2012 IEEE Sixth International Conference on Semantic Computing

The Pursuit of Power and its Manifestation in Written Dialog

Swabha Swayamdipta
Department of Computer Science
Columbia University
New York City, New York, USA
Email: ss4173@columbia.edu

Owen Rambow
Center for Computational Learning Systems
Columbia University
New York City, New York, USA
Email: rambow@ccls.columbia.edu

***Abstract*—In this paper we explore the written dialog behavior of participants in an online discussion for automatic identification of participants who pursue power within the discussion group. We employ various standard unsupervised machine learning approaches to make this prediction. Our approach relies on the identification of certain discourse structures and linguistic techniques used by participants in the discussion. We achive an F-measure of 69.5% using unsupervised methods.**

I. Introduction

In this paper, we consider written interactions in discussion forums in which there is a common goal, but no given hierarchical structure among participants. In any discussion forum where a single meaningful conclusion is aimed at, there are bound to be individuals who are more strongly motivated to impose their views on the whole group than the rest. These participants try to get their opinions heard and put into effect. They tend to assert themselves repeatedly in order that their opinion is accepted by other participants. It is also seen that these participants, more often than not, get into conflicts with others in the group, who have a similar pursuit. Regardless of whether they succeed or not, they can be understood to be in **pursuit of power** in that group.

Formally, we define[1] a person to be in "pursuit of power" if:

- He or she tries to gain control of the actions of the group.
- He or she repeatedly pushes forward his agenda.
- His or her actions cause tension in the group.

All three clauses of the above definition need to be satisfied for a person to be labeled as in pursuit of power.

Detecting power seeking individuals in online conversational situations will help identify both new ideas and their promoters. This could also help in the identification of "trouble-makers" in a group. More generally, this study will contribute to understanding how we use language to achieve our communicative and social goals.

Our task is to try to automatically predict the participants in an online written discussion forum who are in the pursuit of power. Our hypothesis is that pursuit of power will be reflected in a recurrent and coherent set of linguistic behaviors which we can identify automatically. Our method employs transductive and inductive unsupervised machine learning approaches that exploit lexical as well as discourse features to this end. We apply and test this approach on a few hundred written discussion threads from Wikipedia, each featuring a few participants who discuss a single event. Our results show us that a Gaussian Mixture Model performs better than models that make stronger assumptions, like the Naive Bayes assumption. We also show that some of the unsupervised models perform as well as a supervised model.

This paper is organized as follows. After reviewing related literature, we present our data. We then describe the standard models and feature space that we made use of in our task. Finally we focus on our experiments and results that the various systems gave us. We conclude with a discussion of our results and a note about our future work.

[1]This definition is the result of a discussion among the participants of the IARPA SCIL research program, and has been used by several research groups.

II. Related Literature

It has long been established that there is a correlation between dialog behavior of a discourse participant and how influential he or she is perceived to be by the other discourse participants [1], [2], [3], [4], [5]. Specifically, factors such as frequency of contribution, proportion of turns, and number of successful interruptions have been identified as being important indicators of influence. Reid and Ng (2000) [6] explain this correlation by saying that "conversational turns function as a resource for establishing influence": discourse participants can manipulate the dialog structure in order to gain influence. This echoes a starker formulation by Bales (1970) [7]: "To take up time speaking in a small group is to exercise power over the other members for at least the duration of the time taken, regardless of the content." Simply successfully claiming the conversational floor represents a feat of power. The work just discussed was done entirely on spoken dialog. In this paper, we show that the core insight — conversation is a resource for influence and power — carries over to written dialog, and that we can detect not only power, but the pursuit of power by studying the structure of the dialog.

We now turn to the computational literature. We know of no work that discusses specifically the pursuit of power, and we discuss here computational work that attempts to discover various types of power relations. Several studies have used Social Network Analysis [8], [9], [10] or email traffic patterns [11] for extracting social relations from online communication.

978-0-7695-4859-3/12 $26.00 © 2012 IEEE
DOI 10.1109/ICSC.2012.49

These studies use only meta-data about messages: who sent a message to whom when. For example, Craemer *et al.* (2009) [10] find that the response time is an indicator of hierarchical relations. Using NLP to deduce social relations from online communication is relatively a new area which has been studied only recently [12]. Bramsen *et al.* 2011 [12] address the problem of identifying social power relationships from online written communication. They use the Enron email corpus for their experiments. Using knowledge of the actual organizational structure, they create two sets of messages: messages sent from a superior to a subordinate, and *vice versa*. Their task is to determine the direction of power (since all their data, by construction of the corpus, has a power relationship). In contrast, our task is to find those people *pursuing* power among *all* discourse participants, most or all of whom are not pursuing power. Thus, they approach the task as a text classification problem and build a classifier to determine whether the set of all emails (regardless of thread) between two participants is an instance of up-speak or down-speak. In contrast, our data unit is a thread, and a thread may or may not include a person who is in pursuit of power (or more than one person). We construe the problem as a classification task on participants in the thread. Finally, they use only lexical and part-of-speech tag features based on the content of messages in the communication. In contrast, our study focuses on the structure of the dialog (which we can do since our unit is a thread, as opposed to a single message or an arbitrary aggregation of single messages).

