Author: Shivani Singh
s.singh-16@student.tudelft.nl

Explaining XAI Models for Fact-Checking

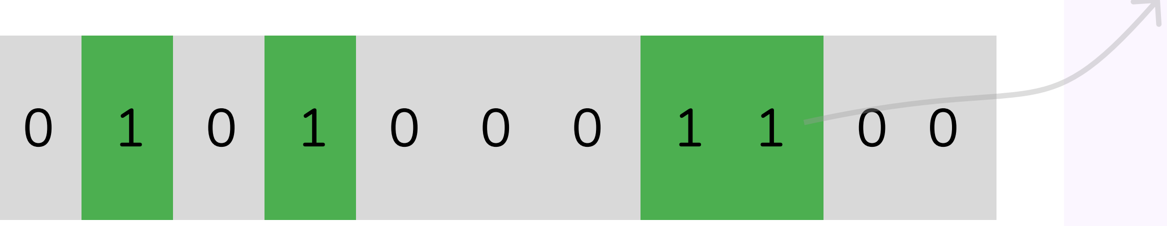
Supervisors: A. Anand, L. Corti & L. Lyu

Background

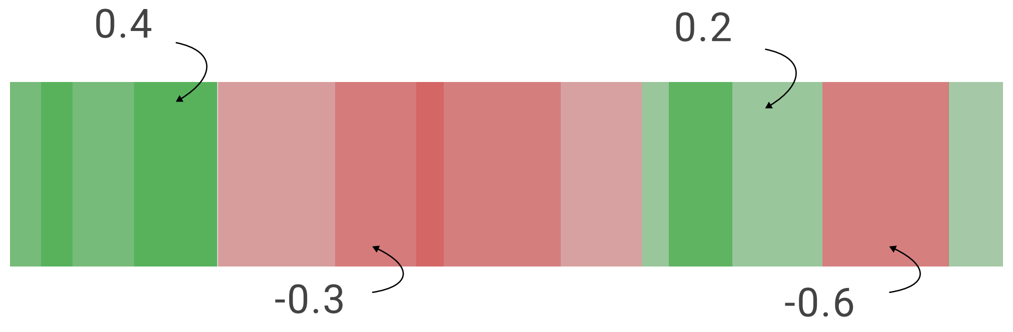
Fact checkers automate the detection of misinformation and provide a credibility check.
With the new target user group of Artificial Intelligence including non-experts, there is a need for explanations that are both **accurate and understandable for the user.**

This study focuses on presentations of the following three **explainable AI explanation method outputs**

Interpretable by Design:
ExPred [1]



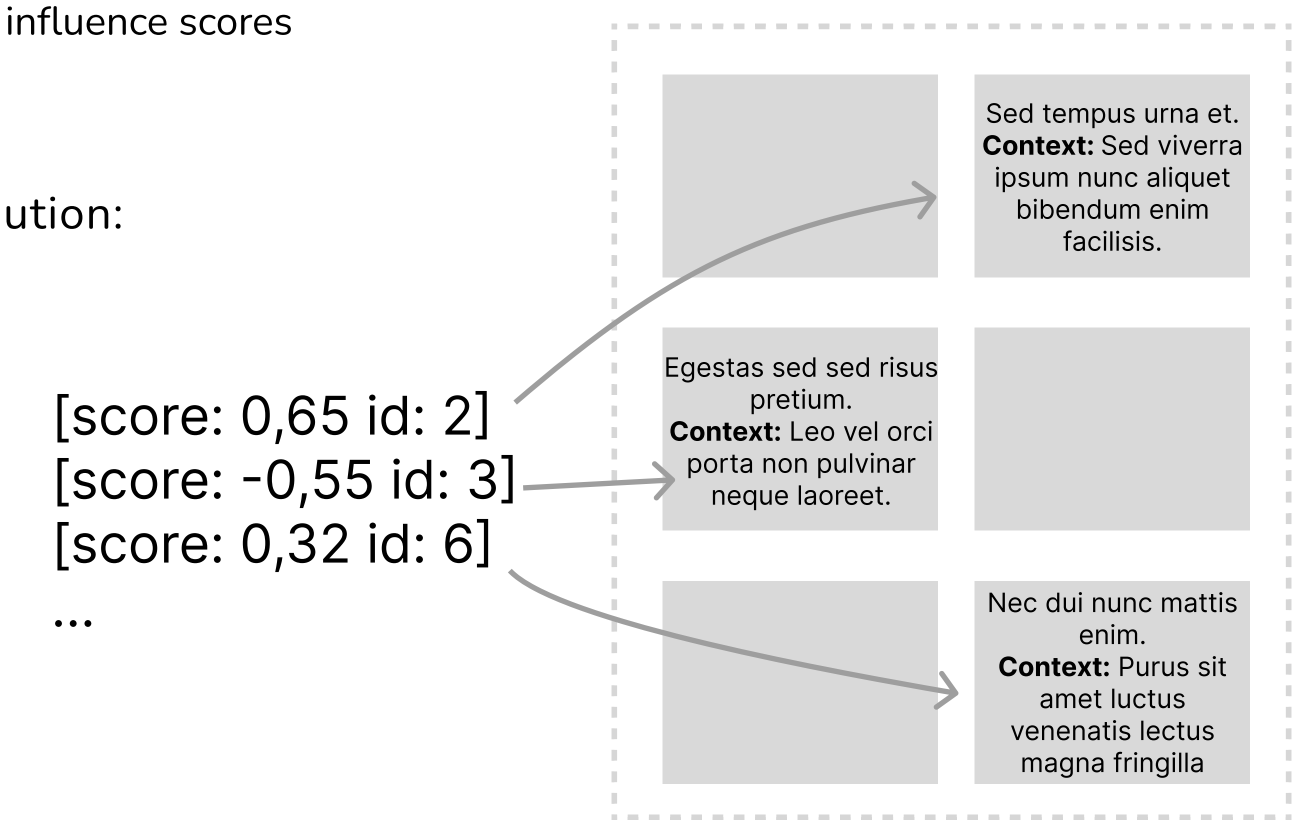
a classification and binary explanation array



