Uncovering Hidden Sentiment in Meetings

Gabriel Murray

University of the Fraser Valley, Abbotsford, BC, Canada
gabriel.murray@ufv.ca
http://www.ufv.ca/cis/gabriel-murray/

Abstract. The sentiment expressed by a meeting participant in their face-to-face comments may differ from the sentiment contained in their private summary of the meeting. In this work, we investigate whether we can predict the sentiment score of a participant's private post-meeting summary, based on multi-modal features derived from the group interaction during the meeting. We describe several effective prediction models, all of which outperform a baseline that assumes the sentiment score of the summary will be the same as the sentiment score of the participant's comments during the meeting.

Keywords: sentiment detection, subjectivity, multi-modal interaction

1 Introduction

Being able to predict group members' positive or negative sentiment based on their interaction in a meeting could be valuable for improving group efficiency, productivity, and social cohesion. However, there are obstacles to being able to accurately predict the sentiment held by meeting participants. For example, a group member might refrain from making highly negative comments during the meeting even though they have negative opinions about items under discussion. Or a group member might make little vocal contribution during the meeting, despite having strong positive or negative opinions.

In this research, we study meeting data in which participants have been asked to write a short, private summary after each meeting. The summaries can also include any problems or issues that occurred during the meeting. The private summaries are not seen by the other participants. We show that we can predict the sentiment scores of these private summaries, based on multi-modal features from the meeting itself. These prediction models outperform a baseline in which it is assumed that the sentiment score for a participant's private summary will be the same as the sentiment score for their comments during the meeting.

The structure of the paper is as follows. In Section 2, we describe related work on sentiment detection in meetings and on meeting analysis in general. In Section 3 we describe our sentiment prediction system, including the dataset and features, sentiment scoring method, and the prediction models. The results are presented in Section 4 and we conclude in Section 5.

2 Related Work

Closely related work has aimed to detect meeting sentences containing positive or negative sentiment. Raaijmakers et al. [1] and Murray and Carenini [2] both use multi-modal features to classify whether dialogue acts segments (sentence-like units in meetings) contain positive or negative subjectivity.[1] Our work differs from theirs in two ways. First, we are predicting the sentiment score of post-meeting participant summaries rather than the sentiment of meeting sentences. Second, our score prediction is a regression, rather than classification, task.

Several recent books survey the more general field of sentiment analysis, including detection of opinions and emotions [3–5].

Much work has been done on studying multi-modal interaction in meetings more generally [6], including the use of machine learning models to learn about and improve group efficiency and productivity in meetings [7, 8]. There has also been a rich vein of research on modelling group interaction and small group dynamics, including phenomena such as dominance and influence [9–14]. Much of that work has focused on non-verbal cues, while we incorporate both verbal and non-verbal features in these experiments.

To our knowledge, this is the first work to use participant summaries to analyze sentiment amongst group members. The only other work we have seen that uses participant summaries is by Kim and Shah [15], who use self-reported summaries to assess whether a group has achieved "consensus of understanding."

3 Hidden Sentiment Prediction

The goal of our system is to predict the sentiment that will be contained in the private post-meeting summary written by a participant, based on the meeting and the participant's interaction in the meeting. The participant summaries must therefore be scored according to their sentiment. We rely on the sentiment lexicon supplied by Taboada et al. [16] as part of their SO-Cal sentiment detection system. The lexicon contains lists of sentiment-bearing adjectives, adverbs, nouns and verbs, each of which is associated with a positive or negative score. Positive scores range from 1 to 5, and negative scores range from -1 to -5.

