

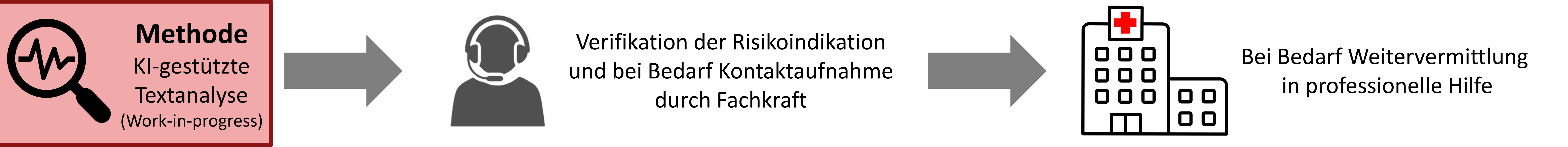
Über Sucht red(dit)en: Früherkennung von Risiken der Glücksspielsucht in Reddit-Posts mit kleinen und großen Sprachmodellen

Hintergrund: Jugendliche und junge Erwachsene befinden sich in einer Lebensphase mit erhöhter Vulnerabilität für die Entstehung von Süchten und anderen psychischen Belastungen und kommunizieren viel über diese Belastungen in öffentlichen Räumen des Internets. Diese explorative Studie vergleicht Methoden zur Erkennung von riskantem Glücksspiel in Reddit-Posts.

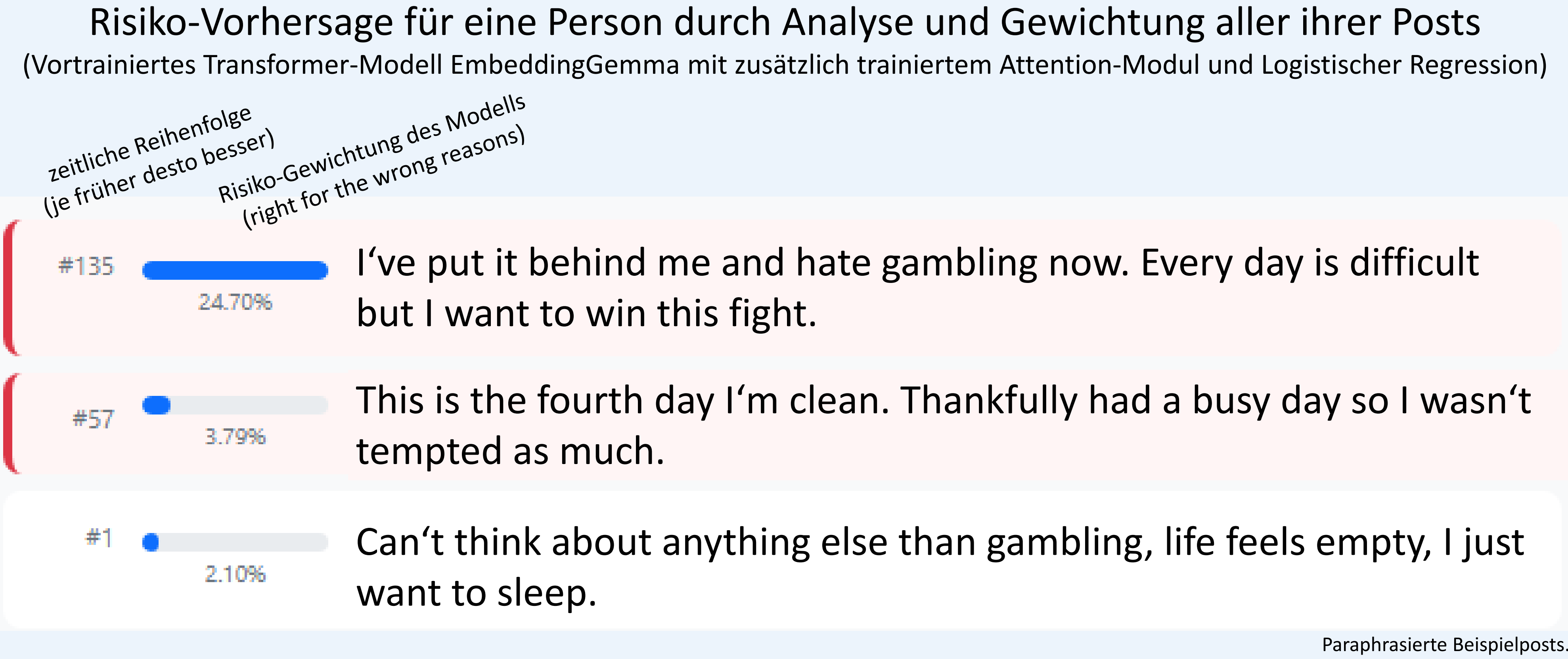
Hypothese: Sprachmodelle können im Inhalt und in der Formulierung von Beiträgen („Posts“) im Online-Forum Reddit hilfreiche sprachliche Signale identifizieren, die zum Screening von potentiell hilfebedürftigen Glücksspielenden genutzt werden können.