Feature attribution:
LIME [2]

an array of token influence scores

Instance attribution:
kNN [3]



[score: 0,65 id: 2]
[score: -0,55 id: 3]
[score: 0,32 id: 6]
...

Sed tempus urna et. **Context:** Sed viverra ipsum nunc aliquet bibendum enim facilisis.

Egestas sed sed risus pretium. **Context:** Leo vel orci porta non pulvinar neque laoreet.

Nec dui nunc mattis enim. **Context:** Purus sit amet luctus venenatis lectus magna fringilla

sorted list of influence scores of the training data

Research Question

How do different explanation presentation strategies of feature and data attribution techniques affect non-expert understanding?

Methodology

For each of the explanation methods we developed two presentation strategies to present to the users.
We compared a high context and textual presentation with a low context and visual presentation:

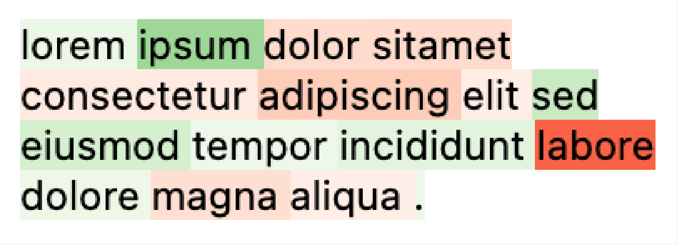
'ipsum aliquet ante metus'

only influential tokens

(a) ExPred FreeText

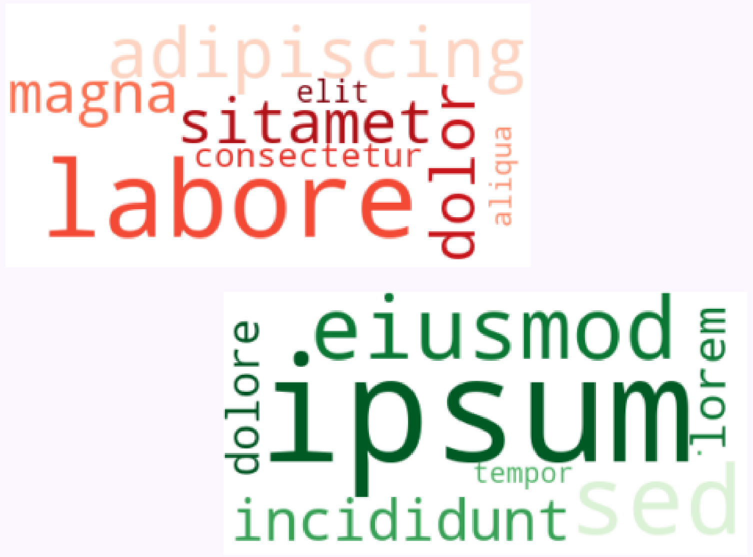
lorem **ipsum** tincidunt **aliquet** id risus feugiat **ante metus** id dictum

