# Predicting Team Performance from Communications in Simulated Search-and-Rescue

Extended Abstract

Ali Jalal-Kamali
University of Southern California
Los Angeles, USA
jalalkam@usc.edu

Nikolos M. Gurney
University of Southern California
Los Angeles, USA
gurney@ict.usc.edu

David V. Pynadath
Rice University
Houston, USA
pynadath@rice.edu

## ABSTRACT

Understanding how individual traits influence team performance is valuable, but these traits are not always directly observable. Prior research has inferred traits like trust from behavioral data. We analyze conversational data to identify team traits and their correlation with teaming outcomes. Using transcripts from a Minecraft-based search-and-rescue experiment, we apply topic modeling and clustering to uncover key interaction patterns. Our findings show that variations in teaming outcomes can be explained through these inferences, with different levels of predictive power derived from individual traits and team dynamics.

## KEYWORDS

Prediction; Team performance; Topic modeling; Clustering

**ACM Reference Format:**
Ali Jalal-Kamali, Nikolos M. Gurney, and David V. Pynadath. 2025. Predicting Team Performance from Communications in Simulated Search-and-Rescue: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

## 1 INTRODUCTION

Autonomous agents have begun leveraging artificial intelligence to improve human teamwork through automated assessment and assistance during task performance [8, 10, 12]. However, their usefulness is limited by an agent's ability to understand the people it wants to help. When interacting with a *team* of humans, the agent must model not only the multiple individuals but also their relationships and interactions. When multiple people work together, communication becomes integral to their behavior across domains [2, 7, 9]. This communication provides valuable information about team characteristics and performance [11]. However, the value depends on the agent's ability to understand this naturally occurring communication, which isn't oriented toward AI systems.

## 2 EXPERIMENTAL TESTBED DATA

We analyze data from Study 3 of DARPA's Artificial Social Intelligence for Successful Teams (ASIST) program [4]. The study used a Minecraft-based urban search and rescue task [1, 3, 5] where teams of three participants worked together. Team members had distinct roles: the *medic* treats victims, the *engineer* clears obstacles, and the *transporter* efficiently moves victims.

We use the following components in our study

- Communication transcripts: teams communicated via audio, which was transcribed. Our analysis uses only these transcripts, ignoring simulation logs and AI advisor data.
- Pre-trial team profiles: The dataset includes eight Background of Experience, Affect, and Resources Diagnostic (BEARD) variables such as anger, anxiety, etc. The BEARD variables measure the team characteristics before trials.
- Dynamic effectiveness Diagnostic (TED) measures: The ASIST testbed also contains variables that measure different aspects of team effectiveness throughout the trials, without any interactions with the team. The TED variables generate a dynamic stream of team processes measures.

## 3 METHODOLOGY AND RESULTS

Our analysis aims to identify teams needing intervention and determine how early to intervene based on the communication patterns.

### 3.1 BEARD Profiles

Linear regression of BEARD variables against performance revealed several significant relationships

- anger: showed a strong negative correlation.
- social perceptiveness: demonstrated a positive correlation.
- transporting skill: showed an unexpected negative correlation, though this may be due to overconfidence effects.

### 3.2 TED Measures

Once we had the impact of the pre-trial variables, we turned our focus to the variables that measure different aspects of team effectiveness throughout the trials, without any interactions with the team members. Since some of these measures inform other ones, e.g., process-effort-s informs process-effort-agg, to exclude all inter-dependencies of TED variables, we only included variables that are aggregates, time measures, and communication-based.

A linear regression of TED variables and scores indicates

- process-effort-agg: has a positive coefficient.
- comms-total-words: has a positive coefficient.
- process-skill-use-agg: has a negative coefficient.

### 3.3 Pre-Processing of Transcripts

Each transcript contains text for a complete experimental session with teams performing the task twice (two trials). We removed administrative text, split files into individual trials, and eliminated

*Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.* 

**Table 1: Most probable words for the 12 topics.**

| *Topic* | *Top Most Probable Words* |
|---|---|
| 1 | one, critical, see, right, go |
| 2 | meeting, one, critical, two, management |
| 3 | need, come, go, yeah, okay |
| 4 | patient, one, critical, engineer, patients |
| 5 | green, blue, hallway, victims, red |
| 6 | yeah, okay, oh, just, right |
| 7 | transporter, engineer, medic, victim, victims |
| 8 | go, one, okay, ahead, think |
| 9 | victim, room, victims, critical, see |
| 10 | engineer, one, just, get, right |
| 11 | critical, victim, engineer, victims, type |
| 12 | right, critical, just, like, going |

redundant trials, resulting in 222 unique trial transcripts. We created document-term matrices using standard preprocessing: lowercase conversion, punctuation/number removal, and stopword filtering.

### 3.4 Intra-team Communication Analysis

After pre-processing, we investigated the content of the conversations to inform our process about the team's performance. Topic modeling is an unsupervised method of extracting potential topics from the text, where such topics are representative of the main content of a document. But first we need to find the best number of topics.We applied Latent Dirichlet Allocation (LDA) topic modeling using textmineR [6]. For each topic count (2-20), we ran LDA 100 times per count, evaluating the average probabilistic coherences; based on which, we selected 12 as the topic count. Table 1 shows the most probable words per topic. While some topics share terms due to the common search-and-rescue context, others (like topics 2, 3, and 5) are distinct, revealing different communication patterns.

