ACARIS: Improving Conversational AI and Human Social Skills using User Embeddings

Simon Slamka

June 1, 2023

Abstract

In this paper, we propose ACARIS, the Advanced Communication Augmentor and Relational Insights System, a system utilizing a novel method to analyze emotional state, intent, and interest of text communication parties. ACARIS is being built with the goal of improving social skills of humans, while also improving the performance of human-facing AI systems. We go over our approach, including the initialization of user embeddings from message features, concatenation of user embeddings with word embeddings, modifications of the BERT architecture, and the training and evaluation processes. We also go over the results of our experiments, which demonstrate the effectiveness of our method.

1 Keywords

ACARIS, Conversational AI, Social Skills, User Embeddings, BERT, DistilBERT, Sentiment Analysis, Intent Classification, Emotion Recognition, Interest Recognition

2 Definitions

- **ACARIS** - the Advanced Communication Augmentor and Relational Insights System
- **NLP** - Natural Language Processing - a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data
- **Deep Learning** - a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input
- **SVM** - Support Vector Machine - a supervised machine learning model that uses classification algorithms for two-group classification problems

- **DT** - Decision Tree -
- **LogReg** - Logistic Regression - a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist
- **RF** - Random Forest - an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees
- **TF-IDF** - Term Frequency - Inverse Document Frequency - a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus
- **the Transformer** - a deep learning model introduced in 2017, used primarily in the field of NLP
- **BERT** - Bidirectional Encoder Representations from Transformers - a transformer-based machine learning technique for natural language processing pre-training developed by Google
- **DistilBERT** - DistilBERT is a smaller, faster, cheaper version of BERT developed by HuggingFace
- **Vector** - a quantity that has both magnitude and direction
- **Vector Space** - a collection of vectors, which may be added together and multiplied ("scaled") by numbers, called scalars
- **Embedding Space** - a vector space with a coordinate for each word in the vocabulary, such that words that share common contexts in the corpus are located close to one another in the space
- **Word Embedding** - A vector representation of a word's meaning
- **User Embedding** - A vector representation of a user's personality, emotional state, intent, and interest
- **ReLU** - Rectified Linear Unit - an activation function that returns the input value if it is positive, otherwise it returns zero
- **Loss** - a number indicating how bad the model's prediction was on a single example
- **Loss Gradient** - the gradient of the loss with respect to the model parameters
- **Ablation Study** - a study in which specific components of a system are removed to analyze their impact on the overall performance

3 Motivation and Premise

I've always had an issue with interpersonal relationships. I could never fully understand the way people work on a social level. I understand the technical and biological fundamentals, and superficially, how the mind works, but not how all this becomes so much more complex when actually talking and dealing with people. Sometimes, I feel that another person and I broadcast on different frequencies, because we are unable to either understand one another or maintain a (romantic) relationship for an extended period of time. I have tried for many years to find a technical solution to this problem, but only after coming to Denmark, I acquired the core knowledge to try and accomplish this ambitious goal by really getting into it and self-studying as much as I could.
I firmly believe that all things in nature are governed by rules and these rules can be observed, measured, and then, based on that data, predicted. This includes human behavior and emotions. We're nothing more than complex electrobiochemical machines driven by electrical impulses and hormones. Apart from that, even if we don't understand our own minds yet, we still have many, many years of knowledge about how human personalities work and how we make decisions based on what happens to or around us. On a fundamental level, this doesn't change. Sure, they say that we're all different. However, the core remains. We all share many attributes that make us human. I believe that we, given enough data, can use these attributes to predict human behavior, reactions, emotional state, and intent.

4 Hypothesis

Given enough conversational data per person, human behavior (emotional state, intent) in text communication can be predicted with a high level of confidence ($> 80\%$) due to the fact that humans are, on a fundamental level, very similar to one another.

5 Introduction

Interpersonal communication has always been an integral part of human lives. With the rise of the Internet and, subsequently, social media and other forms of online text communication, the manner in which we talk has changed dramatically. This has led to a drop in social skills in humans, particularly those of the last generation. ACARIS attempts to adjust for this by providing a way to analyze the emotional state, intent, and interest of text communication parties, providing them with a way to improve their social skills, while also improving the performance of conversational

AI systems, which are becoming increasingly prevalent in our society, and their ability to understand human emotions, intent, and interest is becoming more and more important, especially in AI systems that directly interact with humans, such as digital assistants, chatbots, and others.