(b) ExPred Highlighted

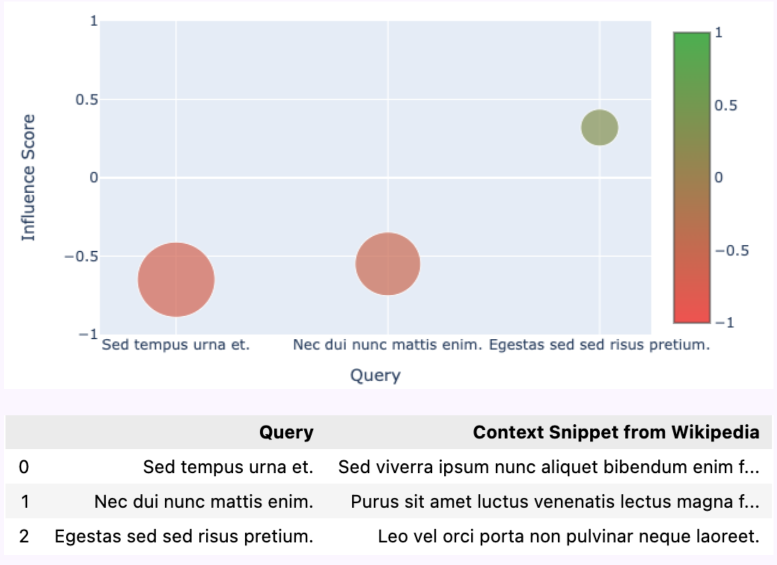


lorem ipsum dolor sitamet consectetur adipiscing elit sed eiusmod tempor incididunt labore dolore magna aliqua .

(c) LIME Heatmap

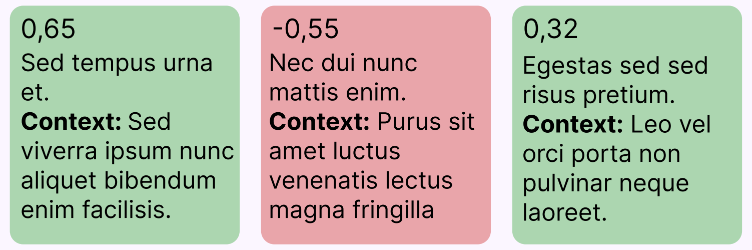


(d) LIME Wordclouds



	Query	Context Snippet from Wikipedia
0	Sed tempus urna et.	Sed viverra ipsum nunc aliquet bibendum enim f...
1	Nec dui nunc mattis enim.	Purus sit amet luctus venenatis lectus magna f...
2	Egestas sed sed risus pretium.	Leo vel orci porta non pulvinar neque laoreet.

(e) kNN Graph



0,65 Sed tempus urna et. **Context:** Sed viverra ipsum nunc aliquet bibendum enim facilisis.

-0,55 Nec dui nunc mattis enim. **Context:** Purus sit amet luctus venenatis lectus magna fringilla

0,32 Egestas sed sed risus pretium. **Context:** Leo vel orci porta non pulvinar neque laoreet.

(f) kNN Boxes

We conducted semi-structured interviews about these prototypes.
With a sample of 20 participants from a technical university, and no issues with colour blindness, we evaluated the prototypes with the following criteria:

Visually Appealing   Easy to Understand
Informative   Convincing   Useful
[4]

[1] Zijian Zhang, Koustav Rudra, and Avishek Anand. Explain and predict, and then predict again.,2021
[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier., 2016
[3] Nazneen Fatema Rajani, Ben Krause, Wengpeng Yin, Tong Niu, Richard Socher, and Caiming Xiong. Explaining and improving model behavior with k nearest neighbor representations., 2020
[4] Gionnieve Lim and Simon T Perrault. Explanation preferences in xai fact-checkers., 2022

Results

Key takeaways from Thematic Analysis:

Formatting: use correct spacing, "fluff" such as icons, fonts and labels, intuitive colours
Context: necessary for the users' understanding and focus.
Data: plays a crucial role in the user's understanding of the fact checker.
Type of Explanation: most useful when directly related to the claim.
Amount of details: not too much, but the explanation should stand alone.

