

A Study of Sentiment and Gender Influence on Negotiation Outcome in Electronic Negotiations

Vivi Nastase¹ and Jelber Sayyad Shirabad²

¹ EML Research
Heidelberg, Germany
`nastase@eml-research.de`

² School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada
`jsayyad@site.uottawa.ca`

Abstract. We study whether sentiments we can detect in messages exchanged in the course of electronic negotiations conducted with the Inspire e-negotiation system are predictive of the negotiation outcome. We use lists of sentiment tagged words to assign sentiment scores to messages exchanged. We then combine such scores with information on negotiator gender obtained from Inspire pre-questionnaires. We study whether sentiment scores for individual messages and global negotiation scores are predictive of the outcome.

Key words: sentiment analysis, classification, electronic negotiations

1 Introduction

Electronic negotiation systems, such as Inspire [3] make possible large scale data analysis, through the recorded message and offer exchanges in numerous on-line negotiations. Data collected thus far has been extensively studied from several perspectives – of the messages [12], offers [17] [8], personal negotiator information from pre-negotiation questionnaires [4]. In this paper we propose another way of looking at messages exchanged in the course of negotiations – from the perspective of sentiment analysis.

Sentiment analysis has enjoyed an increased interest within the Natural Language Processing (NLP) community in the past years, also motivated by commercial interest in detecting people’s opinions about products and services through analysis of messages and blogs. The research is diverse, and takes the form of text classification problem [2], [18], [16], [9], [5], subjectivity identification [18], [19], “empathic” graphic user interfaces [6] or corpus analysis for detecting what makes people happy or sad in their everyday lives [7]. Interest in this domain has led to the development of several lexical resources for sentiment detection, which we will use to detect affect in negotiation messages.

WordNet Affect provides an added level of annotation on top of WordNet 1.6, with emotion, mood and other sentiment information [15]. We use the fine grained affect

information, which added sentiment annotation to WordNet synsets (synonym sets) from the following set: { *joy, surprise, anger, fear, disgust, sadness* }.

The Linguistic Inquiry and Word Count application (LIWC) focuses on analysis of text and computing statistics along 82 language dimensions [10]. Part of these dimensions are dedicated to sentiment analysis, and a dictionary of tagged words provides the necessary information.

The Balanced Affective Word List project (ANEW), originated in 1994, developed a list of words annotated along three psychological dimensions: arousal, dominance and valence [1]. Apart from median values, the data contains standard deviation for each dimension, and separate ratings for males and females, as well as overall scores.

Textual messages exchanged in electronic negotiations have been mined to detect strategies and tactics expressed through specific linguistics patterns – combinations of personal pronouns, modal, volition and mental verbs, and words expressing temporal dimensions. Such patterns can be indicative of negotiators’ attitudes and emotional involvement in the negotiation process [13].

In the experiments presented here we look within negotiation messages for words that carry specific sentiment information, as indicated by the various lexical resources with sentiment annotations. We will use statistics that capture sentiment evidence in messages to predict the outcome of negotiations. Because psychological studies, such as those that led to the creation of the ANEW word list, have shown that men and women have different perceptions of certain sentiment-related word dimensions, we also incorporate gender information in our representation. We will see that this leads to increased predictive accuracy.

The paper follows with a description of the Inspire data, and the datasets we generate for the experiments presented here. We then discuss the experimental set-up, and the results obtained. We compare the results obtained with previous research on Inspire data.

2 Data

The Inspire data we work with consists of recorded negotiations from 1997 to 2004. It contains 8968 messages exchanged in 3063 negotiations. The negotiations have three possible outcomes: successful, failed, or one sided. For our analysis we focus on negotiations that have recorded some interaction between negotiators, and the outcome is successful or failed. This leaves us with 8416 messages from 2006 negotiations.

The WordNet Affect provides us with WordNet 1.6 synonym sets annotated with sentiments from six categories: joy, fear, surprise, anger, disgust and sadness. After collecting the words in these synsets and removing duplicates, we obtain the counts presented in Table 1.

LIWC provides a dictionary with words annotated along several dimensions. We choose positive and negative labels, and extract the two corresponding lists of positive and negative words.

