Analyzing Conflict Narratives to Predict Settlements in EBay Feedback Dispute Resolution

Xiaoxi Xu, David Smith, Thomas Murray and Beverly Park Woolf

Abstract— **We explore the possibility of predicting settlements in online disputes by performing text-analysis on conflict narratives from disputant parties. The experiment domain is eBay Motor vehicles, in which disputants try to resolve complaints, possibly working with online human mediators. The conflict discourse is analyzed based on the divergence of topic distributions in a generative model extending Latent Dirichlet Allocation (LDA) to include role information. A set of distance schemes and a heuristic are designed for various negotiation scenarios to predict settlements. We analyze the quality of discovered topics in terms of topic coherence and evaluate settlement classification and prediction power for settlements on unseen data. Experimental results show that this unsupervised model outperforms a state-of-the-art supervised learner on precision, recall, and F-measure. The performance of a supervised learner with derived features from this model outperforms that using bag-of-features in terms of precision and efficiency.**

I. Introduction

This research focuses on the ability to predict whether two online disputants will reach a settlement based on analysis of their conflict discourse. Automating the process of prediction in online disputes is challenging, in part, because it requires understanding of the discourse in negotiation. We developed a latent variable topic model for modeling negotiation and prediction of whether a settlement will result between the two participants. The model has multiple levels hierarchy to represent cases and back-and forth exchanges within each case. Moreover, the model represents both topics of disputes and topic usage by each type of disputant. Ultimately, we hope to design an automated dispute resolution process in which the model can identify interests and positions of disputants and assess their priorities from their negotiations. The present model is based on the assumption that if topics used by disputant parties are aligned, it is likely that a settlement can be reached. Thus we measure the divergence of topic distributions to make predictions about the possibility of a settlement.

This model is tested in the domain of eBay Motors vehicles feedback. Through the gracious generosity of collaborators, including eBay and Net Neutrals [1], we acquired over 4,000 online exchanges among two eBay participants involved in sales of automobiles and primarily directed at removing negative feedback, see Table 1. Experiments with this data show that the new dispute model outperforms a state-of-the-art supervised learner on precision, recall, and F-measure. Recall is important for this task because the goal removing feedback is to remove unwarranted feedback. A mistakenly removed feedback can always be added back on eBay by users, but a delayed unfair feedback will not only mislead other peoples buying decisions, it can also ruin ones reputation and economic future.

This research makes two contributions: development of a generative model for online dispute discourse and analysis of a set of distance schemes and a heuristic to analyze conflict narratives and predict possible agreement. The organization of the paper is as follows. In Section II, we introduce the concept of online dispute resolution and the experimental domain. In Section III we describe the generative model and its Gibbs sampler. Section IV introduces the experimental setup followed by experiment results in Section V. We discuss related work in Section VI and conclude with future plans in Section VII.

II. EBay Feedback Disputes

People doing business at online auction markets (e.g., eBay) are inevitably anxious about their transactions. Buyers and sellers usually engage in one-shot deals meaning that they have no prior relationship before the transaction and do not anticipate any future commercial relationship [1]. "Relationshipless" disputes reduce the trust between two parties which is the root of their anxiety. In order to solve this public anxiety problem, eBay puts in place a reputation system for buyers and sellers to build trust, that is, the feedback mechanism. The use of feedback rating and comments is a way for buyers and sellers to judge the conduct of the other party for any transaction. Feedback is visible to all users and therefore would influence on sellers' or buyers' future business. Although acquiring a positive feedback is important, avoiding a negative one requires exercising more care. This is because if sellers ignore the negative feedback, they run risks of harming future online life.

A. Dispute Process

NetNeutrals is an Online Dispute Resolution (ODR) program that manages disputes or disagreements online. The company has been contracted by eBay to review Motors feedback disputes. Nearly all the disputes are about negative feedback placed on a sellers website by the buyer. Neutrals are trained, independent professionals with automotive service experience. As an online dispute resolution program, NetNeutral offers eBay users two types of voluntary service to resolve customer disagreements. Direct Negotiation is a

Xiaoxi Xu, David Smith, Thomas Murray and Beverly Park Woolf are with the Department of Computer Science, University of Massachusetts, Amherst, MA 01003, USA (phone: 413-545-3444; email: {xiaoxi, dasmith, murray, bev}@cs.umass.edu).

[1]http://www.juripax.com and http://www.netneutrals.com/

free dispute resolution process in which two disputant parties work together to come to a resolution on their own without the help of a third party. The independent Feedback Review (IFR) is a dispute resolution process that costs $100 where a third party, a human mediator, determines whether a rating should be descored. The human mediator evaluates evidence provided by buyers and sellers and offers comments based on eBays guidelines, including did the member demonstrate a good faith effort to complete the transaction? was the feedback submitted in a reasonable timeframe? is the transaction-related information factually inaccurate? did the member make an attempt to extract excessive value from the other party?

III. A Generative Model for ODR

This research reduces the online dispute into a binary classification problem and presents a language model to predict settlements of disputes based on disputants narratives. Such an automated process should be advantageous over that of a human reviewer in that it would be more consistent in the manner of judgment, more impartial, efficient, and cost-saving. In future work we hope to enhance the model so that it also recognizes participants interests and positions, assesses priorities from their negotiations, provides interventions at proper time, and computes resolutions that may provide each side with more than they themselves might be able to negotiate [1].