6 Literature Review and Related Work

6.1 Studied Literature

6.1.1 Human Emotions

–

6.1.2 Sentiment Analysis

- **Machine Learning with PyTorch and Scikit-Learn** by Yuxi Liu, Vahid Mirjalili, Sebastian Raschka
- **Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow** by Aurélien Géron

7 Initial Attempts

Our initial attempts of implementing ACARIS included the use of SVMs, DTs, LogReg, and RFs, each paired with TF-IDF in the preprocessing stage. None of these methods have proven to be accurate enough (barely reaching 50% accuracy) for large amounts of complex language, especially casual language that often includes slang, sarcasm, emoji, and similar elements. The primary issue was the lack of these methods' ability to capture semantical and contextual aspects of our dataset. Another issue was TF-IDF's lack of word order understanding. Knowing that the Transformer architecture is capable of capturing these aspects due to its positional encodinds and self-attention mechanism, we decided to use it as the basis for ACARIS.

8 Methodology

In this section, we provide a detailed description of our approach, including the initialization of user embeddings, concatenation with word embeddings, modifications to the model architecture, and the training and evaluation processes.

8.1 User Embeddings Initialization

For each user u, we create an initial user embedding e_u of size d_e, where d_e is the dimensionality of the user embeddings. We initialize e_u with random values sampled from a Gaussian distribution:

$$\begin{aligned} &\texttt{from torch import randn} \\ &e_u = \texttt{randn}(d_e) \end{aligned} \tag{1}$$

8.2 Concatenation of User and Word Embeddings

Given a message w (converted to a word embedding) with its word embedding e_w (of size d_m) from a BERT-like pre-trained model such as DistilBERT, we concatenate the user embedding e_u to the word embedding e_w to create a joint representation e_{joint} (of size $d_e + d_m$):

$$\begin{aligned} &\texttt{from torch import cat} \\ &e_{\text{joint}} = \texttt{cat}([e_u, e_w], \text{dim} = -1) \end{aligned} \tag{2}$$

8.3 Model Architecture Modification

To accommodate the concatenated embeddings, we create a custom input layer with the new input size $d_e + d_m$ within a BERT-flavored model. An alternative could be to use a separate fully-connected layer (with ReLU activation) to map the concatenated embeddings to a compatible size:

$$\begin{aligned} &\texttt{from torch import randn} \\ &\texttt{from torch import relu} \\ &\texttt{from torch import matmul} \\ &W1 = \texttt{randn}((d_e + d_m, d_m)) \\ &b1 = \texttt{randn}(d_m) \\ &e_{\text{map}} = \texttt{relu}(\texttt{matmul}(e_{\text{joint}}, W1) + b1) \end{aligned} \tag{3}$$

Then, pass e_{map} through the existing model architecture.

8.4 Training with User Embeddings

We train the model with the concatenated embeddings, which involves modifying the training loop to include user embeddings for each message. For each training step, we fetch the corresponding user embeddings, concatenate them with word embeddings, and pass the result to the model. The model learns to associate user-specific patterns with the message content.

8.5 Updating User Embeddings

To update user embeddings over time, we backpropagate the gradients from the model loss to the user embeddings. This requires a custom training loop to ensure that the gradients are properly propagated. Alternatively, we can consider methods like moving averages or other heuristics to update the embeddings based on new interactions.

For example, if using gradient descent, the user embedding update could look like this:

$$e_u = e_u - \text{lr} \cdot \frac{\partial L}{\partial e_u} \tag{4}$$

where $\frac{\partial L}{\partial e_u}$ is the gradient of the loss with respect to the user embedding e_u and lr is the learning rate.

9 Evaluation

We evaluate our approach using benchmark datasets for various conversational AI tasks, such as sentiment analysis and intent classification. We compare the performance of models trained with and without user embeddings to demonstrate the effectiveness of our method. We also perform ablation studies to analyze the impact of various components, such as the user embedding size and update methods.

10 Results

NaN

11 Conclusion

NaN