NLP for Human Collaboration

Josie Zvelebilova
Collaborative Social Systems Lab
Northeastern University | DREAM Program
zvelebilova.j@northeastern.edu

Abstract

This paper presents a natural language processing (NLP) model designed to detect subject matter and sentiment in real-time messages sent between human team-members and generate intelligent interventions for improved collaboration. Development of the model included trials of NLP techniques ranging from simple dictionary matching to transformer-based models, with DistilBert providing the highest accuracy. To pilot test the NLP model in a team communication environment, we integrated the NLP into our larger team communication chat system. We used Flask to create a RESTful API, enabling the transfer of message data from the client to the NLP model and back. The resulting system offers a practical and effective solution for real-time entity and sentiment detection in user chat conversations.

Keywords: natural language processing, human-computer interaction, machine learning

Introduction

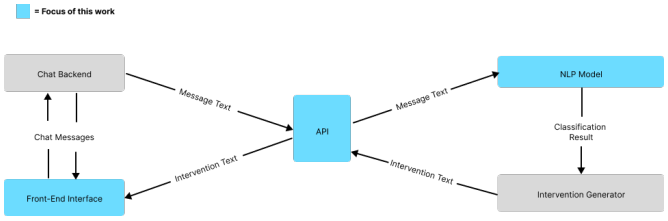
Effective communication is crucial for successful collaboration and teamwork, especially in fast-paced and complex environments. In recent years, chat applications have become increasingly popular as a means of communication between team-members, both in the workplace and in personal settings. Organizations rely heavily on teamwork, making efficient and effective communication a priority, but certain limitations and biases in human reasoning result in suboptimal collaboration. Artificial intelligence offers a potential solution to this problem.

Recent advances in AI have brought about the possibility of human-AI collaboration as a tool for optimizing human-human collaboration. However, developing agents that can interact with multiple, heterogeneous humans in cooperative communication remains challenging. In an extension of a previous study by Westby and Riedl[1], we draw on cognitive science findings to develop an AI "teammate" that aids humans in collaborative decision making.

Theory of mind (ToM) describes the ability of humans to infer the state of mind of others. The degree to which people can do this accurately is a decent predictor of the effectiveness of their collaboration and their collective intelligence as a team. Research on collective intelligence indicates that a common failure point for human teams is the inability to fully integrate information known by only some team members[1,3]. Human-AI teams hold the potential to surpass the performance of human-only teams by overcoming human biases and limitations in information processing.

Westby and Riedl suggest a generative, agent-based process theory based on the free energy principle that provides a computational approach to cognition. This computational approach facilitates the development of explainable agents for human-AI collaboration. The model developed in Westby & Riedl 2023 and constructed in this project is a theory of mind agent that forms mental models of the humans in its team based exclusively on real-time observations of human communication.

This model is based on a toy example of the kind of problem a human team might struggle to solve effectively. To develop the model, we use messages from an experiment hosted on the Volunteer Science platform in which human team members were asked to deduce the details of an art heist. This scenario provided a simple example to work with, as the available information was limited to three dimensions (the culprit, the location, and the day) with five possible answers each. The model we developed is one part of an AI collaboration system that would take each message sent in a team chat, classify it, analyze it in the context of the other messages to determine if an intervention is necessary, and then return the intervention.



We foresee that in a culture rapidly adapting to the presence of artificial intelligence, it is essential to understand the effect of AI on human social dynamics and design AI agents accordingly. This paper presents the details of our NLP model, including the methodology, implementation, and design. In particular, we describe the various NLP techniques that were trialed and the implementation details of the current model, including the front-end interface and deployment on a Heroku server.

Background

Human-AI teaming has a rich and evolving history that has kept pace with technological advancements. One critical aspect of effective team interaction is the ability to form accurate mental models about teammates' actions. There is a significant gap on the subject of how AI can form "mental models" of humans. To create an autonomous AI teammate that enhances human team performance, it is necessary to design agents that are capable of operating in a team of unfamiliar teammates without prior knowledge.