Strzalkowski *et al.* (2010) [13] are also interested in power in written dialog. However, their work concentrates on lower-level constructs called *Language Uses* which will be used to predict power in subsequent work.

III. Data

Our data set consists of documents from discussion forums from Wikipedia, across different genres.

Each article on Wikipedia has a discussion forum (called a "Talk Page") associated with it that is used to discuss edits for the page. Each forum is composed of a number of threads with explicit topics, and each thread is composed of a set of posts made by contributors. A Wikipedia thread is used to open a discussion on either an edit to an existing page or the creation of a new one. Each thread contains a set of participants, their respective posts (with unique message IDs) in chronological order, such that the intended discourse tree (which reflects who responds to whom) is preserved. See Table I. The discourse tree is defined as follows: each post should have a single parent post, which is either an earlier post or a reference post, addressed to the entire thread and no participant. The first post,which is typically the introduction to the discussion thread is a child to this reference post .

We have a total of 741 threads, containing a total of 4,678 data points, where each data point corresponds to a unique participant in a thread. For the purpose of evaluation, 70 of these threads were annotated for pursuit of power. These 70 threads contain a total of 390 data points. Each data point was annotated as either in pursuit of power or not. Each thread may have any number of participants in pursuit of power. According to our annotations, 28.2% of our data points were labeled as being in pursuit of power.

The threads were annotated by two graduate students. These students had no prior training in linguistics or sociolinguistics. The annotators were given the full definition from the introduction along with detailed labeled examples. They were asked to list the participants that they thought were in pursuit of power along with a justification indicating why they thought so. The justifications had three parts, one for each clause in the definition. Additionally, the participants in pursuit of power were further annotated as successful or unsuccessful in achieving their intended goal based on the annotators' understanding of the entire discussion. The inter-annotator agreement between the two annotators on whether or not a participant is in pursuit of power (given by Cohen's Kappa) is 0.56. The inter-annotator agreement given in terms of F-measure is 0.715. We are continuing the annotation effort and are hoping to increase the inter-annotator agreement through further training.

IV. Models

The models we employ for our task are standard unsupervised models used for classification. The motivation behind using an unsupervised approach as opposed to a supervised one is the lack of enough labeled data. To estimate the parameters under these models, we shall use the Expectation Maximization (EM) algorithm, which is a parameter estimation algorithm for models with latent variables. In each of our models, the observed variables are the data points representing the participants in the thread and the latent variables are the class labels for each data point that specify whether a data point is in pursuit of power or not.

There are two different approaches to unsupervised learning. Under the inductive approach, the parameters are learnt on data points which are exclusive of the test data points. In contrast, the transductive approach makes use of test data (unlabeled, of course) during learning of the parameters. Both approaches are EM based. We hypothesize that the transductive approach will help us classify the test data points better, as the parameters have been learnt using these data points.

We incorporate the following three models in an increasing order of complexity.

A. Naive Bayes Model (NB)

Naive Bayes is a simple probabilistic model for classification. It makes a rather strong assumption that individual features when conditioned on the class value are independent of each other. Such an assumption drastically reduces the parameter space of the model, but also makes the model "naive".

We have a set of unlabeled data points $x^{(i)}$ for $i = 1...n$ where i represents a unique participant in a thread of n participants. $x^{(i)} \in \{0,1\}^d$ is a d dimensional vector of binary features for the participant i. We wish to predict a label $y^{(i)}$

TABLE I

A WIKIPEDIA DISCUSSION THREAD TITLED "GAP(CLOTHING RETAILER):SUSPICIOUS RANDOM DELETES". PARTICIPANT ANTONY IS IN PURSUIT OF POWER. REPLIES ARE INDICATED BY INDENTATION (E.G. P3 IS A REPLY TO P1) AND THE "TO" LABEL. WHEN THE POST ADDRESSES NO ONE IN PARTICULAR, WE USE THE NOTATION "TO ALL".

