A Microtext Corpus for Persuasion Detection in Dialog

Joel Young
Naval Postgraduate School
jdyoung@nps.edu

Craig Martell
Naval Postgraduate School
cmartell@nps.edu

Pranav Anand
University of California Santa Cruz
panand@ucsc.edu

Pedro Ortiz
United States Naval Academy
portiz@usna.edu

Henry Tucker Gilbert IV
Naval Postgraduate School
henrytuckergilbert@gmail.com

Abstract

Automatic detection of persuasion is essential for machine interaction on the social web. To facilitate automated persuasion detection, we present a novel microtext corpus derived from hostage negotiation transcripts as well as a detailed manual (codebook) for persuasion annotation. Our corpus, called the NPS Persuasion Corpus, consists of 37 transcripts from four sets of hostage negotiation transcriptions. Each utterance in the corpus is hand annotated for one of nine categories of persuasion based on Cialdini's model: *reciprocity*, *commitment*, *consistency*, *liking*, *authority*, *social proof*, *scarcity*, *other*, and *not persuasive*. Initial results using three supervised learning algorithms (Naïve Bayes, Maximum Entropy, and Support Vector Machines) combined with gappy and orthogonal sparse bigram feature expansion techniques show that the annotation process did capture machine learnable features of persuasion with F-scores better than baseline.

1 Introduction

Detecting persuasion in text helps address many challenging problems: Analyzing chat forums to find grooming attempts of sexual predators; training salespeople and negotiators; and developing automated sales-support systems. Furthermore, ability to detect persuasion in social flows such as SMS and chat forums can also further enable targeted and relevant advertising. Prior to this effort, there was virtually no work published on automated persuasion detection. A critical problem inhibiting this research has been the lack of persuasion labeled data to learn from.

To support research in this area we decided to develop a microtext corpus focused on persuasion. Our initial survey of traditional microtext sources such as Twitter, SMS, and chat rooms found limited occurrences of directly persuasive attempts. Using these sources for learning to detect persuasion would require annotating vast quantities of text to find enough instances of persuasion. The resulting corpus would be difficult to work with because of the extreme class imbalance between persuasive and non-persuasive text. Our solution was to turn to a domain which is focused on persuasion with published text: hostage negotiation. Even in hostage negotiation, less than twelve percent of utterances are persuasion attempts.

In this paper, we introduce our persuasion model, the hostage negotiation corpus, outline the annotation process and codebook used, and examine initial results showing that the persuasion signal in our corpus is learnable through automated supervised learning.

Table 1: Four lines from the Rogan transcriptions.

Transcript	Line	Type	Speaker	Utterance
Rogan_beta	221	Other	ON80	Yeah but that fiddler isn't gonna cost so much if you walk out easy
Rogan_beta	223	Other	ON80	come on <HT01> you're just making it worst on yourself
Rogan_charlie	641	None	PNl	Alright [both_hang_up]
Rogan_charlie	691	Commitment	HT1	Bring <Wife_First_Name> and Ill come out

2 Building The Corpus

The NPS Persuasion Corpus (Gilbert 2010) contains four sets of transcripts containing a total of 18,847 utterances: a set of FBI and police negotiations gathered by Rogan et al (2002); a set of police negotiations gathered by Taylor (2008); one set of transcripts from the Waco, Texas stand off; and a San Diego Police negotiation (referred to as the Rogan, Taylor, Waco, and San Diego Police transcription sets respectively). The quality varies across transcripts as each was transcribed from audio tapes with varied use of punctuation, capitalization, and capture of emotional and environmental features. Punctuation was removed, named entities were replaced with place holders such as <HOSTAGE_TAKERS_FIRST_NAME>, and transcriber comments were replaced with square-bracketed tokens. Table 1 shows some sample utterances.

2.1 Persuasion Model

There are many questions to be answered when dealing with persuasion in conversation. What is persuasion? Is the detection of persuasion in conversation innate or are there specific types of markers for persuasion attempts? If persuasion detection has specific markers, can these markers be learned and identified by annotators? We addressed these questions through annotation of a corpus of 37 hostage negotia-

tions with persuasion tags based on the social-psychological model of James Cialdini (2007).