Taboada et al., citing Boucher and Osgood [17], note that many texts seem to have a positive bias, with positive words being much more frequent than negative words. That is certainly the case with meeting transcripts, where negative sentences are relatively rare [18] and difficult to detect [2]. This may be due to participants refraining from stating negative opinions in face-to-face interactions, particularly in meetings where the participants do not know each other, as is the case in the corpus we describe below. This could also be due to the use of euphemisms, where mildly negative words are indicative of strong negative sentiment. Whatever the underlying cause for the imbalance, Taboada et al. assume that negative words carry more cognitive weight and they found

[1] The terms *subjectivity* and *sentiment* are very closely related, and we use the latter.

that increasing the sentiment weights of negative words by 50% improved their sentiment prediction performance in comparison with gold-standard sentiment labels. We carry out the same 50% adjustment of negative word scores in this work.

Having carried out the negative score adjustment just described, the sentiment score for a document is then the average sentiment score for all sentiment-bearing words in the document.

3.1 The AMI Meeting Corpus

The meeting data and associated participant summaries are from the AMI meeting corpus [19]. We use the scenario portion of the corpus, where participants are role-playing as members of a company designing a remote control. Each group consists of four members, assigned the roles of project manager, user interface designer, industrial design expert, and marketing expert. Each group goes through a series of four meetings, wherein they discuss different phases of design, finance, and production. After each meeting, the participants were asked to write individual summaries of what happened during the meeting, including any problems that occurred.

Below we show a sample of the types of comments participants make in these post-meeting summaries:

- "We have no feel for the strengths and weaknesses of the team and what our particular roles are for this project."
- "Lack of familarity with each other personally and socially as a team."
- "A lack of direction in the meetings."
- "I was not convinced myself that some of the trends were desirable to incorporate, and the group confirmed this."
- "Industrial Designer, Alima, who was originally frustrated because she could not find enough information, presented a very coherent explanation of how the remote works."
- "We decided to focus on fashion, usability, and simplicity in our design."

Figure 1 shows the distribution of sentiment scores for participant comments in meetings and for participant summaries. Each meeting is treated as four separate documents, each document consisting of a single participant's comments. One surprising finding is that when the meetings and summaries are scored in the manner described above, meetings tend to be more negative while the summaries tend to be more positive.

Figure 2 shows a scatterplot, where each point corresponds to the sentiment score of an individual participant's meeting comments and the sentiment score of their subsequent private summary. We can see that a participant's sentiment in the meeting is not always a good predictor of their sentiment in the corresponding summary. In many cases, they are relatively neutral in the meeting but positive in the summary, and in a few cases they are positive in the meeting but relatively negative in the summary.

