KU LEUVEN

FACULTEIT WETENSCHAPPEN

Illuminating the black box

Explaining machine learning algorithms that detect fake news on Twitter

Kenan EKICI

Thesis supervisor:
Prof. dr. K. Verbert

Assessors:
Prof. dr. T. Guns
M. Millecamp

Mentor:
J. Ooge

Proefschrift ingediend tot het
behalen van de graad van
Master of Science in de Toegepaste Informatica (AI)

Academiejaar 2020-2021

Preface

The academic year 2020-2021 has been both a wonderful and a terrifying one. Wonderful, because I had the opportunity to work on my thesis and learn the most I have ever had in my life. Terrifying, because of the COVID-19 pandemic which has left many in awe and devastation. The silver lining, I guess, was the opportunity it created for me to work on a very relevant and significant problem. Ironically, I was infected with COVID-19 at some point during my thesis which means I had to battle COVID-19 in two different ways this year.

I can not for the life of me imagine what this thesis would have resulted in without the guidance of my advisor Jeroen Ooge. I want to thank him sincerely for keeping me focused, for his attention to detail, guidance, availability, extreme patience, and most importantly, for being transparent with me. Moreover, I want to thank Prof. dr. Katrien Verbert for her efforts and taking her time to supervise my research. I want to thank all the participants for their participation in the think-aloud study and the user study. Thank you Shotallo Kato, and the members of the HCI department as well, for your feedback on my work.

Last but not least, I want to thank my family and friends. For their support and mostly their presence. I can usually support myself, but I can not perform without the presence of my loved ones. A special thanks goes out to Miray Muratoğlu. Her support was indispensable. Moreover, her legal and Twitter expertise have in many times helped me to better understand the Twitter environment and some of the jurisdictional problems behind black box Artificial Intelligence models.

Abstract (EN)

COVID-19 was not the only major concern during the "infodemic", as the World Health Organization coined. It was the massive spread of misinformation that followed on Online Social Networks. Fortunately, the detection of misinformation or "fake news" has gained traction within the Artificial Intelligence literature. Machine and Deep Learning models are typically trained to detect fake news in social media data. However, deadly viruses and complex AI models have one thing in common. They are both black boxes. In our work, we trained a model for detecting COVID-19 fake news on Twitter data based on techniques from related work. Moreover, we used eXplainable AI (XAI) techniques such as LIME and example-based explanations to explain the models. A human-centered approach was taken to iteratively develop a prototype for presenting the explanations to non-expert end users. The prototype was evaluated through think-aloud studies. Furthermore, we measured factors such as trust and task performance to assess the explanations. We found that LIME explanations lead to higher trust in the AI than example-based explanations. The thematic analysis further revealed that participants felt that the LIME explanations better justified the prediction. Unfortunately, the overall trust score for both explanation methods was low due to disagreements with features and cases of counterfactual reasoning. We also found that both explanation methods helped participants increase their accuracy in classifying COVID-19 misinformation. Due to the limits of the study however, it was hard to verify whether this was solely due to the explanations.

Abstract (NL)

COVID-19 was niet de enige grote zorg tijdens de door Wereldgezondheidsorganisatie benoemde “infodemic”. Het was de massale verspreiding van misleidende informatie die volgde op online sociale netwerken. Gelukkig heeft de detectie van misleidende informatie of “fake news” meer grip gekregen in de literatuur over artificiële intelligentie (AI). Machine- en Deep Learning-modellen kunnen ontwikkeld worden om fake news te detecteren in data afkomstig uit online sociale netwerken. Helaas hebben dodelijke virussen en complexe AI-modellen één ding gemeen. Het zijn beide zwarte dozen. In ons werk hebben we een model ontwikkeld voor het detecteren van COVID-19 fake news op Twitter-gegevens door middel van technieken uit gerelateerd werk. Bovendien hebben we eXplainable AI (XAI)-technieken zoals LIME en example-based methoden gebruikt om de modellen uit te leggen. Vervolgens hebben we gebruik gemaakt van een “human-centered” aanpak om iteratief een prototype te ontwikkelen voor het presenteren van de uitleg aan non-expert eindgebruikers. Voor de evaluatie van de prototypen hebben we de think-aloud studie toegepast. Verder hebben we factoren zoals vertrouwen en taakprestatie gemeten om de uitleg van de AI te beoordelen. We hebben ontdekt dat LIME tot meer vertrouwen in de AI heeft geleid dan de example-based methode. Uit de thematische analyse bleek verder dat de deelnemers vonden dat LIME de voorspelling beter rechtvaardigde. Helaas was de vertrouwensscore voor beide methoden eerder laag wegens onenigheden met de methoden en gevallen van contrafeitelijke redeneringen. We hebben ook ontdekt dat beide methoden de deelnemers hielpen hun nauwkeurigheid te vergroten bij het classificeren van COVID-19 fake news. Door de beperkingen van het onderzoek was het echter moeilijk te verifiëren of dit uitsluitend aan de verklaringen lag.

Contents

Chapter 1

Introduction

In the information age, it is possible for misinformation to spread like wildfire on Online Social Networks such as Twitter. The COVID-19 "infodemic" as the WHO coined, may have been the worst example of this to date. Fortunately, there exists a wide range of Artificial Intelligence (AI) models for the detection of fake news in the literature. However, most of these models can be considered black boxes. This is when the inner workings of a model are difficult, if not impossible to understand. XAI is one of the domains concerned with explaining or in other words "illuminating" such black boxes. In our work, an explainable machine learning model was developed for detecting fake news on Twitter data. The model was explained with local explanation techniques which were presented to end users. The explanations and the interface were evaluated using a human-centered approach

In the next chapter, the literature study, we start by introducing the black box problem, the desiderata behind model explanation, and the trade-off problem. We then strive to understand what an explanation on itself means and the social process behind it. The field of Social Science overlaps with the field of XAI and some of the right terminology should thus be understood. Afterwards, we follow up with an introduction to the field of XAI and discuss the methods that exist to solve the black box problems including some of the proposed explanation techniques. Additionally, we cover some of the defined goals and measures used in XAI literature to assess the results of the explanation methods. Since the aim is to explain machine learning models used for classifying fake news on Twitter, we provide a brief literature study on misinformation in Online Social Networks. The literature study ends with a related work section covering some of the useful work on explainable misinformation classifiers and a table of the literature that was used in this work. In Chapter 3, we propose the methods from the literature of how we developed, explained, and evaluated a prototype for classifying misinformation using a human-centered approach. We also present research questions and the methodology behind the final user study. The training phase of the misinformation classifier will be discussed in Chapter 4. In Chapter 5, we present the implementation behind the prototype during many of its iterations. In Chapter 6, we present the results of the user study. In Chapter 7, we answer the research questions, discuss the results and the limitations of our work. We conclude our work in the final chapter.

Chapter 2

Literature study

2.1 Introduction

In this chapter, we will discuss what it means for AI models to be black box models and why this is a problem. We will elaborate on the need for explainable models from a social, ethical and jurisdictional perspective. Furthermore, we will strive for an understanding of what constitutes to a "good" explanation when considering a human-centered approach. This will be followed by an overview of techniques from XAI literature that are used to explain black box models. This is by no means meant to be an exhaustive overview. The techniques presented either aim to help end users interpret the model itself, or its decisions. In our work, we will mostly focus on the latter as it is a more feasible explanation method for non-expert users who will face the outcome of black box models. In order to put the black box problem into context, we will take a closer look how black boxes are currently being used in the domain of social media. We will zoom in on the black box classification of social media misinformation. In light of current events, we target COVID-19 misinformation.

2.2 The black box problem

A black box system is one whose internal functionality is impossible to interpret or to make sense of. There are two reasons why this could be the case. Either the internals are simply not open for inspection, i.e. hidden. Thus, there is a clear lack of transparency towards its users. Or, the internals are not interpretable for the user who is looking to understand the system at hand. The latter could be the case when the representation of the systems inner function is too large to interpret, or encoded in such a way that intuition alone will not be enough for interpretation. In other words, that either too much time or expert knowledge would be required. In this text, we will refer to this inner functionality as reasoning. When we want to understand a black box system, we want to know how it reasons given a problem. And when we say system, in the context of machine learning,

we usually mean model. A model that takes in a specific input, and returns a specific output. If we understand and follow the line of reasoning, we will understand how an input of the model relates to its output. For example, consider a mathematical model for solving a class of predictive problems such as classification and regression. When we want to understand such a model, it suffices to understand the mathematics behind the model in order to understand the reasoning. However, there are two ways in which we can do this. That is, globally and locally.

Note that throughout the text "understand" and "interpret" will be used interchangeably. The word "transparency" may also be used. However, transparency in this text has a separate meaning. If a black box model is physically impermeable (hidden internals), we can make it transparent by revealing the internals. However, just because it is now "transparent" does not imply it is actually interpretable/understandable.

2.2.1 Global vs. local understanding

First, we can understand a model globally. Let us consider a linear model that given your age x predicts which age you would be on your next birthday. Say $f(x) = x + 1$. In order to understand this model globally, it would be adequate to understand the function. That is, we add a one to the input x. We now know and understand what happens to the input. More importantly, not a specific input but any given input. And thus we understand how the model works globally. Second, we can understand this model locally by inspecting the output for a specific input and the output around that input. In this way, we can inductively try to infer how the model operates globally. Consider the following equation:

$$\{(x, f(x))|x \in \{22, 23, 24\}\} = \{(22, 23), (23, 24), (24, 25)\}$$

It is clear what function f does locally around $x = 23$. Now, if we truly understand the model, we should understand what happens for $x = 25$. Imagine that you do not have a global understanding of the model (which is the function f) and the model for $x = 25$ actually returns 500. This would be frustrating as our intuition and mental model that was built on the local understanding is now worthless. This is why achieving a local understanding of a model can become a tedious task, especially if the model is sensitive to sudden changes and thus not robust. And note that while this is rather intuitive for the mentioned function f, the problem takes a steep turn when we expand in dimensionality, or say that f suddenly becomes a non-linear function. In other words, when the complexity of the model increases. Aside from a mathematical model, consider a hypothetical rule-based model that predicts whether someone has COVID-19 or not. Say that the model contains the following rule:

IF coughs=true AND hasfever=true THEN hascovid19=true

In order to globally interpret this model, it suffices to understand the logic and for most people this could be deemed as interpretable. In order to locally approach this model, say when we do not have access to the logic, we would have to do so inductively. Explaining this model also becomes fairly easy, whether it is to explain the model reasoning or its individual predictions.

2.2.2 Interpretation vs. explanation

Now note the critical difference between explaining a model and the model being interpretable. The difference is that interpretation of a model requires no other explanation. Explanation in that sense, can be referred to as "post-hoc interpretability" [37]. Interpretability thus required a clear understanding of the reasoning. In the logic-based model above, that was certainly the case. When that is not the case, we would need explanations in such a way that the explanation further provides (a global or local) understanding in the model. The explanation should then suffice and should need no further explanation. In other words, the explanation should be interpretable [23]. However, while reading this, the inevitable question might arise. Why would we bother explaining? Why would we go through all of the trouble and not train an interpretable model right away?

2.2.3 Trade-off between accuracy and interpretability

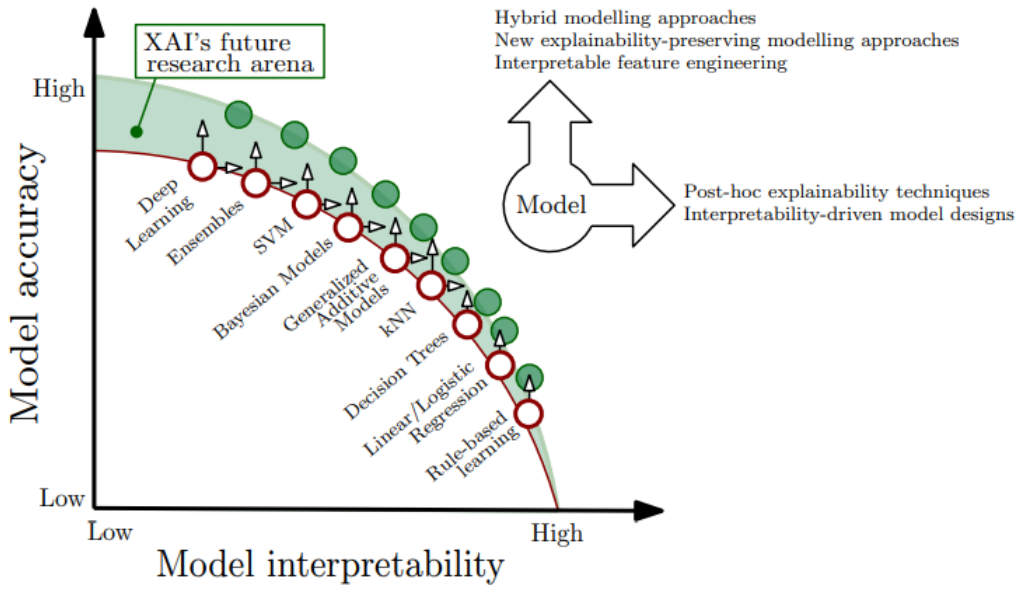


Figure 2.1: The trade-off curve between model accuracy and model interpretability with Deep Learning models currently being in the upper left corner [4].

In the last few years, there has been a surge of machine and deep learning models that exert groundbreaking accuracy and performance when applied to tasks such as image recognition and natural language processing. Unfortunately, this also seems to coincide with the increasing difficulty of model interpretation, meaning that as model accuracy increases, so does model complexity [4]. Figure 2.1 visualizes this. The term model complexity may however ultimately depend on the model itself. For Decision Trees and Rule based systems, this could be length of tree or the amount of rules [23]. Though some models are considered inherently complex. An example of such a model is the Deep Neural Network [4]. Rather than feature engineering, these types of models often

require none and instead learn the features in an iterative manner, that is through the minimization of a loss function. In order to learn both low and high level features, they employ interconnected layers of "neurons" that learn a large number of parameters called weights and biases. It may be infeasible to achieve understanding purely from the sub-symbolic model representation or the network parameters. What we can essentially infer from the trade-off problem is that there is a reason why black box models such as Deep Neural Networks are complex. That is, its representation power contributes to its accuracy. Is it therefor a futile attempt to train an interpretable model for problems that would need the representative power of black box models in order to perform well? In [47], Rudin et al. mentions that the trade-off problem is not always necessarily true and proposes the idea that there could exist models that are inherently interpretable and as accurate as black box models.

> The belief that there is always a trade-off between accuracy and interpretability has led many researchers to forgo the attempt to produce an interpretable model. [47]

Rudin et al. argues that the trade-off figure (such as Figure 2.1) is merely a belief and is not based on actual data. The opposite is proposed in that more interpretation could lead to more accuracy. A reason for this is that the pursuit for more interpretation, for example by means of iterative data processing and searching for meaningful features, can generally increase performance. In conclusion, Rudin et al. advocates to not explain black box models in case of important decisions and instead to create models that are interpretable and understandable in the first place [47].

2.2.4 The need to explain

From a social science perspective, several reasons exist as to why people ask for human explanations, such as curiosity, persuasion or to assign blame. However, the primary reason for explanation or why people explain in the first place, is to facilitate learning. Explanations can help us to improve our own understanding of something and to learn and derive a mental model. This model can in return be used to make certain predictions or to assess control. People also tend to ask for explanations in order to find meaning, in other words, to find inconsistencies in their own knowledge. A final reason would be to control interaction socially, to "share meaning" i.e. to influence beliefs or to affect actions [37].

Persuasion was mentioned above to be one of the reasons why people ask for explanations. That is particularly true if the goal of the explainer is to increase trust. This becomes important when the explainer is not a human, but for example an AI [37]. Moreover, when the AI is used for making decision that could have severe consequences. In the medical field, the consequences could be human lives. In the financial market, this could be financial loss. It is evident that utilizing an AI that is not interpretable in such cases can be disastrous. The following story that is can perhaps best demonstrate this. The

story explains that at one point the military had trained a classifier in order to recognize enemy tanks, as opposed to friendly ones. Although the classifier showed high accuracy on the test set, when it was deployed in real life it performed terribly. It was later discovered that the photos of the enemy tanks which the classifier was trained on were taken on cloudy days and that the friendly ones had been taken on a sunny day. The classifier was thus biased by the weather (and its effect on the pictures), rather than learning the difference between friendly and enemy tanks [23]. This story captures the phenomenon of data leak, which causes the classifier to perform too well. Unfortunately, only because certain unwanted features in the data correlate well with the target feature and thus overrule the decision [46].

In some cases, the data can even hold ethnic biases or prejudices, which may cause a classifier to become discriminating towards certain ethnic groups. In that case, we can not simply hold the AI accountable or responsible for these biases as it only mirrors the biases in its learning behavior. This can be another convincing reason to make AI more interpretable, in this case even "responsible". In this way, we can detect biases before (heavy) consequences can take place. There are also several examples in literature that show instances with black box AI models that are able to classify extremely well yet fail when introduced with minor perturbations in their inputs [23]. Although there is probably no fatal consequence to a "dog" being classified as a "potato" in such cases, the problem becomes much more salient when deployed in tasks such as self driving cars. This is currently the case with Tesla's self-driving cars which employ several black box AI's. There are many debates as to who is liable in a car accident when it is due to a mistake made by AI. Although it is said that the company assumes no liability in most cases, it can still be held liable in case of software errors [27]. In such cases, having an explainable AI could provide more insight and thus minimize errors.

Another need for explainability stems from jurisdiction. In 2018, the European Union (EU) proposed an ethics guideline aimed towards "Trustworthy AI". These guidelines were documented by High-Level Expert Group on Artificial Intelligence (AI HLEG) and consist of seven requirements which AI systems must meet in order to be considered "Trustworthy AI". Although it is currently a guideline, it only proves that there is an increasing need to illuminate black boxes and increasing urgency to do so [20].

2.2.5 How to explain

From the previous subsection we know that there is an increasing need to explain black boxes. The related work in this domain shows that ongoing research is currently tackling this desiderata. However, Miller et al. argue in [37] that researchers mostly use their own intuitions in order to define what a "good explanation" is. They also argue "that this may lead to failure" if current research on XAI does not "build on frameworks of explanation from social science" [37].

> The very experts who understand decision-making models the best are not in the right position to judge the usefulness of explanations to lay users. [37]

Why is this important? Generating explanations is a human social behavior. This behavior and the evaluation of the explanation quality has been thoroughly researched for the past decades by psychologists and experts in social science. Thus, we can not approach the challenge of "how to explain" in a human-centered fashion merely through computer science. We have to consider and combine work that has been done in domains such as Human-Computer Interaction, Social Sciences, and AI [37] as seen in Figure 2.2.

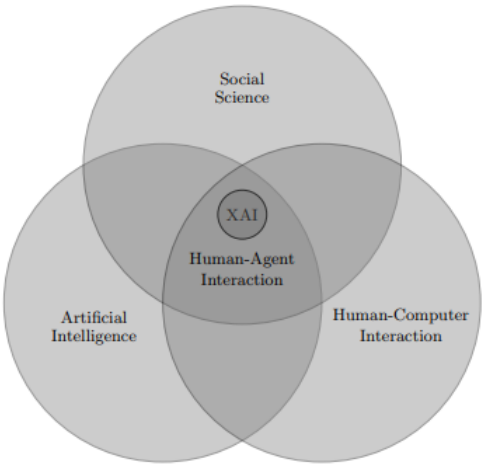


Figure 2.2: The different scopes that are involved in XAI research [37].

There is a lot of (philosophical) debate to what actually constitutes as "explanation". However, most of it points out that explanation refers to causes. Thus, in order to explain an event we have to understand the causes of why that certain event happened. There are many definitions of causality and what constitutes to "useful causal information". It is said that two events A and B occurring at the same time may not give useful information, as in that A causes B. Instead, we have to consider the counterfactual model of causality. That is, if A causes B then hypothetically, if A does does not occur, then B would also not have. It is said that humans use counterfactual simulations to make causal judgments [37].

Miller et al. also elaborates that explanation is a twofold process. First, it is a cognitive process where the causes are to be identified. People do not have immediate access to causes and thus have to infer these from prior information or observation (causal inference). Afterwards, they select a subset of these in order to come up with an explanation (causal selection). This subset will be based on what the goal is of the explanation and some other criteria such as robustness. A cause is deemed robust by people if the effect would still have occurred if conditions other than the cause were slightly different. It is said that causes during causal inference are inferred through abductive reasoning and this is used by people to determine what is considered to be the best cause. Abductive reasoning is a form of reasoning different from induction and deduction in that a hypothesis is typically derived to explain a certain event. In case of explainable AI, an AI would not require abductive reasoning because it would be certain of the causes that relate to a decision or event. And finally, explanation is social process. The resulting explanation is transferred between the explainer and explainee [37].

Explainees also go through a process called "explanation evaluation" in which they decide whether an explanation was satisfactory. The main criteria is whether the explanation allowed the explainee to understand the cause. However, because of cognitive bias, people might have preference to explanations of a certain type. People will accept certain explanations more likely if they conform to prior belief. Another criteria is simplicity. In a study, people happened to prefer simple explanations that contain less causes more than complex explanations [37]. Kulesza et al. [31] discuss explanation soundness and completeness. The former refers to what degree the explanation is truthful in explaining the system and the latter refers to whether the explanations do not omit information that is key to describing the system [31]. They concluded the importance of completeness over soundness and listed several important principles of explainability. That is, to be complete, sound and not to overwhelm [31, 37].

Millers findings in [37] also point out that explanations are contrastive. This means that rather than asking why a certain event X (the fact) occurred, people ask why that event X occurred rather than another event Y (the foil). Thus, people seem to find contrastive explanations a lot more intuitive. However, it is infeasible to leave Y up to the user. Instead, abnormalities can be used to infer Y by using HCI techniques such as eye gaze or gestures [37]. It was also proven that showing probabilities as a form of explanation was not deemed effective. Although they are an important part of the explanation, referring to them was found to be less effective than referring to actual causes. Moreover, it is incorrect to assume that most probable causes are always the best explanation [37]. We refer to an example by Miller et al.

> ...a student coming to their teacher to ask why they only received 50% on an exam. An explanation that most students scored around 50% is not going to satisfy the student. Adding a cause for why most students only scored 50% would be an improvement. Explaining to the student why they specifically received 50% is even better, as it explains the cause of the instance itself. [37]

We can also expect users to require less explanation the more they interact with the system. The reason given is that users will construct a mental model and will be able to generalize (and thus learn) the behavior of the system better over time. Additionally, Miller et al. explains that interactions do not necessarily have to be verbal. Interaction can happen through conversation or in a visual way. Regardless of media, it is argued that in order for agents to explain themselves well, the explanations that they generate must be interactive. An example that is provided is what an explainer (agent) should do the explainee does not agree with the explanation [37]. We finish this section with an aspect of explanation that according to Miller et al. is insufficiently researched. That is, the social aspect, as we mentioned that explanation is a social process. When the explainer presents an explanation to an explainee, both are engaged in a conversation. In that case, the conversation is subject to basic rules of human conversation. This means that we can not simply identify and present causes as many of them may not be relevant to the question. And similarly to conversation, linguistic structure of an explanation may also be an important way to manage impressions, especially for personalized explanations [37].

2.3 Illuminating black boxes

In the previous section we have emphasized that there is a need for explainability of black box AI systems. In this section, we will zoom in on techniques that currently exist in order to explain such AI. When doing so, we should simultaneously consider the expectations that humans have in terms of explainability and try to use this as a filter of what constitutes to useful explanation. The reason for this is because we have to keep in mind the argument that researchers and experts in this area of research sometimes use their own intuition of what makes explanations useful or effective [37].

2.3.1 Explainable AI (XAI)

The term XAI is frequently used in the literature to describe post-hoc explanation methods and techniques for AI systems.

Before explaining some of the existing XAI methods, we have to understand that some of the methods are more suited for users with certain levels of domain expertise. As in [53], users of XAI can be classified into three groups:

- AI novices: people that might use AI daily yet have none to little expertise in machine learning or AI [53]. These types of users will be our primary target when it comes to XAI as we previously mentioned in our human-centered approach.
- Data experts: people that utilize machine learning for research and analysis. These could be data analysts and data scientists with very specific domain expertise such as medicine or finance. These profiles are likely to use visualizations in order to gain insights in data. However, they may lack an expert insight and understanding in the machine learning methods utilized [53].
- AI experts: people that design (complex) machine learning algorithms and develop techniques for XAI. They possess expert insight and understanding [53].

Mohseni et al. [53] also defines a set of XAI goals and measures suited for each group of users with certain experience levels. It is said that these levels and the goals of the users can overlap among the defined groups [53]. Later, we will provide an overview of some of these goals and measures. More importantly, goals that we find important in our human-centered approach and measures that best capture these. First, we will provide a general overview of techniques that exist on how to "illuminate" black boxes. Guidotti et al. [23] defines four ways in which we can solve the black box problem.

- **Black Box Model Explanation:** in this problem, a white box model will try to mimic the black box behavior such that the resulting model is globally interpretable. Here, it is important that the fidelity of the white box model is as high as possible such that it is an accurate representation of the black box behavior. The white

box model in question could be a interpretable Linear model, Rule Set or Decision Tree [23]. These models that mimic the black box are also called global surrogate models. We will not discuss this method any further, as even the interpretable surrogate models can be challenging for humans that have no prior experience with machine learning. The next method is more suited for that purpose [46].

- **Black Box Outcome Explanation:** this problem consists of finding a method to provide an explanation, i.e. the reasoning, as to how the black box came to a particular outcome or decision. This coincides with local explanations which we previously defined that explains a black box on instance level [23]. There are many ways to achieve this. In [46], a model agnostic technique is proposed called Local Interpretable Model-agnostic Explanations (LIME) [1]. Here, a linear model is used to locally approximate the complex decision region of a black box. This is done by sampling multiple points around a single instance and weigh them using a proximity function. Instances that are closer to the selected instance will thus receive a higher weight. The loss function will ensure that the resulting model is locally faithful and has a low enough complexity such that it is interpretable by humans [46]. Refer to the example in Figure 2.3 in which LIME is used to explain why a mushroom instance was predicted as "poisonous".