### 3.5 Categorization Abstraction

In order to have an abstraction over the categories that may be present among the trial conversations and to find potential subgroups that could indicate various performances, we can perform a clustering over the topic probability distributions for the trials using theta matrix as the variables. Using gap statistics, we determined 8 as the optimal number of clusters. K-means clustering revealed strong differentiation between first and second trials (Table 2), despite not using performance data.

To investigate how these clusters relate to performance, we examined the score distributions per cluster. Linear regression showed significant relationships between cluster assignment and performance (p=0.0008), indicating a strong relation between cluster assignment and trial outcomes. The analysis revealed

- Cluster 5: Lowest performance (coefficient: -196.30)
- Cluster 2: Second-lowest (coefficient: -147.753)
- Cluster 3: Third-lowest (coefficient: -115.095)

BEARD logistic regression for cluster 5 showed:

- Negative correlation with social perceptiveness
- Positive correlations with spatial ability and game skills

**Table 2: Trial-one vs. trial-two separation using clustering**

| *Cluster* | *Trial One* | *Trial Two* |
|---|---|---|
| 1 | 29% | **71%** |
| 2 | **84%** | 16% |
| 3 | **84%** | 16% |
| 4 | 14% | **86%** |
| 5 | 53% | 47% |
| 6 | 11% | **89%** |
| 7 | **91%** | 9% |
| 8 | 29% | **71%** |

Team Effectiveness Diagnostic (TED) variables revealed distinctive patterns across clusters

- Cluster 5 (lowest performing): high inaction rates, low process effort/coverage/triaging, minimal workload distribution, poor communication balance.
- Cluster 2: negative correlation with communication equity, moderate process coverage, uneven task distribution.
- Cluster 3: negative correlation with process workload, inconsistent team coordination, mixed communication patterns.

These patterns suggest that effective teams maintain balanced communication and workload distribution, while struggling teams show more fragmented interaction patterns.

### 3.6 Early Prediction and Intervention Pipeline

For early predictions, we analyzed transcripts portions

- with the first 1/10 of each transcript, we can predict which cluster the trial belongs with 47% accuracy.
- at 1/3 of each transcript, the accuracy reaches 76%.

The analysis and processes above is used in this pipeline for an autonomous agent to identify the low performing trials

(1) with 10% of the trial, the agent predicts trial's cluster.
(2) if the cluster is low-performing, the agent uses the BEARD variables to decide to intervene.
(3) at 30% of the trial, the agent predicts the trial's cluster again.
(4) if the cluster is still low performing and the TED measures have not improved since the 10%, the agent intervenes again.
(5) repeat steps 3-4 at 50% and 70% if the team needs more help.

Our pipeline allows an agent to predict the team performance early on and take appropriate action, which is quite effective to have a system that allows for such predictions to happen as early as 10 to 30 percentage of the trial.

## ACKNOWLEDGMENTS

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

## REFERENCES


[1] Christopher C Corral, Keerthi Shrikar Tatapudi, Verica Buchanan, Lixiao Huang, and Nancy J Cooke. 2021. Building a synthetic task environment to support artificial social intelligence research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 65, 1 (2021), 660–664.

[2] Stephen Emmitt and Christopher Gorse. 2006. *Communication in construction teams.* Routledge.

[3] Jared T Freeman, Lixiao Huang, Matt Woods, and Stephen J Cauffman. 2021. Evaluating artificial social intelligence in an urban search and rescue task environment. In *AAAI Fall Symposium on Computational Theory of Mind for Human-Machine Teams.*

[4] Lixiao Huang, Jared Freeman, Nancy Cooke, John "JCR" Colonna-Romano, Matt Wood, Verica Buchanan, and Stephen Caufman. 2022. Artificial Social Intelligence for Successful Teams (ASIST) Study 3. https://doi.org/10.48349/ASU/QDQ4MH

[5] Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The Malmo platform for artificial intelligence experimentation. In *Proceedings of the International Joint Conference on Artificial Intelligence.* 4246–4247.

[6] Jones, T. 2021. textmineR Package. https://cran.r-project.org/web/packages/textmineR/index.html. Accessed: 2024-08-10.

[7] Shannon L Marlow, Christina N Lacerenza, and Eduardo Salas. 2017. Communication in virtual teams: A conceptual framework and research agenda. *Human resource management review* 27, 4 (2017), 575–589.

[8] Sangwon Seo, Lauren R Kennedy-Metz, Marco A Zenati, Julie A Shah, Roger D Dias, and Vaibhav V Unhelkar. 2021. Towards an AI coach to infer team mental model alignment in healthcare. In *Proceedings of the IEEE Conference on Cognitive and Computational Aspects of Situation Management.* 39–44.

[9] Joachim Stempfle and Petra Badke-Schaub. 2002. Thinking in design teams-an analysis of team communication. *Design studies* 23, 5 (2002), 473–496.

[10] Gita Sukthankar, Katia Sycara, Joseph A Giampapa, Chris Burnett, and Alun Preece. 2007. Towards a model of agent-assisted team search. In *Procceedings of the First Annual Conference of the International Technology Alliance in Network and Information Science.*

[11] Judith Tiferes and Ann M Bisantz. 2018. The impact of team characteristics and context on team communication: An integrative literature review. *Applied Ergonomics* 68 (2018), 146–159.

[12] Sheila Simsarian Webber, Jodi Detjen, Tammy L MacLean, and Dominic Thomas. 2019. Team challenges: Is artificial intelligence the solution? *Business Horizons* 62, 6 (2019), 741–750.