To model the negotiation process among disputants and predict case resolutions, we propose a disputant negotiation model (DNM) that extends LDA [2] to include role information. The model predicts dispute resolutions based on evaluating the divergence of disputants' topic distributions. The new DNM model does not not have a label node that represents case resolutions, since we are exploring how to represent the divergence of topic distributions from the perspective of a generative process, which is a challenging yet unexplored research problem itself and label information about case resolutions is not necessarily available in the real world.

A. Disputant Negotiation Model (DNM)

The graphical representation of DNM is shown in Figure 1. In DNM, the outermost plate denotes a dispute case or session. Each session contains a number of exchanges among disputants. DNM assumes the following generative process for our dispute corpus:

1. For every topic ϕ out of K, draw a word distribution $\phi_k \backsim$ Dirichlet (β).

2. For each disputant r, draw a topic proportion $\theta_r \backsim$ Dirichlet (α).

3. For each exchange m in each case d,

(1) Observe the disputant that generates the exchange.

(2) For each word,

(a) Draw $Z_{d,m,n} \backsim$ Multinomial ($\theta_{r_{d,m}}$).

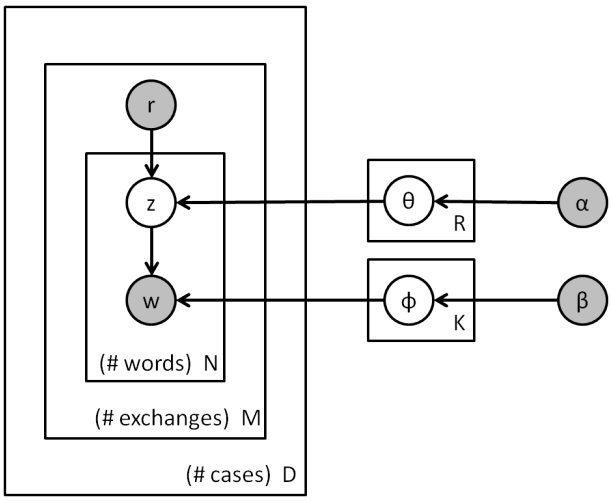

Fig. 1. Disputant Negotiation Model

(b) Draw $W_{d,m,n} \backsim$ Multinomial ($\phi_{z_{d,m,n}}$).

B. Gibbs Sampling for DNM

We use collapsed Gibbs sampling [3] to estimate the posterior distribution of hidden variable z given the input variables $\mathbf{w}$, and $\mathbf{r}$, and model parameters, α and β.

$$P(\theta, \phi, z|\mathbf{w}, \mathbf{r}, \alpha, \beta) = \frac{P(\theta, \phi, z, \mathbf{w}, \mathbf{r}|\alpha, \beta)}{P(\mathbf{w}, \mathbf{r}|\alpha, \beta)}$$

Note that we use symmetric Dirichlet priors α, β, in this work, and it is easy to adapt to use asymmetric Dirchlet priors in our model.

Using Gibbs sampling, we construct a Markov chain that converges to the posterior distribution on z and then use the results to infer θ and ϕ. The transition between successive states of the Markov chain is achieved from random sampling z from its distribution conditioned on all other variables, summing out θ and ϕ. By derivation, we get:

$$P(z_i|\mathbf{z_{-i}}, w, r) \propto \frac{N_{k|r} + \alpha}{N_r + K\alpha} \cdot \frac{N_{w|k} + \beta}{N_k + V\beta}$$

where the subscript z_{-i} denotes all topic assignments excluding the ith word. $N_{k|r}$ is the number of times that topic k is assigned to disputant r, excluding the current instance, and $N_{w|k}$ is the number of times that word w is assigned to topic k, excluding the current instance.

After the Gibbs sampling process, the model parameters in DNM can be obtained as follows:

$$\phi_{w|k} = \frac{N_{w|k} + \beta}{N_k + V\beta}$$

$$\theta_{k|r} = \frac{N_{k|r} + \alpha}{N_r + K\alpha}$$

where $\phi_{w|k}$ is the probability of using word w in topic k, and $\theta_{k|r}$ is the probability of using topic k by disputant r.

TABLE I

PROPERTIES OF THE DATA SET

of 327 Cases; 3982 Exchanges
792 Disputants; Average of 12 posts/case
eBay offers over six million goods and services for sale every day and assumes little or no responsibility for the transactions. When problems arise, members write negative feedback about the other party. eBay handles *40-60 million online disputes* per year. Bad feedback primarily hurts sellers. We examined some of the 20% of the online disputes that require human facilitators.