P1. Antony to all: I didnt realise wikipedia was only limited to those who dont work for GAP Corporate, however I don't think sections should be removed. However wikipedias purpose is not just to company bash, therefore balance needs to be considered. GAP is an example of a company who has responded to labour practices and the reputation it gained, recognition of that is as useful as still highlighting the issues that continue to exist in the apparel market today.
P2. Duncan to all: I would like to remind users to take it to the talk page if they have a problem with any of the content.
P3. Duncan to Antony: – the above was posted under my text from a GAP corporate staff member (confirmed by IP Address)
P4. Bianca to Duncan: A contribution I added a while back (00:24, 10 February 2005) is no longer there. This is the text of that contribution: Despite this, The Gap has received mounting criticism over working conditions in its factories. During the spring of 2003 The Gap, along with 21 other companies, was involved in a class action lawsuit filed by sweatshop workers in Saipan. The allegations included "off the clock" hours, where workers were not paid for working overtime, unsafe working conditions, and forced abortion policies. A settlement of 20 million dollars was reached but The Gap contends that the allegations were without merit, saying that lumping the companies together in one lawsuit was unfair.Perhaps this one such edit. Considering that there is no longer any criticism section on the page.
P5. Bianca to Bianca: Gap is indeed participating in the editing of this article. Additional information was placed in the Labor section of the article, with the author referring to Gap as "our". I edited the addition, keeping relevant information and reformatting the grammar so it sounded like an encyclopedia article. Still, I urge people to watch this page and prevent Gap from deleting information. I'll be posting something to the editors user page warning him to watch his step.
P6. Antony to Bianca: An interesting threat, but as previously pointed out GAP employees are entitled to update Wikipedia as much as anyone else. I have never deleted anything, without clearly pointing out the reasons (which of course on wiki are debatable in this forum). Furthermore it is the interest of the wikipedia that everyone has the opportunity to add and where necessary correct information or as you have done to ensure its independence from bias (whether supportive or critical)
P7. Ephesus to Antony: Wal-Mart Stores, Inc. was accused of whitewashing its own article on Wikipedia about a year ago, and the media made quite a big deal about it. Now we know that Gap edits its article because the ip address is registered to the organization.
P8. Antony to all: I don't live my life based on what the media do or do not make a fuss about. However I do update the GAP article as I want it to be accurate, is that not what we are all after? As long as the guidelines are followed (which I do), then there is no issue. GAP employs 160,000, are they all suddenly not allowed to update this article? Wikipedia is about accuracy not a focus for attacks on the GAP organisation that are unfounded, or as the previous author suggests whitewashing.
P9. Casca to Antony GAP employs 160,000, are they all suddenly not allowed to update this article?
P10 by Ephesus to Casca: Not when they're paid to. There are many people here who edit the articles of the companies they work for when they're off the clock, and I see nothing wrong with it that way. Wikipedia editors are paid with respect, not money, for voluntarily contributing their time here.
P11. Antony to Ephesus: That's nonsense, I am not paid by GAP to update this page. It is not in my job description, not a part of my role. It is updated in my own time and I occasionally use a work computer (which I also have at home). So I am as much a volunteer as you, with an interest in the article being accurate. So as you suggest I am off the clock.

for each $i = 1...n$ where $y^{(i)} \in \{1, -1\}$, meaning that person i is or is not in pursuit of power (PoP).

$$y^{(i)} = \begin{cases} 1 & \text{i is PoP} \\ -1 & \text{otherwise} \end{cases}$$

The joint probability of the user represented by the feature space $x^{(i)}$ belonging to the class $y \in \{-1, 1\}$ under this model is given by

$$p({x_1}^{(i)}...{x_d}^{(i)}, y) = p(y)p({x_1}^{(i)}...{x_d}^{(i)}|y)$$

The Naive Bayes model makes the assumption that the value of the random variable x_j is independent of all other attribute values, $x_{j'}$ for $j' \neq j$ which makes

$$p(x_1^{(i)}, x_2^{(i)}...x_d^{(i)}|y) = p(x_1^{(i)}|y)p(x_2^{(i)}|y)...p(x_d^{(i)}|y)$$

The joint probability could now be written as

$$p({x_1}^{(i)}...{x_d}^{(i)}, y) = q(y)\prod_{j=1}^{d} q_j({x_j}^{(i)}|y)$$

where $q(y)$ and $q_j(x|y)$ for $j = 1...d$ are the parameters of the Naive Bayes model.

The standard EM estimates of these parameters are calculated as follows

$$q(y) = \frac{1}{n}\sum_{i=1}^{n}\delta(y|i)$$

$$q_j(x|y) = \frac{\sum_{i:{x_j}^{(i)}=x}\delta(y|i)}{\sum_{i=1}^{n}\delta(y|i)}$$

where $\delta(y|i) = p(y|x^{(i)})$

Once these parameters are estimated, the probability of a data point $x^{(i)}$ belonging to the class $y = 1$ is given by

$$p(y^{(i)} = 1|x^{(i)}) = \frac{q(1)\prod_j q_j({x_j}^{(i)}|1)}{\sum_{y\in\{-1,1\}} q(y)\prod_j q_j({x_j}^{(i)}|y)}$$

Using the above, the user i can be assigned a class y^*, according to the following inference equation.

$$y^* = argmax_y p(y|x^{(i)})$$

B. Gaussian Mixture Models (GMM)

If the feature space were to contain real-valued features, we would need to model our data with a Gaussian Mixture Model(GMM). A GMM is formed by taking linear combinations of simple Gaussians. It belongs to the class of mixture distributions and is a widely used probabilistic model. Note that binary features are also real-valued, so the GMM can model them accurately.