  In [36], Lundberg and Lee propose SHAP which stands for Shapley Additive Explanations. Simply put, SHAP combines methods such as LIME and Shapley values from game theory in order to explain the decision of any black box model. It does this by calculating feature importance scores for a single prediction and shows how these scores affect the output of this prediction against a baseline output. Using SHAP [2], we can visualize which features have most affected the output of the model for a single instance.

  In [30], Krause et al. argue that local explanations may not scale well for humans as they would have to see explanations on all the data points to understand the model. This would thus not be feasible or realistic. Their paper proposes the technique of providing local explanations in aggregation and demonstrate that doing this can help users detect biases in data.

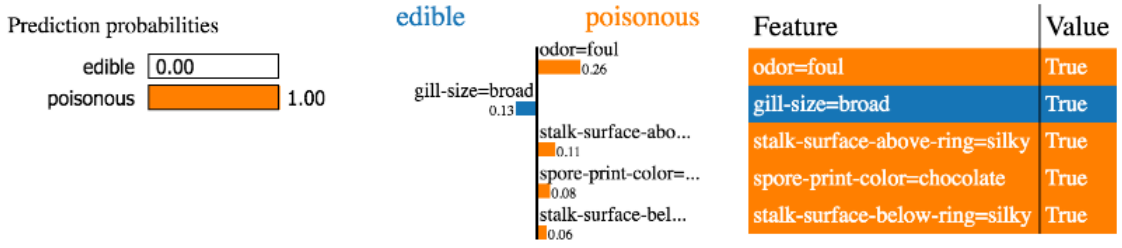


Figure 2.3: The LIME tabular explainer showing a bar chart which explains through its attributes why a mushroom instance was predicted as poisonous. It also shows which attributes push towards the opposite prediction.

[1] https://github.com/marcotcr/lime
[2] https://github.com/slundberg/shap

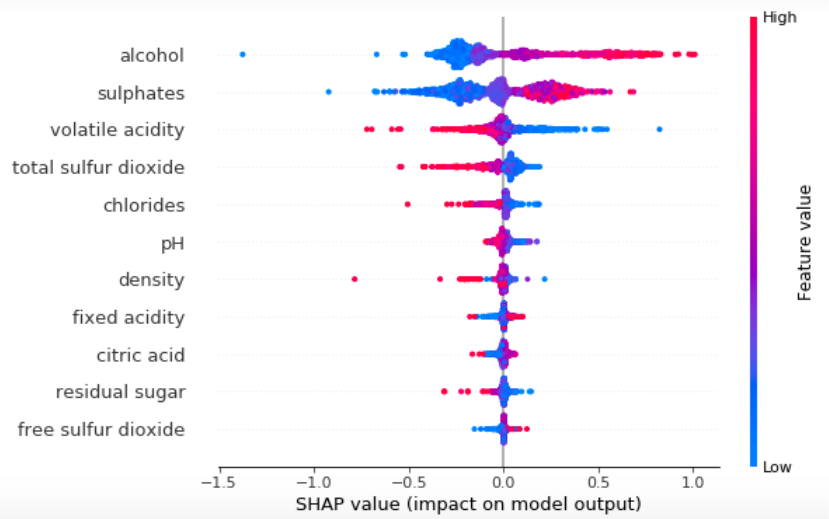


Figure 2.4: A bar chart of SHAP values on wine data features with wine quality as target feature. The features are arranged in descending order of importance. In this example, the feature alcohol has the strongest impact on wine quality and is rather positively correlated, which means that higher values of alcohol contribute to higher wine quality. The feature volatile acidity for example is negatively correlated which means that lower acidity contributes to higher wine quality [16].

- **Black Box Inspection Problem:** this problem consists of providing either a visual or textual representation in order to understand how the black box works. This representation could also provide insights as to why the black box is more likely to output certain predictions [23]. Partial dependence plots (PDP) can be used for model inspection as a way to analyze partial dependence. That is, to understand what the relation is of a certain input feature and the corresponding output feature. PDP can also be used to observe how the values of the output feature change for multiple input features. The Scikit library [3] contains examples on how to utilize this. An example of PDP on house prices and house features is shown in Figure 2.5. Another way to do model inspection is to plot feature importance. This plot can provide an understanding to the relation between input features and the output feature and show which features most affect the output. In the previous problem, SHAP by Lundberg and Lee was used to gain a local understanding by visualizing feature importance. However, SHAP can also be used to plot feature importance on a global level. An example on wine quality data is illustrated in Figure 2.4.

[3] `https://scikit-learn.org/stable/modules/partial_dependence.html`

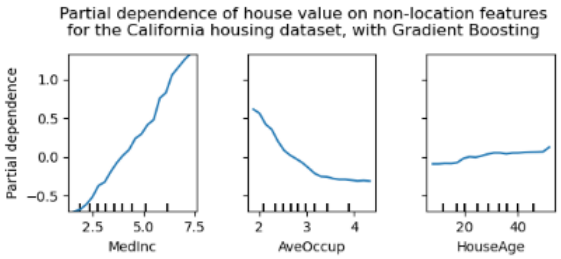


Figure 2.5: Partial Dependence Plots of house prices.

- **Transparent Box Design Problem:** for this problem, we try to design a globally and locally interpretable white box model, rather than designing a black box model at all. Ultimately, we might not achieve the same performance as we would with a complex black box model. This is due to the trade off problem that we previously mentioned [23]. Similar to the first problem, this technique might not be feasible given that even the interpretable models may not be understandable to most humans.

- **Example-based explanations:** this technique aims to explain a prediction for an instance by presenting similar looking or representative examples (prototypes) from the dataset [1]. These examples are also referred to as "neighbors" of an instance. There are two types of neighbor instances that could be presented namely factual or counterfactual neighbors. The former are examples that are similar to the instance to be explained. These may be less informative than counterfactual examples which are examples that show what should be changed in an instance to achieve the opposite prediction [29]. An example to illustrate both techniques: if an applicant were not granted a loan, then in case of factual neighbors, applicants would receive examples showing similar applicants that were also refused a loan from the bank. In case of counterfactual examples, applicants would be presented similar looking applicants that were granted a loan such that the applicant requesting for a loan could determine what to change in order to be granted a loan [29]. The latter technique may thus perhaps be more informative than simply showing factual neighbors. Researchers have argued that counterfactuals are more informative explanations than local explanations such as LIME and that counterfactual explanations are compliant with GDPR [29]. Model agnostic libraries such as DiCE [4] exist to generate counterfactual examples [40].

2.3.2 Goals and measures

There are certain goals that we can aim for when we want design an XAI system. Similarly, there exist measures to evaluate these goals. In this section, we list several high-level goals from [53] that we can generally aim for when developing XAI systems for novice users. We will also list some important measures inspired by Figure 2.6, which is a conceptual model that is proposed in both [26] and [24]. This model is a visualization of a more brief

[4] `https://github.com/interpretml/DiCE`

and general overview of the explanation process, unlike the low level process that we have discussed in previous sections. We can assume that most AI or XAI researchers do not have a pure background in Social Sciences or Psychology. That much is clear after studying Miller's work in [37]. Diagrams such as these can appear in literature to help researchers understand the explanation process. The conceptual model in Figure 2.6 combines a few important measures and their interactions. That is, explanation goodness/satisfaction, mental model, and trust/mistrust.

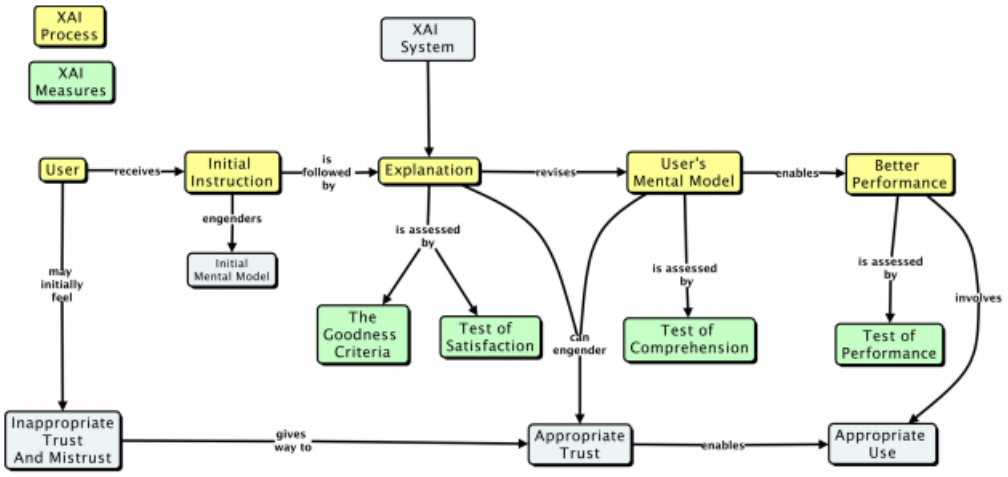


Figure 2.6: A conceptual model of an XAI system [26]

.

Mohseni et al. has bundled some high-level goals and measures in a framework that XAI researchers can use in order to iteratively develop explainable AI. We provide a non-exhaustive list of these goals:

- Interpretability: this is an evident goal with the aim of providing a local or global understanding. We have already discussed that global interpretability might be an unattainable goal for novice users. If we do wish to attain this however, we might have to give up most of the performance for the sake of interpretability because of the trade off problem [53]. The question then becomes whether the resulting model will even remain useful, given that much of the performance was traded in. Needless to say, it is goal that is definitely worth exploring. In fact, some knowledge-based models in AI such as belief-desire-intention models alleviate folk psychology in order to provide interpretability. The "Folk Psychology" framework in question is typically used by humans to explain behavior to other humans [37].

- Trust: perhaps the most important goal to aim for in settings where the models decision may lead to certain consequences or when critical decision making is required. Ribeiro et al. even argues that users will not use a model that they do not trust and defines trust in two ways. That is, trust in the individual model prediction or trust in the model [46]. As mentioned before, simplicity, generality and coherence should be considered important when providing explanation. Miller et al. mentions that these should be considered especially if the goal is to generate trust between

the explainer and explainee [37]. Although the high-level definition and notion of trust is usually used in most contexts, trust is multi-dimensional and complex [7]. It can be captured through a variety of factors such as user confidence, reliability, predictability, perceived accuracy and more [26].

- Bias mitigation: in the previous section we have seen that data can be subject to biases. A goal of explainable is thus to reveal (and remove) these biases and avoid the models to learn and eventually use them in their decision making [53]. We have also explained that users may use certain cognitive biases in order to select explanations. In that case, mitigating biases can be rather infeasible as these are typically innate or developed. Instead, it could be convenient to use these to our advantage when generating explanations [37]. An example of this is the use of "placebic" explanations. These are explanations that convey no information and can essentially be used as a baseline to compare with actual explanations [53]. In [55], Wang et al. proposes a theory driven driven framework to build user centric XAI systems with the main goal of mitigating human cognitive biases. The conceptual framework proposed for XAI researchers considers both the philosophical and psychological theories with respect to human cognition.

We also provide a non-exhaustive list of measures based on the conceptual model that may constitute to a good explanation:

- Goodness, satisfaction and usefulness: there are both implicit/explicit and objective/subjective ways to measure whether explanations are "good", "useful", and whether users will be satisfied with the explanations. Explicit ways follow qualitative methods such as surveys or user ratings. Implicit ways can be to measure user response time or checking whether users have interacted with the explanations [53]. In [26], Hoffman et al. refers to the goodness criteria as an a priori measure of explanations and refers to satisfaction as a posteriori measure. They propose a variety of Likert scales on both explanation goodness and satisfaction [26]. Other more high-level (Likert) scales exist such as NASA's TLX questionnaire [10] to qualitatively measure the perceived workload of a task on users and SUS scale [8] to measure the usability of a tool. These could be applied to an XAI system or adapted to fit in the context as they are originally meant to measure most software systems and tools.

- Mental model: we know that explanations can help to facilitate learning and that users will try to develop a mental model of the system [37]. There exist qualitative ways to measure mental models such as interviews and questionnaires. Self-explanation and think-aloud studies also are used in where users are directly asked to elaborate their mental model of the model decision [53]. Mental model can also be measured by task performance. This is based on the hypothesis that users who have a good mental model due to the explanations will perform better on a task (and thus have a higher performance) than when explanations are not provided [26]. In [28], Jhaver et al. proposes a study on whether providing explanations of content removal (on Reddit) can change user behavior. That is, whether users will adapt their behavior to a more productive and acceptable behavior. Their study reveals

that providing explanations actually reduces the chance that future post will be removed and thus that removal explanations can be useful to educate users on the rules and social norms of social media platforms.

- Trust: establishing trust between explainee and explainer is one of the primary goals of XAI. It is safe to assume that the trust levels will be low in the beginning and might evolve over time [37]. There are both quantitative and qualitative factors of trust that can be measured such as perceived accuracy, user agreement, perception of control/transparency. These can be used to measure levels of justified or unjustified trust and mistrust that people can have towards the machine and its prediction. More general and qualitative notions also exist which aim to capture a broader sense of user trust and mistrust in AI. These factors must be chosen carefully as some of them can be highly context dependent [26]. In [26], Hoffman et al. recommends a trust scale based on the Cahour-Fourzy scale. The aim of the scale is to capture multiple factors of trust such as reliability, efficiency, predictability, etc... The scale does however assume that the participants already have some experience with the XAI system. Refer to Appendix C for the recommended trust scale [26].

2.4 Misinformation in Online Social Networks

Online Social Networks have widely adopted AI on their platforms for a variety of tasks such as generating recommendations, automatization, marketing, computer vision, and more. Unfortunately, many of these features are currently hidden behind a black box which makes it impossible for users to understand how they work. There are multiple reasons for the lack of transparency. The first one is rather evident. The social media platform in question may not want to provide transparency for competitive reasons or to avoid users "gaming the system". The latter refers to the practice of users trying to figure out how the black box works in order to manipulate it to gain an advantage. On social media, this is typically done to gain more exposure [45]. A second reason why social media platforms would not want to provide transparency is simply because it is not trivial to do so. The black boxes that are employed behind the scenes are based on sophisticated algorithms and models such as Deep Neural Networks which employ many latent factors and complex architectures [13]. Fortunately, some of the social media platforms are in the process of illuminating their black boxes, possibly to abide with increasing jurisdictional requirements. In this section, we are going to focus on Twitter. More specifically, we will discuss the spread of fake news which haunts many Online Social Networks to this day and discuss some of the proposed black boxes used to tackle this problem. Moreover, we will try to assess whether the deployment of these black boxes is justified given the inevitable requirement of XAI.

2.4.1 Spread of misinformation

Platforms such as Twitter allow information to be spread rapidly and reach large amount of audiences all around the world. The information or media that is spread on Twitter

specifically, can take many different forms (text, videos, photos,...) and consist of opinion, satire, debate, facts, and more. There are multiple reasons why users use social media. The obvious reason, as the term social media suggests, is to socialize. While most users do indeed use social media for this purpose, some users unfortunately do for harmful manipulative reasons. These users benefit from the speed and scale at which information spreads in order to push a specific agenda, or to mislead and misinform audiences [35]. This can be problematic given that most of Twitter users primarily use Twitter for the consumption of news (based on statistics of the US, in 2019). This is not only true for Twitter. In 2018, it was surveyed that 68% of adults in America occasionally consume news from social media [51].

The spread of misinformation is not a new phenomenon. In fact, there are many terms that were used to label misinformation throughout history, with "fake news" being a recent and popular one. Although this term explicitly contains news, it can refer to much more such as opinion, conspiracy, accusations and propaganda. As long as its primary intention is to deceive or mislead public opinion. The people that are responsible for this may have personal, political or financial reasons. They may also use automation such as social bot accounts in order to spread fake news more effectively [35]. Social bots are algorithms that are deployed on social media for either benign reasons or as mentioned above, with harmful intentions. The fact that these accounts are anonymous and easy to create in large volumes makes it an effective tool for the spread of misinformation. These bots can also be used to manipulate and inflate "likes" and "followers" to make it seem as if a specific user is trust worthy [21]. Finally, these bots can be used to "retweet" fake news in order to accelerate its spread [35]. The recent COVID-19 pandemic (see Figure 2.7), has proven that the phenomenon of fake news and social bots is still an ongoing and alarming issue [5]. In fact, the study mentioned in [42] shows that more than half of the tweets collected on COVID-19 originates from social bots [42]. In recent years, many solutions have been researched and proposed to tackle the spread of fake news. Before providing an overview of the related work, we must understand why humans (intentionally and unintentionally) spread misinformation in the first place. Moreover, we will try to find out how we can explain misinformation to humans such that they will be able to understand what distinguishes information from misinformation.

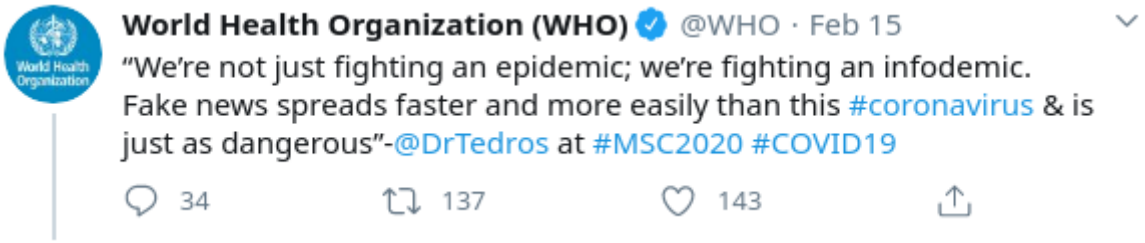


Figure 2.7: WHO's tweet on the pandemic.

2.4.2 Explaining misinformation

Several concepts exist that can explain why humans are susceptible to fake news. The first one is motivated reasoning. People are more likely (and thus motivated) to believe in something that confirms their own thoughts and opinions. If you dislike raisins for example, any article that expresses its hate for raisins (whether it is true or not) is likely to win your trust. Another concept is that of naïve realism which is the tendency to discredit anything that does agree with prior belief. This is also the reason why some people are likely to label information as fake news, simply because it does not agree with their ideology or perception of reality [56]. We might be tempted to think that fake news only affects "irrational" people or people that are unable to think analytically. The following study [44] proposes that analytical thinking may in fact worsen motivated reasoning thereby lowering the ability for people to judge news accurately. This is because analytical thinking may lead to more rationalization and justification of prior beliefs. The same study suggests that (in a political context) "lazy thinking" makes humans more susceptible to fake news, rather than bias [44]. It is worth noting that not only human susceptibility to misinformation is to blame. Most algorithms on social media select content with the primary goal of increasing user engagement. However, they might accidentally end up reinforcing cognitive biases by presenting its users with information that merely confirms their prior belief. And even worse, contribute to the manipulation of public opinion by potentially spreading manipulated content to already susceptible humans [12].

We now understand some of the reasons why people fall for fake news. The question is how we could possibly counter this on Online Social Networks with AI, and how we can explain the model in a human-centered way. A possible way is to present the abnormal factors and features in which fake news can be clearly distinguished from factual news. In [37], Miller et al. mentions studies which confirm that people select unusual and abnormal events to come up with plausible explanations. It is also argued that abnormality plays a role in interpretability and is a useful way for people to select explanations. That is, to statistically distinguish normal behavior from anomalous behavior [37]. This method could thus be applied to develop an explanation interface for a black box model (explainer) such that it can explain to users (explainee) why it has classified information as fake or factual.

Fortunately, there is no need to reinvent the wheel, as misinformation is an emerging topic within AI literature. The related work, which we will cover in the next section, clearly shows that the battle against misinformation is an ongoing issue. In the next section, we will explore a variety of concrete solutions that propose explainable misinformation classifiers. Refer to the work of Chen et al. for a more general and extensive survey that focuses more on Visual Analytics and social media [11].

2.5 Related work

	XAI					AI				Measures								Visualizations									Study		Metadata
	Global	Local	Surrogate	White box	Survey	Fake news	Social bot	Dataset	Recommender	Trust	Perceived accuracy	Mental model	Satisfaction	Goodness	Usability	Computational	Bias mitigation	Task performance	Charts	Saliency	Qualitative	Interactive	Dashboard	Example-based	Survey	Qualitative	Quantitative	Mixed methods	Citations	Year	Keyword
Ribeiro et al.		x							x			x	x				x	x	x						x	x		3948	2016	LIME
Guidotti et al.				x																				x				800	2018	Survey XAI
Davis et al.		x					x								x			x										511	2016	BotOrNot
Schnebly and Sengupta	x						x								x													5	2019	RF Twitter bot
Hadeer et al.					x										x			x		x								46	2018	Opinion spam
Buntain et al.					x		x								x													83	2017	CREDBANK
Pennycook and Gordon					x					x	x																x	225	2018	Lazy, not biased
Zhou et al.					x																			x				5	2020	Survey fake news
Mohseni et al.					x				x	x	x							x	x		x	x	x		x	x		0	2020	Trust evolution
Gu et al.					x	x																						5	2017	Trend Micro
Lundberg and Lee	x	x									x		x		x			x	x							x		1719	2017	SHAP
Mohseni et al.				x																				x				45	2018	Survey XAI
T. Miller				x																								776	2017	Social science
Adadi et al.				x																				x				581	2018	Survey XAI
Kulesza et al.		x									x						x	x		x	x	x	x		x			284	2015	Explanatory Debugging
Arrieta et al.				x																				x				235	2019	Survey XAI
Ferrara et al.						x																						1255	2016	Social bots
Petre et al.					x	x		x																	x			13	2019	Gaming the system
Covington et al.								x							x													1166	2016	Youtube
AI HLEG									x																			/	2018	Trustworthy AI
Müller et al.																									x			256	2016	Survey AI expert opinion
Horne and Adali					x		x																					270	2017	Fake news
Liao et al.				x					x	x	x	x	x	x		x								x	x			57	2020	XAI question bank
Keane et al.		x																					x					17	2020	Case-based XAI
Mothilal et al.		x																					x		x	x	x	116	2019	DiCE
Rittman et al.																												14	2004	Determining subjectivity
Yang et al.		x							x					x									x			x		20	2020	UCI leaf dataset
Mohseni et al.				x	x																							10	2019	Research opportunity
Jakob Nielsen														x											x			/	2012	Think-aloud study
Henderson et al.																									x			4	2019	Measure Twitter use
Hoffman et al.									x			x	x	x			x										x	78	2018	Metrics XAI
J. Brooke													x	x	x										x			>5000	1996	SUS
Cao et al.													x		x		x								x			>5000	2009	NASA TLX
Gunning et al.				x																				x				146	2019	DARPA
Berkovsky et al.		x						x	x											x							x	35	2017	Trust factors
Wang et al.				x										x		x								x				94	2019	User-centric XAI
Zhou et al									x									x			x				x	x		11	2017	Uncertainty and Cognitive Load
Krause et al.	x								x							x		x							x	x		10	2018	Aggregating Explanations
Berendt et al.						x									x							x						0	2020	FactRank
Yang et al.	x	x				x									x			x	x	x	x	x						10	2019	XFake
Jhaver et al.		x							x								x									x		23	2019	Reddit moderation
Shu et al.		x				x									x		x		x	x						x		49	2019	dEFEND
Chen et al.																								x			x	38	2017	Social media VA
F. Alam						x		x																				9	2020	C19 fake news tweets
L. Cui						x		x																				7	2020	CoAID
S. Lee						x		x																				/	2020	C19 fake news
S. Cresci							x	x																				154	2017	Social bot data
Rudin et al.				x	x																			x				552	2019	Arguments for strict whitebox approach
GK. Shahi et al.						x		x																				12	2020	C19 fake news Tweets
GK. Shahi et al.						x		x																				6	2020	C19 fake news

Figure 2.8: Table of literature study.

Currently, solutions for the automated classification of fake news range from fact-checking methods [6], to unsupervised, and supervised black box models [2, 9]. Additionally, numerous black box solutions exist to detect social bots on Twitter [17, 48]. Unfortunately, most of these solutions have one common problem: they are not interpretable. In [57], Yang et al. propose XFake, an explainable fake news detector. XFake uses an ensemble of three different frameworks to generate both the prediction and explanations. A complete overview of the literature explored is presented in the Figure 2.8.

The first framework "MIMIC" consists of a teacher model (embedding + deep neural networks) that can learn from the attributes of news articles such as speaker, subject, context, etc,.. It also consists of a student model (tree ensembles) that tries to mimic the performance of the teacher. With such an architecture, MIMIC can benefit both from the black

box performance and the explainability of the tree ensemble. The explanation is then generated by "calculating the node importance of each attribute in the ensemble trees". This allows the calculation of feature importance for each (embedded) news instance [57]. The second framework "ATTN" will analyze the news articles on its semantic information. It employs a word embedding, a convolutional neural network, and an attention mechanism to capture the relationship between the words in the article. Instance-level explanations can be generated through 1,2 and 3-gram analysis. The third framework "PERT" learns from the linguistic attributes of the news articles such as adjective ratio, verb ratio, sentiment score, etc... It trains an XGBoost model on these attributes. Feature importance is calculated through a perturbation-based way technique [57].