In general terms, persuasion is the ability of one party to convince another party to act or believe in some desired way. When defining a persuasion attempt in a conversation corpus, it is simply the agreement between two or more annotators that an utterance is persuasive based on a social model. In other words, if annotators can agree that some utterance of a conversation is meant to be persuasive, then it is. The key for annotators is to use a social model that allows them to have consistent agreement.

For persuasion to be present in a situation, one party must be unwilling or unlikely to perform an act or to believe an idea unless they are influenced by an outside force. This force might be another person, an advertisement, or current social norms and practices. Cialdini identifies six main principles of persuasion:

1. *Reciprocity*: One party becomes indebted to another, and that debt must be repaid.
2. *Commitment and Consistency*: If a person makes a commitment to perform an act or support an idea, that person is obligated to fulfill the commitment. It is *consistent* to keep ones *commitments*.
3. *Scarcity*: The person being influenced must believe that if they do not act in a certain amount of time they will miss a valuable opportunity.
4. *Liking*: People are influenced by similar people or items that bring satisfaction.
5. *Authority*: People are influenced by thoughts, words and actions of authority figures. Authority can be embodied in both individuals and organizations.
6. *Social Proof*: Expectations and behavior are both influenced by social norms.

To ease training and codebook development, we split *commitment* and *consistency* into separate categories.

Our initial annotation attempt using the above categories was not successful. For most of the transcripts, Cohen Kappa scores (a statistical measure of inter-annotator agreement) were below 0.4. To determine if the low scores were a training or model issue, we took a 600 utterance transcript and had the annotators tag the first 300 utterances together, justifying with each other their reasons for tagging each utterance as persuasive. They then tagged the remaining 300 utterances individually.

This process revealed problems with the categories selected. First, there were utterances annotators felt were persuasive, but did not fit into any clear category. In addition, we found that annotators reliably identified some utterances as persuasive but couldn't identify a category. As a result, we added the *other* category to represent persuasive utterances that do not fit into one of the above categories. These eight persuasion categories and *no persuasion* resulted in a total of nine categories.

2.2 The Codebook

Based upon the above refinements, we drafted a new codebook for the annotators. Here are some highlights:

1. Reciprocity
   (a) Look for instances where one party tries to make the other feel indebted to them
      i. The favor done should not be part of a previously established agreement, otherwise reminding the person of the agreement would fall under consistency
      ii. In negotiations, common examples include instances where the negotiator conveys to the hostage taker how hard he is working for him but needs something in return
      iii. This is not a *commitment* (see below) of the form "If you do this, I'll do that"
   (b) Look for cases where the negotiator asks for a favor, which is then rejected, but then follows with a request for a smaller favor. In cases like this one, do not tag the original request with reciprocity, but only the subsequent smaller requests.
2. Commitment
   (a) Look for any kind of deal making ("If you do this, I'll do that...")
   (b) May have to be taken in context as in the following excerpt (Texas State University 1993):

1	JIM	As soon as we get these kids hooked up, I'm going to go back and talk to these commanders –
2	STEVE	All right. Okay.
3	JIM	–about that perimeter motion, okay?
4	STEVE	Right.
5	JIM	And, and some of these issues we've discussed.
6	STEVE	Right, okay.
7	JIM	I've got to round them up. There's a bunch of them in here.
8	STEVE	Okay.
9	JIM	And –
10	STEVE	And call me back then?