The ANEW data we work with contains 1034 words annotated along three dimensions - arousal, dominance and valence. Unfortunately, we could not exploit this resource, because there was no overlap with Inspire messages data.

resource	class	count
WordNet Affect	anger	240
	disgust	48
	fear	134
	joy	364
	sadness	187
	surprise	70
	total	1043
LIWC	negative	345
	positive	265
	total	610

Table 1: Word-sentiment counts from WordNet Affect and LIWC

In order to obtain message sentiment scores, we tokenize every message, and count the number of words appearing in the 6 WordNet Affect and 2 LIWC lists. In Table 2 we provide additional statistics about the negotiations data used in this paper. The table shows absolute and average scores per message, negotiator and negotiation. Each negotiation has two negotiators. The difference between 2*2006 and 3376 (the number of negotiators in Table 2) comes from the fact that not all negotiators have exchanged messages with their negotiation partners. The negotiators have the option to send offers only, so the negotiation may unfold with no textual message exchange from one or both of the participants.

sentiment	count	avg/message	avg/negotiator	avg/negotiation
anger	159	0.0189	0.0471	0.0793
disgust	15	0.0017	0.0044	0.0075
fear	226	0.0268	0.067	0.1127
joy	6176	0.7338	1.83	3.0788
sadness	1339	0.1591	0.3966	0.6675
surprise	1269	0.1508	0.3758	0.6326
negative	2147	0.2551	0.636	1.0703
positive	17597	2.091	5.2127	8.7727
messages length	493743	58.67	146.25	246.13
	tokens	messages	nr. of negotiators	nr. of negotiations
totals	493743	8416	3376	2006

Table 2: Sentiment counts and message length information

The features that describe each message, or all messages sent by one negotiator when we work with aggregated data, are the following:

gender: the gender of the message sender;

partner gender: the gender of the recipient of the message;

message length: the length of the message (or total length of messages sent by a negotiator);

negative: the count of tokens appearing in the list of negative words from LIWC;

positive: the count of tokens appearing in the list of positive words from LIWC;
anger: the count of tokens appearing in the WordNet Affect list of anger-expressing words;
disgust: the count of tokens appearing in the WordNet Affect list of disgust-expressing words;
fear: the count of tokens appearing in the WordNet Affect list of fear-expressing words;
joy: the count of tokens appearing in the WordNet Affect list of joy-expressing words;
sadness: the count of tokens appearing in the WordNet Affect list of sadness-expressing words;
surprise: the count of tokens appearing in the WordNet Affect list of surprise-expressing words;
outcome – the target attribute – the outcome of the negotiation (successful or failed).

We extract gender information from Inspire pre-negotiation questionnaires. The data is filled in by the negotiators, and they are allowed to skip questions, or indeed the entire questionnaire. Because of this, gender information is not complete, and apart from the *m* (male) and *f* (female) genders, we also have the *not_def* (not defined) value. Table 3 shows the gender distribution in our data. We add gender and partner gender information to each message, and also create an additional gender feature with three possible value: *not_def* if either of the two negotiation partners has as gender value “not_def”, *same* when the negotiators have the same gender and *diff* when they don’t.

We refer to the data set that has one record (or example) per message *Messages_single* data.

gender	message		negotiator		negotiation			
	count	average	count	average	buyer	average	seller	average
female	3192	37.93%	1308	38.74%	680	33.9%	628	31.3%
male	4452	52.9%	1752	51.9%	878	43.77%	874	43.57%
not_def	772	9.17%	316	9.36%	448	22.33%	504	25.13%
totals	nr. of messages 8416		nr. of negotiators 3376		nr. of buyers 2006		nr. of sellers 2006	

Table 3: Gender information distribution

After generating individual message sentiment scores, we create overall scores per negotiator in a negotiation by adding the scores for messages sent by this particular negotiator. This process generates a new data set, called *Messages_negotiator*, which has one example per negotiator and captures aggregated features for that negotiator within one negotiation.

We also built a data set with one example per negotiation. Each negotiation is represented by concatenating the aggregated features generated for each of the two negotiators participating in the negotiation. This data set is called *Negotiation_negotiators*. When one of the negotiators in a negotiation has not sent any messages, the corresponding part of the example vector is filled with default value of 0 for sentiments.

The last data set we created also represents each negotiation as one example. However, an example in this case contains aggregated sentiment scores for all messages ex-

changed in the negotiation, independent of which negotiator has produced them. This data set is called *Negotiations*.

In all four data sets discussed above each example includes the gender information for the two participating negotiators. Table 4 presents an overview of the number of instances and the distribution of classes in each of the datasets we created.

dataset	instances	class distribution	
		successful	failed
Messages_single	8416	5701 (67.74%)	2715 (32.26%)
Messages_negotiator	3376	2220 (65.75%)	1156 (34.25%)
Negotiation_negotiators	2006	1321 (65.85%)	685 (34.15%)
Negotiations	2006	1321 (65.85%)	685 (34.15%)

Table 4: Datasets statistics

When the data set contains vectors with information for each negotiator in a negotiation, these attributes appear twice, once for each negotiator.