In a treatment similar to the NB, the joint probability of a data point $x^{(i)} \in \Re^d$ belonging to a class y under this model is given by

$$p(x^{(i)}, y) = q(y)N(x^{(i)}; \mu_y, \Sigma_y)$$

where $N(x^{(i)}; \mu_y, \Sigma_y)$ is the Multivariate Normal Distribution and the vectors μ_y and Σ_y for $y \in \{-1, 1\}$ are the parameters of the GMM.

The standard EM estimates of these parameters are calculated as follows

$$q(y) = \frac{1}{n}\sum_{i=1}^{n}\delta(y|i)$$

$$\mu_y = \frac{1}{N_y}\sum_{i=1}^{n}\delta(y|i)x^{(i)}$$

$$\Sigma_y = \frac{1}{N_y}\sum_{i=1}^{n}\delta(y|i)(x^{(i)} - \mu_y)(x^{(i)} - \mu_y)^T$$

where $\delta(y|i) = p(y|x^{(i)})$ and $N_y = \sum_{i=1}^{n}\delta(y|i)$

Once these parameters are estimated, the probability of a data point $x^{(i)}$ belonging to class $y = 1$ is given by

$$p(y^{(i)} = 1|x^{(i)}) = \frac{q(1)N(x^{(i)};\mu_1,\Sigma_1)}{\sum_{y\in\{-1,1\}} q(y)N(x^{(i)};\mu_y,\Sigma_y)}$$

Using the above, the user i can be assigned a class y^*, according to the following inference equation.

$$y^* = argmax_y p(y|x^{(i)})$$

Unlike the Naive Bayes, this model additionally offers us the significant advantage that it does not assume features are independent when conditioned on the class.

C. Mixture of GMM and Naive Bayes (GMM + NB)

To exploit advantages offered by both the NB and the GMM, we use a third model that is a mixture of both these models. Since we could have both binary as well as real valued features, the data can be modeled in the following way. The hypothesis behind using this model is that NB might be able to classify binary valued vectors better than GMM.

We suppose the data point $x^{(i)}$ contains d_1 binary features and d_2 real-valued features, and can be written as

$$x^{(i)} = {x_1}^{(i)}...{x_{d_1}}^{(i)}x'^{(i)}$$

where $x'^{(i)} \in \Re^{d_2}$ is the d_2 dimensional real-valued component of $x^{(i)}$.

The joint probability can now be written as

$$p(x^{(i)}, y) = q(y)\prod_{j=1}^{d_1} q_j({x_j}^{(i)}|y)N(x'^{(i)}; \mu_y, \Sigma_y)$$

where $q(y)$, $q_j(x|y)$ for $j = 1...d_1$, the vectors μ_y and Σ_y for $y \in \{-1, 1\}$ are the parameters of the model.

The standard EM estimates of these parameters are calculated as follows

$$q(y) = \frac{1}{n}\sum_{i=1}^{n}\delta(y|i)$$

$$q_j(x|y) = \frac{\sum_{i:{x_j}^{(i)}=x}\delta(y|i)}{\sum_{i=1}^{n}\delta(y|i)}$$

$$\mu_y = \frac{1}{N_y}\sum_{i=1}^{n}\delta(y|i)x'^{(i)}$$

$$\Sigma_y = \frac{1}{N_y}\sum_{i=1}^{n}\delta(y|i)(x'^{(i)} - \mu_y)(x'^{(i)} - \mu_y)^T$$

where $\delta(y|i) = p(y|x^{(i)})$

Once these parameters are estimated, the probability of a data point $x^{(i)}$ belonging to class $y = 1$ is given by

$$p(y^{(i)} = 1|x^{(i)}) = \frac{q(1)\prod_{j=1}^{d_1} q_j({x_j}^{(i)}|1)N(x'^{(i)};\mu_1,\Sigma_1)}{\sum_{y\in\{-1,1\}} q(y)\prod_{j=1}^{d_1} q_j({x_j}^{(i)}|1)N(x'^{(i)};\mu_y,\Sigma_y)}$$

Using the above, the user i can be assigned a class y^*, according to the following inference equation.

$$y^* = argmax_y p(y|x^{(i)})$$

V. Features

A crucial question here is how to define the features in our data, starting with just the discourse structure and textual content of posts in the threads. We hypothesize that significant prediction of power could be made using simple features of two kinds: dialog and lexical features.