   In this case, Steve wants a commitment from Jim to call him back after he speaks with his commanders. Utterance 10 should be tagged with *commitment*, even though there is no specific "if-then" phrase.
   (c) Commitment can also be one party emphasizing an agreement has been made. Getting someone to recognize a commitment has been made is different from reminding them of a previous agreement, which is covered in the consistency tag.
3. Consistency
   (a) When one party (party A) makes a reference to a previous commitment by Party B in order to persuade Party B into some action or request.
   (b) When one party (Party A) makes reference to a want or need by Party B in order to influence some kind of belief or action.
4. Scarcity
   (a) Any utterance which implies explicitly or implicitly that time is a factor

i. A hostage taker setting a deadline for demands to be met

ii. A negotiator claiming that a situation may get worse in the future unless the hostage taker acts now

(b) Any time "secret" information is used to influence another party's decision. For example a negotiator says

> let you all in on a little secret okay? These guards out here, they're pushing me to get something done and I am trying to hold them back because I know you all are up to your word

during a prison riot situation in one of the Taylor transcriptions.

5. Liking

(a) Any kind of compliment in order to influence decisions. Most times, annotators agree that influential compliments are given from the negotiator to the hostage taker, however there can be cases where the hostage taker uses *liking* to influence the negotiator.

i. Can be brown-nosing and insincere

ii. Words and phrases like "buddy" and "friend"

(b) Any reference to similar life experiences

(c) Any expression of affection towards one party; phrases like "I like you" and "it's important to me that you make it out of this" are examples.

6. Authority

(a) Appealing or making reference to a higher authority or expert in order to persuade or influence one party's beliefs or actions

i. A negotiator making reference to his boss's needs in order to influence the hostage taker

ii. A negotiator making reference to an authority figure in the hostage taker's life like a parent or older sibling

(b) Any request for action or belief to a hostage taker from an authority figure. If an utterance like "Please put the gun down and come outside" comes from the negotiator, then that utterance is not persuasive. However, if the same utterance were said by the hostage taker's mother, then the utterance would be persuasive, because the mother is an authority figure. This requires the annotators to understand the context.

7. Social Proof

(a) Any reference to what is normal or customary in situations (a social norm). The negotiator might make reference to what a judge would normally do in order to influence the hostage taker.

(b) Any appeal to what a group thinks the person should do. A negotiator might make reference to the hostage taker's friends or family, claiming they all think he or she should give up. In the following example from the Rogan transcript, the negotiator (PN60) is trying to convince the hostage taker (HT01) to give up.

Table 2: Transcript Kappa scores for revised codebook

Transcript	Utterances	Kappa
Taylor6	2093	0.707
Waco_11B_3	228	0.853
Waco_13_2	312	0.712
Waco_16_1	218	0.855

1	PN60	Suppose you got a healthy body and a healthy mind, right?
2	HT01	[Laughs] I wouldn't bet on that. [Laughs]
3	PN60	Well hell that's quite a bit just that one right there. What? Well I don't know what to tell you know, you got all '- [?]
4	HT01	Huh?
5	PN60	– your friends talkin' to ya and tryin' to give you advice and these people who know you and like you.

In Utterance 5, the negotiator uses peer pressure by making reference to the hostage taker's friends and their opinion that he should give up.

8. Other: The "other" category is a catchall for any utterance that annotators view as persuasive but does not fit the above. Here are some examples agreed on by the annotators:

(a) An appeal to the hostage taker to think about their children

(b) An emphatic plea by the negotiator using words like "guarantee," "absolutely," etc., in order to gain trust

(c) Reasons why certain actions should be performed (justifications)

With this codebook, we were able to dramatically improve inter-annotator agreement. Table 2 shows the resulting Cohen Kappa scores. Note that four previously unseen transcripts were used for these tests to help avoid bias. These results show that persuasion is detectable by annotators using specific markers.

3 Supervised Learning

In order to demonstrate that the above annotation process captured a learnable model of persuasion, we explored several supervised learning methods trained on our corpus. While no work has previously been published addressing automated persuasion detection in naturalistic data, similar tasks have been addressed. A number of researchers have recently investigated the feasibility of determining the side of an issue a participant is on. This work has explored a variety of genres, including congressional floor speeches (Thomas, Pang, and Lee 2006), political opinion pieces (Lin et al. 2006), and online forums (Somasundaran and Wiebe 2010). In addition, Palau et al (2009) has investigated the detection of argumentative passages in legal texts. These efforts demonstrate that detecting social aspects of human behavior is feasible through straightforward supervised learning methods.