3 Experiments

We apply machine learning algorithms to the datasets produced, to test whether sentiment scores and gender information can predict accurately negotiation outcome. We use Weka 3.4.10 [20], in particular J48, an implementation of decision tree-based learning. J48 ran with the default configuration and we performed 10 fold cross-validation.

We present in Table 5 a selection of the most interesting results obtained. We include precision, recall, F1-score¹ and accuracy results for successful and failed negotiations. Best results, if better than the majority class baseline, are highlighted.

With respect to gender features, we have noticed that the gender of the sender, or the gender of the receiver, alone are not helpful in the classification by themselves – the classifier built is the trivial one, for all examples the predicted class is the majority class. But when both of them are present, the learner builds a non-trivial model, with good predictive performance for single messages and overall negotiation scores. The feature built by combining the two gender features is, surprisingly, not useful, leading to poorer performance than when separate gender features are used.

The best results in terms of accuracy and precision of predicting successful negotiations were obtained by using single messages. . It is a surprising result, because, while there may be messages that are indicative of the outcome, we don't expect most of them to be. Or rather, from all messages in one negotiation, we expect a few of them to be good outcome predictors. For the case of failed messages, where we have obtained high precision but low recall, this phenomenon may be even more pronounced.

¹ We use the common formula:

$$\text{F1-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

features	successful				failed		
	accuracy	precision	recall	f1-score	precision	recall	f1-score
Messages_single							
baseline	67.74	67.7	100	80.8	0	0	0
all features	69.64	69.6	97.9	81.4	70.1	10.3	17.9
no gender features	67.74	67.7	100	80.8	0	0	0
only gender features	69.99	69.6	98.8	81.7	79.3	9.5	16.9
Messages_negotiator							
baseline	65.75	65.8	100	79.3	0	0	0
all features	67.62	67.8	96.7	79.7	64.9	11.9	20
no gender features	65.43	66	97.9	78.8	43.4	3.1	5.8
only gender features	68.45	67.9	98.8	80.5	81.4	10.2	18.1
Negotiation_negotiators							
baseline	65.85	65.9	100	79.4	0	0	0
all features	62.41	66.8	85.2	74.9	39.3	18.4	25
no gender features	61.76	65.8	87.4	75.1	33.7	12.4	18.1
only gender features	68.19	67.7	99	80.4	82.2	8.8	15.8
Negotiations							
baseline	65.85	65.9	100	79.4	0	0	0
all features	66.69	68.5	91.5	78.4	53.5	18.8	27.9
no gender features	64.7	66.1	95.5	78.1	38.1	5.4	9.5
only gender features	67.94	67.5	98.8	80.2	78.4	8.5	15.3

Table 5: Results for predicting success or failure in negotiations using affect and gender information

We may increase performance in predicting a failed outcome if for one negotiation, we make a prediction based only on few, possible one, message. We will try to isolate these “special” messages by selecting messages from certain time points in the course of a negotiation, to find whether there is a point in the negotiation when the outcome becomes clear. On the other hand, results on aggregate representations generate better F-scores for the failed class. This seems to contradict the previous hypothesis. But in the case of the aggregate representations, the precision is much lower. These two types of evidence – from single message and aggregate representations – point to two phenomena in our data: (i) there is one (or very few) message at which the negotiation takes a turn for the worse, (ii) there is an accumulating negative feeling throughout the negotiation that leads to its ultimate failure. We plan to explore both these phenomena in future work.

An analysis of the model which has generated the highest accuracy, revealed that when the gender of both negotiators is known, the outcome of the negotiation is predicted as *successful*. When only the gender of one of the negotiators is known the outcome is predicted as *failed*. Finally, when the gender of both negotiators is unknown the predicted outcome is *successful*. In this particular experimental set-up there is no difference in predicted outcome when negotiator genders are known (and they are either the same or different). This leads to a potentially interesting observation about negotiators’ attitudes, revealed by the existence or lack of gender information. Those negotiators that have completed the questionnaires may be more interested and more

motivated to perform the task successfully, whereas those who haven't – for which we do not have gender information, so the value is *not_def* – may be more disinterested. When two (possibly) disinterested negotiators work together, they may decide to wrap up the negotiation quickly, to finish the task sooner. This hypothesis can be tested by performing an analysis of frequency and length of messages exchanged, and of numerical information exchanged through offers.