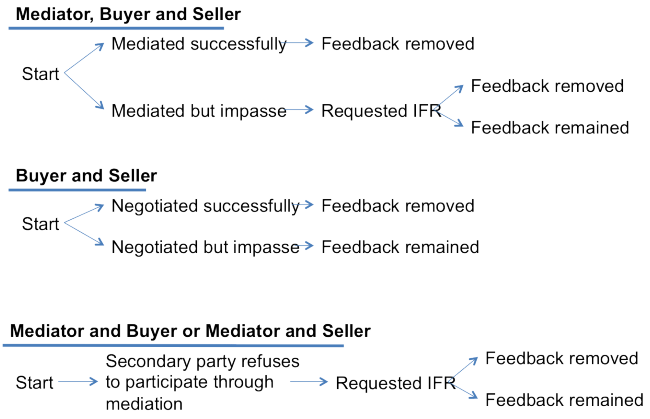

Fig. 2. An illustration of negotiation processes for various scenarios. All participants are human, including mediator (M), buyer (B), seller (S), and independent feedback review (IFR).

IV. EXPERIMENTAL SETUP

In this section, we describe the eBay Motors feedback dispute data set and how we devised distance schemes to measure topic distributions under various scenarios.

A. Data Set

The eBay Motors data set is a collection of discourses of 2-3 people in conversation around removing negative feedback from 2005 to 2008. Table I summarizes the properties of this data set. Each of the 327 cases falls into one of the following 4 scenarios.

- **Scenario 1: Mediator, Buyer and Seller**
- **Scenario 2: Buyer and Seller (no mediator)**
- **Scenario 3: Mediator and Buyer (Seller did not participate)**
- **Scenario 4: Mediator and Seller (Buyer did not participate)**

The negotiation process associated with each scenario is shown in Figure 2. We further provide data statistics for various scenarios in Table II. The cases that include Mediator, Buyer and Seller represents 42.20% of all the cases, Buyer and Seller (no mediator) represent 17.43% of the cases, Mediator and Buyer represent 0.61% and Mediator and Seller represent 39.76% of the cases. Furthermore, 64.53% of the cases in this data are successfully settled, while 35.47% of the cases remain unsettled.

TABLE II

DATA STATISTICS WITH VARIOUS SCENARIOS (MEDIATOR (M), BUYER (B), AND SELLER (S))

Scenarios	Feedback Removed	Feedback Remained
M, B, S	88	50
B, S	12	45
M, B	1	1
M, S	110	20
Total	211	116

B. Distance Schemes for Various Scenarios

The idea of using the divergence of topic distributions through text analysis to predict a resolution to a dispute is based on the following assumption: *Lower divergence correlates with increased possibility of a resolution (which means feedback removal in the case of eBay disputes).*

Note that an IFR may be requested to evaluate the situation when disputants reach an impasse, and then a mediator will inform disputant parties of the outcome. This means that the content of discourses from mediators has the information of dispute outcomes, which will provide supervisory information for the model. We thus do not use topic distributions from mediators for the settlement prediction task.

We now provide three distance schemes (DS) and one heuristic for the four scenarios provided in the previous section.

DS1 for Mediator, Buyer and Seller

D_1 = MIN (x, y)

where x = Div(Buyer's topic distribution, Seller's topic distribution), y = Div(Mean(Buyer's topic distribution, Seller's topic distribution), guideline's topic distribution), and Div is a divergence metric that will be introduced later.

For scenario 1 (Mediator, Buyer and Seller), the case resolution can be either mediated successfully or mediated but remain at impasse. We develop two distance measures corresponding to these two situations. The distance used for predicting settlement will take the minimum. For the cases that are mediated successfully, only the divergence of the buyer's topic distribution is compared against the seller's topic distribution. For the cases that are mediated but result in an impasse, the average of topic distributions from the two disputant parties is used and compared with the topic distribution of the eBay feedback guidelines.

DS2 for Buyer and Seller

D_2 = Div(Buyer's topic distribution, Seller's topic distribution)

In scenario 2 (Buyer and Seller), negotiations always occur between disputants, regardless of the outcome. Therefore, we only need to evaluate the divergence between buyer's topic distribution and seller's topic distribution.

DS3 for Mediator and Buyer or Mediator and Seller

D_3 = Div(Buyer's or Seller's topic distribution, guideline's topic distribution)

We also describe a heuristic:

Heuristic for Mediator and Buyer or Mediator and Seller

# of exchanges (posts) in a case

As explained before, the data from scenarios (Mediator and Buyer) and (Mediator and Seller) have missing information. To deal with the missing data issue, we designed a heuristic in addition to a distance measure that is not reliable when used alone. The distance measure evaluates the divergence between buyer's or seller's topic distribution and the topic distribution of the eBay feedback guidelines. The heuristic was developed based on an assumption that uses the structure of the negotiation process (i.e., the number of interactions/posts): *More interactions lead to settlement (or feedback removed in eBay disputes).* Note that the heuristic of using post numbers for prediction is only used for scenarios ((Mediator and Buyer) and (Mediator and Seller).

We used two different methods to measure distributional similarity: symmetric Kullback Leibler divergence [4] and Jensen-Shannon divergence [5]. Assume that P and Q are two topic distributions.

The symmetric Kullback Leibler divergence is given by:

$$SKLD(P||Q) = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2}$$

where $D_{KL} = \sum_i P(i) log \frac{P(i)}{Q(i)}$.

The Jensen-Shannon divergence based on Kullback Leibler divergence is given by:

$$JSD(P||Q) = \frac{D_{KL}(P||M) + D_{KL}(Q||M)}{2}$$

where $M = \frac{P+Q}{2}$.