Feature vectors are often built using frequency counts of unigrams, bigrams, or larger n-grams over either words or

Table 3: Extracted features from line 2124 (non-persuasive) of the Taylor5_LP transcript.

Feature	Token-Count Pairs
Unigram	WELL-1 I-1 CAN-1 SEE-1 HOW-1 YOU-1 DID-1
Bigram	start_WELL-1 WELL_I-1 I_CAN-1 CAN_SEE-1 SEE_HOW-1 HOW_YOU-1 YOU_DID-1 DID_end-1
Gappy	start_WELL-1 start_I-1 start_CAN-1 start_SEE-1 start_HOW-1 WELL_I-1 WELL_CAN-1 WELL_SEE-1 WELL_HOW-1 WELL_YOU-1 I_CAN-1 I_SEE-1 I_HOW-1 I_YOU-1 I_DID-1 CAN_SEE-1 CAN_HOW-1 CAN_YOU-1 CAN_DID-1 CAN_end-1 SEE_HOW-1 SEE_YOU-1 SEE_DID-1 SEE_end-1 HOW_YOU-1 HOW_DID-1 HOW_end-1 YOU_DID-1 YOU_end-1 DID_end-1
OSB	start_0_WELL-1 start_1_I-1 start_2_CAN-1 start_3_SEE-1 start_4_HOW-1 WELL_0_I-1 WELL_1_CAN-1 WELL_2_SEE-1 WELL_3_HOW-1 WELL_4_YOU-1 I_0_CAN-1 I_1_SEE-1 I_2_HOW-1 I_3_YOU-1 I_4_DID-1 CAN_0_SEE-1 CAN_1_HOW-1 CAN_2_YOU-1 CAN_3_DID-1 CAN_4_end-1 SEE_0_HOW-1 SEE_1_YOU-1 SEE_2_DID-1 SEE_3_end-1 HOW_0_YOU-1 HOW_1_DID-1 HOW_2_end-1 YOU_0_DID-1 YOU_1_end-1 DID_0_end-1

Table 4: Unigrams and bigrams most and least predictive of persuasion

Most Predictive Unigram	Prob	**Least Predictive Unigram**	Prob	**Most Predictive Bigram**	Prob	**Least Predictive Bigram**	Prob
SINCERE	0.88	JESUS	0.04	YOUR-FRIENDS	0.94	YEAH-IM	0.04
HONORABLE	0.87	THANKS	0.03	THAT-GUN	0.93	ME-IN	0.04
ANSWERS	0.86	SHALL	0.03	I-GUARANTEE	0.93	HANG-UP	0.04
CLUBS	0.86	HUH	0.02	YOUR-FAMILY	0.93	NAME-IS	0.03
LEGITIMATE	0.85	SEALS	0.02	YOUR-CELLS	0.92	I-TRIED	0.03
ABOARD	0.83	HELLO	0.02	GET-ALL	0.92	MM-HM	0.03
GUARANTEED	0.83	HI	0.02	YOUR-SAFETY	0.92	MY-WIFE	0.03
BOUT	0.83	CHRIST	0.01	GOOD-JOB	0.92	OF-GOD	0.02
TRUSTING	0.83	BYE	0.01	WHAT-ID	0.92	YOU-DOING	0.02
COOPERATE	0.82	HUM	0.00	GUN-DOWN	0.91	UM-HUM	0.00

characters. As is typical for microtext, each utterance is extremely short and as a result, the feature vectors are extremely sparse. One technique to work around this is to use gappy word bigrams (Bikel and Sorensen 2007). Gappy word bigrams are formed by pairing words within a given distance from each other. The orthogonal sparse bigram (OSB), like the gappy bigram, pairs words within a given distance. Unlike gappy bigrams, OSBs include the distance between words as part of the feature (Cormack, Gómez Hidalgo, and Sánz 2007). While the phrases "the purple dog" and "the big purple dog" both map to the gappy bigram, "the_dog", they map to different OSBs, "the_dog_1" and "the_dog_2" respectively.