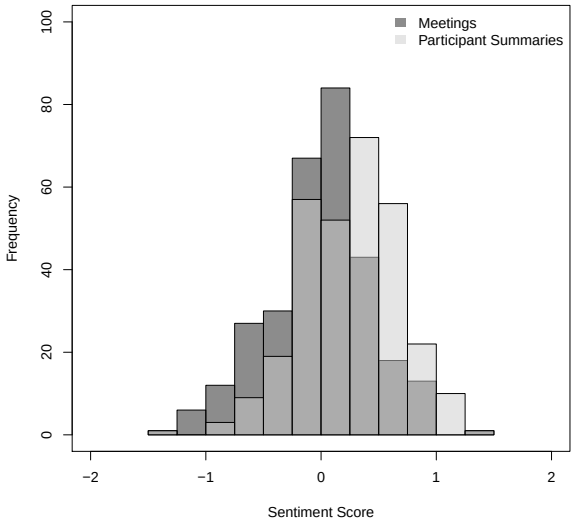

Fig. 1. Sentiment Distribution

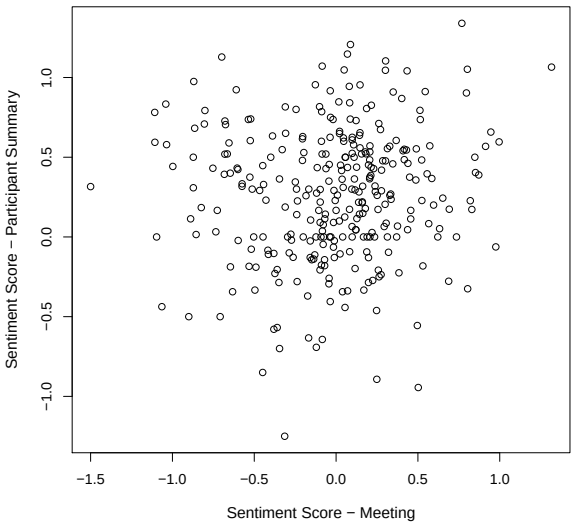

Fig. 2. Sentiment in Meetings vs. Summaries

3.2 Sentiment Features and Models

All of our prediction models use the same set of verbal and non-verbal features derived from the meetings. We group them into the following broad classes:

Sentiment Features

- **posWords,negWords**: Respectively, the number of positive and negative sentiment words used by the participant in the meeting.
- **totalPosSubjScore,totalNegSubjScore**: Respectively, the sum of the positive word scores and negative word scores used by the participant.
- **totalSubjWords,totalSubjScore**: The total number of subjective words used by the participant, and the sum of those word scores.

Activity Features

- **totalDacts**: The number of dialogue act segments by the participant in the meeting.
- **totalTime**: The total speaking time of the participant in the meeting.
- **totalWords**: The number of words uttered by the participant in the meeting.
- **totalFillPause**: The number of filled pauses (*uh,um*, etc.) by the participant.
- **first,last**: Respectively, these features indicate whether the participant was the first person to speak in the meeting or the last person to speak in the meeting.
- **rateOfSpeech**: The rate-of-speech of the participant, in words per second.

Meeting Features

- **meetA,meetB,meetC**: There are four meetings in the series, A-D. The position in the series is encoded using three binary features.
- **allmeetwords,allmeetdacts**: Respectively, the total number of words and dialogue acts in the meeting, across all participants.

Speaker Features

- **PM,UI,ME** There are four assigned roles in the meeting, encoded by three binary features.

We use three prediction models for this task. The first is a multiple linear regression. The second is a multi-layer neural network, with two hidden layers each containing two units, as shown in Figure 3. The third system is a random forest with 500 trees and seven variables tried at each split.

3.3 Experimental Setup

Each meeting yields four datapoints, one for each participant. However, not all AMI meetings contain participant summaries. We ultimately ended up with 302 datapoints. For the multiple regression and neural network predictions, we report results using 10-fold cross-validation. For the random forest regression, we report out-of-bag prediction results.

The evaluation metric used is mean-squared error (MSE).

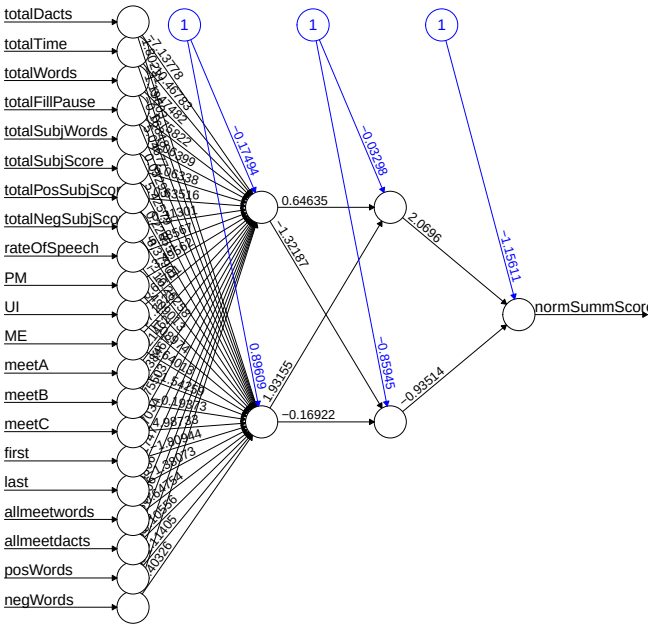

Fig. 3. Neural Network Structure

4 Results

The MSE scores are shown in Table 1. The best-performing predictions were using random forests and multiple regression, which were comparable to each other. The multi-layer neural network did not perform as well, and adding hidden layers and units only resulted in further degraded results. All systems performed better than a baseline prediction that assumes the summary score will be the same as the meeting score.