A hidden profile task describes a problem-solving context in which a) some information is common to all team members, b) every team member also has a piece of information that the other team members don't have, and c) all pieces of information are necessary to reach an optimal solution. Stasser and Titus[2] developed the hidden profile paradigm to test the effect of incomplete information on group decision making. They found that groups are likely to discuss and reinforce information common to all group members rather than share the unique pieces of information held by individuals. This shared information bias results in collaborative decisions based on incomplete information, with suboptimal results.

A possible cause of shared information bias is the tendency of the individual to impute their own knowledge onto others. Thus the group decision making problem becomes a theory-of-mind problem: better mental models of others result in better information sharing. An AI trained to develop accurate mental models of team members could be configured to intervene when it detects a bias towards shared information.

Westby and Riedl[1] used data from a 2016 experiment to develop one such model. In the experiment, 145 individuals in teams of five were asked to solve a hidden profile task framed as a criminal investigation. Teams had to determine the location, the culprit, and the day of an art heist. Each of the five individuals in each team was given one private clue, and one public clue was given to the whole group. The teams were given five minutes to communicate over text-based chat before giving their answer. These chat messages constitute the ground truth data that we used to train our model, extending the work of Westby and Riedl.

Implementation

Using the data gathered from the 2016 experiment, we developed a natural language processing model to classify messages sent between team members. Each message in our training data set was manually coded a Yes, No, Maybe Yes, or Maybe No according to the following coding scheme:

Category	Description	Example
Y	Yes	"I know it was Eagle"
MY	Maybe Yes	"Most heists happen on Friday"
N	No	"It wasn't the gallery"
MN	Maybe No	"Monday is unlikely"

In addition, we labeled each message according to subject, e.g. Monday, Eagle, gallery. Below are some examples.

Classification	Example
MY | Gallery	"The keys to the gallery were stolen"
N | Tuesday	"It wasn't Tuesday"

Due to the inherent complexity of natural language and the nuances in individual expression, we encountered some inevitable areas of ambiguity. To ensure consistency in our training data, we established the following rules:

1. Messages expressing agreement with a prior message (e.g. "agreed" or "right") were coded as neutral in dimension and sentiment.
2. Questions were classified according to dimension, but marked neutral in sentiment.

NLP Model

We began with a "bag of words" approach to classification. In this approach, a document (such as a sentence or an entire book) is represented as a bag (or set) of its individual words, disregarding grammar and word order but keeping track of their frequency of occurrence. This method involves converting the text into numerical values, where each word is assigned a value based on its frequency of occurrence in the text. The frequency of words can then be used to represent the text as a vector, called a "term frequency" vector. The bag of words approach is useful for many NLP tasks, such as text classification and sentiment analysis, but it can be limited in its ability to capture the nuances of language and meaning. Although we anticipated mediocre performance from the bag-of-words model, it provided a useful baseline accuracy on which to improve.

More accurate results were obtained using transformer-based models for natural language processing, as illustrated in the table below. Xlnet, which uses permutation-based training to process ambiguity in natural language, performed

significantly better than our bag-of-words approach, but not as highly as other models. XLM yielded a similar performance. BERT and DistilBERT outperformed the other models we tested, with DistilBERT producing the best results.

To improve the model's ability to generalize patterns in the dataset, we also tested each model with entity details masked. This practice helps prevent the model from simply memorizing specific entity names. Our results improved with masked data, as shown in the table below.