1) Dialog features: The dialog features try to capture the dialog behavior of each participant in the thread. The fifteen dialog features used and the intuition behind each are:

- **MaxPosts**: Binary feature indicating whether or not a participant has the maximum number of posts in the thread. A participant who has the largest number of posts is actively involved in the discussion and is very likely to be in an effort to control the actions of the group and could hence be in the pursuit of power. E.g., Antony, who is the only person in pursuit of power, has the maximum number of posts in the thread in Table I
- **Consecutive**: Binary feature indicating whether or not a participant at any point in the thread posts consecutively without any intervening post. This is seen when a person tries to add on to his last post or does not wait for a response to continue with his proposed agenda.
- **LongestPost**: Binary feature indicating whether or not a participant has the longest post in the thread. Participants who try to push their agenda tend to justify it by providing factual information resulting in longer posts. A participant with the longest post is very likely to be in the pursuit of power.
- **%Posts**: Real valued feature indicating what percentage of a thread's posts can be attributed to a participant. This is related to the MaxPosts feature. A participant who contributes a large fraction of a thread's posts is actively involved in the discussion and is very likely to be in an effort to control the actions of the group and could hence be in the pursuit of power. For example, Antony has the highest percentage of posts in the thread in Table I

- **Initiation**: Binary feature indicating whether a participant is the first person to post in a thread. When a participant begins the discussion, he is usually suggesting an edit or justifying an edit he recently made. As he has a strong agenda, he is very likely to be in the pursuit of power. E.g., Antony starts the thread in Table I.
- **Alternation**: Binary feature indicating whether or not a participant at any point in the thread posts alternatingly with any other participant in a pattern representing "-A-B-A-". Such a setting is seen in an argument or when a participant is actively questioning another, or justifying his stand to another. All of the above could be likely when the individual is in pursuit of power. E.g. Antony posts alternatively in the 6th and 8th posts in the thread in Table I
- **MaxQuestions**: Binary feature indicating whether or not a participant asks the most questions in the thread. Such a participant tends not be in the pursuit of power in the thread because instead of pushing his own agenda he is questioning others' actions.
- **Repetition**: Binary feature indicating whether or not a participant repeatedly posts in the thread. By repitition we mean posting at least twice in a thread. As far as the definition goes, repetition is a requirement for an individual to be in Pursuit of Power. For example, Antony posts repeatedly in the thread in Table I, as do Duncan and Bianca.
- **UnansweredPosts**: Binary feature indicating whether or not a participant has at least one post to which no one replies. This can be detected from the discourse tree structure of the thread. If no post in the thread has any one of the concerned participant's post as its parent, we can say that the participant has unanswered posts. Such a participant could either not be in the pursuit of power or could be pursuing power but failing at achieving it, because he gets no responses from other participants. In Table I, no participant has unanswered posts.
- **Termination**: Binary feature indicating whether or not a participant is the last person to post in a thread. When a participant ends the discussion, he is likely to have had the last word in it, and has silenced all the other participants, which indicates that he has been successful in achieving power in the discussion. However, it is also possible that no one replies to him because the other participants lost interest. In this case, it is not clear whether or not the participant was in pursuit of power. In the example in Table I, Antony is indeed the last one to post.
- **PointJoined**: Real-valued feature indicating at what point in the lifetime of a discussion thread a participant joins in. Participants who join in early have more chances to be in the pursuit of power, where as participants who join later are generally seen to support others, not to push their own agenda from scratch.
- **LongestBranch**: Binary feature indicating whether or not a participant starts a post that the most number of other participants reply to. If the participant is the author of such a post, it indicates that the content of his post has either caused some tension in the group or addresses an issue many participants have an opinion about. Such people generally could be in the pursuit of power.In the example in Table I, Duncan has the longest branch.
- **Inquisitiveness**: Binary feature indicating whether or not a participant asks questions in the thread. This can be detected by looking for the symbol "?" in the participant's posts. It is generally seen that people who have a clear agenda do not ask others' questions, but generally answer others questions. However, participants could question suggestions or edits by others and hence pursue a conflicting agenda. E.g., Antony asks questions in the thread in Table I
- **QuestionsFreq**: Real-valued feature indicating at what frequency a participant asks questions in the thread. This can be detected by dividing the number of questions asked by the participant by the total time for which the participant is active in the thread, which is obtained by the timestamp associated with a post. It is generally seen that people who have a higher frequency of asking questions, are generally not pursuing power.
- **%Questions**: Real-valued feature indicating what percentage of the posts in a thread by a participant are questions. This is obtained by dividing the number of posts by a participant that contain a "?" by the total number of posts. Again, participants not in the pursuit of power have a higher percentage of questions.

2) Lexical features: Apart from dialog features, the content of the post by a participant could be an indicator of whether the participant is pursuing power or not. We model the content simply by using a bag of words (or more specifically, a bag of lemmas).