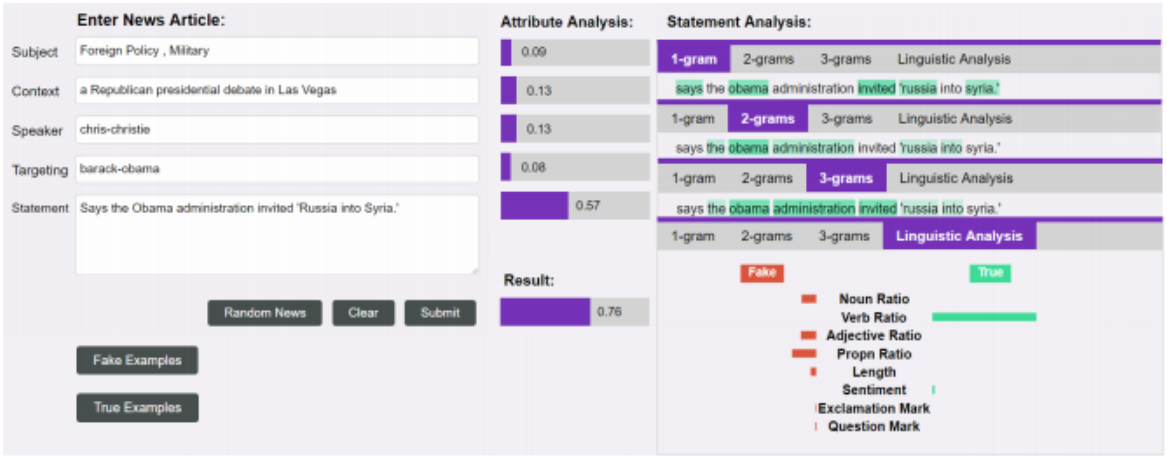


Figure 2.9: The dashboard of XFake, which consists of the news instance attributes that users can fill in and the different components that explain the prediction for each instance [57].

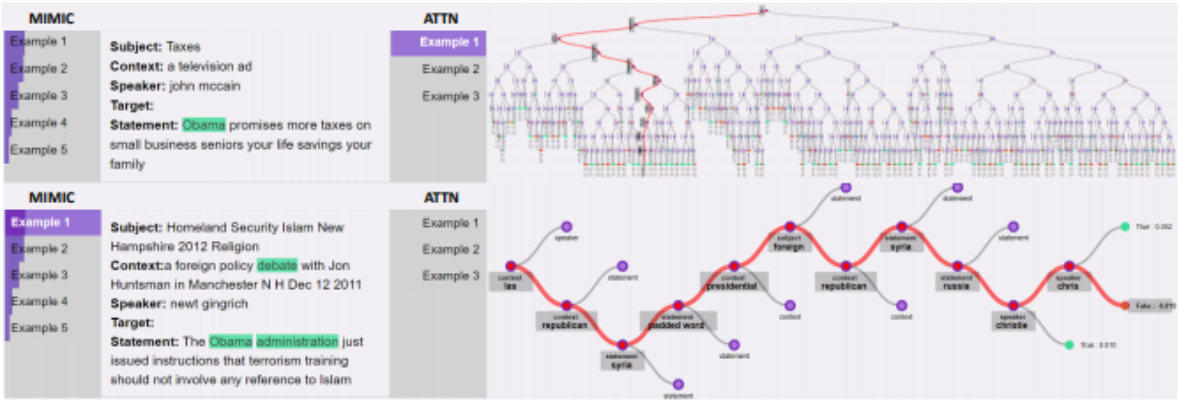


Figure 2.10: Supporting examples for a news instance on the XFake dashboard. The interactive visualization can show which path each tree in the ensemble took to predict the output [57].

The final prediction of XFake is calculated by the weighted sum of each framework's prediction. As explanation, XFake also uses supporting instances which look similar to the

input news instance. The different components of the explanations are visualized using D3.js and is illustrated in Figure 2.9 and Figure 2.10. For each news article instance, XFake predicts the output probability that the article is fake news. For each prediction, it shows which attribute was the most important in determining this prediction. It also includes the n-gram analysis of the news article statement which shows the words that were considered most important for the classification. The histogram below shows which of the linguistic attributes contributed to the prediction and which ones were leaning towards the opposite prediction. Additionally, XFake can interactively show supporting examples and visualize the path of the ensemble trees and the path that each tree takes to predict the final label. A user study with 147 participants was conducted to demonstrate the usefulness of the explanation and to assess user's understanding of XFake. Participants were given two tasks, namely, fact checking and prediction guessing. Accuracy and prediction time were used as evaluation metrics. Most of the details of the results are missing but the paper indicates that even though "explanation does help users to better understand and predict system behavior", the explanation takes more time to inspect in order for it to show benefits [57].

In a later work [39], Mohseni et al. propose a more recent assistant tool to detect fake news in an explainable manner. An ensemble of four classifiers is used to predict fake news. The first model is an LSTM network trained on headlines which can provide explanations in the form of word heatmaps. The second model can predict fake news "based on article set representation constructed using hierarchical attention at sentence level and article level". A third model is used to generate feature importance for features of the article such as claim, article and source. The last model which is a bidirectional LSTM will return related articles for a given news article and provide a heatmap of words as explanation. The interface is shown in Figure 2.11 which consists of the news article headline and a heatmap of words (A), the prediction and confidence (B), more explanation for the confidence (C), tools to navigate and share or report the article (D), related news articles based on the headline (E), bar chart to show feature importance as explanation for the prediction (F), and a supporting donut chart showing feature importance (G) [39].

The goal of the user study in [39] was to research which effect providing interpretability had on trust, and more importantly how it evolved over time. During the study, "perceived accuracy" was quantitatively measured. There were many different design conditions in the study such as showing no explanation and prediction, only showing predictions and showing only certain types of explanations. During the study, a slider was used in which subjects were prompted multiple times for their perceived accuracy, both mid-task and through a questionnaire at the end. The user interface contained a feed in which a balanced amount of fake news and factual news articles were present. The task that was given to users was to choose which news article they would share. The proposed hypotheses were split into five clusters based on how the trust would evolve (increase, decrease, constant, overestimate or underestimate). The study was a between-subjects study with 160 participants recruited. The results (tested with Pearson Chi-square test) showed a significant correlation between the design conditions and trust. That is, users confronted with both predictions and explanations continuously gained trust in the AI [39].

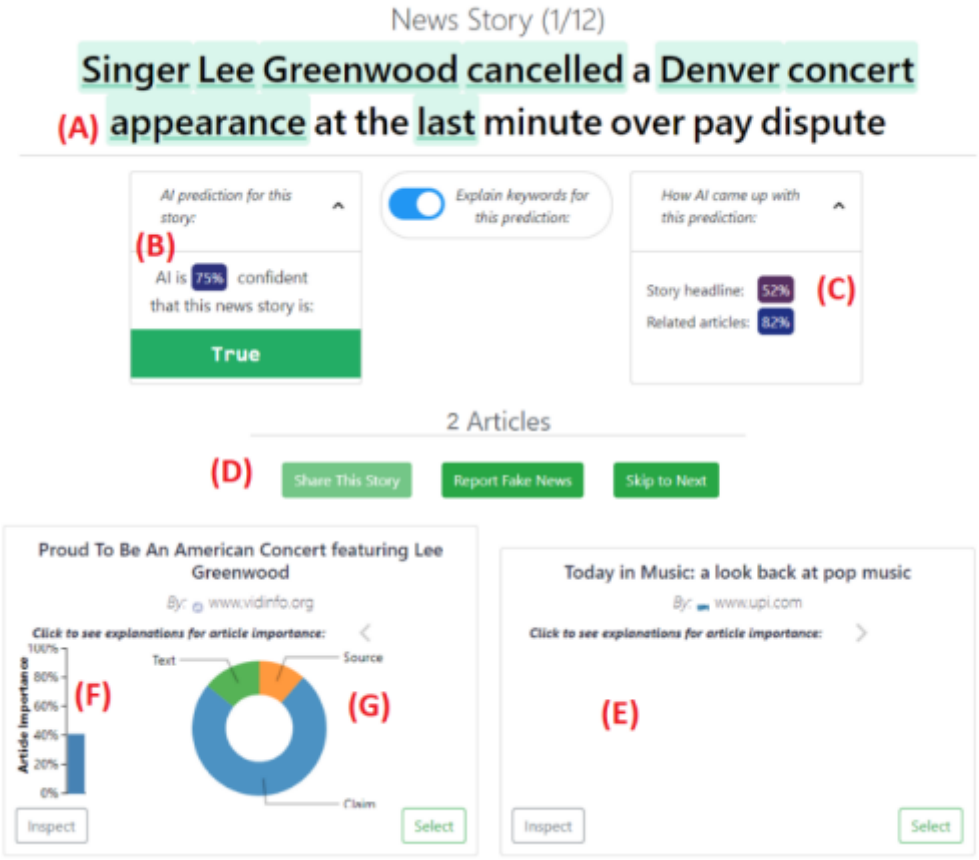


Figure 2.11: The explainable user interface in which the prediction of the AI is explained.

In [52], Shu et al. also propose an explainable fake news detector "dEFEND". In [59] Zhou et al. provides an extensive overview of methods and related work to classify fake news. Moreover, a research opportunity is proposed to make such classifiers explainable to end users.

In Twitter, there currently exist several precautions to tackle the spread of misinformation. Although it usually relies on its users to report tweets, it also employs automated AI systems to detect tweets that are against the Twitter rules and policies [18]. Users might have already noticed during the recent United States presidential election and COVID-19 pandemic that some of the tweets contained warning labels as shown in Figure 2.12. Twitter has mentioned in [22] that this is a feature that must "encourage more thoughtful consideration" before "amplifying" and thus spreading misinformation [22]. The warning label itself provides a reference to the rule that was violated in the Twitter rules and policies. Unfortunately, the amount of transparency that these warning labels provide is rather questionable. Additionally, the following study proposes that the labels only slightly reduce the perceived user accuracy of false headlines [44]. And although Twitter is constantly coming up with initiatives to provide more transparency for users (see [54]), there are currently no initiatives to provide more transparency in the automated classification of misinformation.

Figure 2.12: A warning label on former United States President Donald J. Trump's tweet regarding COVID-19.

2.6 Conclusion

In the literature study we discussed what black boxes are, why these are problematic and which techniques are applicable to "illuminate" these. We also covered the problem of misinformation in Online Social Networks and how explainable models in the related work were used to predict misinformation and explain the classifiers and their prediction.

To the best of my knowledge, there currently does not exist an explainable COVID-19 misinformation classifier on Twitter data. In this thesis, we take this unique opportunity and ones proposed in [38, 59] to develop such an explainable classifier using machine learning and XAI methods from the literature.

Chapter 3

Methodology

3.1 Introduction

A unique opportunity was identified in the related work to develop an explainable machine learning classifier to predict COVID-19 misinformation in tweets. The importance of our work lied in the explanation of this classifier to end users using methods from XAI.

In this section, we propose the methodology used to realize this opportunity. First, we present the data that was collected to train the machine learning classifiers and propose methods to retrieve, engineer and analyze the data. Moreover, we present the classifiers that were trained on the data. Second, we present the explanation techniques used to explain the predictions of the trained classifiers. Third, we discuss the method for iteratively developing and evaluating both the low and high-fidelity prototype. Finally, we present the research questions and the methodology behind the final user study to evaluate the explanations of the explainable classifier.

3.2 Data and classifiers

We collected labeled tweet data in order to develop a supervised machine learning classifier for predicting COVID-19 misinformation on Twitter. The collected data consisted of both unstructured (the tweet text) and structured data (the metadata). The unstructured data was extended with additional articles and claims from websites other than Twitter. We chose to combine three sources of labeled datasets. Refer to Table 3.1 for a summary of the data collected for training a misinformation classifier.

In the literature study, we have seen that social bot users were prone to spread misinformation on Twitter. We thus collected labeled Twitter data in order to develop a classifier capable of classifying these types of users. Refer to Table 3.2 for a summary of this data.

Table 3.1: A summary of the raw data sources [3], [15] and [33]. Note that these numbers are prone to change in the future as some tweets or Twitter users may be deleted on the platform.

	Fake news	Factual news
Website claims	28	338
News articles	1416	3303
Tweets and Twitter users	7635	133674

Table 3.2: A summary of the raw data source [14]. Note that these numbers are prone to change in the future as some tweets or Twitter users may be deleted on the platform.

	Social bot users	Real users
Twitter users	2805	2815

The first source [3] consists of COVID-19 related tweets that have been extensively labeled by a community of Twitter users [3]. The data itself consists of tweet ID's, the tweet text and columns that were manually filled in during labeling. The ID is a reference to the data on Twitter. A reason why only the ID is available is due to compression which is done for privacy reasons. This means that the actual data can only be retrieved with a Twitter developer account. We registered for such an account and were granted authorization to retrieve and use the Twitter data for research purposes. Retrieving the uncompressed Twitter data (in JSON format) through their corresponding ID's will also require a third party library such as Twarc. This process is referred to as "rehydration".

The second source "CoAID" [15] contains both factual and misleading COVID-19 claims from websites and news articles. It also contains tweet ID's of COVID-19 related tweets. The factual articles and claims have collected from reliable media outlets and websites such as WHO's official website. The fake news articles have been collected from fact-checking websites such as PolitiFact. The tweets were queried based whether they contained news titles of misleading or factual news and thus labeled accordingly. The paper in which CoAID is proposed, also presents important insights on the CoAID dataset such as the frequency of the hashtags used in both fake and real news and a list common claims. For our dataset, we collect only the labeled claims, news articles, and tweets (both news and claims) but not the user replies.

The third source [33] contains factual and fake news articles collected by senior data scientist Susan Li. We used this dataset mainly for extending the unstructured column (tweet text) of the previous datasets. The last source [14] contains ID's of users that were labeled as either social bot or real accounts. It was labeled through CrowdFlower, a crowdsourcing platform. This is not a public dataset and requires permission from its authors. We were granted permission to use this dataset. Again, "rehydration" was used to retrieve all of the data.

Table 3.3: The features selected and engineered for developing the misinformation classifier.

Column identifier	Description	Source	Type	Mined?
Favorite_count	Amount of times tweet was favorited	[9]	Int	
Retweet_count	Amount of times tweet was retweeted	[49]	Int	
N_nnp	Amount of proper nouns used in tweet	[1]	Int	x
Length	Amount of characters used in tweet	[9]	Int	
N_question	Amount of questions marks used in tweet	[9]	Int	x
N_exclaim	Amount of exclamation marks used in tweet	[9]	Int	x
N_hash	Amount of hashtags mentioned in tweet	[49]	Int	x
N_uppercase	Amount of uppercases used in tweet		Int	x
Has_link	Does the tweet contain a link?	[49]	Bool	
Subjective	Is the tweet written subjectively?	[9]	Bool	x
Possibly_sensitive	Does the tweet contain sensitive content?		Bool	
Source	Source from where tweet was posted	[49]	String	
Expanded_url	Sources mentioned in tweet	[49]	String	
Age_account_days	The amount of days user has spent on Twitter	[9]	Int	
Statuses_count	The amount of statuses that the user posted	[9]	Int	
Followers_friends_ratio	Amount of followers user has divided by amount of friends	[9]	Float	x
Likes_age_ratio	Amount of favorites of user divided by age_account_days	[9]	Float	x
Default_profile_image	Does the user have the default profile image?		Bool	
Default_profile	Is the user a default profile?		Bool	
Protected	Is the user a protected profile?		Bool	
Verified	Is the user verified on Twitter?		Bool	
Cluster_id	Which cluster does the tweet belong to		Int	x
Has_bad_tag	Does the tweet contain hashtags frequently mentioned in fake news?		Bool	x
Full_text	The tweet text		String	
Is_tfidf_fake	The prediction score of the TFIDF misinformation classifier		Float	x
Has_bad_source	Whether the tweet contains an URL of an untrustworthy media source		Bool	x

Table 3.4: The features selected and engineered for developing the social bot classifier.

Column identifier	Description	Source	Type	Mined?
Followers_friends_ratio	Amount of followers of a user divided by amount of people user follows	[48]	Float	x
Likes_age_ratio	Number of favourites divided by age_account_days	[48]	Float	x
Favourites_count	Number of favourites of user	[48]	Int	
Statuses_count	Number of statuses posted by user		Int	
Bio_length	Amount of characters in user's bio	[48]	Int	x
Username_length	Amount of characters in user's username		Int	x
Name_length	Amount of characters in user's full name		Int	x
Age_account_days	Amount of days user has spent on Twitter	[48]	Int	
Listed_count	Amount of lists that user is added to		Int	
Followers_count	Amount of people following user	[48]	Int	x
Tweet_age_ratio	Statuses_count divided by age_account_days	[48]	Float	x

Feature selection was performed to retrieve useful columns from the tweet and user data. Additional features were extracted from these columns through data mining and data engineering to extend the amount of useful features. Data mining is a technique in which new knowledge is inferred from existing data. The columns to be retrieved and mined were mostly based on previous work on misinformation (Table 3.3) and social bot classification (Table 3.4). We had to reduce the amount of entries per class, in order to have a balanced dataset. Data analysis was performed on the retrieved columns to limit the amount of features to be used to train the classifiers. This was to reduce high dimensionality which can reduce the performance of the classifiers [9]. The analysis was done by visualizing the distribution of the continuous features and testing for correlation of the input features with the output labels. We used Spearman correlation for the continuous variables and Jaccard similarity for the boolean variables. We replaced outliers with the median value whenever necessary using the Interquartile Range Method. We used this method as we could not assume a Gaussian distribution of the data. A factor of 1.5 was used to define the lower and upper bound.

The unstructured data (tweet text) was preprocessed and analyzed separately according to methods in [2]. More specifically, we trained a separate classifier on the features that were extracted from the unstructured data. The boolean output of this classifier served as input to the misinformation classifier, in addition to the structured data. We used n-grams to model the unstructured data. This is a frequently used language modeling technique in NLP. The n-gram represents a sequence of words or characters with n the length of the sequence. We opted for the word n-gram and will further refer to each tweet text as "documents" [2]. Similarly to the structured features, the unstructured data (text) was preprocessed. Popular text preprocessing methods include stemming where words are reduced to their stem (e.g. walked and walking become walk), stop word removal and removing certain characters such as punctuation marks.

We used Term Frequency Inverse Document Frequency (TFIDF) technique to extract the important n-gram features from the data. This is an information retrieval technique to extract important n-grams from a large corpus of documents. It is the product of the Term Frequency (TF) and the Inverse Document Frequency (IDF). The IDF is necessary to scale down the TF score of words that occur in many documents. These are typically stop words that carry less important when classifying text [2].

Let $|d|$ denote the number of terms in document d, $|D|$ denote the number of documents in corpus D and $n_t(d)$ denote the number of times term t appears in document d [2]. The TFIDF of a term t is in a document d with corpus D is then calculated as follows:

$$\text{TFIDF}(t)_{d,D} = \text{TF}(t)_d \cdot \text{IDF}(t)_D$$

$$\text{where TF}(t)_d = \frac{n_t(d)}{|d|} \quad \text{and} \quad \text{IDF}(t)_D = 1 + \log\left(\frac{|D|}{|\{d : D|t \in d\}|}\right)$$

The resulting vectors will be high-dimensional depending on the amount of word chosen to represent the vectors. The dimensions had to be lowered in order to visualize the vectors in a three dimensional feature space. The Singular Value Decomposition (SVD) technique was used to reduce the high dimensional feature space. Concretely, TruncatedSVD [1] from Sklearn was used. TruncatedSVD does not center the data and works well with sparse matrices typically returned from TFIDF vectorizers.

In [2], several classifiers were trained on the n-gram features that extracted through TFIDF. This was done to compare the performance of the classifiers on different counts of n. Some examples of the trained classifiers include Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Decision Tree (DT). In this work, both a Decision Tree and its ensemble version, the Random Forest classifier, were trained.

3.3 Explanation techniques

We chose for the following explanation methods from the literature to explain the machine learning models:

- The "Black Box Outcome Explanation" method [23]. This method seems appropriate for end users that have little to no technical expertise. In other words, AI novices. Multiple solutions exist for the outcome explanation problem. We opt for the model-agnostic technique LIME to explain the classifiers locally.
- The "Transparent Box Design" method [23]. This method involves a machine learning classifier that is generally considered interpretable such as a Decision Tree. The model should thus be self-explanatory in which no post-hoc explanation should be required for interpretation [23]. In case a Decision Tree is trained for example, then the tree itself can be visualized as explanation.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html`

- Example-based explanations. More specifically, the factual example-based explanation method. With this method, examples from the dataset that look similar to the predicted instance will be presented to the end user [1].
- Feature explanation. The features presented in the previous section will also require further explanation. More specifically, the TFIDF features extracted from the tweet text. In [2], a horizontal bar chart was used to display the most frequent used n-grams both fake and factual news [2]. It could be challenging to explain TFIDF to users, instead, the chart can show how the features are represented (n-grams) and that frequency plays an important role.

3.4 Design process

Two prototypes were developed in this work to present the tweet instances, the predictions and the explanations. Namely, a low and high-fidelity prototype. We opted for a human-centered approach in which end users were involved during the evaluation of the prototypes. The low-fidelity prototype was developed with Google Slides which allowed us to simulate clickable user interfaces. For the high-fidelity prototype, both a a front end and an API were developed using React and Flask respectively. The front end served as the interface for the data. The API served as storage for interface data, and the endpoint to log user interface interactions and questionnaire answers.

A think-aloud study was conducted at the end of each design phase to identify important usability issues and to iteratively improve the explanations and the user interface. The think-aloud protocol requires participants to think aloud while performing a given task. This can help us to document and evaluate the usability of a user interface. Moreover, it can help us understand the participant's interaction with the interface. It is a cheap and easy method that does not require many participants [43]. Participants were sampled through Snowball sampling. The inclusion criteria for the think-aloud studies were some familiarity with Twitter. After verbal agreement, participants were sent a link to the platform where the interface could be tested. The study was conducted online through an agreed upon video streaming platform. Participants were asked to share their screen during the study. First, participants received practical information about the study. Then, participants were given tasks to perform on the provided user interface. Finally, participants were asked a set of open questions. This allowed us to come up with potential research questions or any other insight that could help with the design of the final user study.

A few important design practices were implemented during each iteration that were proposed in [34]. In this work, Liao et al. develops a question bank that represents user needs for user-centered XAI systems. This was achieved by interviewing UX practitioners with prior experience on several AI products. This work was selected mainly because it follows a interdisciplinary approach motivated by Miller's work in [37].

The results of each iteration are presented in Chapter 5, as well as the final version of the high-fidelity prototype.

3.5 Final user study

A research goal of this thesis was to compare two explanation methods from the literature that were applied on the misinformation classifier, namely LIME explanations and example-based explanations. Trust and task performance were the metrics we chose to compare the explanation methods. In the context of predicting misinformation, task performance refers to the accuracy of users in identifying misinformation. A high task performance would mean that the user was successful in identifying misinformation. The hypothesis is that if the classifier explanations can help users to understand why tweets are classified as misinformation, then users should become better at this task [26].

We propose two research questions:

> RQ1: Do example-based explanations lead to higher trust in a COVID-19 tweet misinformation classifier in comparison to LIME explanations?

> RQ2: Do example-based explanations help users to better predict COVID-19 misinformation on Twitter in comparison to LIME explanations?

There were two design conditions in the user study that was designed to answer these questions. That is, the design with the LIME explanations (interface A) and the design with the example-based explanations (interface B). We opted for a within-subjects study and implemented counterbalancing in order to minimize the learning effect between groups. In the first group, we first presented interface A and then presented interface B. For the second group, we performed the opposite.

Participants were recruited through Snowball sampling. The link to the study which included a short introduction was shared through social media platforms such as Twitter, LinkedIn and Reddit. Upon clicking the link, participants were redirected to Heroku, a platform where the prototype was hosted. Prior to starting the study, participants were shown information regarding the study and the researcher. There were a variety of inclusion criteria for the study. Participants were required to be familiar and have experience with Twitter. Participants also were required to be least 18 years old as we expect cognitive maturity. The latter was also important as participants were to face misinformation. Participants were required to agree with the informed consent in order to proceed with the study. The informed consent included the inclusion criteria mentioned above as well as addition criteria and information as recommended by *Sociaal-maatschappelijke Ethische Commissie* (SMEC). It was further recommended that participants could download the informed consent in the beginning and during debriefing of the study. Prior to the study, we chose to collect demographic data. Participants were asked to report their Twitter use and their expertise of AI. For the former, we used the 7-point scale from [25] and for the latter, we used the 10-point scale from [41] but resized it to a 7-point scale for consistency.

In order to answer the first research question, we needed a measure of trust. We chose for the validated trust scale presented in [26]. In this paper, Hoffman et al. proposes a scale to measure trust factors using eight questions each containing a 5-point Likert scale. The Wilcoxon signed rank test was chosen to analyze the Likert scale data as the samples were dependent and normality could not be assumed. Additionally, we chose to collect the perceived accuracy per tweet instance shown. This was done through a slider similarly to [39]. But instead of a 0-100 slider, we chose for a 7-point slider with higher values corresponding to a higher perceived accuracy. We took the average of the perceived accuracy for each instance presented per design condition and used ANOVA to analyze this data.

For the second research question, we had to measure task performance. Participants were given a task in order to measure this. Similar to the prototypes, we asked participants to classify tweets that contain fake or factual news. The accuracy for each participant could thus be calculated as we know the ground truth of the presented tweets. The amount of time that users spent in total for each design condition was also recorded. Both variables were analyzed using ANOVA. Additionally, we designed a set of open questions (Appendix C) primarily for surveying user trust. The aim was to retrieve important insights that might further explain the quantitative results. We also collected interaction amounts of the elements in the design conditions. We analyzed these counts using ANOVA. Nextcloud was used to securely store the research data.

To sum it up, we collected the following data for each participant:

- Demographics: age, Twitter use and AI expertise
- Classification per tweet instance before and after seeing explanation (fake/factual)
- Perceived accuracy per tweet instance (0-6)
- Amount of interaction with explanation interface and “do not agree” buttons
- Answers to the trust questionnaire and open questions
- Time spent at each interface (in seconds) and data/time at which study took place

This user study with reference number G-2020-2879-R3(MIN) was approved by *Sociaal-maatschappelijke Ethische Commissie*.

The results of the user study are presented in Chapter 6. Refer to Appendix D for the about, informed consent form, and survey of participant demographic. Refer to Appendix C, for the trust factor and open questionnaire. Refer to Appendix E for the additional information that participants could use to close their knowledge gap on how the AI works.

Chapter 4

Training explainable models

This chapter covers the development of the machine learning classifiers which includes data retrieval, processing and analysis. It also covers the application of the explanation methods, and the development of the API that served the explanations.