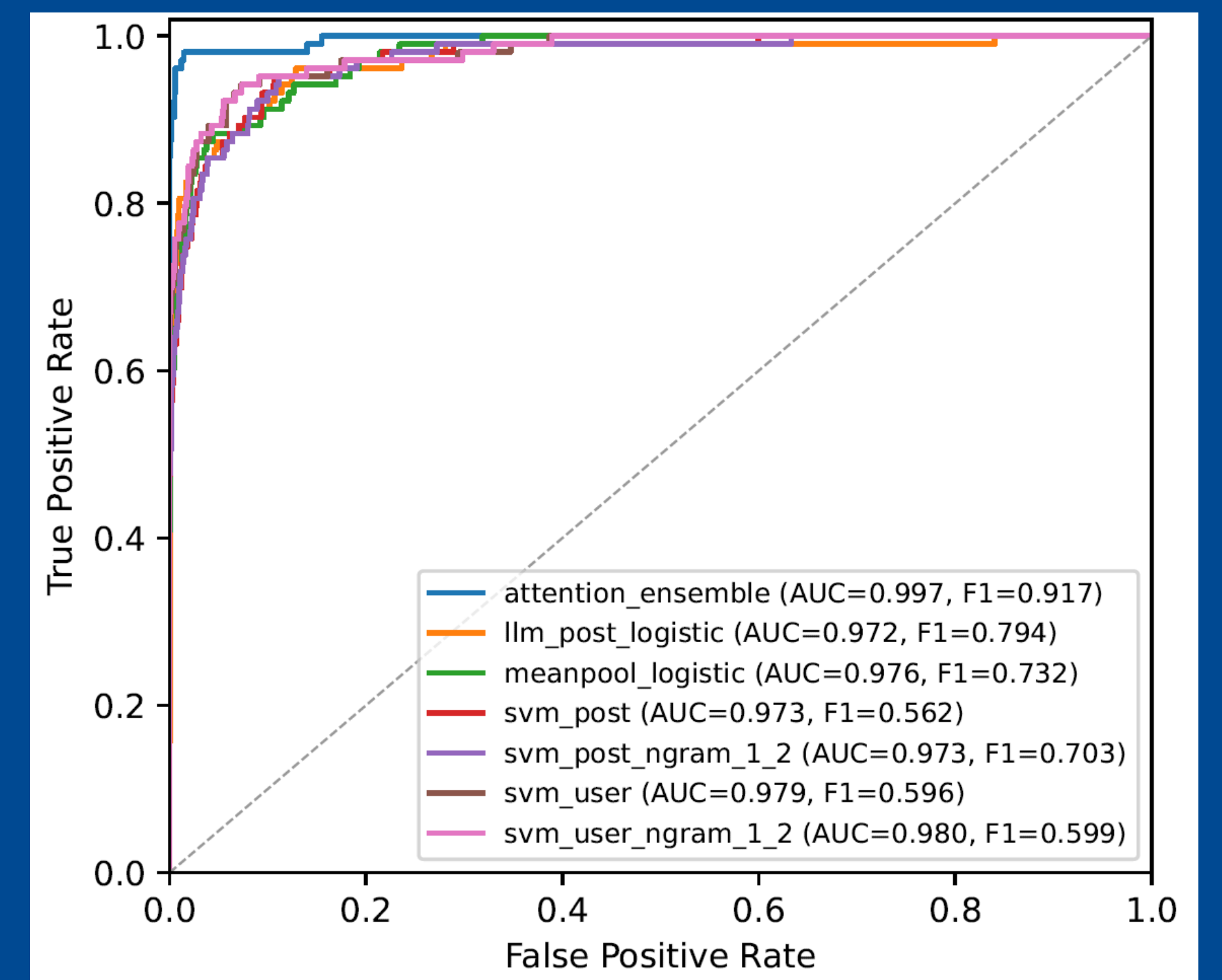
r/GamblingAddiction



Großes Sprachmodell



Quantitative Auswertung Receiver-Operator Characteristic (ROC)