When sentiment information is added to the mix (all features row for the *Messages_single* data set), the decision trees show a combination of gender and affect features. But, when message length is removed from the dataset, the only feature apart from gender that appears in the decision tree is *sentiment_negative*, and it comes into play when one negotiator has no gender information, the other one is male. Interestingly, the model built in this situation shows that if there is at most one negative word exchanged (*sentiment_negative* value is less or equal to 1), the negotiation outcome is *failed* in 81.81% of the cases when the male is the receiver of the message (63 out of 77 situations), and 85.56% when the male is the sender (83 out of 97 situations). When some negative sentiment exists (*sentiment_negative* value is greater than 1), the negotiation outcome is always *successful* (9 out of 9 cases). This may indicate that when under pressure, a disinterested negotiator can step up to the plate.

4 Comparison with previous results

There is quite an extensive analysis of textual Inspire data (see for example [12],[13]). Because of the particular nature of textual data, these studies focus on negotiations on the same topic – sale/purchase of bicycle parts (which is the most frequent topic in Inspire negotiations to date). From the various machine learning algorithms used, the best results in these experiments were also obtained with a decision tree implementation (C5.0). The best accuracy reported was 74.5%, while the simple baseline² was 55%.

Kersten and Zhang [4] analyze a combination of personal information (extracted from pre- and post-negotiation questionnaires) and numerical data of offers recorded in negotiation transcripts. At the time of this study the Inspire data consisted of 1525 negotiations. The best reported accuracy was 75.33%, also obtained with a decision tree algorithm.

Nastase [8] analyzed the concession curve, which shows how offer utility values change in the course of a negotiation. The best reported classification accuracy is 76.44%, while the simple baseline was 74.9%, a higher imbalance than the data used in textual analysis.

Overall accuracy does not tell the full story. We take a look at precision and recall scores for predicting negotiation outcome. Table 6 includes precision, recall and F1-score results for negotiation outcome classification from [8] and [12], and from some of the experiments presented in this paper.

We notice very high precision for predicting failed negotiations, with a close recall value, and comparable performance on the successful negotiations. This indicate that combining gender and affect features with other negotiation features – either textual or numeric – the prediction performance may improve.

² In the discussion that follows we always consider the simple baseline to be the accuracy of classification when an example is assigned the majority class.

method	successful			failed			accuracy
	precision	recall	f1-score	precision	recall	f1-score	
concession curve	78%	95%	85.6%	59%	20%	29.87%	76.44%
textual data <i>strategy words</i>	72.5%	87.6%	79.25%	–	–	–	74.5%
Messages_single <i>gender & partner gender</i>	69.6%	98.8%	81.7%	79.3%	9.5%	16.9%	69.99%
Messages_negotiator <i>gender & partner gender</i>	67.9%	98.8%	80.5%	81.4%	10.2%	18.1%	68.45%

Table 6: Detailed classification results

5 Conclusions

We have presented experiments that use gender and affect features to predict negotiation outcome. We have obtained results that indicate interesting combination possibilities between these features, and others extracted from Inspire messages and exchanged offers.

The experiments presented worked on raw Inspire data. Sokolova et al. [14] shows a thorough analysis of the language in the Inspire messages, which is more problematic to work with than corpora or documents, as it contains a large amount of spelling errors and shorthand expressions (such as *u* for *you*). In future work we plan to work with a cleaner version of this data, in which the spelling errors that can be confidently resolved will be corrected. We also plan to use syntactic analysis, to detect negations that modify affect words. The results obtained with raw Inspire data show interesting potential for exploration of gender and sentiment influence on the outcome of negotiations, and we plan to explore this further with cleaned and syntactically processed data.

Much of the research done on textual data can be attempted with this sentiment-based representation. For example, Sokolova and Szpakowicz [13] show that textual data from messages exchanged during the first half of a negotiation is as predictive of negotiation outcome as the complete message exchange. It would be interesting to see whether the same phenomenon occurs when considering only sentiment aspects of negotiation messages. We are also interested in performing a comparative analysis of sentiment-loaded words, to detect patterns in successful and failed negotiations.

Another interesting aspect is to separately analyze the linguistic behaviour of buyers and sellers. Sokolova et al. [11] shows that buyers and sellers have distinctive language usage patterns, which can be exploited to predict negotiation outcome with high accuracy. We are interested to make this a hypothesis for affect and gender data as well.

Finally, we would like to analyze in more detail the relation between negotiators that fill in questionnaires and those who don't. Statistics on the textual messages and the numeric offers exchanged may show that there is negotiation behaviour difference between these two categories of negotiators, and the way they approach the task.