4.1 Tools

The implementation concerning data retrieval, engineering, analysis, training and explaining machine learning classifiers was implemented and documented in a Python notebook on Google Colab [1]. This notebook also includes the implementation of the pipeline for exporting the tweet instances and explanations for the user interface of the high-fidelity prototype.

The Python notebook contains the following list of libraries used:

- Data retrieval: Twarc
- Data processing and storage: Pandas, Numpy
- Visualization: Seaborn, Matplotlib, OpenCV2, GraphViz
- Natural Language Processing: NLTK, textblob
- Machine Learning and metrics: Sci-kit Learn, Scipy
- Explanation: LIME

[1] `https://colab.research.google.com/drive/1diuKRoEbQtvaTwv97GnPF9pDA8vlH6-N?usp=sharing`

4.2 Data retrieval

Raw data was collected from the first source [2]. The data was structured with the first two columns containing the tweet ID and the tweet text and the rest of the columns containing answers to the following questions:

- Q1: Does the tweet contain a verifiable factual claim?
- Q2: To what extent does the tweet appear to contain false information?
- Q3: Will the tweet's claim have an effect on or be of interest to the general public?
- Q4: To what extent does the tweet appear to be harmful to society, person(s), company(s) or product(s)?
- Q5: Do you think that a professional fact-checker should verify the claim in the tweet?
- Q6: Is the tweet harmful for society and why?
- Q7: Do you think that this tweet should get the attention of a government entity?

Tweets were chosen based on whether they answer the following: "yes" to question 1, "yes, probably or definitely contains false info" to question 2, and "yes, either urgent or not urgent" to question 5.

The raw data from the second source [3] and third source [4] were collected. Only the labeled claims, news articles, and tweet ID's (both news and claims) were selected from the second source but not the user replies. This dataset contains data from various dates. We chose folders "05-01-2020" and "07-01-2020" that were available at the time of implementation. From the third source, both the content and the title of the articles were chosen and combined in one column. The data was further preprocessed in a similar fashion as done in the article [33] written by Susan Li, the author of the dataset.

Finally, we collected the raw data from the last source [5]. This labeled data contains Twitter user ID's of both social bots and real users.

The ID's were then used to retrieve the actual data from Twitter using Twarc, a Python library used for rehydration. This process resulted in an array of JSON objects. Feature selection was performed to retrieve the columns of the tweet and user object. Feature extraction and data mining were performed to extract new columns from existing ones. Refer to the methodology for the list of all the features selected. An additional feature determining whether a text was written subjectively or objectively, was mined by counting

[2] https://github.com/firojalam/COVID-19-tweets-for-check-worthiness

[3] https://github.com/cuilimeng/CoAID

[4] https://raw.githubusercontent.com/susanli2016/NLP-with-Python/master/data/corona_fake.csv

[5] http://mib.projects.iit.cnr.it/dataset.html

Table 4.1: A summary of the preprocessed dataset for developing a misinformation classifier.

	Fake news	Factual news
Website claims	28	338
News articles	1416	3303
Tweets and Twitter users	7635	7635

Table 4.2: A summary of the preprocessed dataset for developing a social bot classifier.

	Social bot users	Real users
Twitter users	2805	2815

the number of adjectives and adverbs. The implementation was based on [32]. The JSON data retrieved and selected previously, were parsed to Pandas dataframes as they are more convenient to work with. The amount of tweets (and their users) had to be limited to 7635 for each label. This was done to achieve a balanced dataset containing similar amounts of fake news and factual news tweets. Refer to table 4.1 and 4.2 for the preprocessed result of the dataset.

For the TFIDF analysis in the next section, we combined the unstructured data (website claims, news claims and tweet texts) in one dataset. This resulted in a dataset of 18158 documents consisting of a 50-50 split of fake news and factual news. This dataset was then preprocessed. It was cleaned from punctuation marks and URLs. The documents were not stemmed as this would result in the TFIDF features being stemmed, which could complicate interpretation of these features. The unstructured documents were then converted in vectors of TFIDF features using the TFIDFVectorizer from Sci-kit Learn. We enabled the option for the vectorizer to do stop-word removal during vectorization. We picked 1000 as the as the maximum amount of features that the vectorizer should learn. This was a trial and error estimate. A high number of features could lead to overfitting as it would start picking unimportant TFIDF features with a low TFIDF value. It would also produce high dimension vectors which is not ideal for visualization of the vectors. We then picked 2 and 4 as the n-gram range for each feature. An n-gram where where $n > 1$ includes multiple neighboring terms. This was done specifically to ensure that the resulting vector features would include claims (that typically occur in both fake and factual news) rather than single terms. In fact, it would be hard to use the single n-gram terms as explanation. After vectorization of the documents, the vectorizer picked the top 1000 features with the highest TF score. It simultaneously converted the documents of our dataset in a vectors of size 1000 where each entry (terms or combination of terms) corresponds with the TFIDF score of that term.

4.3 Data analysis

Two sets of features were analyzed, namely features for detecting misinformation in tweets and social bots in Twitter users. Refer to Table 3.3 and Table 3.4 for a complete overview of these features.

4.3.1 Misinformation

We started with the analysis of the "social impact" features, namely ***favorite_count*** and ***retweet_count***. Each tweet in the dataset was plotted using a scatterplot with the *favorites_count* as y-axis and the *retweets_count* as the x-axis. Spearman correlation was used (in Pandas) to search for correlation between the label fake or factual news and either *favorites_count* and *retweets_count*. See Figure 4.1 for the scatterplot and Figure 4.5a for the correlation matrix. The matrix shows a rather weak positive correlation between the "social impact" features and the label.

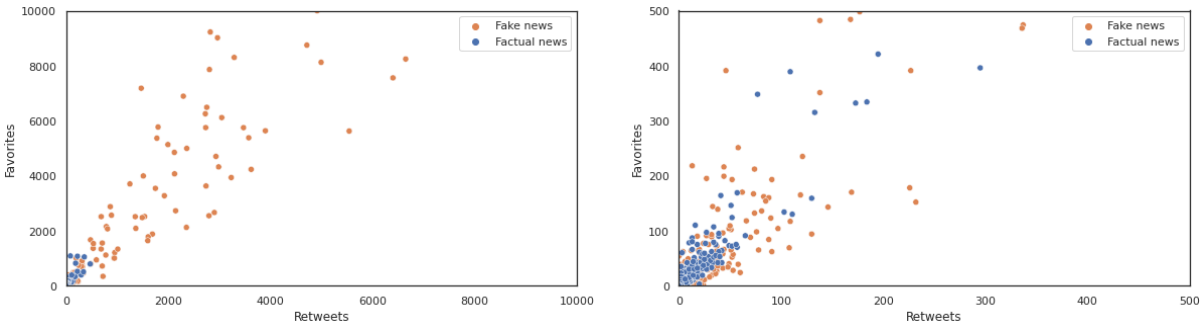


Figure 4.1: A scatterplot of the "social impact" features (left). We see that it becomes harder to separate fake and factual news instances when we zoom in to the plot (right).

The distributions of the linguistic features ***length***, ***n_question*** ***n_exclaim***, ***n_hash***, ***n_uppercase*** and ***n_nnp*** for both fake and factual news are plotted in 4.2. See Figure 4.5c for the correlation matrix. The results show skewed and multimodal distributions with many outliers. There is however some significant discrepancy between the distributions of the *n_nnp* feature. This is reflected on the correlation matrix by a strong positive correlation between the label and the *n_nnp* feature.

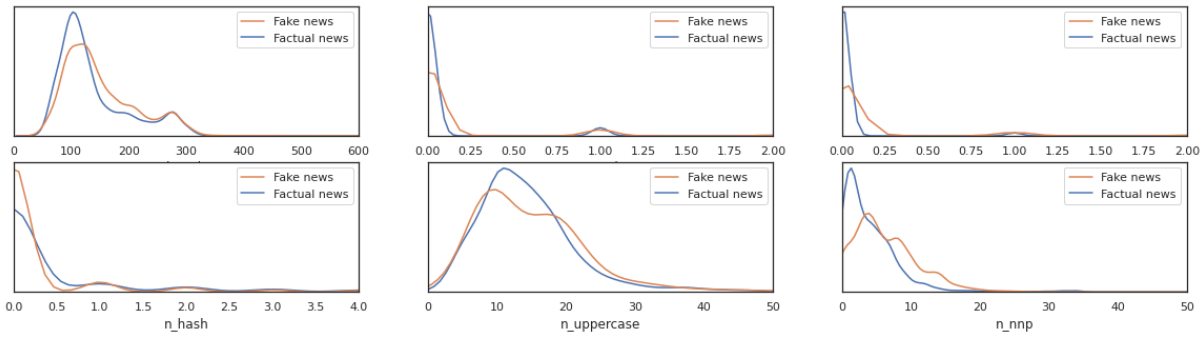


Figure 4.2: Distribution of the tweet text features.

The Jaccard similarity scores of the boolean features ***has_link***, ***subjective***, ***possibly_sensitive*** against the label fake and factual news resulted respectively in (rounded to 2 decimals): 0.46, 0.15, 0.03. A low similarity score implies that the boolean features are not informative. Conversely, a high similarity score does not imply informative features. For example, if the labels were to co-occur with boolean values by chance. A significance test could be used instead.

Both the platform which tweets were posted from and the source of the URL used in the tweet were analyzed. This was done by creating a list of platforms for both fake and factual tweets. We then plotted a bar chart for both labels. There was no significant discrepancy between fake and factual news given the platform that a tweet was posted from. This was confirmed by plotting a list of the platform sources for both fake and factual news. Although users who posted misinformation were more likely to post their tweet from a mobile application, the difference was not significant. Both types of tweets were posted from a variety of different sources and platforms. A list of the URL sources was created in a similar fashion for fake and factual news and plotted on a bar chart. However, shortened URLs such as “bit.ly” or “ow.ly” were cleaned from the lists beforehand. The lists of URL sources were also sorted from most to least frequently used. In Figure 4.3, the URL sources are plotted for both fake and factual news. An additional feature ***has_bad_source*** was mined from these lists which indicates whether a tweet contains an URL that occurs in the list of URLs frequently used in fake news. This feature was then tested with the label using Jaccard similarity. The *has_bad_source* feature of each tweet showed a significant Jaccard score of 0.75 (rounded to 2 decimals).

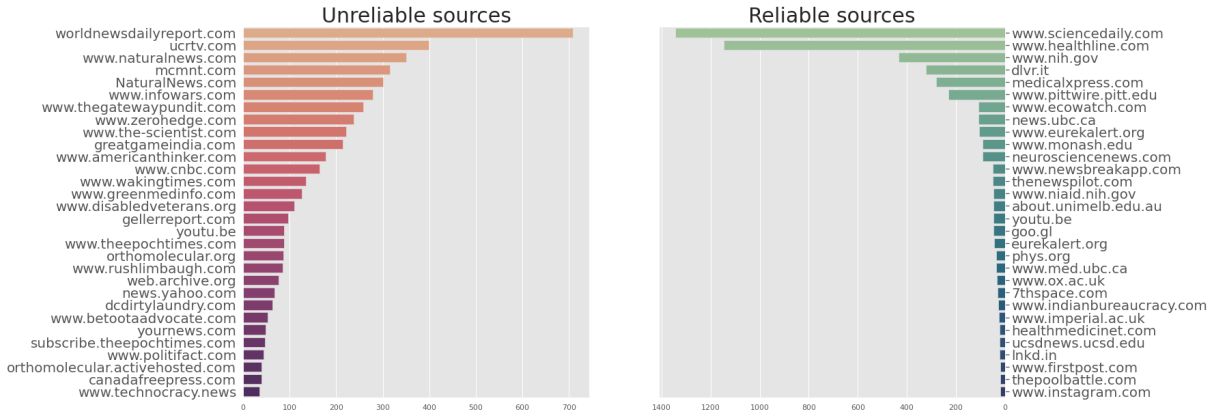


Figure 4.3: A bar chart showing the media sources mentioned in both fake and factual news.

We also analyzed the features of the users that posted the tweets. See Figure 4.4 for the distributions of the user features ***age_account_days***, ***statuses_count***, ***followers_friends_ratio***, ***likes_age_ratio*** and Figure 4.5b for the correlation matrix. Some of the outliers were replaced with the median value for the features *likes_age_ratio* and *n_nnp*. See Figure 4.6a and 4.6b for the resulting boxplots of the features. The Jaccard similarity scores of the boolean features ***has_url***, ***default_profile_image***, ***default_profile***, ***protected***, ***verified***, ***has_location*** against the label fake and factual news resulted respectively in (rounded to 2 decimals): 0.20, 0.05, 0.35, 0.0, 0.04, 0.40.

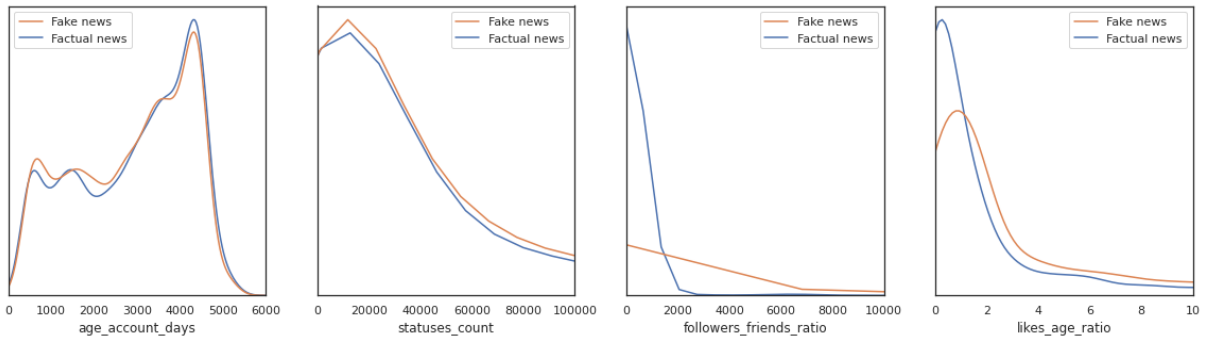
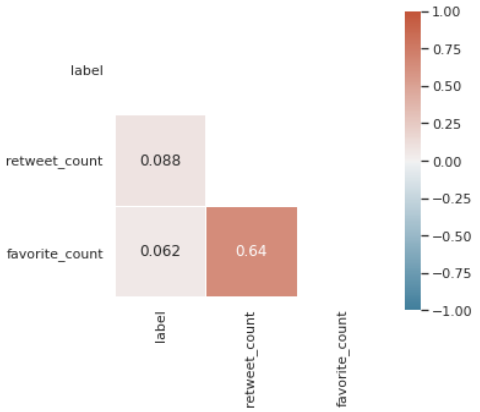


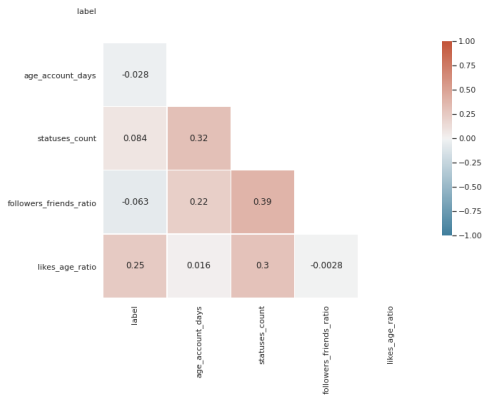
Figure 4.4: Distribution of the user features.

In [2], opinion spamming is described as the phenomenon of users spamming one piece of information. The purpose of this is to spread misinformation faster in order to more effectively manipulate public opinion. One way this can be achieved on Twitter is when a lot of users retweet one specific news article or claim. In our implementation, we analyzed how many tweets are actually unique and how many have been retweeted or duplicated.

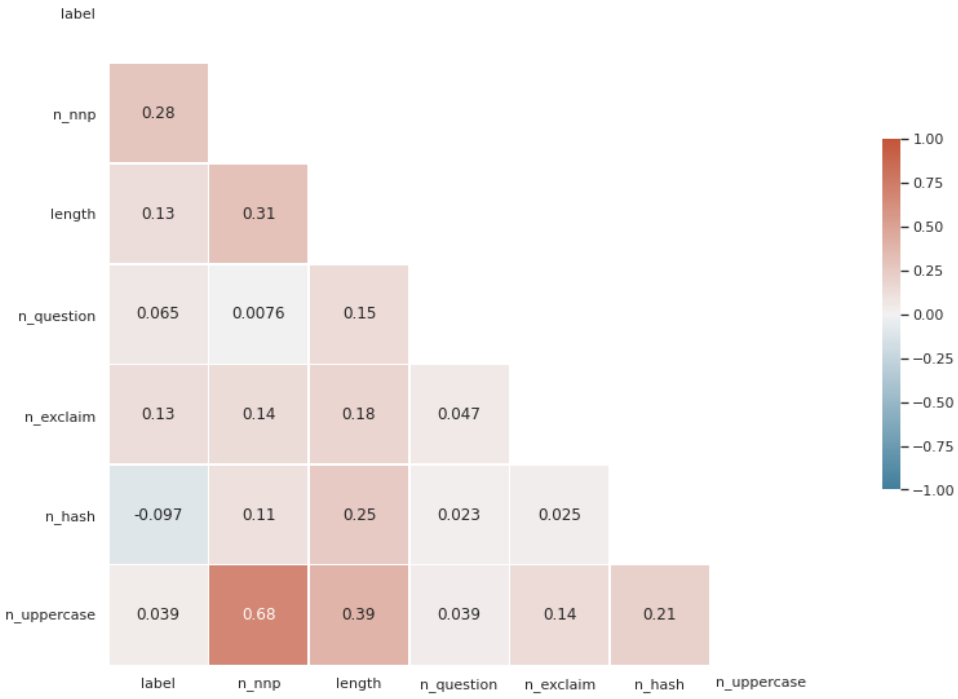
We did this naively by list-wise comparing each tweet content using cosine similarity. A high cosine similarity will usually indicate whether two tweets contain similar content. We used cosine similarity rather than hashing because the content of the tweet is likely to be transformed when it has been retweeted. That is, new piece information may be added. By lowering the cosine similarity threshold, we could relax the similarity distance.



(a) *Retweet_count* and *favorite_count*.

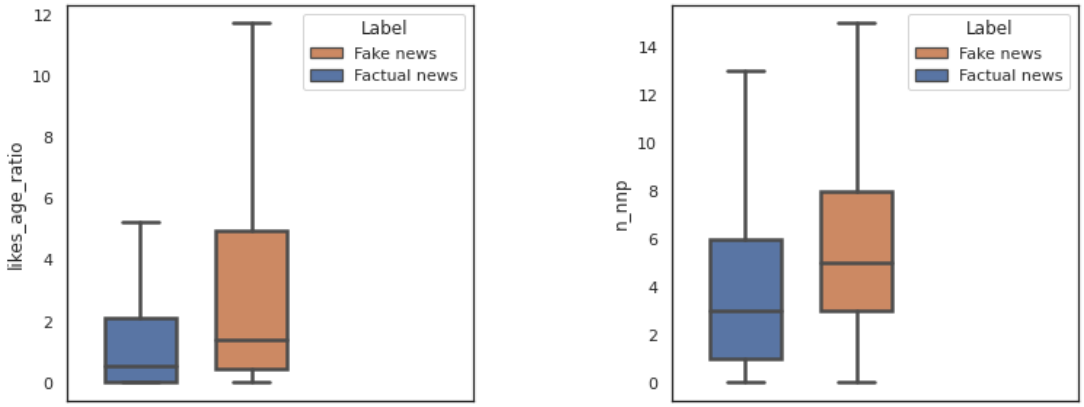


(b) The user features.



(c) The tweet text features.

Figure 4.5: The correlation matrix of features against the label using Spearman correlation.



(a) Boxplot of likes_age_ratio. (b) Boxplot of n_nnp.

Figure 4.6: Boxplot of useful features after replacing outliers with median.

However, if the threshold was too low (or too high), two tweets might be dissimilar and the results will not be accurate. The hypothesis to confirm was whether misinformation tweets spread more. In other words, were there less total amount of unique fake news tweets compared to factual news tweets?

A naive algorithm was designed to cluster tweets that are similar. A dictionary was used to store the clusters of tweet ID's. The length of the resulting dictionary of clusters will determine roughly how many unique tweets each dataset (fake news and factual news) is composed of. If the fake news tweets contain less unique tweets than factual news, then the hypothesis could be confirmed. The cosine similarity will have a significant effect on the clustering in that there will be a trade-off between the cosine similarity and accuracy. If the cosine similarity is too high, then there will be smaller and more clusters and thus "many" unique tweets. This will be an overestimate of the ground truth. If the cosine similarity is too low, then there will be less clusters but larger in size. In that case, we will have an underestimate of amount of unique tweets. Similar to choosing the amount of K's in KNN algorithm, it is hard to define a good cosine threshold. Because the algorithm is really slow, it is impossible to run many trials and pick a threshold accordingly. By limited trial and error, we picked a threshold of 0.9. After building a dictionary of tweet clusters, we added the cluster information to each tweet in the dataset. In this way, we could use the dictionary to find tweets that were clustered together with a specific tweet.

When clustering the tweets with the naive algorithm that we developed in the previous section, we have seen a reduction of 57% and 55% for fake news tweets and factual news tweets respectively. This means that there is not a significant difference and that only half of tweets for both the fake and factual news/claims were actually unique. This is a useful insight because when a certain tweet is labeled as misinformation, every other similar looking tweet can also be labeled the same. In this way, a large portion of tweets can be labeled very efficiently.

The hashtags of each tweet were also analyzed. We used word clouds to visualize the hashtags used in the entire dataset. Although word clouds are not desired way to visualize qualitative data, in this case they sufficed for testing purposes. Frequently occurring

hashtags were collected for fake and factual news. A new boolean feature ***has_bad_tag*** was mined which denotes whether a tweet contains a hashtags that frequently occurs in fake news. We can see in Figure 4.7, that some hashtags overlap in both sets, but there are still a fair amount of hashtags unique to each set. The feature *has_bad_tag* that was generated using the list of hashtags did not produce a high Jaccard similarity score when compared to the label.

Figure 4.7: Hashtags that prominently occur in fake (left) and factual news tweets (right).

Finally, we analyzed the TFIDF features. This was done in several ways. First, a bar chart was plotted showing 20 of the most important the TFIDF features, sorted in descending order of TFIDF value. See Figure 4.8. We picked the color red for the fake news features and green for the factual news features. The size indicates how large the cumulative sum of the TFIDF value of a feature is in the respective set of documents. Second, we used TruncatedSVD [6] from the Sci-kit Learn library to decompose the resulting TFIDF vectors such that the features could be plotted on a 3D feature space. The resulting 3D plots captured from different angles are shown in Figure 4.9.

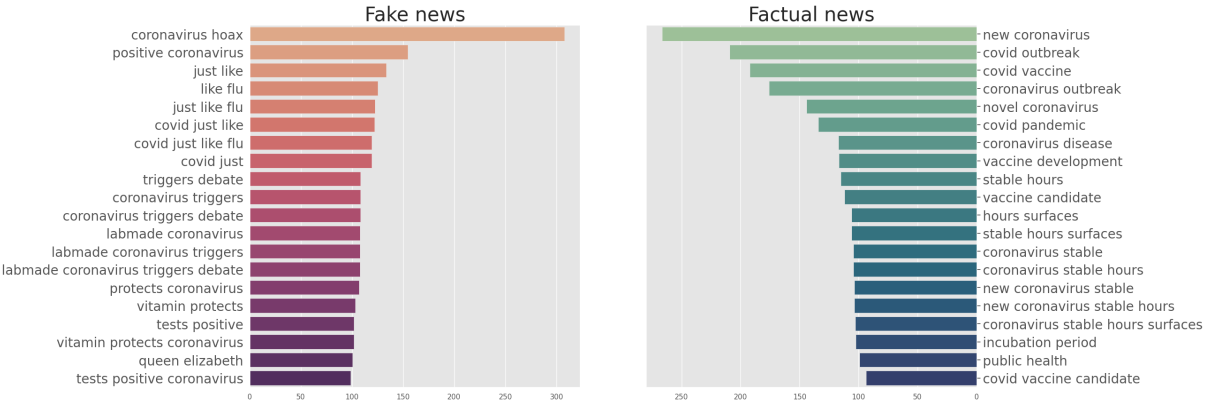


Figure 4.8: Bar chart of important n-grams as found by TFIDF technique for fake news and factual news respectively.

The n-grams on the bar chart show significant n-gram features representing strong claims such as “coronavirus hoax” and “labmade coronavirus” and other n-grams such as “triggers debate” or “queen elizabeth” that do not make much sense when not used or explained within context.

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html`

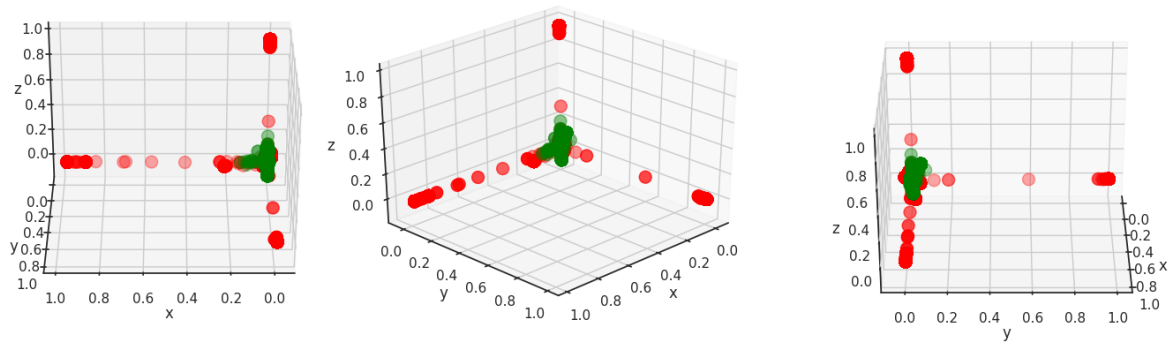


Figure 4.9: The high-dimensional TFIDF vectors reduced to the first three principal components (we called them x,y and z) with the highest variance. The red dots (projections) represent the TFIDF vector instances labeled as fake news and the green dots represent the factual news vectors.