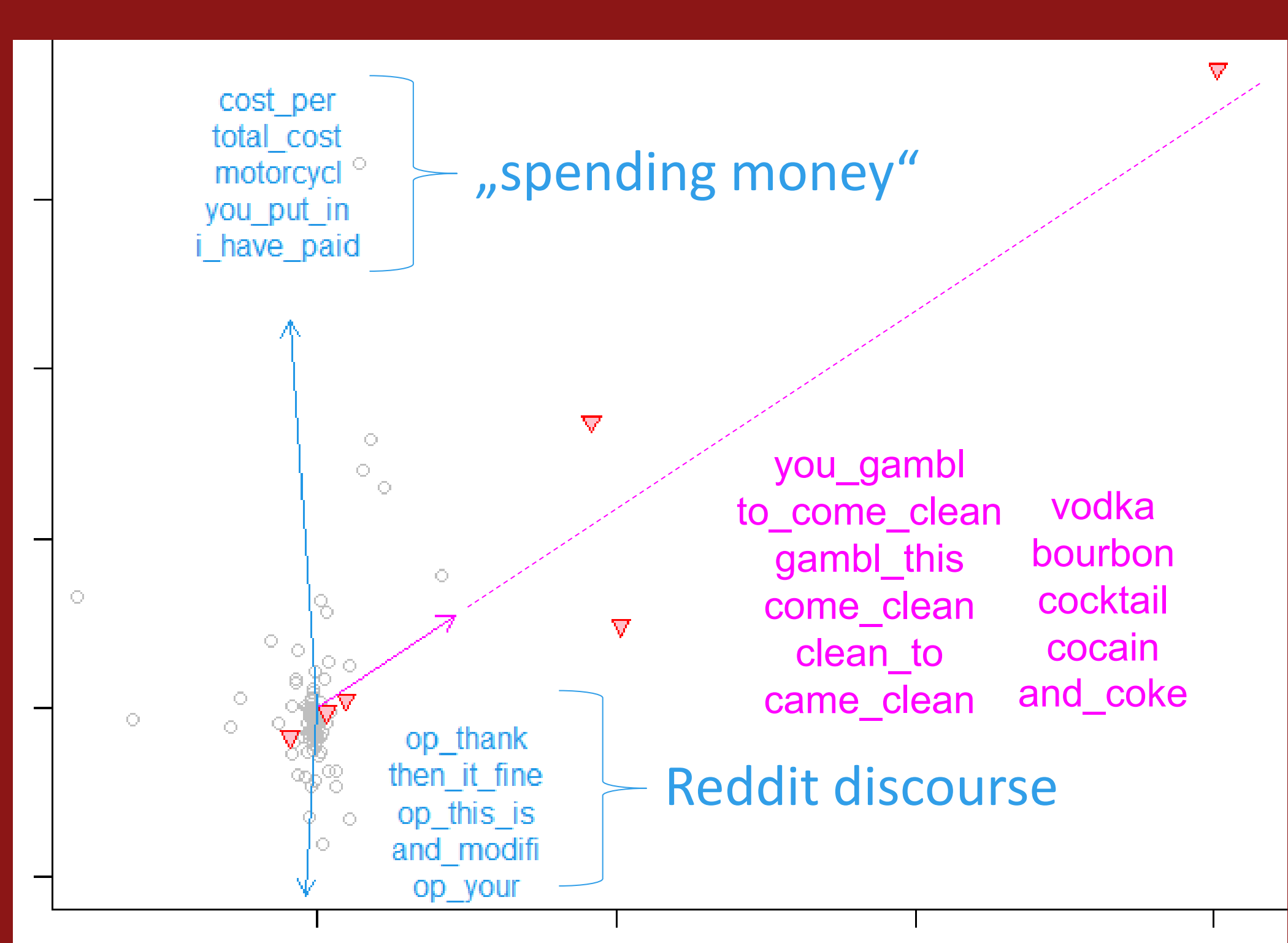
Kleines Sprachmodell

Risiko-Vorhersage für einen Post durch Zählen und Gewichten von Wortfolgen
(Support Vector Machine, SVM)

Die Methoden wurden auf den „Pathological Gambling“ Daten von Parapar et al. (2023) vorläufig evaluiert.
Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F. (2023). Overview of eRisk 2023: Early Risk Prediction on the Internet. In: Arampatzis, A., et al. Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2023. Lecture Notes in Computer Science, vol 14163. Springer, Cham. https://doi.org/10.1007/978-3-031-42448-9_22

Zwischenfazit: Die quantitative Auswertung legt zwar zunächst nahe, dass alle getesteten Modelle riskant glücksspielende Personen sensitiv und spezifisch erkennen können ($\text{ROC-AUC} \geq 0.972$) und dass **die Gewichtung signalstarker Beiträge** den positiven Vorhersagewert deutlich verbessert (mit: $\text{F1} = 0.917$; ohne: $\text{F1} \leq 0.794$). Die in der qualitativen Auswertung des kleinen SVM-Modells als relevant identifizierten Vektorraum-Komponenten sind jedoch **auch leicht mit Substanzkonsum sowie stark mit dem Überwinden der Sucht assoziiert**. Ersteres deutet auf latente komorbide Symptome hin, die für eine effektive Früherkennung Hilfsbedürftiger idealerweise explizit modelliert werden sollten. Letzteres ist vermutlich ein **Störfaktor in den verwendeten Daten**, der die Früherkennung deutlich erschwert und zukünftig kontrolliert werden sollte.

Qualitative Auswertung Principal Component Analysis (PCA)



Vorhersagen von auf diesen Daten trainierten Modellen laufen sonst Gefahr „right for the wrong reasons“ zu sein und in der Praxis zu scheitern.