References

- [1] Bradley, M. M. and P. Lang (1999): "Affective norms for English words (ANEW): Instruction manual and affective ratings", working paper Technical Report C-1,

- The Center for Research in Psychophysiology, University of Florida.
- [2] Hatzivassiloglou, V. and K. McKeown (1997): “Predicting the semantic orientation of adjectives”, in: *Proceedings of the 35th ACL/8th EACL*, Madrid, Spain, pp. 174–181.
 - [3] Kersten, G. and S. Noronha (1999): “WWW-based negotiation support: Design, implementation and use”, *Decision Support Systems*, 25(2), pp. 135–154.
 - [4] Kersten, G. and G. Zhang (2003): “Mining Inspire data for the determinants of successful Internet negotiations”, *Central European Journal of Operational Research*, 11(3), pp. 297–316.
 - [5] Liu, H., H. Lieberman, and T. Selker (2003): “A model of textual affect sensing using real-world knowledge”, in: *Proceedings of the 7th International Conference on Intelligent User Interfaces*, pp. 125–132.
 - [6] Ma, C., H. Prendinger, and M. Ishizuka (2005): “Emotion estimation and reasoning based on affective textual interaction”, in: *Proceedings of the First International Conference on Affective COmputing and Intelligent Interaction*, Beijing, China, pp. 622–628.
 - [7] Mihalcea, R. and H. Liu (2006): “A corpus-based approach to finding happiness”, in: *AAI Spring Symposium on Computational Approaches to Weblogs*.
 - [8] Nastase, V. (2006): “Concession Curve Analysis for Inspire Negotiations”, *Group Decision and Negotiations*, 15(2), pp. 185–193.
 - [9] Pang, B. and L. Lee (2004): “A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts”, in: *Proceedings of the 42nd ACL*, Barcelona, Spain, pp. 271–278.
 - [10] Pennebaker, J. W., R. J. Booth, and M. E. Francis (2001): “Linguistic Inquiry and Word Count: LIWC”, [Computer software]. Mahwah, NJ: Erlbaum.
 - [11] Sokolova, M., V. Nastase, and S. Szpakowicz (2004): “Language in Electronic Negotiations: Patterns in Completed and Uncompleted Negotiations”, in: *Proceedings of 3rd International Conference on Natural Language Processing, ICON 2004*, Hyderabad, India, pp. 142–151.
 - [12] Sokolova, M. and S. Szpakowicz (2005): “Analysis and Classification of Strategies in Electronic Negotiations”, in: *Proceedings of the 18th Canadian Conference on Artificial Intelligence, CAI 2005*, Victoria, British Columbia, Canada, pp. 145–157.
 - [13] Sokolova, M. and S. Szpakowicz (2006): “Language patterns in the learning of strategies from Negotiation Texts”, in: *Proceedings of the 19th Canadian Conference on Artificial Intelligence, CAI 2006*, Quebec City, Quebec, Canada, pp. 288–299.
 - [14] Sokolova, M., S. Szpakowicz, and V. Nastase (2004): “Automatically building a lexicon from raw noisy data in a closed domain”, working paper INR 01/04, INTERNEG working papers.
 - [15] Strapparava, C. and A. Valitutti (2004): “WordNet-Affect: an affective extension of WordNet”, in: *Proceedings of the 4th International Conference on Language Resources and Evaluation - LREC 2004*, Lisbon, Portugal, pp. 1083–1086.
 - [16] Turney, P. and M. Littman (2003): “Measuring praise and criticism: inference of semantic orientation from association”, *ACM Transactions on Information Systems*, 21(4), pp. 315–346.

- [17] Vetschera, R. (2006): “Preference Structures of Negotiators and Negotiation Outcomes”, *Group Decision and Negotiations*, 15(2), pp. 111–125.
- [18] Wiebe, J. (2000): “Learning subjective adjectives from corpora”, in: *Proceedings of the 7th National Conference on Artificial Intelligence and the 12th Conference on Innovative Applications of Artificial Intelligence*, pp. 735–740.
- [19] Wilson, T., J. Wiebe, and R. Hwa (2004): “Just how mad are you? finding strong and weak opinion clauses”, in: *Proceedings of the 19th National Conference on Artificial Intelligence*, San Jose, CA, USA, pp. 761–769.
- [20] Witten, I. H. and E. Frank (2005): *Data Mining: Practical machine learning tools and techniques*, 2nd edition edition, Morgan Kaufmann, San Francisco.