4.3.2 Social bot

Similarly as we did for the features to train a misinformation classifier, we analyzed the continuous features ***followers_friends_ratio***, ***likes_age_ratio***, ***favourites_count***, ***statuses_count***, ***bio_length***, ***username_length***, ***name_length***, ***age_account_days***, ***listed_count***, ***followers_count***, ***tweet_age_ratio*** for correlation with the label social bot user and real user. See Figure 4.10 for the plot of the distributions and Figure 4.11 for the resulting correlation matrix. Some of the features show a very strong negative correlation with the label.

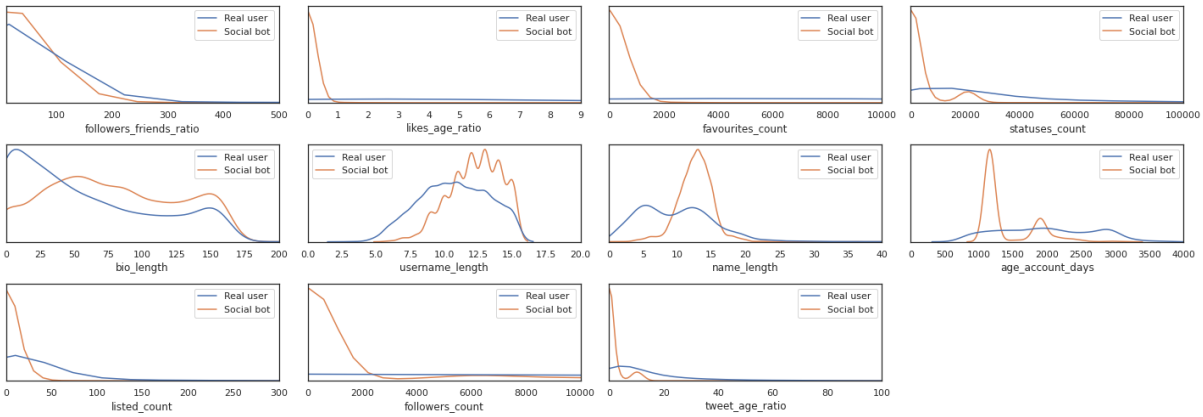


Figure 4.10: Distribution of the continues features for social bot account and real users.

We analyzed the boolean variables ***has_url***, ***protected***, ***default_profile***, ***default_profile_image***, ***has_location***, ***verified***. The Jaccard similarity against the label social bot and real user resulted respectively in (rounded to 2 decimals): 0.54, 0.0, 0.06, 0.0, 0.19, 0.0.

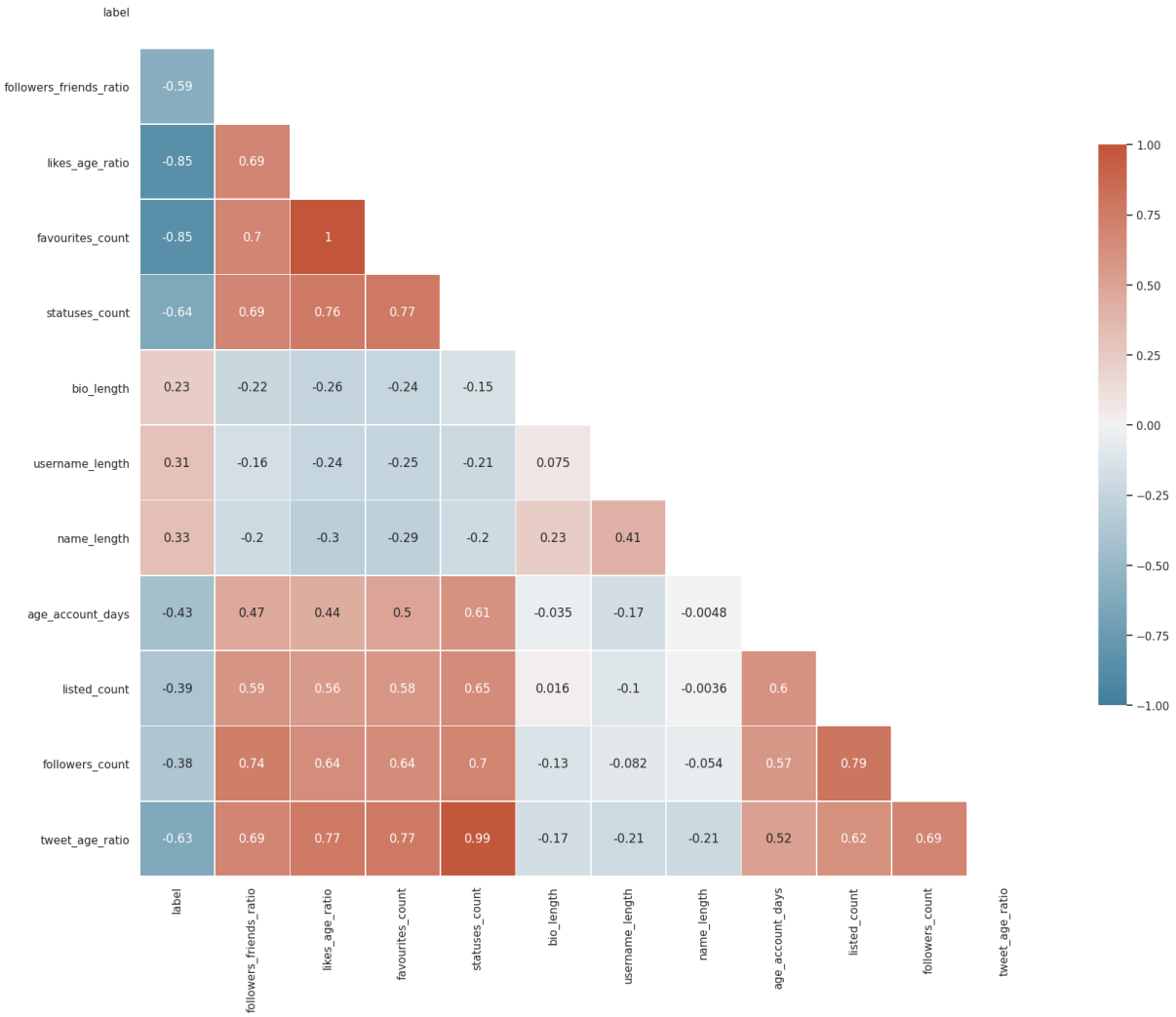
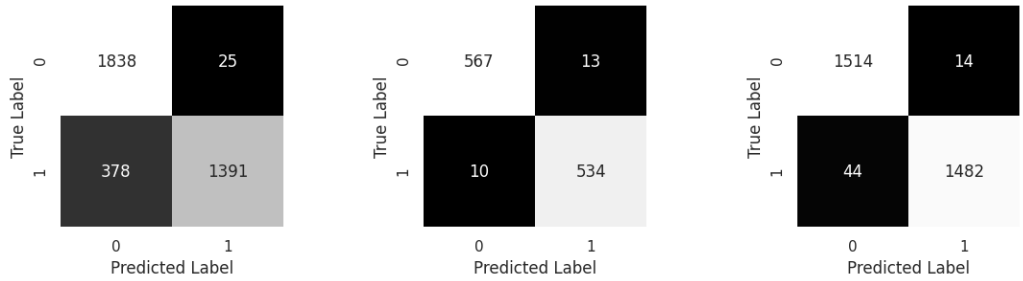


Figure 4.11: Correlation matrix of the continuous features against the label social bot and real user using Spearman correlation.

4.4 Training classifiers

Multiple classifiers were trained during this phase. For each classifier, we calculated the F1-score (from the Sci-kit Learn library) to show the performance on the test set. We also visualized the confusion matrix using Seaborn. First, two classifiers were trained on the TFIDF features to learn to distinguish fake and factual news from the tweet text. The data was first split in a training and test set. We then trained two classifiers on the training data, namely a Decision Tree and a Random Forest classifier. The Decision Tree only had a depth of 3. The depth was limited in order for the tree to be visualized as explanation (refer to the next section) and thus made easier for interpretation. For the Random Forest classifier, we picked 20 estimators and a max depth of 90. The depth and amount of trees were limited to avoid overfitting. The trained classifiers scored an F1-score (rounded to 2 decimals) of 0.54 and 0.88 on the test set, respectively. Refer to



(a) (b) (c)

Figure 4.12: Confusion matrix of the TFIDF classifier (a), the social bot classifier (b), and the misinformation classifier (c).

Figure 4.12a for the confusion matrix of the Random Forest classifier on the test set. The output of the Random Forest classifier for each tweet was used to generate the feature ***is_TFIDF_fake***. Essentially, this feature denotes whether the tweet text of a certain tweet was detected to contain TFIDF features related to fake news. This feature was used as input to the misinformation classifier.

Second, we trained a classifier on ***followers_friends_ratio***, ***likes_age_ratio***, ***favourites_count***, ***statuses_count***, ***username_length***, ***tweet_age_ratio***, ***has_url***, ***has_location***. We first split the data in a training and test set. We then trained both a Random Forest classifier (with 10 estimators and max depth of 10) and a KNN classifier (with K=10) from the Sci-kit Learn library. The choice of estimators and K was rather arbitrary. This Random Forest classifier scored an F1-score (rounded to 2 decimals) of 0.98 on the test set. Refer to Figure 4.12b for the confusion matrix. The reason why we trained the KNN classifier was to find similar for a given Twitter user. This was required for generating example-based explanations in the next subsection.

Finally, we train a misinformation classifier on ***retweet_count***, ***favorite_count***, ***is_tfidf_fake***, ***has_bad_source***, ***subjective***, ***n_nnp***, ***n_exclaim***, ***n_hash***, ***n_uppercase***. We split the data in a training and test set and trained a Random Forest classifier on the training data. The trained classifier scored an F1-score (rounded to 2 decimals) of 0.98 on the test set. Refer to Figure 4.12c for the confusion matrix.

4.5 Explaining classifiers

We generated classifier explanations in addition to the plots that were generated during analysis. We used Graphviz to visualize the Decision Tree that was trained on the TFIDF features. The output was a model that could be globally interpreted. Figure 4.13 illustrates the Decision Tree that was trained on the n-gram TFIDF feature vectors where n is strictly greater than one. The Decision Tree can be interpreted as follows. For each node, we have to traverse to the right or left branch depending on the condition of that node. That is, if it possible for the condition to be evaluated considering that the input might not fulfill any conditions of the Decision Tree. For the first node, we move to the

right branch if the TFIDF value of the n-gram "positive coronavirus" exceeds 0.154 in the tweet text and we move to the left branch otherwise. This is repeated until there is no more condition left to fulfill.

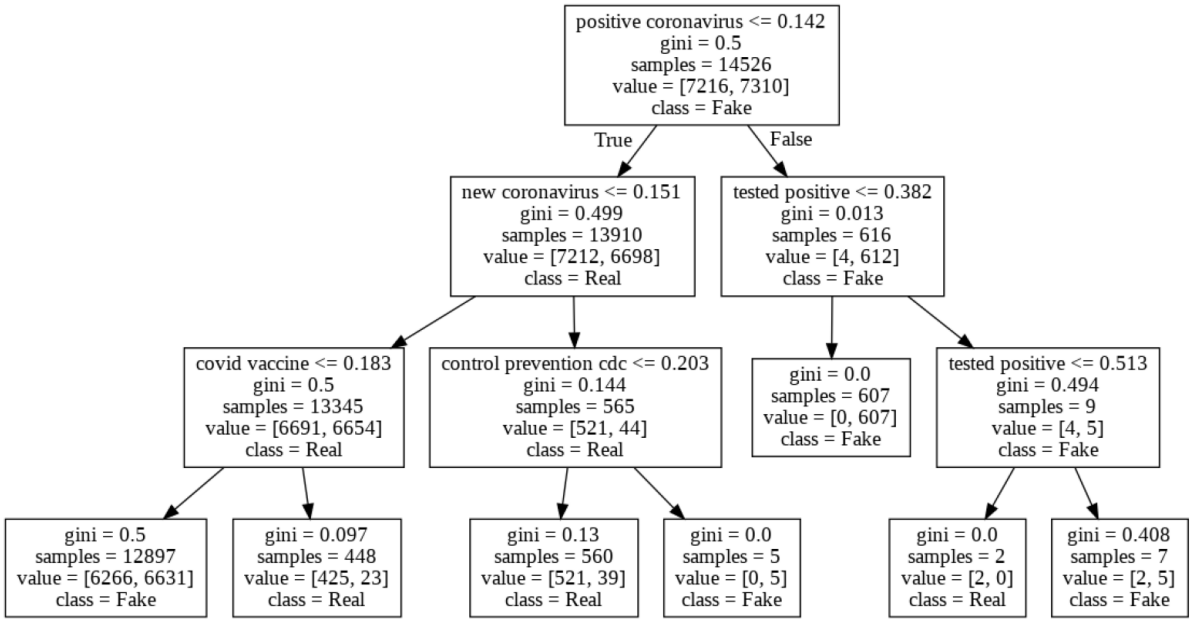


Figure 4.13: The Decision Tree for predicting misinformation in the tweet text visualized with GraphViz.

The LIME Python library was used to generate a heatmap explanation for the Random Forest classifier trained on the TFIDF features. However, due to restrictions of the library for n-gram features where n is larger to one, an explanations could not be generated. We did however generate LIME explanations (on tabular data) for both the social bot Random Forest classifier and the misinformation Random Forest classifier. Refer to Figure 4.14 and Figure 4.15 for the respective explanations.

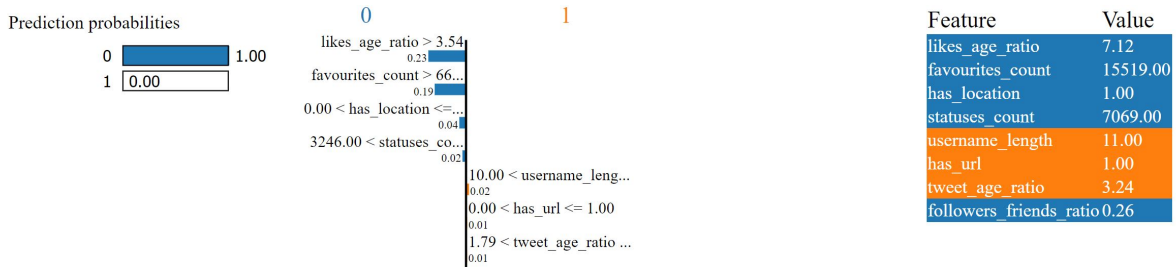


Figure 4.14: A Twitter user instance whose ground truth (real user) was correctly predicted. The LIME explanation shows that the *likes_age_ratio* and the *favourites_count* are some of the important factors that contributed to why the user was predicted to be real.

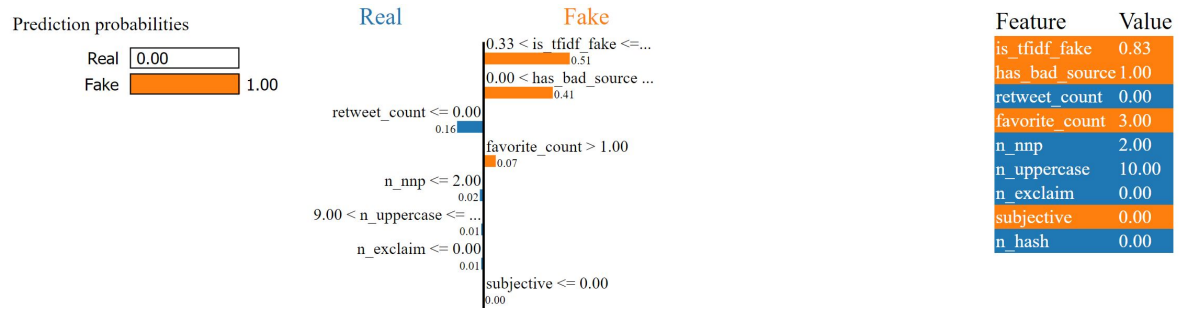


Figure 4.15: The classifier predicted the instance correctly with 100% certainty. The *has_bad_source*, *is_tfidf_fake* and "social impact" features had the most impact on the decision of the misinformation classifier. The tweet contained an unreliable source, contained n-gram features indicating misinformation, and had some amount of likes.

We also generated example-based explanations. This was done manually for the misinformation classification. Three types of examples were generated. Namely, examples with similar tweet text (TFIDF), similar source and similar article. It was trivial to find examples for the tweet text as we had clustered similar looking tweets together in a dictionary. Finding large amounts of example tweets in the dataset with a similar source and article was also trivial. That is, if the source or article had a frequent occurrence. The examples were limited to 30 instances. Additionally, similar examples for the explanation of the social bot classification were generated. The KNN classifier trained earlier was used to retrieve 10 neighbors for any given Twitter user.

An explanation pipeline was built in order to automate the process of generating explanations for any set of tweets and users of the dataset, as was required for the user interface of the high-fidelity prototype. We chose four tweets for each design condition of the user study. Each set of tweets contained two true positives and two true negatives. The pipeline utilized previously trained classifiers and the LIME explanation models in order to generate explanations. The tweets for the high-fidelity prototype were chosen based on the explanations that would have been generated. For the explanations, we picked tweets where the explanations pass face validity of goodness. We also ensured that the instances contained at least one social bot and we only picked tweets that contained claims and other news elements. For each instance, the pipeline generated a LIME explanation for why a tweet was predicted as fake news or factual news and why the user that posted the tweet was predicted as a real user or social bot. It also included the terms that overlap with the features which the TFIDF vectorizer has picked, such that a heatmap could be manually generated on the high-fidelity prototype. Finally, the LIME explanations, example-based explanations, the predictions, tweet and user level data were exported as a JSON file. We also exported additional explanations to close the knowledge gap. This includes a plot of the untrustworthy and trustworthy sources, a list of correlation factors for each of the features that the social bot classifier was trained on, a plot of the important n-gram features of the TFIDF vectorizer, and a plot showing the significance of the favorites and retweets count. The list of correlation factors was used to explain users which features are most important in determining whether a Twitter user is real or a bot.

Chapter 5

Design process

In this chapter, we will explain the iterative development behind the two prototypes, namely a low-fidelity and a high-fidelity prototype. We used a human-centered approach in which end users were involved in the evaluation of the prototypes through think-aloud studies. The overview of the design process is shown in Figure 5.1.

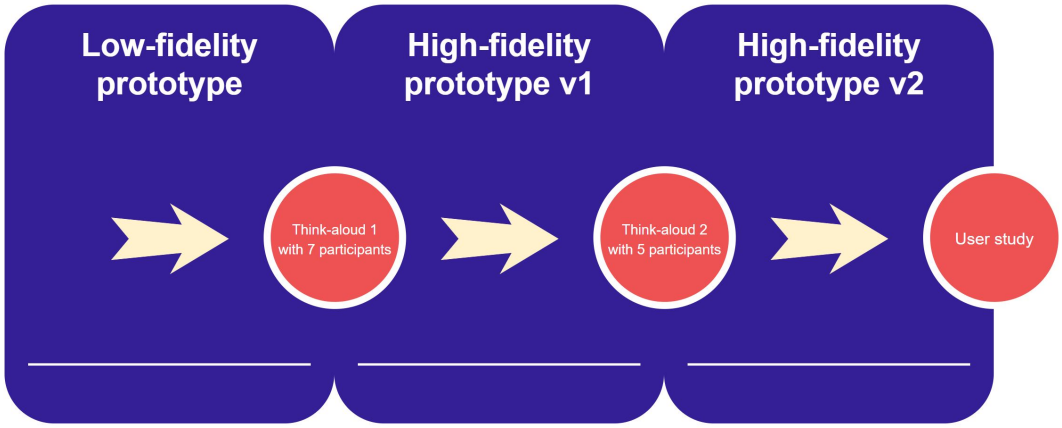


Figure 5.1: Low-fidelity prototype showing tweet level explanations.

5.1 Low-fidelity prototype

The low-fidelity prototype was developed in Google Slides. The prototype served as the interface for presenting the prediction and the explanation of the misinformation classifier as well as the related tweet and Twitter user instances. Note that the low-fidelity prototype was implemented in the first iteration. The classifiers and explanation techniques were not fully implemented in that phase. Only a limited amount of features and explanations were presented on the low-fidelity prototype as a result. Each slide of the prototype was dedicated to one tweet/user instance. In this way, users were guaranteed to see one instance at a time and the same interface could be recycled for other instances.

Figure 5.2 illustrates the most important elements of the low-fidelity interface. The largest part of the interface was dedicated to showing the tweet instance and the Twitter user metadata. We limited ourselves to the metadata that were necessary for the classifier explanation and for completion of the task. In the left bottom corner, the prediction of the AI were shown along with the confidence score. The color changed from either red or green depending on the prediction "fake" or "real". The term "real news" was changed to factual news in the high-fidelity prototype.

A "why" button was provided to reveal the buttons "explain tweet" and "explain user". The explanations were thus split in two parts, the tweet level explanations and the user level explanations. This was done to form a logical split for the explanations. Warning labels would be revealed after pressing either buttons. These could be clicked to show the specific explanation that the warning label referred to. The positioning of the labels were intentional as they were aligned with the feature to be explained. In Figure 5.2 for example, we can see a heatmap of words that the classifier considered important for the decision. The warning label to the left of the heatmap would reveal further explanation.

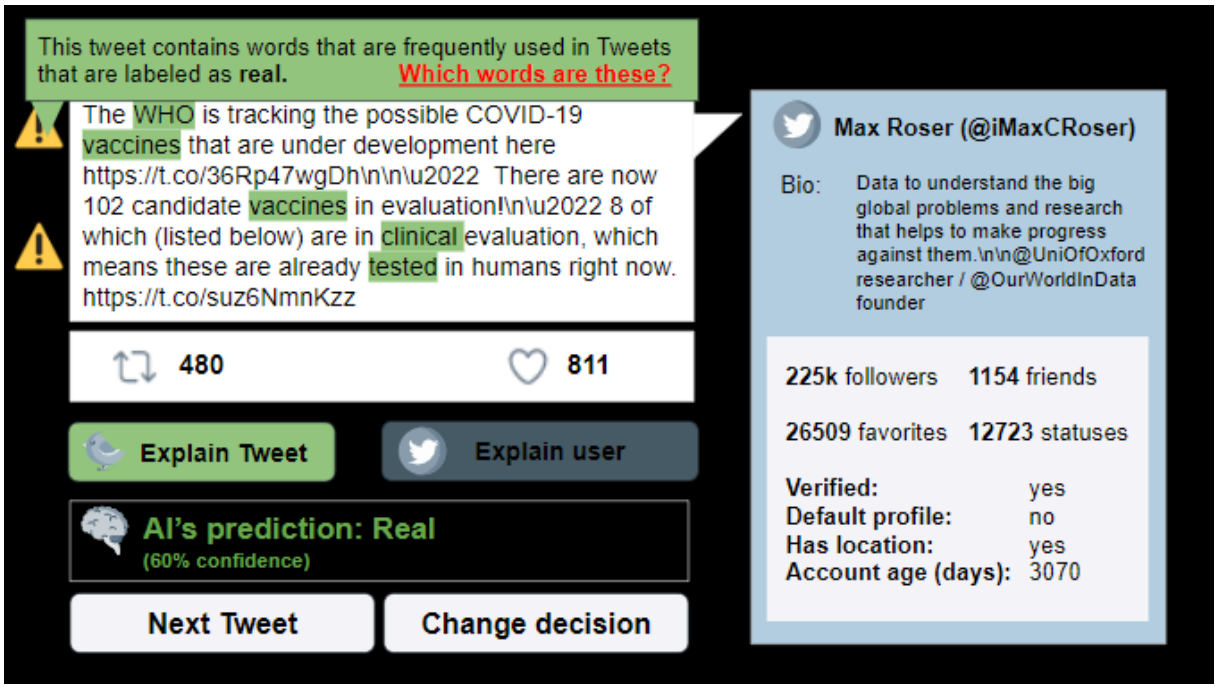


Figure 5.2: Low-fidelity prototype showing tweet level explanations.

Most of these explanations were visualized in natural language. Interactivity was provided in the sense that users could dive deeper in the explanations if their understanding was not complete or if the provided explanation was considered too shallow. Figure 5.2 illustrates this. For example, users could click on "which words are these?" to reveal a bar chart of the important TFIDF terms. No additional explanation was provided next to this bar chart. Similar to the tweet level explanations, users could view additional explanations regarding the user metadata. In Figure 5.3, we are shown the explanation of why the user was detected to be a social bot. Similarly as before, the warning labels would reveal more explanations to provide a deeper understanding of the classification.