After normalizing for case, gappy bigrams and orthogonal sparse bigrams were extracted using five neighboring tokens (gap of four); that is, each word was paired with each of the five closest words. No stemming was done so as to preserve potentially discriminative morphological features (e.g., tense and number). Table 3 shows the resulting entries for a single utterance.

Due to the sparseness of persuasive utterances (less than twelve percent), we grouped all of the persuasion types together into either persuasive or not. This also simplified our learning task as we only needed to address the binary classification of "persuasion" versus "not persuasion." Table 4 shows the most and least persuasive unigram and bigram features after filtering out features not occurring in both persuasive and non-persuasive utterances. Also shown is the conditional probability of persuasion given we observe the feature in an utterance.

3.1 Experimental Design

Two cross-validation approaches were used. First we performed a six-fold cross-validation in which the 36 longest transcripts were randomly grouped into six blocks of six. Six tests were done training on five blocks and testing on the sixth. Next, to test if our persuasion detection efforts were generalizable, we used leave-one-transcription-out validation. That is, we trained on three sets of transcriptions and tested on the fourth.

Table 5: Dataset wide prior probabilities of each class by segmentation type

Naïve Bayes	
Class	Utterances
Persuasive	0.116
Not Persuasive	0.884

3.2 Supervised Learners

We ran our experiments with three binary classification techniques that are often effective for feature vector models: Naïve Bayes with add-one smoothing, Maximum Entropy (Berger, Pietra, and Pietra 1996), and Support Vector Machines (SVMs) (Cortes and Vapnik 1995). Parameters for the above techniques were tuned through grid-search over a set of ten-fold cross-validation experiments.

For Naïve Bayes, the prior probability of persuasion was incrementally increased by 5%. The resulting set of experiments included one set with the prior probability proportional to the probability of the class in the training set and 19 experiments with the prior probability of persuasion set to multiples of 5%, starting at 5% and ending at 95%. For each experiment the prior probability of "not persuasion" was set to $1 - p(\text{persuasion})$.

Our Maximum Entropy experiments were executed with the MegaM system (Daumé III 2004). We tuned the Gaussian prior (λ) with the initial value was set to 2^{-10} and incrementally increased by a power of two up to $\lambda = 2^{10}$. Higher values of λ result in smoother fitting distributions.

SVM experiments were conducted using the LIBSVM package (Chang and Lin 2001) using a radial basis kernel for all experiments. We tuned both cost (penalty for misclassification) and γ (flexibility of the hyperplane). High γ allows the hyperplane to more closely fit the data. The initial value of cost was set to 2^{-5} and increased by two powers-of-two until reaching a maximum value of 2^{15}. For each value of cost, the initial value of γ was set to 2^{-15} and increased by two powers-of-two until reaching a maximum value of 2^5 (Hsu et al. 2003).

For both Naïve Bayes and Maximum Entropy models, we also experimented with tuning the amount of high-entropy terms removed from the data set. Experiments were run removing the highest 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45% and 50% of the terms in the feature set from both the training and test data. These experiments followed the Naïve Bayes protocol previously outlined. For each feature type (e.g. unigrams), 200 grid-search experiments were conducted over six folds.

Unsurprisingly given the large class imbalance, we found that increasing the prior probability of persuasion from the baseline shown in Table 5 increased F-scores. This change was driven by an increase in recall with a simultaneous small decrease in precision. In tuning by removing high-entropy features, the highest F-scores were achieved using no reduc-

Table 6: Tuned Naïve Bayes, Maximum Entropy and SVM parameters

	Naïve Bayes		MaxEnt		SVM	
Features	Prior	Reduction	Lambda	Reduction	C	γ
Unigrams	0.15	0.00	2^{-2}	0.00	2^{15}	2^{-15}
Bigrams	0.45	0.10	2^{-2}	0.00	2^{7}	2^{-7}
Gappy	0.95	0.10	2^{-10}	0.10	2^{7}	2^{-11}
OSBs	0.95	0.00	2^{-8}	0.05	2^{13}	2^{-11}