Under this feature space, each data point i can be defined as the complete collection of posts $x^{(i)}$ by a unique user in a thread where $i \in 1...n$. Each x^i is a binary vector of dimension $|\vartheta|$, where ϑ is the lemma vocabulary.

For $v \in \vartheta$

$$x_v{}^{(i)} = \begin{cases} 1 & \text{if lemma v has been used by i} \\ 0 & \text{otherwise} \end{cases}$$

VI. Experiments and Results

Our goal is to evaluate the overall performance of unsupervised learning of parameters in order to classify participants in a discussion thread into two classes - PoP and not PoP. We ran experiments to compare various models under an unsupervised approach, to see how the size of data influences the results under these different models, to see how robust each of the different models is, to see how these compare with a supervised model, to compare transductive learning with inductive learning in an unsupervised setting, and finally to see how lexical and dialog features affect the performance of the models. The following sections talk about each of these experiments in more detail.

TABLE II
COMPARISON OF F- MEASURES OF UNSUPERVISED MODELS AND THE RANDOM BASELINE ACROSS VARIOUS DATA SIZES

# Data Points	GMM	GMM + NB	NB	Random Baseline	Positive Baseline
4288	0.6765	0.5941	0.3284	0.2844	0.4402
3988	0.6909	0.5822	0.3283	0.2774	0.4402
3620	0.6785	**0.6026**	0.3865	0.2831	0.4402
3125	0.6828	0.5868	0.3510	0.2777	0.4402
2618	**0.6950**	0.5676	0.3509	0.2820	0.4402
2293	0.6839	0.5865	0.3283	0.2770	0.4402
1805	0.6868	0.5699	**0.4066**	0.2810	0.4402
1026	0.6760	0.5108	0.3566	0.2759	0.4402
704	0.6862	0.5427	0.4009	0.2812	0.4402
390	0.6702	0.5155	0.3737	0.2821	0.4402

A. Comparison of different models for Unsupervised Learning

The following three models with their respective feature space were studied under the unsupervised approach. In each of these, the parameters were initialized from a uniform distribution. For estimating parameters, the EM algorithm was taken till convergence such that if the log likelihood did not increase beyond a threshold of 0.01, convergence was assumed.

- **NB**: This model took only the binary features described in the previous section as input.
- **GMM**: The feature space comprised binary and real-valued features.
- **GMM + NB**: Same as GMM.

We also use two baselines. The first is a **randome baseline** that clusters data points randomly based on a prior distribution of positive and negative examples in the data. The second is a **positive baseline** that classifies all the data points as positive.

Table II shows the results across different models and different data sizes, averaged across 100 runs. For each set of data points, the GMM makes the best predictions, followed by the mixture of GMM and NB. The NB model alone is the weakest. This could be attributed to the fact that it makes a strong assumption about the independence of various features of the input data. The same explanation goes for why a mixture of NB and GMM performs worse than a GMM. Further, even the positive baseline outperforms the NB. However, the GMM and mixture of NB and GMM perform better than the positive baseline and all three models outperform the random baseline.

Figure 1 shows how each of the models perform as we vary the data size. Note that the variations in the GMM across data sizes is minimal whereas the variations in the mixture of GMM and NB are larger, but the NB model has very large variations. This is because NB is extremely sensitive to initialization and converges to the local minima very frequently.

Table III shows the standard deviations of each of the four models under 100 different random initializations. It can be seen that GMM and GMM+NB introduce robustness as compared to the random baseline and to NB.

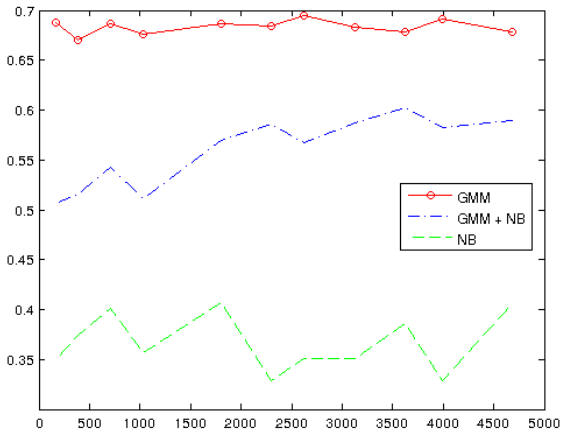

Fig. 1. Comparison of f-measures of unsupervised models across various data sizes

TABLE III
STANDARD DEVIATION OF F-MEASURE ACROSS ALL THREE UNSUPERVISED MODELS AND THE RANDOM BASELINE ON A DATA SET OF SIZE 4678, OVER 100 DIFFERENT INITIALIZATIONS

	SD of F-Measure
Baseline	0.037
NB	0.281
GMM + NB	0.019
GMM	**0.006**

B. Significance Measures

We calculated the significance measures of all three models with respect to both the baselines. For this, we used the p-value calculated with McNemar's test with the continuity correction. The smaller the p-value, the greater the statistical significance of the difference between the two systems being analyzed. From Table IV below, the improvement provided by all our systems over the random baseline is highly statistically significant. However, GMM and GMM+NB have much higher statistical significance as compared to NB against the random baseline, showing that NB is more closely associated with the latter. The difference between the positive baseline and

the NB is statistically significant and that between either the GMM or GMM+NB and the positive baseline is highly so. When compared against each other, the difference between GMM and GMM+NB is less statistically significant. However the difference between either of these and NB is highly statistically significant.