SYSTEM	MSE
Baseline (Score Same as Meeting)	0.416
Multi-Layer Neural Network	0.243
Multiple Regression	0.177
Random Forest	0.175

Table 1. MSE Scores

For analyzing the most useful features, we consider just the best-performing system, random forests. Figure 4 shows two measures of variable importance in the random forest regression. The "%IncMSE" plot shows the percentage that

the MSE increases when the variable is removed. "IncNodPurity" shows the increase in node purity when splitting on that variable. Many of the sentiment features from the meeting are useful predictors of the sentiment in the resultant summary. However, non-sentiment features that relate to the length of the meeting are also very good predictors. To highlight one feature, the number of filled pauses is a very useful indicator according to both metrics.

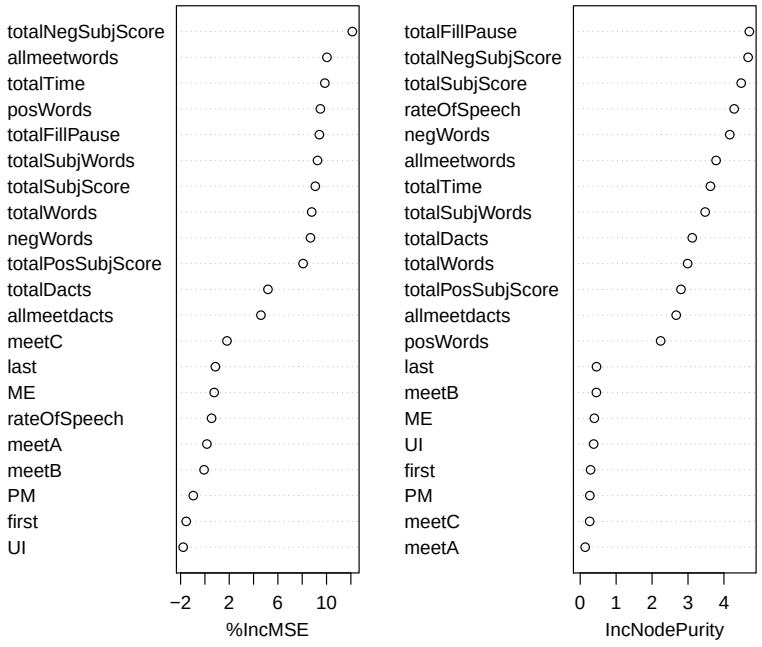

Fig. 4. Variable Importance

Despite the in-meeting sentiment features being amongst the most useful predictors, we can achieve very good performance using only the non-sentiment features for prediction. A random forest regression using the non-sentiment predictors yields only a slightly higher MSE of 0.18, compared with 0.175 for the full feature set.

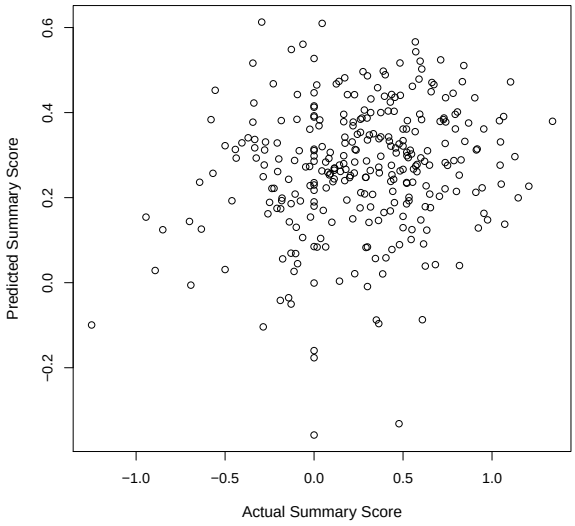

Fig. 5. Actual Summary Scores vs. Predictions

5 Conclusion

In this work, we investigated whether we can successfully predict the sentiment contained in a meeting participant's private summary, based on characteristics of the meeting and the participant's interaction in the meeting. Using a variety of verbal and non-verbal cues, we showed that three prediction models can outperform a baseline where the sentiment of the summary is predicted to be the same as in the meeting. Of the three prediction models, multiple regression and random forests performed the best.