Table 7: Average performance across repetitions for six-fold cross-validation

Naïve Bayes					
Feature	Precision	Recall	F-score	Baseline F-score	Percent Change
Unigrams	0.4996	0.4052	0.4450	0.2018	120.5
Bigrams	0.4572	0.4172	0.4334	0.2018	114.8
Gappy	0.5072	0.4572	0.4772	0.2018	136.5
OSBs	0.5402	0.3712	0.4358	0.2018	116.0
Maximum Entropy					
Unigrams	0.5430	0.3700	0.4376	0.2018	116.8
Bigrams	0.5950	0.3012	0.3960	0.2018	96.2
Gappy	0.5280	0.3126	0.3902	0.2018	93.4
OSBs	0.6042	0.2564	0.3562	0.2018	76.5
Support Vector Machine					
Unigrams	0.4968	0.3582	0.4134	0.2018	104.9
Bigrams	0.5188	0.3406	0.4080	0.2018	102.2
Gappy	0.5516	0.3142	0.3966	0.2018	96.5
OSBs	0.5498	0.2920	0.3770	0.2018	86.8

tion in the feature set for unigrams and OSBs and a 10% reduction for bigrams and gappy bigrams. Table 6 shows the parameter set chosen for all subsequent experiments.

3.3 Performance Analysis

Each experiment is evaluated with the standard classification metrics of precision (of those our learner said were persuasive, what fraction truly were), recall (of all those that really were persuasive, what fraction did our learner discover), and F-score (the harmonic mean of precision and recall). The harmonic mean is used to prevent extreme recall or precision values from unfairly weighting the results. The baseline F-score is calculated by assuming all utterances are persuasive. The relative improvement is calculated as a percentage of the baseline metric.

3.4 Six-fold cross-validation with five repetitions over utterances

Table 7 shows the average performance across five repetitions of six-fold cross-validation for each of our three supervised learners. Using unigrams as features produced the highest F-scores for Maximum Entropy and SVMs, while gappy bigrams scored highest for Naïve Bayes. Since Maximum Entropy and SVMs are discriminative approaches, this is not an unexpected result. Generative models select which class is most likely, while discriminative models indicate which class is most similar. Furthermore, Ng and Jordan (Ng and Jordan 2002) proved that generative models reach their asymptotic error more quickly than discriminative models. Surprisingly, all three techniques resulted in low recall and high precision with OSBs. Based on the success of Cormack et al (2007) when classifying SMS messages, blog comments, and emails summary information, it was expected

Table 8: Naïve Bayes over utterances, trained on three of four transcriptions

Trained with all, except Rogan (13608 utterances, 89.7% not persuasive)					
Features	Precision	Recall	F-score	Baseline F-score	Percent Change
Unigrams	0.520	0.389	0.445	0.262	69.8
Bigrams	0.508	0.417	0.458	0.262	74.8
GBGs	0.559	0.418	0.478	0.262	82.4
OSBs	0.553	0.305	0.393	0.262	50.0
Trained with all, except Taylor (11944 utterances, 89.7% not persuasive)					
Unigrams	0.453	0.373	0.409	0.227	80.2
Bigrams	0.425	0.399	0.411	0.227	81.1
GBGs	0.486	0.410	0.445	0.227	96.0
OSBs	0.502	0.348	0.411	0.227	81.1
Trained with all, except SDPolice (18033 utterances, 88.7% not persuasive)					
Unigrams	0.731	0.560	0.635	0.297	113.8
Bigrams	0.773	0.482	0.594	0.297	100.0
GBGs	0.737	0.596	0.659	0.297	121.9
OSBs	0.851	0.447	0.586	0.297	97.3
Trained with all, except Waco (12986 utterances, 86.0% not persuasive)					
Unigrams	0.270	0.444	0.335	0.119	181.5
Bigrams	0.242	0.530	0.332	0.119	179.0
GBGs	0.244	0.532	0.334	0.119	180.7
OSBs	0.258	0.481	0.336	0.119	182.4