We preprocessed the data by filtering standard English stopwords and tokens less than two characters. We used unigram features and the Porter stemmer [1]. After data preprocessing, we had 134,184 words with vocabulary size 3194. It is not surprising that we have a small size of vocabulary given that the dispute discourse is from one domain. We experimented different configurations of the number of topics and found that three topics provided a good overview of the contents of the corpus. The Dirichlet priors alpha was set to 16, beta to 0.1; the Gibbs sampler was run with 1000 burn-in iterations and 1000 sampling iterations.

V. Results

We performed three sets of experiments to evaluate the proposed model. In the first experiment, we evaluated the topics discovered by DNM, in the second we assessed the performance of DNM on the task of settlement classification and in the third we tested the predictive power of DNM on unseen data. The results below use a single sample from the Gibbs sampler.

A. Topic Discovery and Quality Evaluation

Figure 3 illustrates the three topics learned by the DNM model for the eBay dispute corpus. The topics were extracted from a single sample at the 2000th iteration of the Gibbs sampler. Each topic is illustrated with the top 10 words most likely to be generated conditioned on the topic. The

[1]http://tartarus.org/martin/PorterStemmer/

Topic 0 *Transaction*	
WORD	**PROB.**
feedback	0.0729
post	0.0426
guidelin	0.0355
rate	0.0337
review	0.0303
withdraw	0.0291
case	0.0260
meet	0.0221
transact	0.0208
ebai	0.0189

Topic 1 *Subject Matter*	
WORD	**PROB.**
car	0.0331
vehicl	0.0210
seller	0.0152
buyer	0.0150
state	0.0103
time	0.0093
purchas	0.0083
said	0.0081
ebai	0.0080
item	0.0075

Topic 2 *Mediation*	
WORD	**PROB.**
feedback	0.0251
thank	0.0194
want	0.0162
mediat	0.0160
go	0.0156
pleas	0.0151
know	0.0144
neg	0.0140
ask	0.0136
work	0.0133

Fig. 3. General topics as discovered by DNM in the eBay dialogues and the top 10 words related to those topics.

TABLE III
COHERENCE OF LEARNED TOPICS USING THE 5 MOST SALIENT WORDS

Topics	Scores	5 Most Salient Words
Topic 0	-59.4	feedback, post, guidelin, rate, review
Topic 1	-58.0	car, vehicl, seller, buyer, state
Topic 2	-64.2	feedback, thank, want, mediat, go

first topic is mostly related to *transaction* (e.g., feedback, post, review); the second topic is related to the *subject matter* (e.g., car, seller, purchase); and the third topic is related to *mediation* (e.g., mediate, thank, want).In a closer examination, we found that 30% of the text was categorized as *transaction*, 43% as *subject matter*, and 27% as *mediation*.

1) Topic Coherence: Perplexity[6] is often used for evaluating model performance on unseen data. But practically, we are interested in whether learned topics are coherent, that is, whether words in a topic are semantically related to any other words in the same topic. In this work, we used the topic coherence metric [7] to evaluate the quality of learned topics. The assumption of topic coherence is that pairs of words belonging to a single topic will cooccur within a single document, whereas those belonging to different topics will not. In other words, more words will cooccur in coherent topics; few words will cooccur in random topics.

The topic coherence metric is defined as:

$$TC(k; W^{(k)}) = \sum_{m=2}^{M} \sum_{i=1}^{m-1} log \frac{D(w_m^{(k)}, w_i^{(k)}) + 1}{D(w_i^{(k)})}$$

where $D(w)$ is the document frequency of word w and $D(w, w\prime)$ is the co-document frequency of word v and $v\prime$, and $W^{(k)}$=$(w_1^{(k)}, \ldots, w_M^{(k)})$ is a list of the M most probable words in topic k. A smoothing count of 1 is included to avoid taking the logarithm of zero. The coherence scores of learned topics using the 5 most salient words are shown in Table III, and those using the 10 most salient words are shown in Table IV. Numbers closer to zero indicate higher coherence. As can be seen from Table III, the learned topics are highly coherent.

B. Settlement Classification

In this section, we present the results of settlement classification by our unsupervised model DNM and also compare

TABLE IV

COHERENCE OF LEARNED TOPICS USING THE 10 MOST SALIENT WORDS

Topics	Scores	10 Most Salient Words
Topic 0	-212.1	feedback, post, guidelin, rate, review, withdraw, case, meet, transact, parti
Topic 1	-242.2	car, vehicl, seller, buyer, state, time, purchas, said, ebai, item
Topic 2	-240.6	feedback, thank, want, mediat, go, pleas, know, neg, ask, work

its performance with Support Vector Machine (SVM) [8], a state-of-the-art supervised learner for text classification. The classification performance is evaluated quantitatively in terms of Accuracy (% of correct predictions on resolved cases), Precision (% correct of cases that were settled), Recall (% labeled as settled that were predicted to be settled), and F-measure (the harmonic mean of precision and recall).

As explained earlier, reputation is a precious commodity on eBay. If an automated system such as DNM can achieve high precision and recall then unfair feedback that negatively impacts users can be efficiently removed.