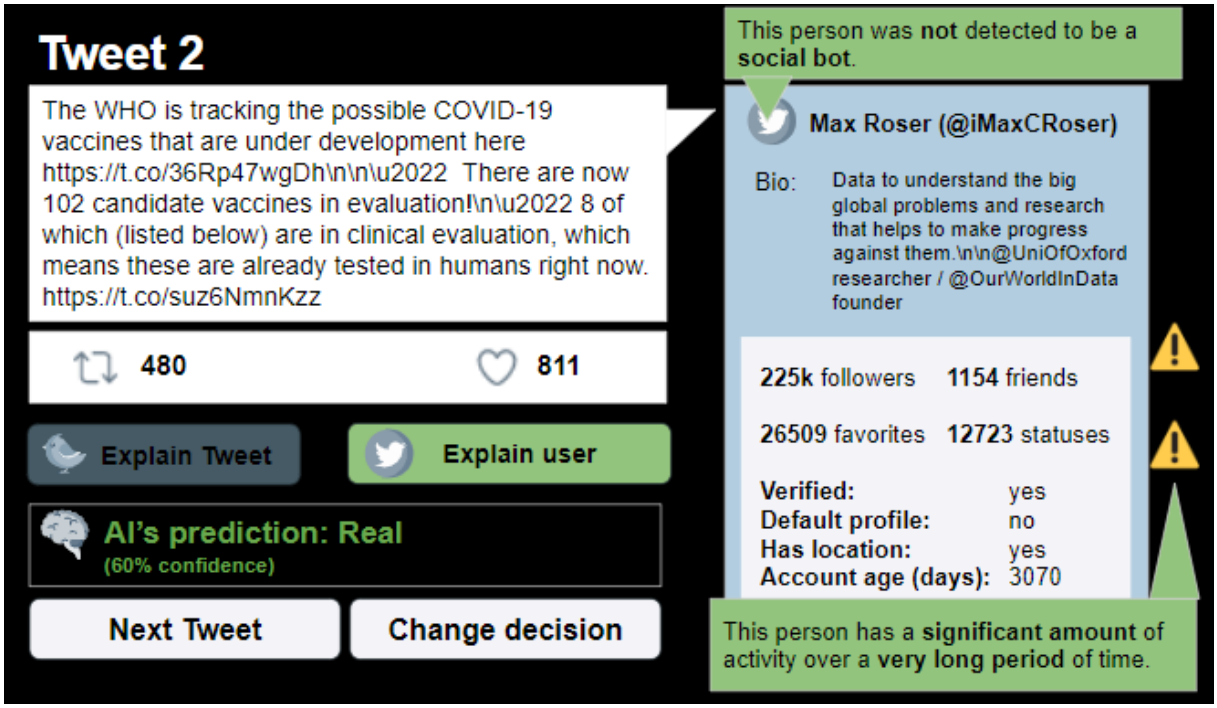


Figure 5.3: Low-fidelity prototype showing user level explanations.

5.2 Think-aloud study 1

Seven (male and female) participants were recruited consisting of friends and family between the ages of 22 and 25. The Twitter use of the participants ranged from rarely, to frequent and even professional Twitter use. The occupation of the participants ranged from white collar (tech, logistics, law) to one blue collar and two master students. A complete table of issues sorted on their occurrences can be found on Appendix A.

The participants were generally good at identifying the elements of the low-fidelity user interface. However, sometimes the participants were not entirely sure who posted the tweet or were confused with the user level metadata. A proposed solution was to make the username more obvious and only show the metadata during the explanation phase as to not to cognitively exhaust the users. Some users did not know that the warning labels that appear after showing explanations were clickable. The sometimes unfortunate placement of these labels also caused confusion to what it exactly was referring to. As a solution, we tried another approach of revealing explanations that do not require the users to click around the screen and avoid bloating the interface as well. There was also some confusion around the tweet content because users expected some of the tweet entities to be visualized similar to Twitter. For example, some of the special characters were not parsed in the low-fidelity interface. The URL's in the tweet were also shown as is. Twitter users would expect a clickable link and a thumbnail underneath the tweet. A solution for these issues was to provide a user interface experience as similar as possible to Twitter. In this way, users can rely more on their intuition and prior Twitter experience.

There were also issues with the explanations. Some of the participants expected interactivity with the bar chart that was provided as explanation. A participant was also confused by the terms that were stemmed by the TfidfVectorizer and was interested in more details about the chart such as the source. Some of the explanations were considered too shallow and lacked detail. We solved this by providing interactivity such that users could reveal more detail if they wished to. Another problem was that users did not view all of the explanations. This is hard to solve as it is realistic for users to stop when fatigued or confused. During the final user study, we ensured that the most important explanations were always shown or prioritized (by providing feature importance) and the less important explanations were somewhat hidden. Finally, the heatmap explanation always made users search for more explanation during the think-aloud study. A list of important n-gram features for both fake and factual news seemed more satisfactory but lacked interactivity and further explanation according to end users.

During the study, we also asked seven questions regarding trust and task performance. We summarize the answers to the questionnaires.

Q1: How difficult was it for you to predict fake or real news?

Participants mainly found the difficulty to depend on the tweet itself. It was mentioned that some tweets "give it away easier" and have typical reoccurring elements. Participants may have also adapted their answer based on their own task performance. That is, they would answer that the task was easy, if their predictions coincided with that of the AI and found it more difficult otherwise. We have seen that some users verbally considered the prediction of AI to be the ground truth if their own predictions overlapped. And when that was not the case, the user would mistrust the AI and consider their own predictions as ground truth. The mistrust was thus unjustified.

Q2: Did you trust the predictions of the AI?

Participants answered with higher levels of trust when their task performance was high and low otherwise. There was also some confirmation bias where users would mainly spend time on explanations that overlapped with their own beliefs. These explanations were mentioned as a pro argument in their answer to this question. The explanations that provided new insights were also positively received. Participants specifically mentioned some of the explanations in their answers that provided new beliefs. One element that played a very significant role in user trust was uncertainty. For example, users deemed the AI less trustworthy when the AI showed "too much" uncertainty in its confidence score. During the study, some participants did not consider looking at the explanations when the AI had a "low" confidence score. There may also have been cases of unjustified trust.

Q3: Did you understand the explanations of the AI?

Most users found the explanations to be clear aside from the previously mentioned issues and the confusion due to UI elements. There was however some dissatisfaction around the bar chart in which users found it to be lacking in interactivity, detail, and context.

Q4: Did the explanations help you with your own prediction?

The answers to this question were divided. A reason for this may be because each participant seemed to interpret the question differently. Sometimes users confused "prediction" with "explanation". In that case, they were corrected. One participant found the explanations to be useful because it taught them something new. Some participants found the explanations to be useful if their classifications did not overlap with that of the AI. A majority of participants associated the "helpfulness" with whether they learned something new. It was thus not clear to participants that the question was more targeted towards task performance.

Q5: How would you rate explanations of the AI?

Participants were generally satisfied with the explanations. Some participants were explicitly asked for dissatisfying aspects of the explanations. The issues mentioned mostly overlapped with the table in Appendix A. Some participants mentioned that some of the explanations were too shallow or lacked interactivity.

Q6: Do you understand how the AI predicts fake or real news?

The answer to this question always seemed to start with "yes". Participants in general seemed to overestimate their beliefs and understanding of how the AI worked. It also appears that the terms and technicality of explanations severely improved with levels of expertise. For example, users with absolutely no expertise based their explanations of the AI on the features and used these as arguments on how they believe they AI worked. More technically skilled users provided answers consisting of terms such as "tables", "certainty" and "percentages". One expert user had a fairly accurate (abstract) understanding of the AI and used terms such as "heuristics". Surprisingly, one participant with no expertise implicitly referred to machine learning concepts such as supervised learning ad bias data without explicitly mentioning or knowing about the terms (this was probed afterwards).

Q7: Would you use this Artificial Intelligence on Twitter?

The possible answers were: A) I would only use its predictions, B) I would use both prediction and explanations, C) I would not use this.

Answers varied greatly in which the majority of participants answered that they would use either prediction or explanations. Some participants mentioned that even though they would not use it, that they understand why it could be useful to someone else. This can be an indication that it is difficult for participants to answer honestly, given the relation between interviewer and interviewee. To mitigate this, participants were repeatedly reminded to answer honestly and not to hold back any opposing thoughts or beliefs. After doing so, some participants revealed that they would not use it because they would not have time to do so, or they simply do not appropriately trust the AI.

5.3 High-fidelity prototype v1

All the relevant information of the tweets were shown in a similar look and feel as Twitter. That is, in a tweet container that contained all the information as shown in Figure 5.4. This included the user that posted the tweet, the tweet content, the date and the amount of likes and retweets. The Twitter user information was limited to the attributes that were relevant to explain the social bot prediction. The user attributes could be opened by clicking on the username which was made observable with a drop down icon. These attributes were presented in a clear and compact manner to make it easier for end users to analyze the data. For each tweet container, we provided the prediction of the AI in a small container next to the tweet. The certainty percentage was omitted as this can negatively affect user trust when the user is under heavy cognitive load [58]. The explanation was placed under the prediction.

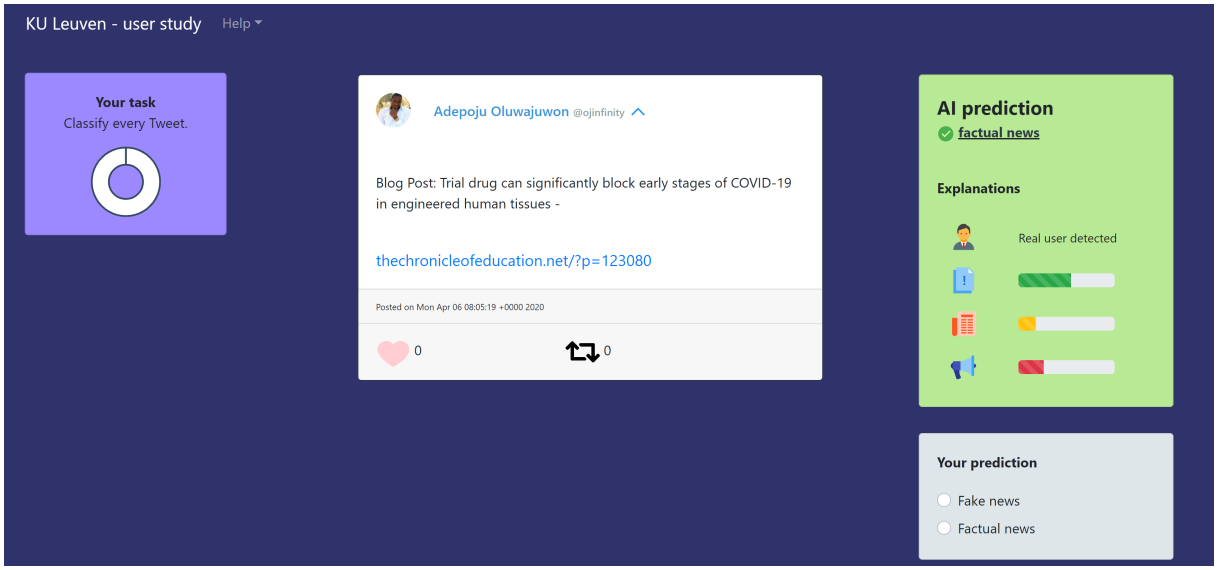


Figure 5.4: The high-fidelity prototype with the important elements of the interface. The tweet and profile of the user who posted the tweet were placed in the center. The prediction of the AI and the LIME explanations for the tweet instance could be found on the right with the task performance survey on the bottom right corner. Finally, the task progression could be found on the left upper corner.

The first explanation interface (A) contained the LIME explanations. These included the following features: social bot prediction, the TFIDF features (and heatmap), the tweet style, the source of the tweet, and the favorites/retweets count. Each of these features were exported in the JSON file with the (normalized) feature impact scores. We presented the features as icons that users could click on to reveal the explanations. Next to icons, a progress bar was placed showing the impact of the respective features. This was an alternative to the warning label approach that was originally implemented in the low-fidelity interface. An explanation container appeared after clicking on one of the icons. This container would appear below the component of the tweet that it was meant to explain. It was always presented in a compact red container regardless of the prediction.

Although the contents of these containers were mostly static, some parts of the text differed depending on the prediction. Note that not all the icons were visible at all times. This was because the LIME explanations may differ per instance and some of the features may be absent.

The LIME explanation interface (A) included five explanation containers in total, that is if all the features were present. We will discuss each explanation container as if they were all present in the respective order that they are presented.

The first container presented the social bot prediction and explained the prediction of the classifier. This container is shown in Figure 5.5. Here, the certainty percentage was again omitted for similar reasons as before. The LIME explanations that were generated during the pipeline were converted here into blocks of predefined explanations. Explanations differed for each users depending on the attributes that determined the prediction, and the prediction value itself. The top most important explanations (with the highest weight) were shown first while the remaining explanations were made invisible unless the user chose to expand these. This was done to ensure that a long list of explanations would not bloat the user interface and fatigue users.

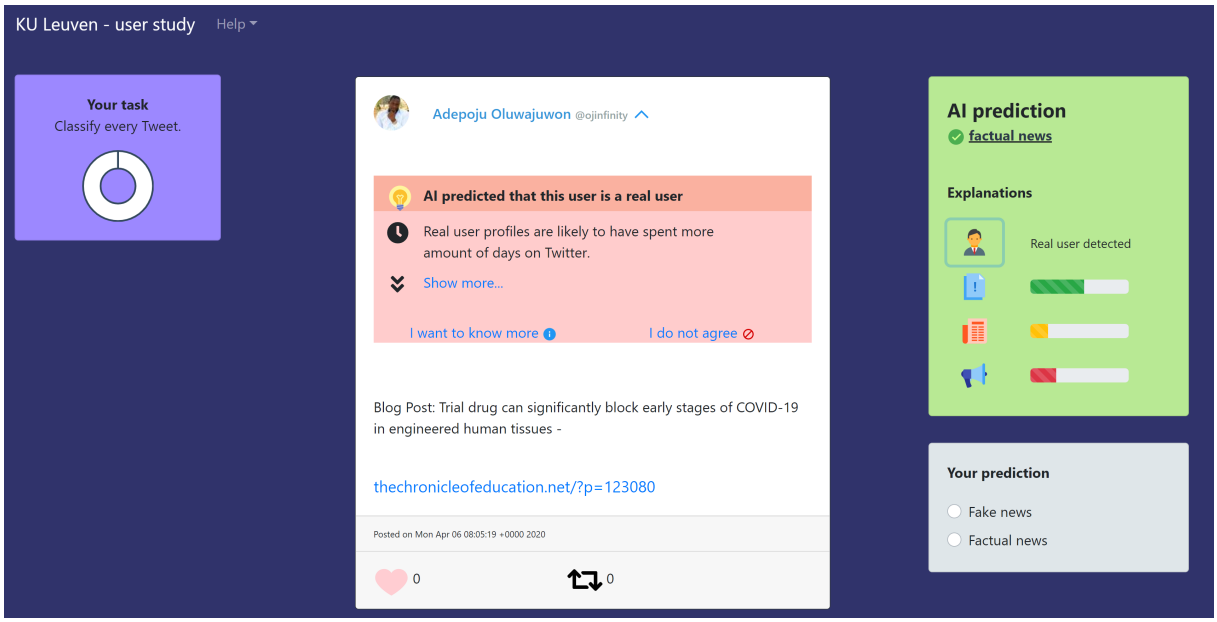


Figure 5.5: The explanations of why the user who posted the tweet was predicted to be a real user.

The second explanation container presented the TFIDF features. This is shown in Figure 5.6. Upon clicking the icon that opens this container, the words or combination of words were highlighted in the tweet text. These were the n-gram features exported with the explanations. The second container also showed an explanation and listed the words as highlighted in the text.

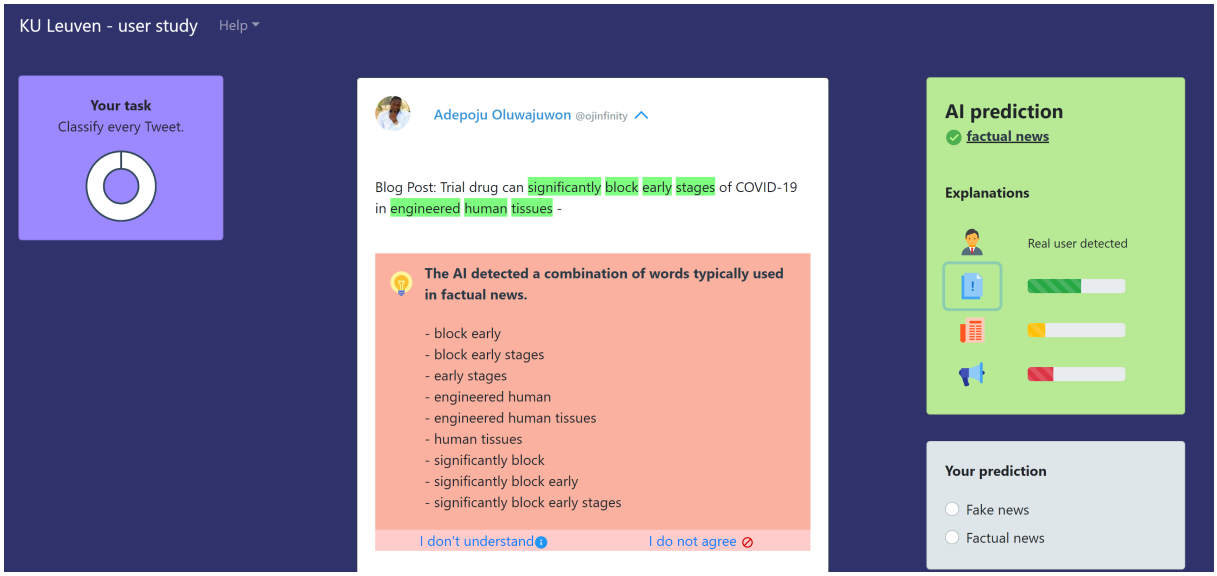


Figure 5.6: Interface A with the TFIDF n-grams as one of the explanations as to why the AI predicted the tweet as factual news.

The third container presented the explanation for the tweet style. This included the explanation itself and a list of style features that were important for the prediction of the instance. The list was dynamic in that the elements would differ based on the given instance. It included the following: whether the tweet was written objectively/subjectively and whether it contained appropriate amount of exclamation marks, proper nouns, uppercase letters and hashtags.

The fourth container presented the source explanation. If this feature was available, then the AI would explain that a source was detected that was either (depending on the prediction) in the list of untrustworthy or trustworthy sources.

The final explanation container included an explanation on social impact. That is, whether the tweet was popular which was typical for fake news. Or not popular at all for the factual news case.

Each explanation container also included a button that would refer to a new page (in a new tab) which provided a more global or "tutorial like" explanation. A reason for doing this was to allow users to close their knowledge gap as suggested in [34]. Refer to Appendix E for these pages. Some of these pages also contained an interactive component to help users with the understanding process additional to the textual explanation. The social bot explanation page for example contained a list of features with their correlations to give users an overview of the important features.

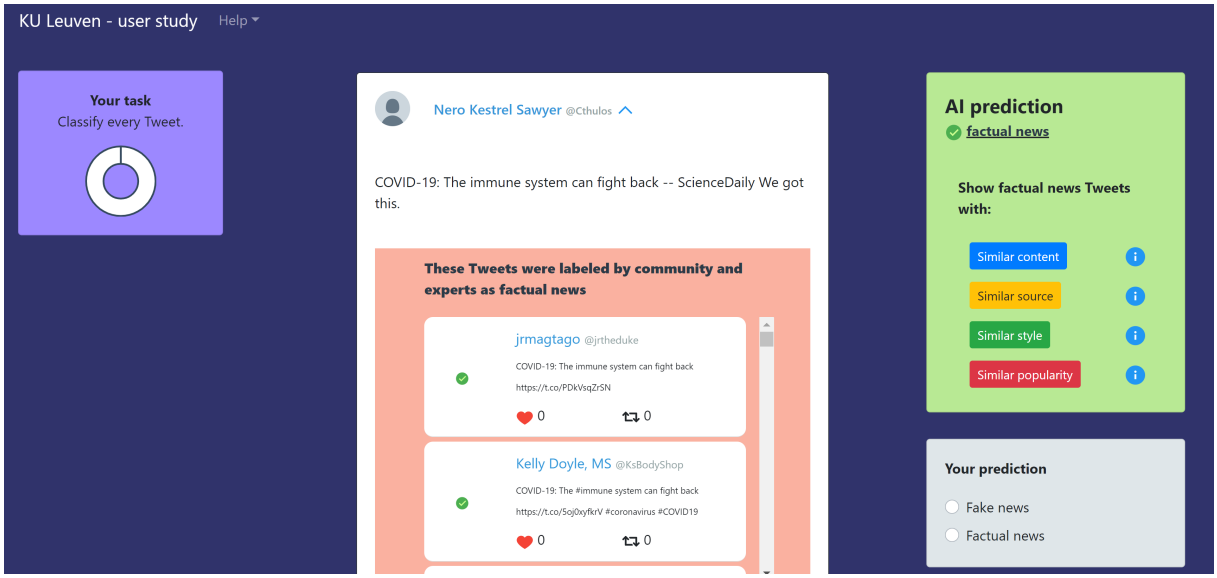


Figure 5.7: Interface B showing the example-based explanations based on the content of the tweet.

The example-based explanation interface (B) had two differences compared to the LIME interface. The first difference was that it contained one button for showing similar looking examples as explanations for the prediction. The second difference was the explanation container. Interface B only contained several explanation containers that could be opened by clicking the buttons under the prediction. These revealed similar looking tweet examples. We defined four types of similarity. That is, similarity based on contents of the tweet, the source, the style (proper nouns, use of uppercase letters,...), and similarity in terms of popularity. These are shown in Figure 5.7. The explanation shown in the container mentions that the similar looking example tweets were labeled by human users. This was important because the user has to understand that it concerns the ground truth. However, to show users that the AI was robust, we also showed with each example tweet, the prediction of the AI for that tweet. We limited the amount of similar looking examples to an arbitrary but small number of 30 tweets.

Each user was given the task to classify the tweets presented on the interface. They had to classify every tweet. Two radio buttons per tweet were provided to enable this action. A floating progress bar on the left showed the task progress in order to remind the user of the task. When finished, the user was reminded to scroll to the bottom of the page to continue to the questionnaire.

5.4 Think-aloud study 2

We recruited five participants (male and female) consisting of friends and family between the ages of 22 and 25. The Twitter use of the participants ranged from rarely, to frequent and even professional Twitter use. A complete table of issues sorted on their occurrences can be found in Appendix B.

Perhaps the biggest usability issue that affected the study was that participants did not know how to interact or interpret the LIME explanations. First, participants did not know that the clickable icons (the features that affected the prediction) were in fact clickable. Second, participants did not understand that the bar next to the icons represented feature importance. An additional issue was that users were mostly focused to the task of predicting tweets as fake or factual news.

After the think-aloud study, we also added example users, i.e. users that are similar (in metadata) to an instance user shown on the high-fidelity prototype. If for example, a Twitter user was detected to be a social bot, then the user was shown similar users from the dataset that were labeled as social bots. This was to assure that the comparison with LIME was fair, as these also contain the user as a feature to explain the outcome of the misinformation classifier.

5.5 High-fidelity prototype v2

Additionally, we received expert feedback from HCI researchers of the KU Leuven. The feedback caused a dramatic change to the flow of the prototype as well as the design. Rather than asking users to classify tweets while the explanation was visible, we chose to first ask users to classify before showing the explanations. The elements of the interface also switched places in order to guarantee that all the space was used efficiently. The navigation element underwent a dramatic change as well to ensure that users could clearly track their progress. It was also necessary to provide a short (static) text that would further explain the explanation. The colors of the explanation containers also changed to a more neutral color such that it would not confuse users. At the same time, it could draw more attention rather than blending in with the background. Feedback and recommendations from the SMEC evaluation were also implemented at this stage. The result of the evaluation mostly required changes to the informed consent form shown to users prior to the study.

The final version of the high-fidelity prototype as presented to users during the final user study is shown in Figure 5.8, Figure 5.9, Figure 5.10, Figure 5.11, and Figure 5.12.

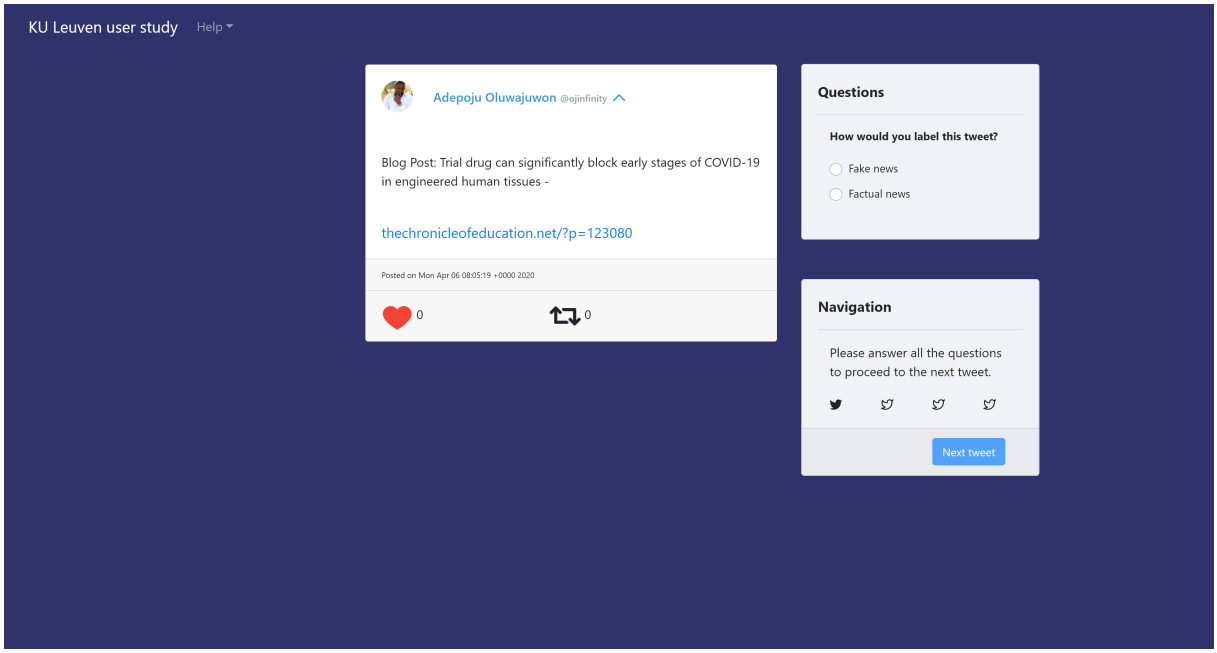


Figure 5.8: Interface A without the AI prediction and explanation. Participants have to classify the tweet presented before the AI explanation will be revealed.

