Enhancing Diabetes Care through AI-Driven Lie Detection

RQ: "How can linguistic indicators from a patient's chat message be used to detect deceit in a diabetes support system?"

Renee van Westerlaak | 5244285
May 17, 2025
TU Delft | CSE3000

Supervisors:
Prof. Catholijn Jonker
J.D. Top, MSc

Sources:
[1] B. de Ruiter and G. Kachergis, "The mafiascum dataset: A large text corpus for deception detection," 2019.
[2] R. Loconte, R. Russo, P. Capuozzo, P. Pietrini, and G. Sartori, "Verbal lie detection using large language models," *Scientific Reports*, vol. 13, 12 2023.

1. Background

- For diabetes patients, a lifestyle intervention can decrease insulin resistance and improve their health
- Research has shown that patients' adherence to such support systems cannot be assumed and should be monitored
- For example, it has been shown that patients do not always truthfully report their glucose levels
- There are linguistic cues that have been found to indicate deception
- ***CHIP*** [figure 1] is a diabetes support system in the making, as a part of the ***Hybrid Intelligence (HI)*** project, in cooperation with ***De Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek (TNO)***
- The ***CHIP*** system contains a ***Software Agent*** which gives recommendations in response to messages from a patient

This study attempts to detect deception in the software agent by analyzing single messages for deceptive cues, with the ultimate goal of enhancing diabetes care

2. Objective

The goal of this research is to find out which linguistic cues can be used to detect deceit in a diabetes support system and how to implement that using a Machine Learning model

Figure 1. ***CHIP*** System with proposed lie-detection module in white

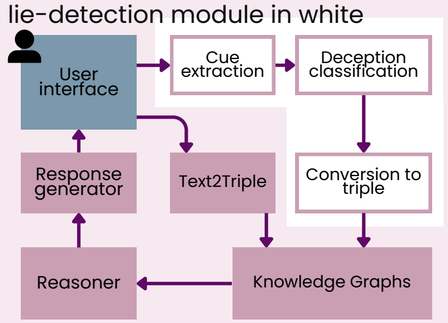

Figure 3. Classification Reports

Class	Precision	Recall	F1-Score	Support
Truthful (0)	0.77	0.80	0.78	744
Deceptive (1)	0.23	0.20	0.21	224
Accuracy			0.66	968
Macro avg	0.50	0.50	0.50	968
Weighted avg	0.64	0.66	0.65	968

3.1 With training and testing dataset split 90-10

Class	Precision	Recall	F1-Score	Support
Truthful (0)	0.77	0.79	0.78	2232
Deceptive (1)	0.22	0.20	0.21	671
Accuracy			0.65	2903
Macro avg	0.50	0.50	0.50	2903
Weighted avg	0.64	0.65	0.65	2903

3.2 With training and testing dataset split 90-10

3. Literature Study

3.1 Methodology

Scientific papers were found and analyzed on the following topics:

1. Linguistic cues to deception
2. Lie detection and the use of ***Machine Learning (ML)*** methods

3.2 Linguistic cues to deception

The linguistic cues found in the literature study can be found in *Figure 2*. Only the bold-face cues were used in this research, since the other cues involve the conversation context, which is not within the scope of this project

3.3 Similar experiments

- An experiment with messages from truthful and deceitful players in a game of Mafia (***Mafiascum Dataset***), that used a ***Support Vector Machine (SVM)*** model trained on linguistic cues to detect deceitful messages, achieved an average precision of 0.39 (chance = 0.26) [1]
- An experiment that used a ***Large Language Model (LLM)*** to detect deceitful texts fabricated by participants in the research achieved an average accuracy of 76% [2]
  - As an ***LLM*** does not use linguistic cues for classification and the reasoning behind its decisions is unclear, this method was not chosen

4. Experiment

4.1 Methodology

Dataset: *Mafiascum Dataset* with over 8000 documents, each containing messages from a player in a game, annotated with either a deceptive or truthful role

Cue extraction: The documents were preprocessed and the linguistic cues (Figure 3) extracted from them using ***SpaCy***: a natural language processing library

Model: An ***SVM***: a supervised classification model, trained on the dataset using the ***Scikit Learn*** library with cues as features.

Testing: The ***SVM*** model was trained and tested with training and testing dataset splits 70-30 and 90-10

Design: The aforementioned steps can be implemented to create a lie-detection module in ***CHIP*** (See Figure 1)

4.2 Results

Classification reports were generated for both scnarios (see Figure 3). The F1-Score for the deceptive class in both scenarios is very low (0.21). The dataset comprises approximately 80% truthful and 20% deceptive instances

Figure 2. Lie detection cues

Cue	Count*
Fewer words used**	3
More sentences, fewer distinct words	1
Fewer exclusive words (but, except)	2
Fewer tentative words (may, perhaps)	1
More negation terms (no, never)**	2
More negative emotion words	4
Fewer first-person pronouns**	4
More second-person pronouns**	1
More third-person pronouns**	3
More motion verbs	2
Fewer insight/cognitive words**	2
Speech errors and disfluencies**	1
Indirect/ritualized speech	1
Self-deprecation	1
Fewer sensory details	1
Fewer details	1
Fewer causation words (only relevant in omission lies)	1

*the amount of papers where supporting evidence was found
**contradicting evidence was found in one or more studies

5. Conclusion

From the experiment results, the conclusion can be drawn that the ***SVM*** with the chosen linguistic cues does not predict deceptive messages more accurately than chance when trained and tested on the ***Mafiascum Dataset***.

5.1 Result comparison

A model that randomly predicts truthful 80% of the time and deceptive 20% of the time would get F1-Scores of approximately 0.80 and 0.20 respectively

6. Limitations

6.1 Limitations to deception detection

- No guaranteed correlation between cues and actual deception
- Cues may differ between people and contexts
- ***Non-transferability:*** an ***ML*** model trained and tested on one dataset, performed worse when tested on a different dataset

6.2 Limitations to *CHIP*

- Single messages only → Reduces the available context
- Textual messages only → Non-verbal cues cannot be measured
- No ground truth to confirm the module's predictions
- No appropriate training dataset for the specific context

6.3 Validity of testing data

In the experiments mentioned in this research, data was either generated by participants in the experiment, or by players in a game. According to research, cues to deception occur due to the impact lying has on a person. When fabricating deceptive accounts, the implications of being caught are different, and as a result, cues might differ from those in a natural setting. In the ***Mafiascum Dataset***, all messages from one player are annotated with the player's role, potentially causing truthful messages from deceptive players to be falsely labeled as deceptive and vice versa.

TU Delft