We experimented with two divergence metrics to measure the divergence of topic distributions and found that the following thresholds work best in the Motors domain: (1) if the symmetric Kullback Leibler divergence (SKLD) of the topic distribution is below 0.1, the case is considered settled; (2) if the Jensen-shannon divergence (JSD) of the topic distributions is below 0.02, the case is considered settled; (3) if the number of exchanges (interactions) in a case is more than 5, the case is considered as settled [1].

Figure 4 shows the classification performance of DNM by using (1) divergence metrics alone (left panel), and (2) divergence metric together with the number of posts (right panel). Please note that the heuristic was only applied to scenarios 3 and 4. The upper left table shows the performance of using SKLD; the upper right table shows that of using SKLD with post numbers. It is expected that the classification performance is boosted by using the heuristic, because it accounts for the effect of applying distance measure on data with missing information. Similarly, the performance of JSD together with the heuristic is better than using JSD alone. When comparing the performance of the use of different divergence metrics (i.e., the upper left table and the lower left table), we found that JSD achieves higher accuracy, recall and F-measure, while SKLD achieves higher precision. Of the four experimental settings, SKLD with post number has greater success for settlement prediction in terms of accuracy and precision, while JSD with post number performs better on recall and F-measure, as highlighted in Figure 4. We also found that, in all of the experimental settings, the proposed model had consistent higher recalls on scenarios that involve a mediator (except for scenario 3 that has only one case) than that without. This is because working with a mediator, disputants tend to have focused discussions on the same

[1]The heuristic of using post numbers for prediction is only used for scenarios (Mediator and Buyer) and (Mediator and Seller).

Symmetric Kullback Leibler Divergence (SKLD)

True Positive = 88	False Positive = 50
False Negative = 123	True Negative = 66

Accuracy = 47.10%
Precision = 63.77%
Recall = 41.71%
F-measure = 50.43%

Scenarios	Precision	Recall
M, B, S	51.13%	69.23%
B, S	83.33%	31.25%
M, B	0%	0%
M, S	30.00%	80.49%

SKLD + Postnum

True Positive = 166	False Positive = 62
False Negative = 45	True Negative = 54

Accuracy = 67.28%
Precision = 72.81%
Recall = 78.67%
F-measure = 75.63%

Scenarios	Precision	Recall
M, B, S	51.13%	69.23%
B, S	83.33%	31.25%
M, B	100%	100%
M, S	100%	85.27%

Jensen-Shannon Divergence (JSD)

True Positive = 117	False Positive = 69
False Negative = 94	True Negative = 47

Accuracy = 50.15%
Precision = 62.90%
Recall = 55.45%
F-measure = 58.94%

Scenarios	Precision	Recall
M, B, S	71.59%	67.74%
B, S	83.33%	25%
M, B	0%	0%
M, S	40%	83.02%

JSD + Postnum

True Positive = 184	False Positive = 81
False Negative = 27	True Negative = 35

Accuracy = 67.00%
Precision = 69.43%
Recall = 87.20%
F-measure = 77.31%

Scenarios	Precision	Recall
M, B, S	71.59%	67.74%
B, S	83.33%	25%
M, B	100%	100%
M, S	100%	84.02%

Fig. 4. Performance of DNM for settlement classification by using (1) divergence metrics alone (left panel), and (2) divergence metric combined with post number (right panel)

topics. Therefore, the DNM model more likely correctly predicts the "settled" cases (i.e.,feedback removal in the case of eBay disputes), resulting in high recall.

We also compared DNM with SVM, a state-of-the-art supervised learner for text classification. The idea of SVM is that input vectors are non-linearly mapped to a high-diminutional feature space where a linear decision surface can be constructed [9]. The Motors data set is unbalanced because the size of the positive labeled data is twice as large as that of the negative labeled data. In order to effectively run SVM, we split the data into 2 subsets and preprocessed the data in a similar way as we did for DNM. The performance of SVM that uses unigram features (term occurrence), linear kernel, with 5-fold and 10-fold cross validations are shown in Figure 5. We also tested other kernels, but found that using non-linear kernels did not improve the performance. This is because the number of features is very large in the Motors data, mapping data to a higher dimensional space would not be necessary and not useful for creating a separating decision boundary. The average performance of applying SVM to the two subsets is presented on the bottom row in Figure 5. DNM outperforms SVM in terms of precision and F-measure, when using SKLD with post numbers. It outperforms SVM in terms of precision, recall, and F-measure, when using JSD with post numbers.

C. Settlement Prediction on Unseen Data

To evaluate the predictive power of DNM, we also carried out experiments to train a classifier (SVM) using derived features from DNM, which we refer to DNM+SVM. Specifically, the derived features include the symmetric Kullback Leibler divergence learned from DNM for each case and a binary feature representing whether the number of posts in a case exceeds the confidence threshold we set. As can be seen from Figure 6, DNM+SVM achieves comparable

SVM (5-Fold CV)	
True Positive = 80	False Positive = 42
False Negative = 26	True Negative = 74
Accuracy = 69.30%	
Precision = 65.57%	
Recall = 75.47%	
F-measure = 70.17%	

SVM (10-Fold CV)	
True Positive = 86	False Positive = 42
False Negative = 20	True Negative = 74
Accuracy = 72.11%	
Precision = 67.19%	
Recall = 81.13%	
F-measure = 73.50%	