TABLE IV
MCNEMAR'S STATISTICAL SIGNIFICANCE MEASURES FOR EVERY PAIR OF SYSTEMS ON A TEST SET OF 390 POINTS

System 1	System 2	p-value
GMM	GMM + NB	0.0432
GMM + NB	NB	≤ 0.0001
NB	Random Baseline	≤ 0.0001
NB	Positive Baseline	≤ 0.0001
GMM + NB	Random Baseline	≤ 0.0001
GMM + NB	Positive Baseline	≤ 0.0001
GMM	Random Baseline	≤ 0.0001
GMM	Positive Baseline	≤ 0.0001

C. Transductive vs Inductive Unsupervised Learning

Both the transductive and inductive learning approaches were studied. For the inductive approach (which is our standard approach for all the experiments), the parameters for all three models were learnt using a data set containing 4288 points. The "test" set containing the remaining 390 points was classified under these fixed parameters.

In contrast, under the transductive approach, the parameters for all three models were learnt using a data set containing all the 4678 data points, which included the test data points.

Table V shows the results of these experiments. It can be seen that all models perform equivalently under both approaches. The transductive approach performs about the same for the GMM and the NB models whereas it performs slightly worse for the mixture model.

TABLE V
F-MEASURES FOR TRANSDUCTIVE VS INDUCTIVE UNSUPERVISED LEARNING APPROACHES FOR ALL UNSUPERVISED MODELS

	Transductive	Inductive
GMM	**0.6787**	0.6765
GMM + NB	0.5890	**0.5941**
NB	**0.4056**	0.3284

D. Supervised vs Unsupervised Learning

The labeled data was also used for supervised learning using Support Vector Machines with cross validation across five stratified folds on a data set of 390 points. The results obtained are compared with the best results from each model under unsupervised learning on a test set of 390 points in Table VI. Interestingly, the SVM performs about the same as the best GMM performance (obtained by learning parameters on data set of 2618 points). It performs better than GMM + NB, and much better than the NB model.

TABLE VI
COMPARISON OF UNSUPERVISED LEARNING WITH GMM, GMM+NB AND NB AND SUPERVISED LEARNING WITH SVM

	F-Measure
SVM	**0.696**
GMM	0.695
GMM + NB	0.602
NB	0.406
Random Baseline	0.284
Positive Baseline	0.440

E. Features

Various experiments were performed in the unsupervised setting with different selections of features. Specifically, for each dialog feature, we performed one experiment where it was removed, but all other features were retained. It was found that the removal of some features, for example, the MaxPosts feature, decreases the performance of each system, by nearly 10%. On the other hand, removal of some features, for example the QuestionsFrequency feature (that measures how frequently a participant asks a question, taking time into account) also improved the performance of all three systems. For all our experiments reported up to now, we have removed the features QuestionsFreq and %Questions from our systems, as the removal of these gave the best system performance.

In Table VII below, we tabulate the f-measures of the GMM in a transductive learning setting over 4678 data points, when each of the fifteen features was removed, one at a time. It can be seen that removing MaxPosts brought about a significant decrease in f-measure while removing %Questions positively affects the system. All comparisons are against the f-measure of the system when none of the features are removed.

TABLE VII
PERFORMANCE OF GMM ON VARIOUS DIALOG FEATURES

Feature Removed	F-Measure of GMM
MaxPosts	0.595
Consecutive	0.611
LongestPost	0.611
%Posts	0.622
Initiation	0.629
Alternation	0.634
MaxQuestions	0.648
Repetition	0.655
UnansweredPosts	0.656
Termination	0.658
PointJoined	0.658
LongestBranch	0.653
Inquisitiveness	0.663
QuestionsFreq	0.666
%Questions	0.679
None	0.661
%Questions + QuestionsFreq	**0.678**

We also experimented with bag of words features added to our best model. The hypothesis behind this was the content used by the participants could be an indicator of whether or not the participant is in pursuit of power. However this turned out to be false. As a proof of concept, we tried to incorporate bag

of words on a small data set containing 67 data points.Table VIII shows the performance of GMM with and without the bag of words features.

A possible explanation could be that the words used by most of the participants in a given discussion are the same, and hence cannot be an indicator of who is in pursuit of power. We continue to investigate this and will use these features on the full size data set in future experiments.