There is still room for improvement in our sentiment predictions, as evidenced by Figure 5 showing the actual summary sentiment scores plotted against the predicted sentiment scores. In particular, there is a positive bias in the predictions, with the predicted scores generally being more positive than the actual scores.

An unexpected finding is that the participant summaries are not more negative than the participants' comments in the meeting. In fact, the summaries tend to be slightly more positive than the corresponding meeting comments. Our assumption that participant's true opinions would tend to be more negative than they indicated in the meeting was not supported by this data.

In future work, we plan to incorporate intensification, diminishment, and negation, which may be improve our sentiment modelling. We also plan to incorporate additional non-verbal features such as prosody and head gestures, in order to improve prediction performance.

References

1. Raaijmakers, S., Truong, K., Wilson, T.: Multimodal subjectivity analysis of multiparty conversation. In: Proc. of EMNLP 2008, Honolulu, HI, USA. (2008)
2. Murray, G., Carenini, G.: Subjectivity detection in spoken and written conversations. Natural Language Engineering **17**(03) (2011) 397–418
3. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Now Publishers (2008)
4. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012)
5. Liu, B.: Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge (2015)
6. Renals, S., Bourlard, H., Carletta, J., Popescu-Belis, A.: Multimodal Signal Processing: Human Interactions in Meetings. 1st edn. Cambridge University Press, New York, NY, USA (2012)
7. Murray, G.: Analyzing productivity shifts in meetings. In: Advances in Artificial Intelligence. Springer (2015) 141–154
8. Kim, B., Rudin, C.: Learning about meetings. Data Mining and Knowledge Discovery **28**(5-6) (2014) 1134–1157
9. Rienks, R., Zhang, D., Gatica-Perez, D., Post, W.: Detection and application of influence rankings in small group meetings. In: Proc. of ICMI 2006, Banff, Canada. (2006)
10. Pentland, A., Heibeck, T.: Honest signals. MIT press (2008)
11. Jayagopi, D., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations from non-verbal activity cues. IEEE Transactions on Audio, Speech and Language Processing **17**(3) (2009) 501–513
12. op den Akker, R., Gatica-Perez, D., Heylen, D.: Multi-modal analysis of small-group conversational dynamics. In Renals, S., Bourlard, H., Carletta, J., Popescu-Belis, A., eds.: Multimodal Signal Processing. Cambridge University Press, New York (June 2012) 155–169
13. Dong, W., Lepri, B., Pianesi, F., Pentland, A.: Modeling functional roles dynamics in small group interactions. IEEE Transactions on Multimedia **15**(1) (2013) 83–95
14. Frauendorfer, D., Mast, M.S., Sanchez-Cortes, D., Gatica-Perez, D.: Emergent power hierarchies and group performance. International Journal of Psychology (2014)
15. Kim, J.H., Shah, J.: Automatic prediction of consistency among team members' understanding of group decisions in meetings. In: Proc. of IEEE SMC. (2014) 3702–3708
16. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics **37**(2) (June 2011) 267–307
17. Boucher, J., Osgood, C.: The pollyanna hypothesis. Journal of Verbal Learning and Verbal Behavior **8**(1) (1969) 1–8
18. Wilson, T.: Annotating subjective content in meetings. In: Proc. of LREC. (2008) 2738–2745
19. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. In: Proc. of LREC 2006, Genoa, Italy. (2006) 181–190