(a) Applying SVM to balanced subset 1 (Pos: 106, Neg: 116)

SVM (5-Fold CV)	
True Positive = 82	False Positive = 44
False Negative = 23	True Negative = 72
Accuracy = 69.71%	
Precision = 65.08%	
Recall = 78.10%	
F-measure= 71.00%	

SVM (10-Fold CV)	
True Positive = 83	False Positive = 45
False Negative = 22	True Negative = 71
Accuracy = 69.68%	
Precision = 64.84%	
Recall = 79.05%	
F-measure = 71.24%	

(b) Applying SVM to balanced subset 2 (Pos: 105, Neg: 116)

SVM (5-Fold CV)
Accuracy = 69.51%
Precision = 65.33%
Recall = 78.58%
F-measure= 71.35%

SVM (10-Fold CV)
Accuracy = 70.90%
Precision = 66.02%
Recall = 80.09%
F-measure = 72.38%

(c) Average performance of SVM on 2 balanced subsets

Fig. 5. Performance of SVM for settlement classification

SVM (10-Fold CV)	
True Positive = 175	False Positive = 65
False Negative = 36	True Negative = 51
Accuracy = 67.06%	
Precision = 72.92%	
Recall = 82.94%	
F-measure= 77.61%	

DNM+SVM (10-Fold CV)	
True Positive = 159	False Positive = 52
False Negative = 52	True Negative = 64
***Accuracy** = 67.89%*	
***Precision** = 75.36%*	
***Recall** = 75.36%*	
***F-measure** = 75.36%*	

Fig. 6. Performance of SVM (left panel) and DNM + SVM (SVM using derived features from DNM, right panel) for settlement prediction on unseen data

performance to SVM on predicting settlement. We feel that DNM+SVM is quite promising because using derived features is much efficient than using bag-of-word features.

VI. Related Work

Previous research has tested the idea that topic divergence distributions can predict whether participants will reach a settlement, as well as the assumption that low divergence in topic distributions will lead to agreement [10]. For example, in a speed dating classification task, the divergence of topic distributions of dialogues from a dating pair is used to predict men and women's decisions about whether they want to meet again. However, no prior research has attempted to analyze dispute dialogue from a corpus with topic models and we are the first to develop a topic model for modeling online negotiation and predicting settlements in dispute resolution.

The author-topic model (ATM) [11] is quite similar to the developed DNM model and both models represent the content of disputes. The difference is that DNM has more levels than does ATM to model the nested structure of cases and exchanges within each case. Additionally, DNM models the topic usage of different types of disputants (i.e., buyers and sellers) rather than that of individual disputant and the role of each disputant is observable at the exchange level (and therefore at the case level). Prior research to extend LDA by incorporating a supervision node in the model, such as [12], [13], and [14], are related to this work. DNM does not have a supervision node partly because we are still exploring how to represent the divergence of topic distributions of disputants from the perspective of a generative process, and partly because the label information is not always necessarily observable in the real world and partially observable data is a good thing for generative models.

Research in Online Dispute Resolution (ODR) uses technology to facilitate the resolution of disputes and has been employed to handle disputes from consumer-to-consumer issues and marital separation to workplace grievance and interstate conflicts [1]. ODR shows great advantages over traditional litigation and has the potential to provide greater flexibility, substantial cost-savings, and higher efficiency. In e-commerce, ODR has gained wide popularity by reducing travel time and providing mediators for those who cannot afford them. Moreover fully automated online services, such as Cybersettle [2], SettlementOnline [3], and ClickNsettle [4], own huge markets for disputes and have had huge commercial success for disputes that are solely over the amount of monetary settlements. Such systems use simple procedures to compare demands with offers and determine settlements if demands and settlements are within a range [1]. For example, Cybersettle alone claims to have handled more than 60,000 transactions during the period between 1998 and 2003, facilitating settlements for more than $350 million [5]. In contrast, other online dispute ventures that are not automated appear to have had more limited success [15]. As Internet usage continues to expand, e-commerce is growing and the number of disputes from e-commerce will also rise. It has become increasingly necessary to design automatic mechanisms for resolving online disputes beyond monetary settlements.

EBay, the largest online auction site, has 83 million users in the U.S. alone in 2009 and millions of sales opening and closing everyday. The eBay reputation system supports sellers and buyers to acquire mutual trust by supporting feedback, ratings and comments, to be left by buyers and sellers for each other. If disagreements about feedback are not settled automatically by disputant parties, then a trained professional may guide participants to reach solutions. Once a fully automated process for reaching settlements has been developed, it will potentially improve on human mediation as it would be wholly impartial, highly efficient, and involve a low cost.

VII. Conclusions and Future Work

In this paper, we proposed a generative model to predict whether a settlement would be reached by disputants in the eBay Motor vehicle corpus. The topics discovered by Dispute Negotiation Model (DNM) were related to transaction, subject mattes, and mediation. The coherence score of each topic using the 5 most salient words showed that the learned topics were highly coherent. In a quantitative

[1] http://en.wikipedia.org/wiki/Online_dispute_resolution
[2] http://www.cybersettle.com
[3] http://www.settlementonline.com
[4] http://www.clicknsettle.com
[5] http://www.cybersettle.com/about/factsheet.asp

evaluation of settlement classification, DNM outperformed SVM on precision, recall, and F-measure. In testing the predictive power of the DNM by using derived features from DNM to train a classifier, DNM + SVM achieved comparable performance to SVM with higher efficiency.