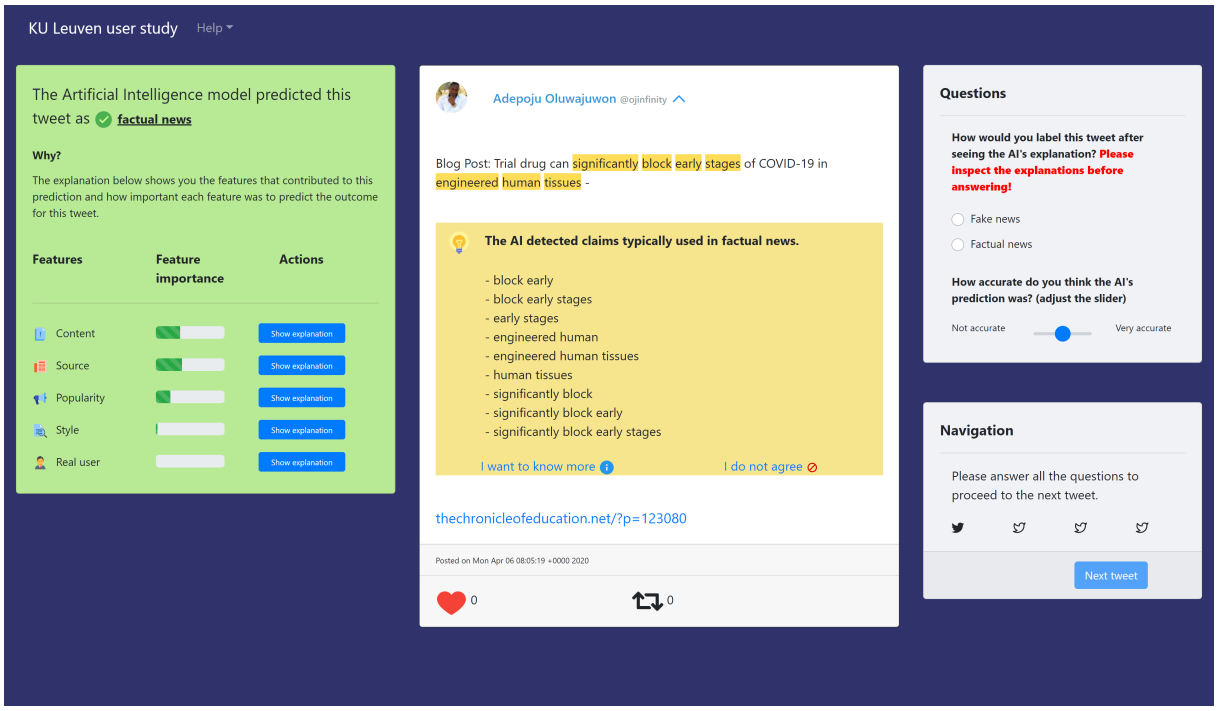


Figure 5.9: The explanation interface is revealed after the participant classifier the tweet instance. In this case, the LIME explanations are shown. The participant at this stage must classify the tweet again and move the perceived accuracy slider in order to proceed to the next instance.

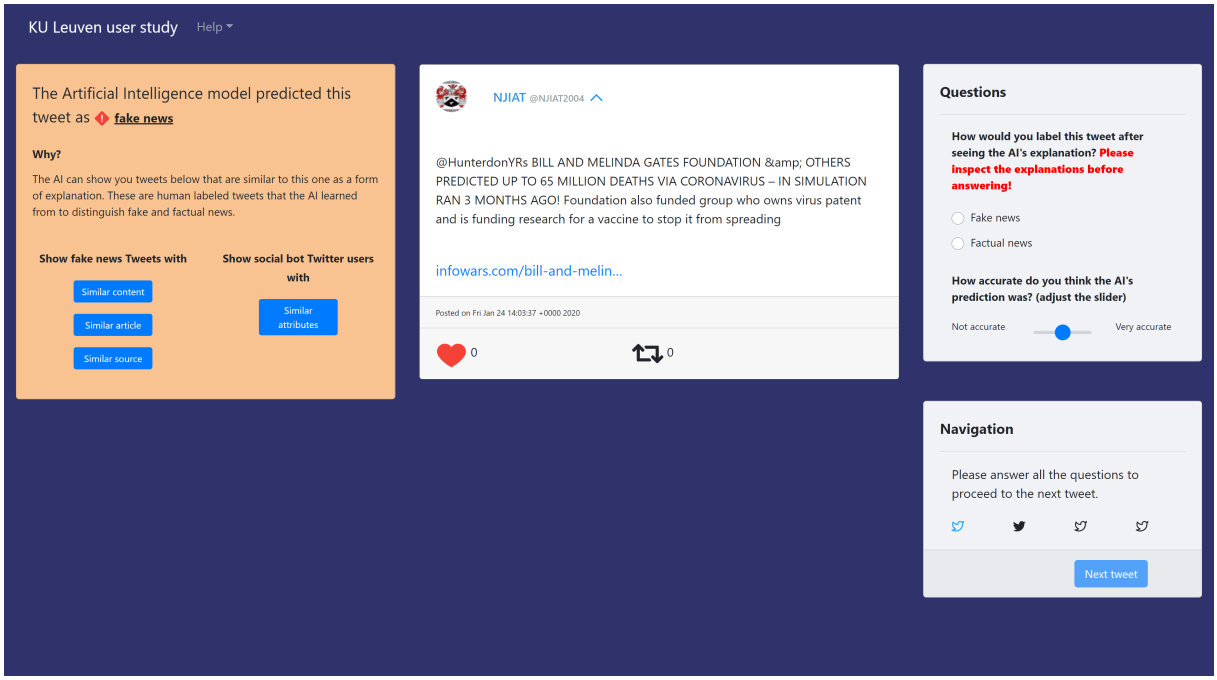


Figure 5.10: Interface B showing the explanation interface with buttons to reveal the example-based explanations.

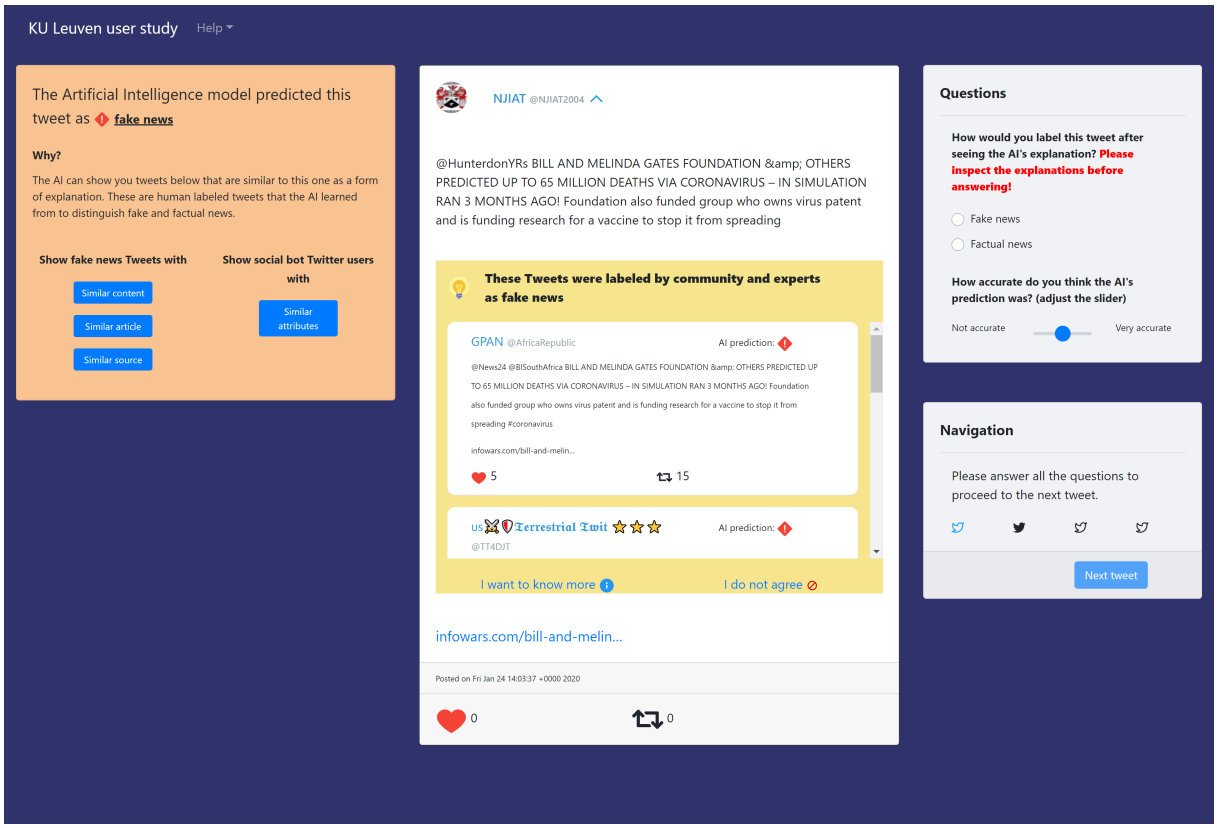


Figure 5.11: Example tweets are shown in the explanation container. In this case, labeled fake news examples with a similar content as the tweet instance proposed. Notice also the AI prediction shown for each of the separate examples.

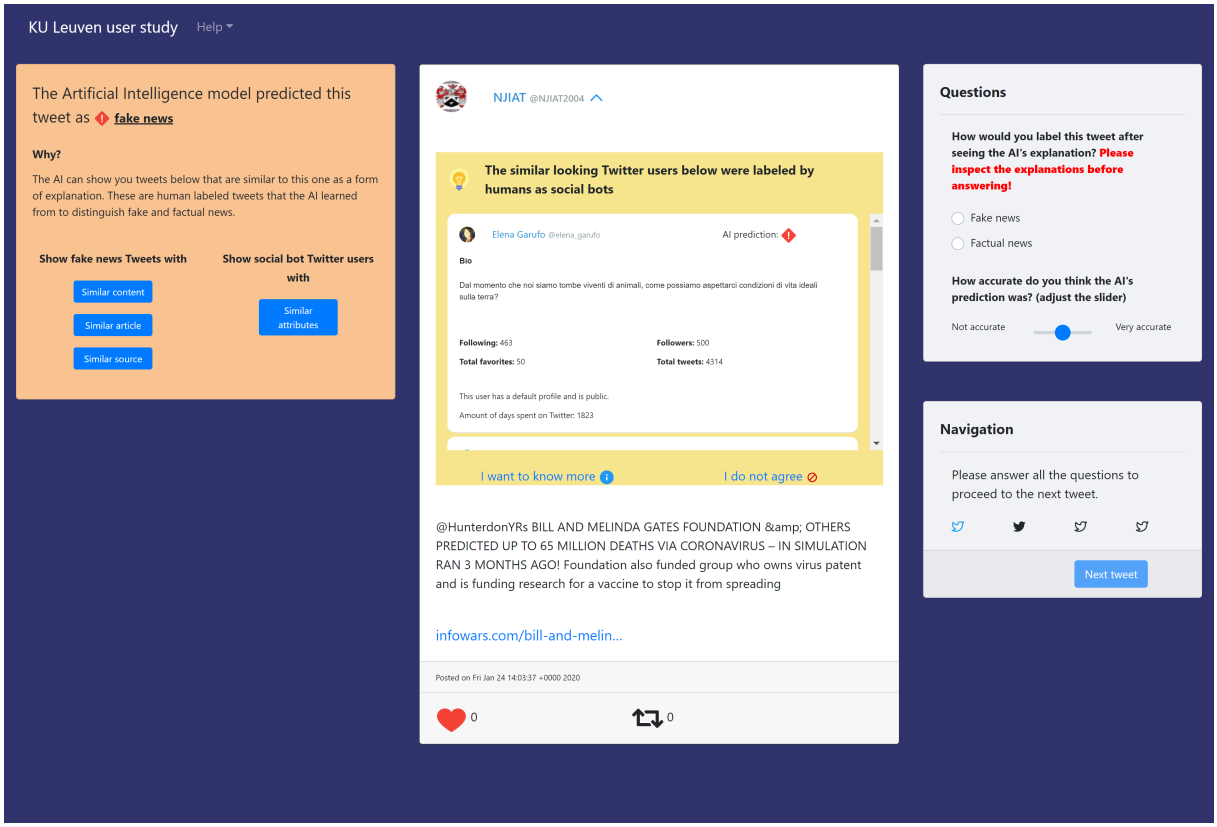


Figure 5.12: Social bot example users shown with similar attributes as the user who posted the tweet as shown on interface B.

Chapter 6

Results

We recruited 65 (male and female) participants in total. Each participant was subject to both interface A with the LIME explanations (Figure 5.9) and interface B with the example-based explanations (5.11). 51 participants remained in the logs after filtering out logs of participants that either A) have not interacted with any of the explanation buttons B) finished the study in less than a minute C) did not finish the study. Refer to Figure 6.1 to an overview of the participant demographics which includes AI expertise, age and Twitter use. The AI expertise surveyed from participants ranged between 0 and 6 (higher value means more expertise). The Twitter use ranged from 0 to 6 as well but presented as follows: 0: never, 1: less often, 2: every few weeks, 3: 1-2 days a week, 4: 3-5 days a week, 5: about once a day, 6: several times a day. We chose not to further use the AI expertise data during the analysis of the results which is explained in the limitations section.

Participants were assigned to two groups. In group 1, participants were shown interface A first, interface B second. In group 2, participants were shown interface B first, interface A second. 28 participants were assigned to group 1 and 23 participants to group 2. Participants were assigned randomly.

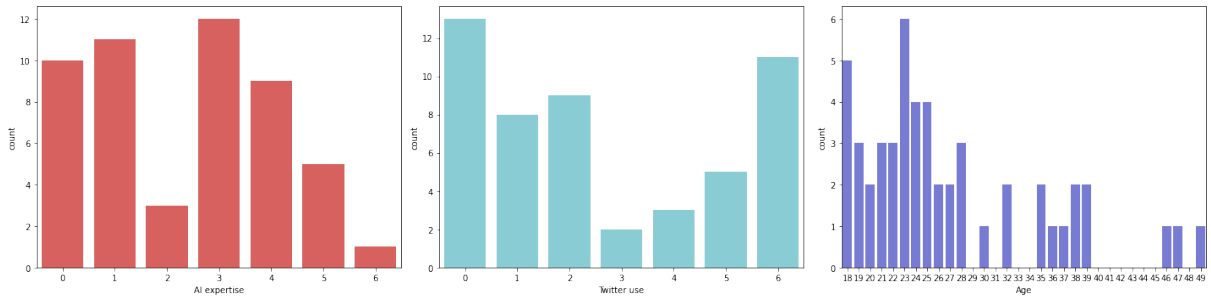


Figure 6.1: Demographics of the participants: self-reported AI expertise, self-reported Twitter use and age.

6.1 Quantitative results

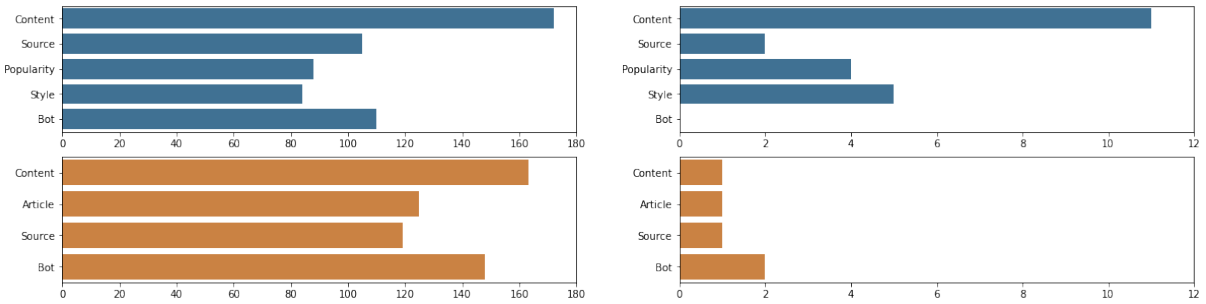


Figure 6.2: Total interaction counts with buttons of interface A (top row) and interface B (bottom row), with each row a plot of the explanation buttons (left) and dislike buttons (right).

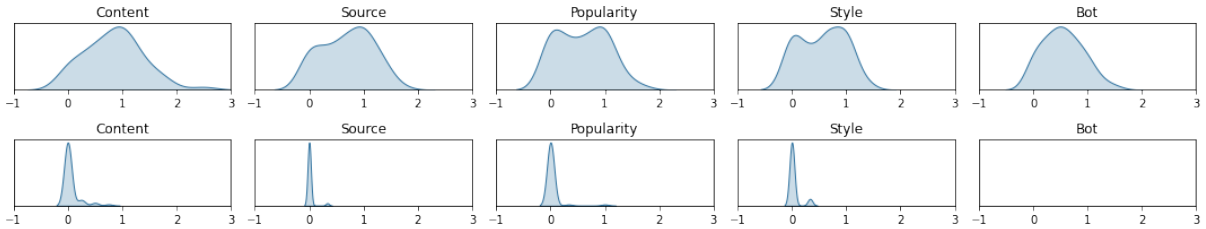


Figure 6.3: The click-through rate for the explanation (top row) and dislike buttons (bottom row) on interface A.

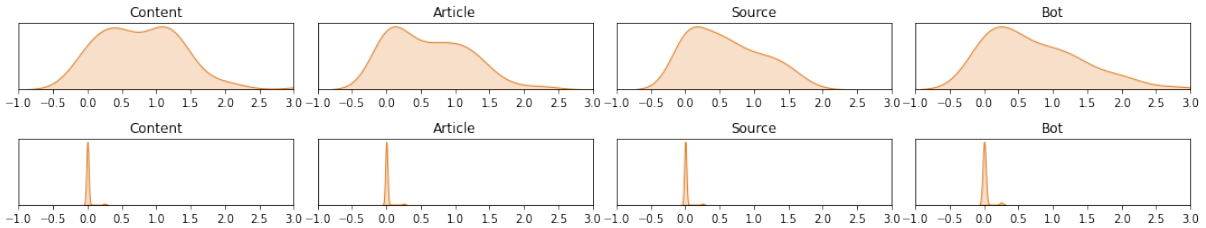
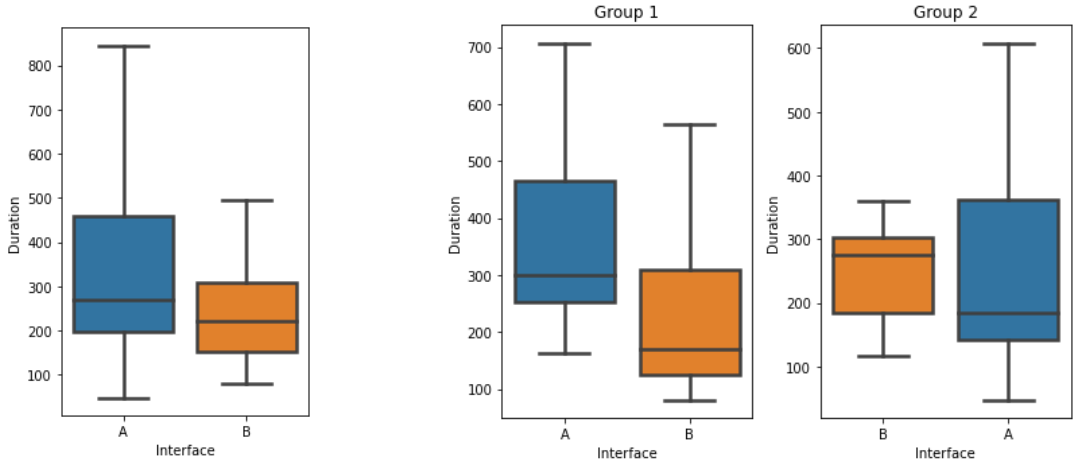


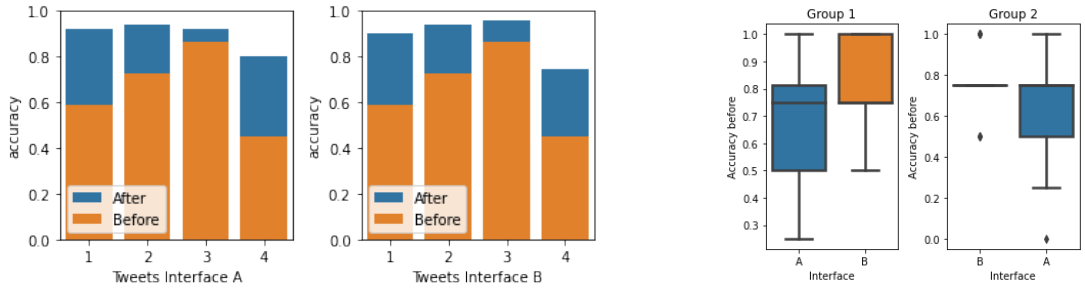
Figure 6.4: The click-through rate for the explanation (top row) and dislike buttons (bottom row) on interface B.

The total amounts of interaction with the explanation and dislike buttons on interface A and B were logged. The total counts are shown in Figure 6.2 and the click-through rate for interface A and B is shown in Figure 6.3 and Figure 6.4 respectively. The click-through rate of the explanation buttons for the LIME interface is higher overall. Some specific buttons of the explanation buttons on interface B have a very low click-through rate. The average click-through rate of the dislike buttons was nearly always zero.



(a) (b)

Figure 6.5: The task performance duration in seconds for each interface including both groups (a) and the task performance duration for each group plotted separately (b).



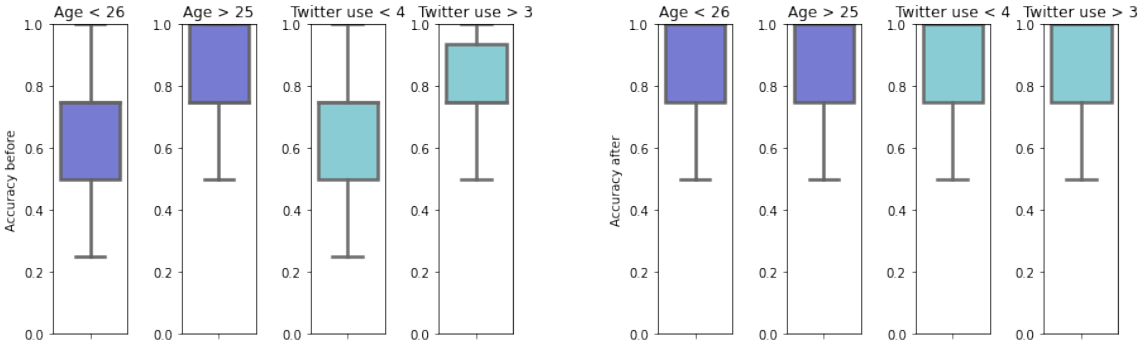
(a) (b)

Figure 6.6: A comparison of the accuracy before seeing the AI explanations (the baseline accuracy) and the accuracy after seeing the explanations for both interfaces (a) and a group wise comparison of the baseline accuracy (b).

The amount of time in seconds that participants spent at each interface was recorded. This is shown in Figure 6.5a. A one way ANOVA test between task performance duration and explanation interface resulted in a p-value of <0.05. The difference in task performance duration between each interface was significant. However, an ANOVA test comparing the duration between interfaces for both group 1 and group 2 resulted in respective p-values of <0.05 and >0.05. The difference is shown in Figure 6.5b.

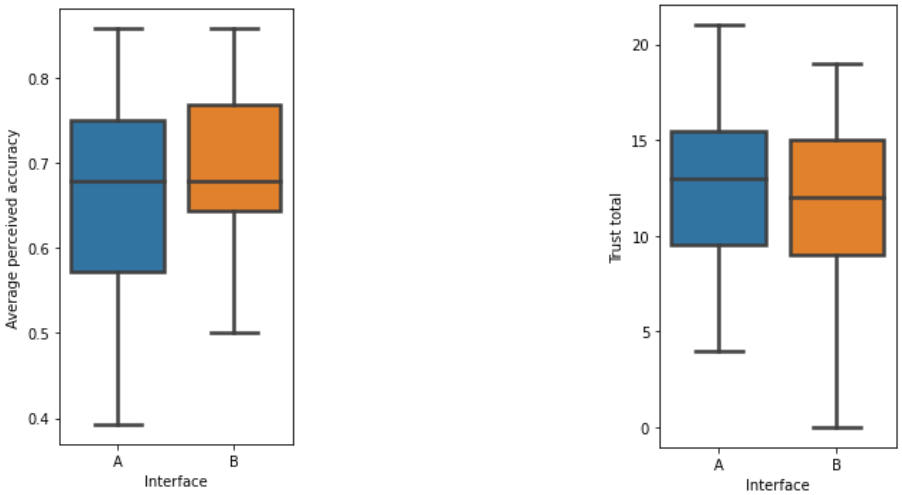
We further compared the accuracy of participants classifying fake and factual news for each tweet between both interfaces. Both the accuracy before (baseline accuracy) and after seeing the explanations were logged as users were asked to classify tweets in both cases. This allowed us to measure whether the accuracy would increase after having seen the explanation. In Figure 6.6a, we see an increase in the accuracy of participants having seen the explanations for both interfaces. We can see that the baseline accuracy between the interfaces does not differ. It does however differ for each tweet instance. There also is a very small difference in the accuracy gain between both interfaces.

The baseline accuracy was also compared group wise to ensure that the increase was not due to the learning effect. In other words, that the higher baseline accuracy in a certain interface was not simply due to users having dealt with the task beforehand. The comparison is shown in Figure 6.6b. The baseline accuracy is not strictly increasing implying that the learning effect might not been the reason why the baseline accuracy increased after seeing the first interface. We also compared the baseline accuracy against the collected demographic information. For this, we split the data in two different age groups (younger/older participants) and two different groups of Twitter users (less to no use/frequent users). The comparison is shown in Figure 6.7b.



(a) (b)

Figure 6.7: A box plot of the the baseline accuracy (a) and the accuracy after seeing the explanations (b) for different age groups and Twitter uses.