TABLE VIII
PERFORMANCE OF GMM WITH AND WITHOUT BAG OF WORDS(BOW) FEATURES

	GMM with BOW	GMM without BOW
Precision	0.3283	0.6666
Recall	1.0000	0.7272
Accuracy	0.3283	0.7910
F-Measure	0.4943	0.6956

VII. CONCLUSION AND FUTURE WORK

We have studied the importance of dialog structure in determining power relations in written online discussions. Identification of pursuit of power in a setting where the data is largely unlabeled turns out to be a hard task and simplistic models perform badly. The performance improves with more sophisticated models. Our best model achieves an F-Measure of 69.5%, which we observe to be at par with a supervised model (trained on a very small annotated data set), and close to the inter-annotator agreement. Lexical features turn out not to contribute to the prediction.

There is a lot of scope of future work. We intend to investigate semi-supervised models, as well as completely supervised models. More sophisticated lexical and dialog features will be employed. NLP-based features like part-of-speech tags will be incorporated to aid learning. We are currently annotating more data to allow for a deeper study of supervised models, and we will also test the models presented in this paper on entirely unseen test data. Finally, we will extend our corpora to include other genres so that our models can be genre-independent.

ACKNOWLEDGMENTS

We would like to thank Michael Collins for suggesting the use of the EM algorithm, and two anonymous reviewers for comments that have improved this paper. We would also like to thank all of our colleagues on the SCIL project for providing an exciting intellectual environment in which to develop the ideas presented in this paper. This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

REFERENCES

[1] R. F. Bales, Strodtbeck, M. F. L., T. M., and M. Roseborough, "Channels of communication in small groups." *American Sociological Review*, pp. 16(4), 461–468, 1951.

[2] K. R. Scherer, "Voice and speech correlates of perceived social influence in simulated juries," in *H. Giles and R. St Clair (Eds), Language and social psychology*. Oxford: Blackwell, 1979, pp. 88–120.

[3] M. Brook and S. H. Ng, "Language and social influence in small conversational groups." *Journal of Language and Social Psychology*, pp. 5(3), 201–210, 1986.

[4] S. H. Ng, D. Bell, and M. Brooke, "Gaining turns and achieving high in influence ranking in small conversational groups." *British Journal of Social Psychology*, pp. 32, 265–275, 1993.

[5] S. H. Ng, M. Brooke, , and M. Dunne, "Interruption and in influence in discussion groups." *Journal of Language and Social Psychology*, pp. 14(4),369–381, 1995.

[6] S. A. Reid and S. H. Ng, "Conversation as a resource for in influence: evidence for prototypical arguments and social identification processes," *European Journal of Social Psychology*, pp. 30, 83–100, 2000.

[7] R. F. Bales, *Personality and interpersonal behaviour*. New York: Holt, Reinhart, and Winston, 1970.

[8] J. Diesner and K. M. Carley, "Exploration of communication networks from the enron email corpus," in *In Proc. of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*, 2005, pp. 21–23.

[9] J. Shetty and J. Adibi, "Discovering important nodes through graph entropy the case of enron email database," in *Proceedings of the 3rd international workshop on Link discovery*, ser. LinkKDD '05. New York, NY, USA: ACM, 2005, pp. 74–81. [Online]. Available: http://doi.acm.org/10.1145/1134271.1134282

[10] G. Creamer, R. Rowe, S. Hershkop, and S. J. Stolfo, "Segmentation and automated social hierarchy detection through email network analysis," in *Advances in Web Mining and Web Usage Analysis*, H. Zhang, M. Spiliopoulou, B. Mobasher, C. L. Giles, A. Mccallum, O. Nasraoui, J. Srivastava, and J. Yen, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 40–58. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-00528-2_3

[11] G. M. S. Namata, Jr., L. Getoor, and C. P. Diehl, "Inferring organizational titles in online communication," in *Proceedings of the 2006 conference on Statistical network analysis*, ser. ICML'06. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 179–181. [Online]. Available: http://dl.acm.org/citation.cfm?id=1768341.1768359

[12] P. Bramsen, M. Escobar-Molano, A. Patel, and R. Alonso, "Extracting social power relationships from natural language." in *ACL*. The Association for Computer Linguistics, 2011, pp. 773–782. [Online]. Available: http://dblp.uni-trier.de/db/conf/acl/acl2011.html#BramsenEPA11

[13] T. Strzalkowski, G. A. Broadwell, J. Stromer-Galley, S. Shaikh, S. Taylor, and N. Webb, "Modeling socio-cultural phenomena in discourse," in *Proceedings of the 23rd International Conference on COLING 2010*. Beijing, China: Coling 2010 Organizing Committee, August 2010. [Online]. Available: http://www.aclweb.org/anthology/C10-1117