These results are encouraging. The next steps for predicting whether an agreement will be reached by disputants is to design a pair of supervised models for settlement prediction. The first model would have a resolution label upstream pointing to a node representing the topic divergence of disputant parties. This model would be based on the assumption that disputant parties come to a negotiation with a predetermined approach about whether they are willing to agree to the settlement, in this case to withdraw a negative rating. We are also interested in the reverse problem that has the resolution label downstream. In that case, we assume that disputant parties have an approach about topics to be discussed and will wait to see if negotiation can help resolve their conflict.

The ultimate research goal is to design an automated dispute resolution process in the ideal situation where the model can identify the interests and positions of disputants and assess their priorities from their negotiations. In future work we will explore such a model using derived psychological, lexical, and cohesion-based features from Coh-Metrix [16] and LIWC [17] methods. The hope is that using bag of derived features would yield performance gains over the bag-of-word features used in this study.

VIII. Acknowledgments

This research was supported in part by grants from the National Science Foundation (0968536). Any opinions, findings, conclusions, or recommendations expressed in the paper are those of the authors and do not necessarily reflect those of the funding agencies.

References

[1] Ethan Katsh, Janet Rifkin, and Alan Gaitenby. E-commerce, e-disputes, and e-dispute resolution: Learning from ebay and other online communities. *Ohio State Journal of Dispute Resolution.*, 2000.

[2] David M. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 2011.

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[4] S Kullback and R.A Leibler. On information and sufficiency. In *Annals of Mathematical Statistics 22 (1)*, pages 79–86, 1951.

[5] J Lin. Divergence measures based on the shannon entropy. In *IEEE Transactions on Information Theory 37 (1)*, pages 145–151, 1991.

[6] Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *the 26th International Conference on Machine Learning*, 2009.

[7] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *EMNLP*, 2011.

[8] Sergios Theodoridis and Konstantinos Koutroumbas. Pattern recognition. *IEEE Transactions on Neural Networks*, 19(2):376, 2008.

[9] C. Cortes and V Vapnik. Support-vector networks. In *Machine Learning 20(3)*, pages 273–297, 1995.

[10] Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of the North American Association of Computational Linguistics (NAACL 2009)*, 2009.

[11] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, New York, NY, USA, 2004. ACM Press.

[12] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

[13] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, 2009.

[14] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification, 2008.

[15] JW Goodman. The pros and cons of online dispute resolution: An assessment of cyber-mediation websites. *Duke L. and Tech. Rev.*, 2003.

[16] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. Coh-metrix: analysis of text on cohesion and language. *Behavior research methods instruments computers a journal of the Psychonomic Society Inc*, 36(2):193–202, 2004.

[17] James W. Pennebaker, Roger J. Booth, and Martha E. Francis. *Linguistic inquiry and word count (LIWC): A computerized text analysis program*. Erlbaum Publishers, 2001.

IX. Appendix

A. Gibbs Sampler Derivation

Goal: Find posterior distribution over latent variables given the observed variables (omitting hyperparameters).

$$P(\theta, \phi, z|w, r) = \frac{P(\theta, \phi, z, w, r)}{P(w, r)}$$

Graphical model gives us:

$$\begin{aligned} P(\theta, \phi, z, w, r) &= P(\theta)P(\phi)P(z|\theta, r)P(w|z, \phi)P(r) \\ &= \prod_r Dir(\theta_r; \alpha) \prod_k Dir(\phi_k; \beta) \prod_n \theta_{z_n|r_n} \\ &\quad \prod_n \phi_{w_n|z_n} \prod_m P(r_m) \end{aligned}$$

We use collapsed Gibbs sampling to integrate out ϕ and θ, and just sample z. Sample z for $P(z|w, r)$.

$$P(z|w, r) = \frac{P(z, w, r)}{P(w, r)}$$

Numerator:

$$
\begin{aligned}
P(z,w,r) &= \int \mathrm{d}\theta \int \mathrm{d}\phi P(\theta,\phi,z,w,r) \\
&= \int \mathrm{d}\theta \int \mathrm{d}\phi \prod_r Dir(\theta_r;\alpha) \prod_k Dir(\phi_k;\beta) \\
&\quad \prod_n \theta_{z_n|r_n} \prod_n \phi_{w_n|z_n} \prod_m P(r_m) \\
&= \int \mathrm{d}\theta \int \mathrm{d}\phi \prod_r Dir(\theta_r;\alpha) \prod_k Dir(\phi_k;\beta) \\
&\quad \prod_r \prod_k \theta_{k|r}^{N_{k|r}} \prod_k \prod_w \phi_{w|k}^{N_{w|k}} \prod_m P(r_m) \\
&= \int \mathrm{d}\theta \prod_r \left[Dir(\theta_r;\alpha) \prod_k \theta_{k|r}^{N_{k|r}} \right] \\
&\quad \int \mathrm{d}\phi \prod_k \left[Dir(\phi_k;\beta) \prod_w \phi_{w|k}^{N_{w|k}} \right] \prod_m P(r_m) \\
&= A \times B \times \prod_m P(r_m)
\end{aligned}
$$

where $A = \int \mathrm{d}\theta \prod_r \left[Dir(\theta_r;\alpha) \prod_k \theta_{k|r}^{N_{k|r}} \right]$, $B = \int \mathrm{d}\phi \prod_k \left[Dir(\phi_k;\beta) \prod_w \phi_{w|k}^{N_{w|k}} \right]$. Now we are going to expand term A and B.