Table 9: Maximum Entropy over utterances, trained on three of four transcriptions

Trained with all, except Rogan (13608 utterances, 89.7% not persuasive)					
Features	Precision	Recall	F-score	Baseline F-score	% Change
Unigrams	0.594	0.334	0.428	0.262	63.4
Bigrams	0.701	0.271	0.390	0.262	48.9
GBGs	0.628	0.313	0.417	0.262	59.2
OSBs	0.734	0.235	0.356	0.262	35.9
Trained with all, except Taylor (11944 utterances, 89.7% not persuasive)					
Unigrams	0.524	0.365	0.430	0.227	89.4
Bigrams	0.605	0.237	0.341	0.227	50.2
GBGs	0.539	0.266	0.356	0.227	56.8
OSBs	0.632	0.186	0.288	0.227	26.9
Trained with all, except SDPolice (18033 utterances, 88.7% not persuasive)					
Unigrams	0.798	0.560	0.658	0.297	121.5
Bigrams	0.892	0.525	0.661	0.297	122.6
GBGs	0.797	0.447	0.573	0.297	92.9
OSBs	0.890	0.461	0.607	0.297	104.4
Trained with all, except Waco (12986 utterances, 86.0% not persuasive)					
Unigrams	0.293	0.296	0.294	0.119	147.1
Bigrams	0.339	0.306	0.322	0.119	170.6
GBGs	0.286	0.317	0.301	0.119	152.9
OSBs	0.324	0.266	0.292	0.119	145.4

that OSBs would perform much better than they did. This may be a result of sparseness in our training set.

3.5 Leave-One-Transcription-Out Experiments

The results presented in Section 3.4 indicated that it was possible to train weak classifiers using utterances. To further validate, we review the results for the three classifiers using the leave-one-transcription-out tests. Tables 8, 9, and 10 show the results for testing on each transcription set and training on the three others. Looking at raw scores, testing on the San Diego Police transcript boasted the highest scores for each metric across all three supervised learning techniques. However, when we consider percent improvement in F-score over baseline, the Waco transcripts consistently did dramatically better. We expect strong results testing on San Diego Police as this is the smallest transcription (only 824 utterances) and thus symmetrically gives the largest training set. Further research is required to determine why testing on the Waco transcription yielded the strongest results.

4 Results and Future Directions

The annotation study showed that people can be trained to reliably identify persuasion using specific indicators. Furthermore, the above experimental results show that we were

Table 10: Support Vector Machine over utterances, trained on three of four transcriptions

Trained with all, except Rogan (13608 utterances, 89.7% not persuasive)					
Features	Precision	Recall	F-score	Baseline F-score	% Change
Unigrams	0.535	0.319	0.400	0.262	52.7
Bigrams	0.631	0.325	0.429	0.262	63.7
GBGs	0.670	0.300	0.414	0.262	58.0
OSBs	0.681	0.285	0.401	0.262	53.1
Trained with all, except Taylor (11944 utterances, 89.7% not persuasive)					
Unigrams	0.491	0.384	0.431	0.227	89.9
Bigrams	0.527	0.269	0.356	0.227	56.8
GBGs	0.587	0.275	0.374	0.227	64.8
OSBs	0.602	0.241	0.344	0.227	51.5
Trained with all, except SDPolice (18033 utterances, 88.7% not persuasive)					
Unigrams	0.770	0.475	0.588	0.297	98.0
Bigrams	0.818	0.574	0.675	0.297	127.3
GBGs	0.839	0.518	0.640	0.297	115.5
OSBs	0.897	0.496	0.639	0.297	115.2
Trained with all, except Waco (12986 utterances, 86.0% not persuasive)					
Unigrams	0.271	0.285	0.278	0.119	133.6
Bigrams	0.283	0.341	0.310	0.119	160.5
GBGs	0.264	0.293	0.278	0.119	133.6
OSBs	0.287	0.296	0.291	0.119	144.5

able to detect persuasion more accurately than our baseline of simply guessing *persuasion* with several experiments outperforming the baseline F-score by over 120%. This demonstrates that our annotation scheme did produce a signal learnable with straightforward feature vector supervised learning methods. Future work falls into three categories: corpus improvements, feature set improvements, and advances in machine learning. Needed data set improvements include annotating additional negotiation transcripts, adding additional genres such as web pages, blogs, and SMS messages, and augmenting the current data set with additional information such as dialog act tags. It is clear that there is a need for more and larger data sets annotated for belief.

