion Quarterly*, 25(1):57.

J. Lee, F. Dineen, and J. McKendree. 1998. Supporting student discussions: it isn't just talk. *Education and Information Technologies*, 3(3):217–229.

J. Lee. 2010. Vicarious learning from tutorial dialogue. *Sustaining TEL: From Innovation to Learning and Practice*, pages 524–529.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the Twenty-Fourth International Conference on Artificial Intelligence*, pages 185–191.

K. Y. Liu and J. E. Fox Tree. 2011. Eavesdropping on friends and strangers: The influence of perceived familiarity on overhearer discourse comprehension. In *52nd Annual Meeting of the Psychonomics Society*.

Zhe Liu, Yi Wang, Jalal Mahmud, Rama Akkiraju, Jerald Schoudt, Anbang Xu, and Bryan Donovan. 2016. To buy or not to buy? understanding the role of personality traits in predicting consumer behaviors. In *International Conference on Social Informatics*, pages 337–346. Springer.

Joseph A. Luchok and James C. McCroskey. 1978. The effect of quality of evidence on attitude change and source credibility. *The Southern Speech Communication Journal*, 43:371–383.

P. J. Ludford, D. Cosley, D. Frankowski, and L. Terveen. 2004. Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proc. of the SIGCHI conference on Human factors in computing systems*, pages 631–638.

Stephanie Lukin and Marilyn Walker. 2013. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. *NAACL 2013*, page 30.

François Mairesse and Marilyn A. Walker. 2006a. Automatic recognition of personality in conversation. In *Proceedings of HLT-NAACL*.

François Mairesse and Marilyn A. Walker. 2006b. Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 543–548.

Traci Mann, David Sherman, and John Updegraff. 2004. Dispositional motivations and message framing: a test of the congruency hypothesis in college students. *Health Psychology*, 23(3):330.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444.

Amita Misra and Marilyn A. Walker. 2015. Topic independent identification of agreement and disagreement in social media dialogue. In *Proc. of the SIGDIAL 2013 Conference: The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in dialog. In *Proc. of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Amita Misra, Brian Ecker, Theodore Handleman, Nicolas Hahn, and Marilyn Walker. 2016a. Nldsucsc at semeval-2016 task 6: A semi-supervised approach to detecting stance in tweets. In *Proc. of the International Workshop on Semantic Evaluation*.

Amita Misra, Brian Ecker, and Marilyn A. Walker. 2016b. Measuring the similarity of sentential arguments in dialogue. In *Proc. of the SIGDIAL 2015 Conference: The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Huy V. Nguyen and Diane J. Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28.

B. Nonnecke and J. Preece. 2000. Lurker demographics: Counting the silent. In *Proc. of the SIGCHI conference on Human factors in computing systems*, pages 73–80.

W. T. Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. *Journal of Abnormal and Social Psychology*, 66:574–583.

Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. 2015. And thats a fact: Distinguishing factual and emotional argumentation in online dialogue. In *NAACL HLT 2015 Workshop on Argument Mining*, page 116.

Shereen Oraby, Vrindavan Harrison, Ernesto Hernandez, Lena Reed, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proc. of the SIGDIAL 2015 Conference: The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Shimei Pan and Michelle Zhou. 2014. Pplum: A framework for large-scale personal persuasion. In *Proceedings of the 3rd Workshop on Data-Driven User Behavioral Modeling and Mining from Social Media*, pages 5–6. ACM.

R. E. Petty and J. T. Cacioppo. 1986. The elaboration likelihood model of persuasion. *Advances in experimental social psychology*, 19(1):123–205.

Richard E. Petty and John T. Cacioppo. 1988. The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, 46(1):69–81.

R. E. Petty, J. T. Cacioppo, and R. Goldman. 1981. Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, 41(5):847.

J. Preece, B. Nonnecke, and D. Andrews. 2004. The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior*, 20(2):201–223.

Elahe Rahimtoroghi, Thomas Corcoran, Reid Swanson, Marilyn A. Walker, Kenji Sagae, and Andrew S. Gordon. 2014. Minimal narrative annotation schemes and their applications. In *7th Workshop on Intelligent Narrative Technologies*.

Elahe Rahimtoroghi, Ernesto Hernandez, and Marilyn A. Walker. 2016. Learning fine-grained knowledge about contingent relations between everyday events. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 350.

Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *COLING*, pages 1501–1510.

S. Stoyanchev and P. Piwek. 2010. Harvesting reusable high-level rules for expository dialogue generation. In *Proc. of the 6th International Natural Language Generation Conference*, pages 145–154.

Rajen Subba and Barbara Di Eugenio. 2007. Automatic discourse segmentation using neural networks. In *Proc. of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, pages 189–190.

S. V. Suzuki and S. Yamada. 2004. Persuasion through overheard communication by life-like agents. In *Intelligent Agent Technology, 2004.(IAT 2004). Proc. . IEEE/WIC/ACM International Conference on*, pages 225–231. IEEE.

Reid Swanson, Stephanie Lukin, Luke Eisenberg, Thomas Chase Corcoran, and Marilyn A. Walker. 2014. Getting reliable annotations for sarcasm in online dialogues. In *Language Resources and Evaluation Conference, LREC 2014*.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proc. of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Avril Thorne and V. Nam. 2009. The storied construction of personality. In Kitayama S. and Cohen D., editors, *The Cambridge Handbook of Personality Psychology*, pages 491–505.

V. G. Vydiswaran, ChengXiang Zhai, Dan Roth, and Peter Pirolli. 2012. Biastrust: Teaching biased users about controversial topics. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1905–1909. ACM.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of the 26th International Conference on Computational Linguistics*.

Marilyn Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig Martell, and Joseph King. 2012a. That's your evidence?: Classifying stance in online political debate. *Decision Support Sciences*.

Marilyn Walker, Pranav Anand, Robert Abbott, and Richard Grant. 2012b. Stance classification using dialogic properties of persuasion. In *Meeting of the North American Association for Computational Linguistics. NAACL-HLT12*.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012c. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

S. Whittaker. 1996. Talking to strangers: an evaluation of the factors affecting electronic collaboration. In *Proc. of the 1996 ACM conference on Computer supported cooperative work*, pages 409–418.

Nick Yee, Nicolas Ducheneaut, Les Nelson, and Peter Likarish. 2011. Introverted elves & conscientious gnomes: the expression of personality in world of warcraft. In *Proc. of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 753–762, New York, NY, USA.