Note that

$$
\begin{aligned}
&\int \mathrm{d}\theta Dir(\theta_r;\{N_{k|r}+\alpha\}) = 1 \\
\Longrightarrow &\int \mathrm{d}\theta \frac{\Gamma(N_r+\sum_k \alpha)}{\prod_k \Gamma(N_{k|r}+\alpha)} \prod_k \theta_{k|r}^{N_{k|r}+\alpha-1} = 1 \\
\Longrightarrow &\frac{\Gamma(N_r+\sum_k \alpha)}{\prod_k \Gamma(N_{k|r}+\alpha)} \prod_k \int \theta_{k|r}^{N_{k|r}+\alpha-1} \mathrm{d}\theta = 1 \\
\Longrightarrow &\prod_k \int \theta_{k|r}^{N_{k|r}+\alpha-1} \mathrm{d}\theta = \frac{\prod_k \Gamma(N_{k|r}+\alpha)}{\Gamma(N_r+\sum_k \alpha)}
\end{aligned}
$$

$$
\begin{aligned}
A &= \int \mathrm{d}\theta \prod_r \left[Dir(\theta_r;\alpha) \prod_k \theta_{k|r}^{N_{k|r}} \right] \\
&= \prod_r \int \mathrm{d}\theta Dir(\theta_r;\alpha) \prod_k \theta_{k|r}^{N_{k|r}} \\
&= \prod_r \int \mathrm{d}\theta \frac{\Gamma(\sum_k \alpha)}{\prod_k \Gamma(\alpha)} \prod_k \theta_{k|r}^{\alpha-1} \prod_k \theta_{k|r}^{N_{k|r}} \\
&= \prod_r \frac{\Gamma(\sum_k \alpha)}{\prod_k \Gamma(\alpha)} \prod_k \int \theta_{k|r}^{N_{k|r}+\alpha-1} \mathrm{d}\theta \\
&= \prod_r \frac{\Gamma(\sum_k \alpha)}{\prod_k \Gamma(\alpha)} \frac{\prod_k \Gamma(N_{k|r}+\alpha)}{\Gamma(N_r+\sum_k \alpha)}
\end{aligned}
$$

Similarly,

$$
B = \prod_k \frac{\Gamma(\sum_w \beta)}{\prod_w \Gamma(\beta)} \frac{\prod_w \Gamma(N_{w|k}+\beta)}{\Gamma(N_k+\sum_w \beta)}
$$

Denominator $P(w,r) = \sum_z P(z,w,r)$ requires Gibbs sampling. We use the full conditional $P(z_i|z_{-i},w,r)$ to simulate $P(z|w,r)$.

$$
\begin{aligned}
&P(z_i|z_{-i},w,r) \\
&\quad = \frac{P(w,z,r)}{P(w,z_{-i},r)} \\
&\quad \propto \frac{P(w|z,r)P(z,r)}{P(w_{-i}|z_{-i},r)P(z_{-i},r)} \\
&\quad = \frac{P(w,z,r)}{P(w_{-i},z_{-i},r)}
\end{aligned}
$$

We know that $P(w,z,r) = A \times B \times \prod_m P(r_m)$. $P(w_{-i},z_{-i},r)$ is the same except for $N_{k|r}-1, N_r-1, N_{w|k}-1, N_k-1$. Because $x\Gamma(x) = \Gamma(x+1)$, $\frac{\Gamma(x+1)}{\Gamma(x)} = x$. After canceling terms, we have

$$
\frac{\frac{\Gamma(N_{k|r}+\alpha)}{\Gamma(N_r+\sum_k \alpha)} \cdot \frac{\Gamma(N_{w|k}+\beta)}{\Gamma(N_k+\sum_w \beta)}}{\frac{\Gamma(N_{k|r}+\alpha-1)}{\Gamma(N_r+\sum_k \alpha-1)} \cdot \frac{\Gamma(N_{w|k}+\beta-1)}{\Gamma(N_k+\sum_w \beta-1)}} = \frac{N_{k|r}+\alpha}{N_r+k\alpha} \cdot \frac{N_{w|k}+\beta}{N_k+V\beta}
$$

The posterior on θ and ϕ using the fact that the Dirichlet is conjugate to the multinomial.

$$
\phi|z,w,\beta \sim Dir(N_k+\beta)
$$

$$
\theta|z,w,\alpha \sim Dir(N_r+\alpha)
$$

Evaluating the posterior mean of θ and ϕ

$$
\mathrm{E}[\phi_{w|k}|z,w,\beta] = \frac{N_{w|k}+\beta}{N_k+V\beta}
$$

$$
\mathrm{E}[\theta_{k|r}|z,w,\alpha] = \frac{N_{k|r}+\alpha}{N_r+k\alpha}
$$