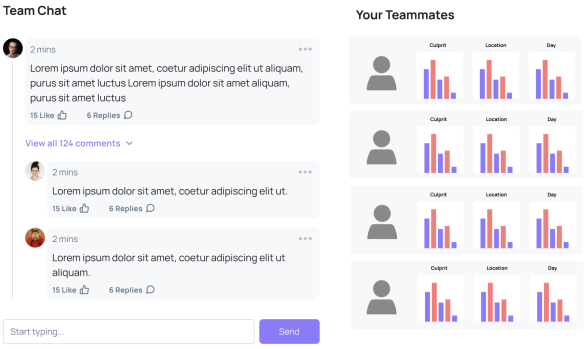
Model Name	Description	F1 Score
Bag of word	Used a predefined set of words	0.46
Bert	By using original classification data	0.83
Xlnet		0.64
XLM		0.7
DistilBert		0.71
Bert	By masking the entity details - eg 'Friday' -> 'target_day'	0.83
Xlnet		0.55
XLM		0.71
DistilBert		0.86

Server and Interface Design

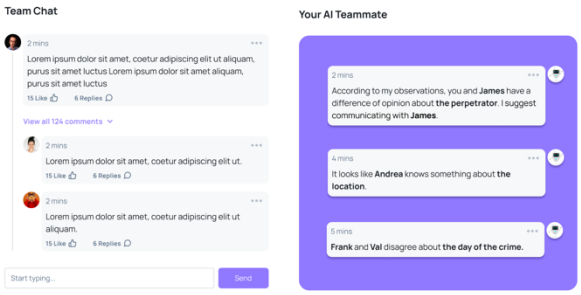
The intended use for this model is to act as an AI "teammate" in human teams to monitor information sharing and generate interventions to help optimize group decision making. To make our model available for use in real-time collaboration, we built a RESTful API using Flask. Our NLP model was written in Python, making Flask a good choice for smooth integration.

Our initial deployment of the model was a test to ensure that our API was functional. It features a simple text box in which a user can enter a message for classification and receive the results. In its eventual form, the API will receive requests from the Volunteer Science platform during experiments with human teams, as chat messages are sent to the NLP model for analysis.

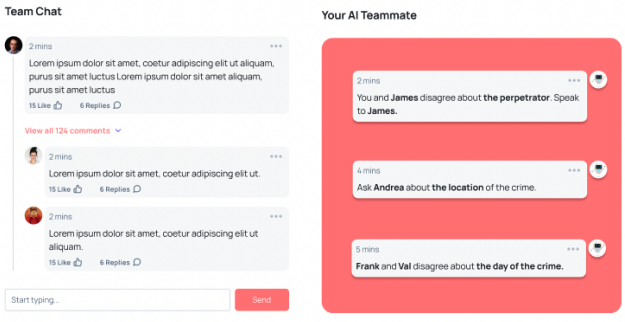
Selecting the method of intervention and designing the interface for the end-user was an iterative process. In one version, charts that represent each team member's confidence in each possible solution are updated continuously in response to the model's classification of the messages.



In another version, we used design elements that placed interaction with the AI teammate on the same level as interaction with human. Messages from the AI teammate appear in text bubbles side-by-side with messages from human teammates.



One technique to encourage users to trust the AI teammate is to use hedging language, which helps convey the impression that the AI is not making demands of the user. For instance, the AI might use phrases such as "perhaps," "it seems," or "maybe" to indicate that its suggestions or responses are not absolute, but rather are offered as possibilities for the user to consider. The use of hedging signals that the AI is not overly confident or authoritative, but rather is open to feedback and willing to adapt to the user's needs. Below is an example of a more assertive AI personality.



Conclusion

With an accuracy score of 0.86, our NLP model is suitable for integration with our AI teammate system, which we believe has potential applications in workplace communication, online forums, and educational settings. Future development of this project will include testing intervention methods and thresholds to determine the most effective kind of intervention for improved teamwork and building this functionality into a user interface.

References

[1] Riedl, C., & Westby, S. (2023). Collective Intelligence in Human-AI Teams: A Bayesian Theory of Mind Approach. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2208.11660

[2] Stasser, G.; and Titus, W. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. Journal of Personality and Social Psychology, 48(6): 1467

[3] Nickerson, R. S. 1999. How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. Psychological Bulletin, 125(6): 737.