Gappy bigrams and OSBs did help in some cases, but OSBs were rarely the best performing feature set. This is likely due to training data sparseness as OSBs have the most possible features. Preliminary work suggests that combining cheap part of speech tagging (Hitt 2010) with the existing feature types may give a performance advantage. Improved persuasion detection on the utterance level may prove difficult as utterances are often quite short. As discussed in the codebook, human annotators often require additional context to properly classify the utterances. While initial experiments in (Ortiz 2010) with text-tiling (Hearst 1997) were not promising, better methods for text segmentation should be explored. In that same work we present results using a simple voting scheme over our three supervised learning techniques. We found that the voting algorithm was only outperformed by Naïve Bayes suggesting that more powerful ensemble techniques should be explored.

This paper introduced a novel microtext corpus and annotation scheme. Experimental results showed that automated persuasion detection can be learned—not well enough to field operational systems, but well enough to justify further machine learning and annotation efforts. Now that we have demonstrated that automated persuasion detection is possible, the critical next step is to revisit the annotation process to produce a deeper and broader corpus.

References

Berger, A.; Pietra, V.; and Pietra, S. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1):39–71.

Bikel, D., and Sorensen, J. 2007. If we want your opinion. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, 493–500.

Chang, C., and Lin, C. 2001. LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cialdini, R. 2007. *Influence: The Psychology of Persuasion*. New York, NY: Collins.

Cormack, G. V.; Gómez Hidalgo, J. M.; and Sánz, E. P. 2007. Spam filtering for short messages. In *CIKM '07: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, 313–320. Lisbon, Portugal: ACM.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20(3):273–297.

Daumé III, H. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper: http://pub.hal3.name#daume04cg-bfgs, Code: http://hal3.name/megam/.

Gilbert, H. T. 2010. Persuasion detection in conversation. Master's thesis, Naval Postgraduate School, Monterey, CA.

Hearst, M. 1997. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64.

Hitt, J. 2010. Implementation and Performance Exploration of a Cross-Genre Part of Speech Tagging Methodology to Determine Dailog Act Tags in the Chat Domain. Master's thesis, Naval Postgraduate School, Monterey, CA.

Hsu, C.; Chang, C.; Lin, C.; et al. 2003. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.

Lin, W.-H.; Wilson, T.; Wiebe, J.; and Hauptmann, A. 2006. Which side are you on?: identifying perspectives at the document and sentence levels. In *CoNLL-X '06: Proceedings of the Tenth Conference on Computational Natural Language Learning*, 109–116. Morristown, NJ, USA: Association for Computational Linguistics.

Ng, A., and Jordan, M. 2002. On Discriminative Vs. Generative Classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems* 2:841–848.

Ortiz, P. 2010. Machine Learning Techniques for Persuasion Detection in Conversation. Master's thesis, Naval Postgraduate School, Monterey, CA.

Palau, R., and Moens, M. 2009. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proceedings of The Twelfth International Conference on Artificial Intelligence and Law*.

Rogan, R., and Hammer, M. 2002. Crisis/hostage negotiations: A communication-based approach. In Giles, H., ed., *Law Enforcement, Communication, and Community*, 229–254. Philadelphia: Benjamins.

Somasundaran, S., and Wiebe, J. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 116–124.

Taylor, P., and Thomas, S. 2008. Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research* 1:263–281.

Thomas, M.; Pang, B.; and Lee, L. 2006. Get out the vote: determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 327–335.