Key takeaways from Quantitative Analysis:

ExPred FreeText: 3.71/5
🏆 ExPred Highlighted: 4.03/5
LIME Wordclouds: 2.31/5
☹ LIME Heatmap: 2.15/5
kNN Boxes: 2.83/5
kNN Graph: 2,63/5

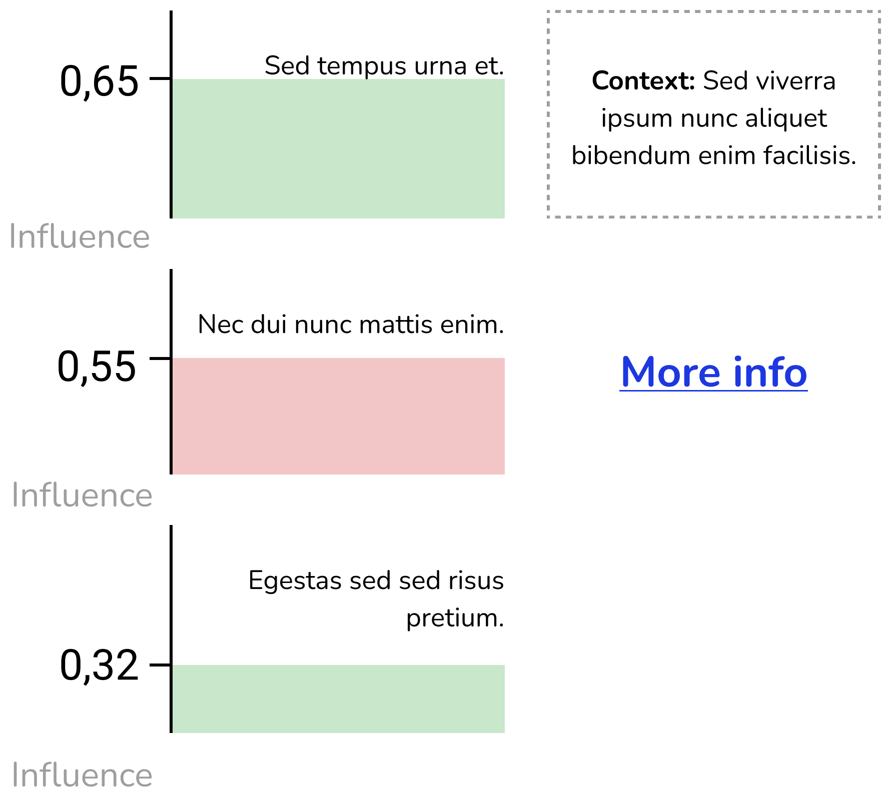
High inter annotator agreement with a Cronbach's α of 0.928.

Recommendations based on Outcomes

ExPred: Show only the influential tokens (FreeText) instead of the whole context, but link the source for verification.

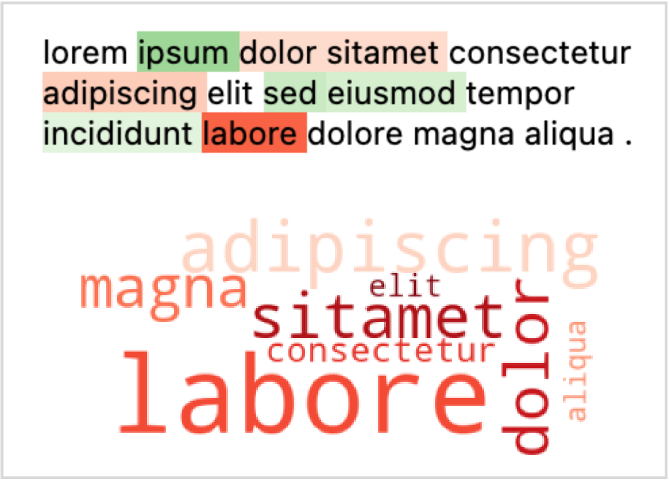
ipsum aliquet ante metus
Lorem ipsum

kNN: Rank the queries from top to bottom presenting the influence as a individual graphs, and only reveal the context upon request.



Context: Sed viverra ipsum nunc aliquet bibendum enim facilisis.

LIME: Combine the two solutions as a compromise between more context and less overwhelming data and introduce neutral values in heat map.



lorem ipsum dolor sitamet consectetur adipiscing elit sed eiusmod tempor incididunt labore dolore magna aliqua .

Conclusion

Participants prefer a simple, structured and textual presentation of all available context and details, rather than visual presentations. Additionally, users should be able to make the same conclusion as the AI with minimal reading effort, and thus understand how the presented data relates to the claim