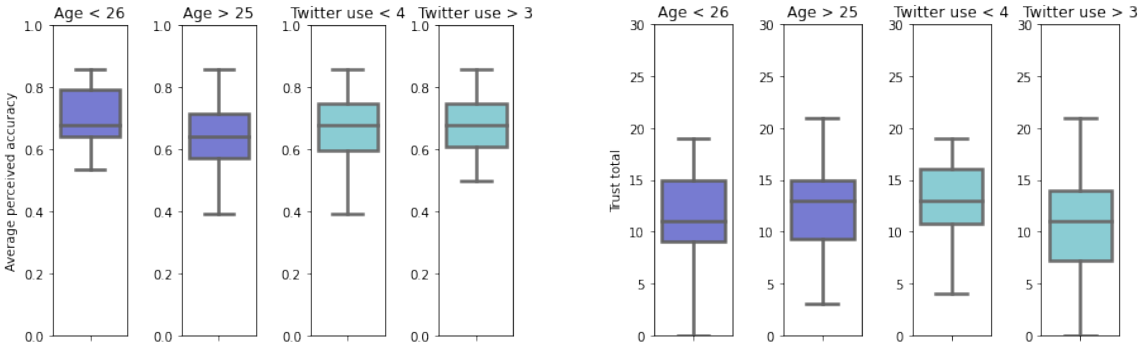
(a) (b)

Figure 6.8: A boxplot of the average perceived accuracy (a) and the total trust scores (b) for each interface.

A Wilcoxon signed rank test of perceived accuracy and interface resulted in a p-value of >0.05. The difference was not significant. However, we see that the mean perceived accuracy on both interfaces is high (± 0.7). This is shown in Figure 6.8a.

The difference between the total trust scores (from the Likert questionnaire) and the interface was tested with the Wilcoxon Signed Rank test. This resulted in a p-value <0.05. The difference was significant and in Figure 6.8b, we see that interface A resulted in a higher median trust score. However, both explanation techniques had a rather low median trust score considering that there are eight, 5-point questions.

Finally, we compared the trust factors against the different age groups and Twitter uses. This is shown in Figure 6.9a for the average perceived accuracy and in Figure 6.9b for the total trust scores.



(a) (b)

Figure 6.9: A comparison of the average perceived accuracy (a) and the total trust scores (b) for different age groups and Twitter uses.

6.2 Qualitative results

Q1: Which explanation method would you prefer for decision-making?

Out of 29 participants that answered this question, 22 answered LIME and 7 answered example-based.

The minority of the participants that preferred the example-based explanations answered found that they prefer seeing examples because it "corroborates with intuition" or because they believe it is more credible as it is labeled by people and experts. A participant also said they would rather look at the examples than spend their time looking at the explanations.

There were a lot of recurring themes as to why participants preferred the LIME explanations. These themes are shown in Table 6.1.

Table 6.1: Thematic analysis of the participant's answers on question 1 who picked LIME as the preferred explanation technique.

Themes	Mentions	Supporting quote
Understanding		
Easier to understand	3	*More easily explain possible outliers or wrongly classified tweets*
Easier to follow	1	
Better understanding of AI	3	*Better provides the technique behind the AI's decision making model and is similar to human decision-making process*
Taught something	1	*The first technique teaches you gradually how to notice fake or factual news*
Clearer	3	
Concrete, to the point	2	*Use of more specific information*
Explanation style		
Use of feature importance/weights/contribution	4	
Use of features/criteria/attributes	4	*I personally prefer the fact that the explanation is based on each "attribute" of the tweet*
Completeness		
More detail	1	*Seems more detailed with the 5 separate categories.*
More insight	1	*Better insights for decision-making*
Visualization		
Use of icons	1	*Icons help build understanding at a glance and catch my attention so I'm more likely to interact with it*
Providing charts	1	*Charts helps me to see which aspects it takes into consideration*

Participants who voted for LIME also used arguments as to why they would not want to use example-based. Opinions were mostly divided. Some participants mentioned that they do not find similarity to be convincing or a productive method for decision making. One participant said that "examples from the dataset can not cover new examples outside the dataset". A participant also found that example-based method was similar to a "hive-mind" approach. The participant may refer to the lack of effort that goes in grouping similar tweets and simply relying on the volume of labeled data to classify rather than actual critical thinking. There were also reasons why some participants would not prefer using LIME for decision making namely that some of the features (mainly popularity) were not deemed factors that should be used to predict fake or factual news. Additionally, a participant found the feature importance bar using in the LIME explanations unclear.

Q2: Which explanation method do you feel that better justified the AI prediction?

Out of 29 participants that answered this question, 25 answered LIME and 4 answered example-based. Again, the majority of participants chose for the LIME explanations and expressed that LIME explanations better justified the AI prediction. We again identified recurring themes in the answers to the question which is shown in Table 6.2. The minority of participants that voted for the example-based explanations found the explanations to be "immediately clear" and found that similarity to be closer to human understanding. Some arguments as to why participants preferred example-based, overlapped with the answers of the previous question. Participants presented arguments as to why they found the example-based explanations to be lacking. Some participants found the approach to be unclear and felt that the method "felt a bit like magic". There were also some technical errors mentioned such as that some presented examples may have deviated from the presented tweet. Again, similarity was not persuading enough for some. A participant found the example-based explanations to be "too much repetition".

Table 6.2: Thematic analysis of the participant's answers on question 2 who picked LIME as the preferred explanation technique.

Themes	Mentions	Supporting quote
Explanation style		
A better, clearer view of feature importance	3	*It gives me a better view of te importance from every features.*
Use of categories, criteria, ranking	5	*Breaking it down into scientific features having it matched against databases of sources and writing gives me personally a better justification than giving me similar tweets that are already judged and verified*
Understanding		
Clearer	4	*It makes it clearer on which features are important*
Easier to understand	2	
Easier to read	1	
Completeness		
Direct, more concrete	1	
More information, more insight	3	*It gives more insight in what an AI does*
Visualization		
Bars	1	*Filled bars give me direct insight*
Sliders	1	*I like the fact that you can see it on a "slider" level instead of only comparing out of similar tweets/users*

Q3: Were there problems or disagreements with either explanation methods that severely affected your trust in the AI?

The majority of participants answered "no" to this question. However, same participants who answered "no", already partially answered this question in the previous questions. We categorized the answers as follows:

Gaming the AI

Some participants answered this question by presenting a "what-if" case in which they were uncertain whether the AI would be able to successfully predict/explain the case. Participants mostly mentioned cases or scenarios were they believed that the AI could be fooled/gamed. For example: "but still I could write you a full capital tweet using the "coronavirus hoax" wording still being a factual one".

Reasoning

The popularity feature was a frequent occurrence of such a feature with a participant saying "Sometimes true news can be a hot topic and be highly popular as well.". A participant found the word frequency approach to be useful but mentioned the lack of human intelligence with this approach. The indicator of objective or subjective was also not found to be useful for the prediction. The lack of human intelligence, or the dissimilarity with human reasoning came up in the answers some times. One participant found that human reasoning can not be mimicked for all areas, implying that not all tasks can be performed by AI.

Feature importance

Some participants disagreed with the feature importance rather than the feature. "The importance of the 'Real user' feature was always low where I would expect that this would be very important" or "What made me a bit wary was that the AI generally gives more weight to content than source. I know that I usually first check if the source is reliable".

Usability

A participant found that the similar attributes button (showing similar users to a given Twitter user) lacked details and should have shown which features were similar. Another participant mentioned that "the words 'factual' and 'fake' both start with the same letters. It's easy for a user to be confused when classifying tweets.". For a few, it was not clear what the feature importance bar stood for.

Chapter 7

Discussion

In this chapter, we first discuss the results of the analysis that was done on the data in Chapter 4, including the resulting classifiers that were trained on this data. We then analyze the results of the user study and answer the research questions. Finally, we discuss the limitations of our work and propose future work.

7.1 Analysis and machine learning

Various continuous and boolean features from the literature [9, 48, 49] were analyzed including TFIDF n-grams extracted from the unstructured data [2]. There were many differences in correlation of features against the labels. For example, frequency of exclamations marks and other characters or the use of certain hashtags could not be considered useful for predicting fake news from the tweet text. Other features such as the amount of proper nouns [1] did however correlate strongly with the label. Splitting the source into reliable and unreliable sources also resulted in useful features. The overall difference in results was expected considering that some of these features considered informative were analyzed on differently populated (non COVID-19 related) datasets. The question rather is, how much of the features were transferable to the COVID-19 fake news detection problem. There were some features that show strong correlations with the label. For social bot detection, most of the features presented in literature can still be considered informative. This shows in the training of the classifier which resulted in a very high F1-score when tested against the test set. Similarly, the TFIDF classifier that learned on the n-grams, also resulted in high performance. However, these results should be interpreted with care as they may not generalize well on other populations.

We list the limitations of our methods:

- There were limited amount of informative features. More data mining could be applied to extract features that have useful correlations with the label. Additionally, other features can be explored from the literature that were proven informative for predicting fake news.

- The sources were split into reliable and unreliable sources. This might not be an ideal way of classifying sources. For example, a tweet could refer to an unreliable source but still present a factual claim. If there are many such examples in the dataset, then a reliable source may unjustifiably be labeled as unreliable. The sources should be collected and evaluated more carefully. It should be considered as a part (in the context) of the tweet content and not as a separate factor.

- For the n-gram features, context is still very important when classifying fake news. A Twitter user may mention "coronavirus is a hoax", but the tweet could rebuke the claim and still be considered factual as there may be other claims (undetected by the TFIDF classifier) present in the text. Furthermore, the extracted n-grams were not always guaranteed to be claims or were sometimes not related to COVID-19. Moreover, the extracted features mostly did not make sense semantically. The AI does not understand the extracted features and it is a challenge to use the extracted features as arguments in order to explain the outcome. To alleviate this, language models such as BERT [19] can be used to do classification on the unstructured textual data.

- There also were a lot of similar tweets in the data which decreased the importance of certain (important) n-grams. As a result, the classifier ignored these n-grams during training. The similar tweets can be filtered out from the dataset with the added trade-off that some of the unstructured data is then lost. The TFIDF feature had a very high feature importance when compared to other features. In order to counter this, we could give less weight to this feature during training to make sure that the classifier generalizes better. Another improvement that we should consider is to remove outliers in the dataset and do extensive preprocessing, especially on the unstructured data.

7.2 Research questions

7.2.1 Question 1

RQ1: Do example-based explanations lead to higher trust in a COVID-19 tweet misinformation classifier in comparison to LIME explanations?

LIME explanations lead to higher trust in the AI than example-based explanations. This was reflected by both the data of the trust questionnaire and the answers of the open question on justification where the large majority of participants preferred LIME. Although there was a significant difference in trust, the trust score that both methods achieved was rather low. The thematic analysis explains how this mistrust in some cases is both justified and unjustified. The average perceived accuracy on the other hand did result in a high median value for both interfaces. Unfortunately, there was no significant difference in average perceived accuracy between both explanation methods to further distinguish them.

Trust is by all means an important and reoccurring factor. We have seen this during the think-aloud studies where participants evaluated the AI explanations. In these studies, we clearly identified cases of justified and unjustified (mis)trust, which is key for interpreting AI predictions [26]. In order to identify these cases during user study, we asked users what negatively affected their trust in the AI. And the responses to this question were very useful to understand how participants reasoned and evaluated the explanations. In the literature study, we have seen that people mostly use counterfactual simulations to make causal judgments [37]. The qualitative study has shown that this is how some participants have evaluated certain features such as popularity, even though we did not imply that these were causes but rather correlations, and that each of the features should be considered in aggregate as contributions and not as a cause apart. For example, we have seen that popularity somewhat positively correlates with fake news. When we explain the fake news outcome of an instance that was very popular, and the explanation shows that popularity contributed to the outcome, it is likely that a participant upon seeing this explanation might reason about the popularity feature in isolation and thus momentarily ignoring the (low) feature importance or other features. When that is the case, the following usually happens as we also have seen during the think-aloud studies and as the qualitative data has shown. If popularity causes fake news then hypothetically, if popularity does not occur, then fake news would also not have. But this evidently is not true because we know that there are non-popular instances that are fake news as well. This reasoning is of course valid, but not when the feature is considered a correlation, and not a cause. It is not clear how much these types of faulty evaluations by participants have impacted their trust in the AI. Because in that case, the mistrust would be unjustified. We provided answers to the "how" question as a fail safe to ensure that users would interpret the explanations as correlations. More effort may be needed to clarify this to end users.

Another factor that may have affected trust was the source and content feature. We argued that the source should not be split into reliable and unreliable sources purely by comparing the source against the label. There are instances in the dataset that may refer to unreliable sources and still be factual and vice versa. We should mind the possible discrepancy between the content of the tweet and the sources used or referred to. Some participants have explicitly mentioned this in the qualitative data as reasons that affected their trust in the AI. We present the following quote by a participant:

> How does the AI respond to cases where a person disagrees with a fake news source? Does it mark the post as fake news? If so then it can be interpreted in both ways: either the user is rebuking with false information or the source is fake news. The opposite is possible as well (user rebuking correct information).

We have also seen that some participants clearly expected to see some kind of "human intelligence" in the AI. In the thematic analysis, intuition or human intuition was sometimes used either to support or to criticize the AI.

7.2.2 Question 2

RQ2: Do example-based explanations help users to better predict COVID-19 misinformation on Twitter in comparison to LIME explanations?

Both explanation methods helped with increasing the accuracy of predicting COVID-19 misinformation. It is hard to verify from the data whether this is solely due to the explanations. The study results do however reveal a high click-through rate on the explanation interfaces which means that there was a significant interaction with the explanation interfaces. The click-through rate on the LIME interaction was higher which is reflected by a higher task performance duration. It is also reflected by the thematic analysis which revealed that participants preferred using LIME for decision-making. However, the difference in task performance duration between the explanation methods was not significant when compared in between groups.

In the results, we only observed a higher task performance duration for the LIME explanations in group 1, where participants were shown LIME explanations first. The participants may thus have spent less time on the second interface due to fatigue. If we do not consider fatigue, it is still difficult to draw a conclusion as we have only measured the total amount of time spent on each interface. It is thus not exactly clear what the participants were doing, when they were spending more time on interface A in group 1. A lower trust in the AI explained by the example-based method could also explain why users spent much time with the LIME explanations. We know from [46] that participants may not want to spend time on or use explanations of AI that they do not trust.

The thematic analysis shows that the LIME explanations were easier to understand, more complete and overall a more valuable explanation for decision-making. It was said to contain more detail, provide more insight and visualization. For the example-based, some participants found this to be a more efficient way for decision-making, but generally a non-preferred explanation method for decision-making. Furthermore, the qualitative data revealed that some participants found the example-based method approach to provide a rather monotone experience (same interface, only some varying elements) and the LIME explanations provide a more interactive/dynamic explanation.

We also observed a significant increase in median accuracy for both interfaces. It would however be dangerous to assume that the increase would solely be due to explanations and not due to automation bias. In the user study, only true positives and true negatives were presented which means that the accuracy of participants would increase if they always agreed with the AI and always followed the outcome. Furthermore, it is also hard to conclude that either interface resulted in better accuracy as the accuracy gain between both interfaces does not differ much. However, tweets presented in a specific interface may have been more predictable. The learning effect was ruled out as the accuracy is not strictly increasing.

Finally, we noticed that participants sometimes gave similar answers to both questions and explicitly mentioned that they have used similar reasoning to answer both questions. This

is not surprising as both reliability (decision-making) and justification are both factors of trust. Although, the first question was aimed towards task performance. Sometimes participants copy pasted their answers. This may have been due to respondent fatigue.

7.3 Limitations

Some participants found that certain tweets were hard to classify. This could be mitigated by considering a dataset with a broader set of labels (satire, opinion, ...) rather than datasets that consist only of binary labels fake/factual news.

We did not use the self-reported AI expertise data in the results. It seems that most participants may have overestimated their understanding in the AI. The survey that was based on one question may be too limited in order to accurately collect "AI expertise" and should be surveyed using better methods.

There were limitations with the measurements. For example, information was lost by measuring the entire session duration. Measuring at specific intervals (time before seeing explanations, time after seeing explanations, hover time) could reveal more insight as to why participants for example spent more time on the LIME explanations. This could also reveal what they do during their time spent (hover, clicking, reading,...). Additionally, open questions could be asked at each tweet instance rather than at the end the session. The inconsistency with the order and the size (0-6 slider vs 5-point Likert) of the scales presented during the study was also a limitation.

Participants had to classify different set of tweets for each interface. It was hard to determine whether one set was easier to predict than the other. This could be alleviated by doing a between subjects study instead where participants are shown identical tweets in each group.

There was no mechanism for understanding why the participants chose to correct their answers. To understand whether the improvement of accuracy is due to the explanations or because participants follow the AI regardless, false positive or false negative instances could be placed between the presented tweets.

We only presented four tweets per interface. This might not be enough in order to assess trust as we know that that trust levels will be low in the beginning (Mohseni et al. [39]) and that the users would require more instances and more interaction with the explanations. Some participants in both the think-aloud studies and the user study have mentioned that explanations that provided new (and possibly confirming) beliefs boosted their trust in the AI. This could somehow be explained by the hypothesis in [26] where explanations can improve mental models which in return can contribute to higher and more justified trust levels.

7.4 Future work

During the think-aloud study, we noticed that task performance was somewhat negatively affected by low levels of trust in the AI. More specifically, that there were instances where the user could have improved their task performance if they relied on the AI's prediction but chose not to because of low levels of trust in the AI. An extreme case of this was a participant predicting the opposite label as the AI simply because of the lack of trust in the AI. An extra research question could be to see whether there is a correlation between trust factors (such as perceived accuracy) and task performance. This question was proposed in [39] as a challenge. Measuring the correlation could for example confirm if the trust factor in the AI is justified or unjustified. We omitted this extra research question due to time constraints.

The factual approach of example-based explanations was used in this work. In future work, researchers could compare LIME with counterfactual explanations rather than factual explanations. Moreover, they could measure factors other than trust or consider measuring more additional trust factors.

We presented some insights showing a difference in baseline accuracy or trust for different groups of participants. In future work, researchers could try to better understand whether Twitter use or age plays a significant role in trusting the AI.

New instances of COVID-19 related Twitter data were added to existing literature [15] or new literature [49, 50]. These could be used in future work to replace or extend the current data set to train the classifiers on more data.

Chapter 8

Conclusion

We started with an extensive literature study on the black box explanation problem and the metrics that exist in the field of XAI to evaluate explanations. Furthermore, we dived into the literature of Miller's work in [37] to better understand explanations and understand how humans evaluate these. In the related work section, we discovered a unique opportunity to build an explainable misinformation classifier to detect COVID-19 misinformation. This introduced many challenges as COVID-19 Twitter data at the time was rather limited and methods to detect misinformation in such a population were not well covered in literature. We based our work on methods used in misinformation classification and combined these with methods used for social bot detection. With the data and the methods, we retrieved, engineered and analyzed the COVID-19 Twitter data that included other external sources as well. The result was a misinformation classifier that could predict fake news based on the most informative features that were selected during analysis.

In addition to the misinformation classifier, we iteratively developed a prototype to present the tweet instances and AI explanations. As we considered a human-centered approach, think-aloud studies were conducted at each iteration in order to continuously evaluate the prototype. We chose to implement two different explanation methods in the prototype, namely LIME explanations as proposed by [46] and factual example-based explanations. We proposed research questions in which we tried to answer which explanation method would result in higher trust in the AI, and which method would improve accuracy of predicting COVID-19 misinformation in Twitter data. A user study was designed to collect the data.

We found that LIME explanations lead to higher trust in the AI than example-based explanations. The thematic analysis further revealed that participants found that the LIME explanations better justified the prediction. Unfortunately, despite the difference, the overall trust score was still low. This was partially explained with the qualitative data which showed that some participants used counterfactual reasoning to assess the value of the explanations. Moreover, disagreements with certain features and approaches may have further decreased trust in the AI.

We also found that both explanation methods helped with increasing the accuracy of predicting COVID-19 misinformation. Due to the limits of the study however, it was hard to verify whether this was solely due to the explanations. On the other hand, the qualitative data and click-through rate showed that participants preferred interacting with and using the LIME explanations for decision-making.

Appendix A

Table think-aloud study 1

Table A.1: Table (1/2) of issues and observations from the first think-aloud study.

Problem	**Type**	**Occurrences (out of 7)**	**Solution**
At first glance, user was not sure who posted the Tweet and did not notice the user information	UI	3	Increase font size and indicate the username. Hide user metadata until explanation
The user was confused by the misplacement of the warning label	UI	3	Put warning labels in correct place
Did not open the bar chart	UI	3	Put call to action
The why button on social bot explanation was ignored	Explanation	3	Increase presence
The confidence was not immediately obvious to the user	UI	2	Increase font size and provide explanation (color)
It was not clear that you could click the warning label	UI	2	Improve call to action
User was not sure whether the data was real or that the AI was actually working (rather than hardcoded)	Other	2	Try to clarify during the tutorial that it concerns real data
The user wanted to interact with the bar chart	Explanation	2	Make the bar chart interactive
The user was confused about the term “friends”, he expected the terms “followers”	Explanation	2	Change terms

Table A.2: Table (2/2) of issues and observations from the first think-aloud study.

Problem	**Type**	**Occurrences (out of 7)**	**Solution**
User did not understand some of the columns meant	Explanation	2	Provide a help button to explain
The user was initially confused by the special characters	UI	2	Parse special characters
The user wanted to click the URL to determine source	Other	2	Put anchor around URLs
The user wanted to close the explanation by clicking twice on the warning label	UI	1	Provide feature
User thought the AI would learn from the actions of the user	Other	1	No fix
The user did not view the user level explanations and jumped to the next tweet.	Explanation	1	Increase call to action
The user wanted to know when the Tweet was posted	Other	1	Include the timestamp
The user did not understand what a social bot was	Explanation	1	Provide information
The user expected the URL to be integrated below the tweet.	Other	1	Integrate the link and display images within the Tweet
The user was confused by the words in the bar chart that got stemmed during preprocessing	Explanation	1	Do not use the stemmed words
The user wanted to know what the source was for the bar chart	Explanation	1	Provide a short description of how the data was collected
The user did not find the explain user explanation to be convincing	Explanation	1	Extend feature

Appendix B

Table think-aloud study 2

Table B.1: Table (1/2) of issues and observations from the second think-aloud study.

Problem	**Type**	**Occurrences (out of 5)**	**Solution**
The user didn't understand that the bar was representing feature importance	Explanation	5	Add feature importance label and provide tooltip
Did not understand what "wary" meant	Other	4	Explain definition
It was not obvious that the LIME explanation buttons could be clicked on	UI	4	Instead of icons, provide actual buttons that call for action
User disagrees with subjective/objective	Explanation	3	/
Close the gap knowledge button was noticed after, but usually not clicked on, user would rather disagree than to close their knowledge gap or improve their understanding	UI	2	Provide easier access to these buttons.
User was too focused/spent too much time on the task	Other	2	Incentivize exploring explanations
Some links did not work	Other	2	Provide Tweets with working links

Table B.2: Table (2/2) of issues and observations from the second think-aloud study.

Problem	**Type**	**Occurrence (out of 5)**	**Solution**
Did not find the style explanation to be very convincing	Explanation	2	Add more style features
User didn't realize that the LIME explanation buttons, were in fact explanations	UI	1	Make buttons larger and change "contributing factors" to "explanations"
User profile expand button not obvious	UI	1	Show in introduction
Button finish not centered	UI	1	Fix
There are duplicate Tweets in neighbors	Other	1	Remove duplicates
More explanation needed with similar examples	Explanation	1	Add close knowledge gap buttons
There was bug in the LIME heatmap explanation	UI	1	Fix bug
User was not sure how to interpret the explanations (correlation) and did not understand the features affected the outcome	Explanation	1	Add a "how to interpret" button?
The social bot explanations sometimes gave contradictory explanations.	Explanation	1	This is mostly due to LIME
User was not sure how the social bot outcome impacted the AI outcome	Explanation	1	Either include the social bot outcome in the AI prediction or clearly indicate that this is only an indicative feature
User found that similar Tweets of popularity and writing style were not convincing	Explanation	1	We could add LIME as a mechanism behind the supporting examples to only show supporting examples when they make sense for the explanation

Appendix C

Questionnaires

Questionnaire

The AI = The Artificial Intelligence model shown and explained in the previous interface

1. I am confident in the AI. I feel that it works well.

○ I strongly agree ○ I agree ○ Neutral ○ I disagree ○ I strongly disagree

2. The outputs of the AI are very predictable.

○ I strongly agree ○ I agree ○ Neutral ○ I disagree ○ I strongly disagree

3. The AI is very reliable. I can count on it to be correct at all the time.

○ I strongly agree ○ I agree ○ Neutral ○ I disagree ○ I strongly disagree

4. I feel safe that when I rely on the AI, I will get the right answers.

○ I strongly agree ○ I agree ○ Neutral ○ I disagree ○ I strongly disagree

5. The AI is efficient in that it works very quickly.

○ I strongly agree ○ I agree ○ Neutral ○ I disagree ○ I strongly disagree

6. I am wary of the AI. (wary = cautious and careful with the AI's outcome)

○ I strongly agree ○ I agree ○ Neutral ○ I disagree ○ I strongly disagree

7. The AI can perform the task better than a novice human user.

○ I strongly agree ○ I agree ○ Neutral ○ I disagree ○ I strongly disagree

8. I like using the AI for decision making.

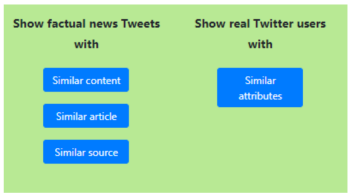
○ I strongly agree ○ I agree ○ Neutral ○ I disagree ○ I strongly disagree

Final questionnaire

Please answer the following questions.

1. Which explanation method would you prefer for decision-making?

Explanation technique 1

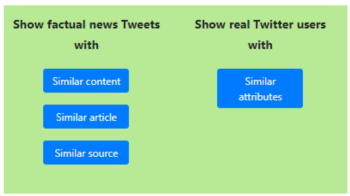


Explanation technique 2

Motivate your choice

2. Which explanation method do you feel that better justified the AI prediction?

Explanation technique 1



Explanation technique 2

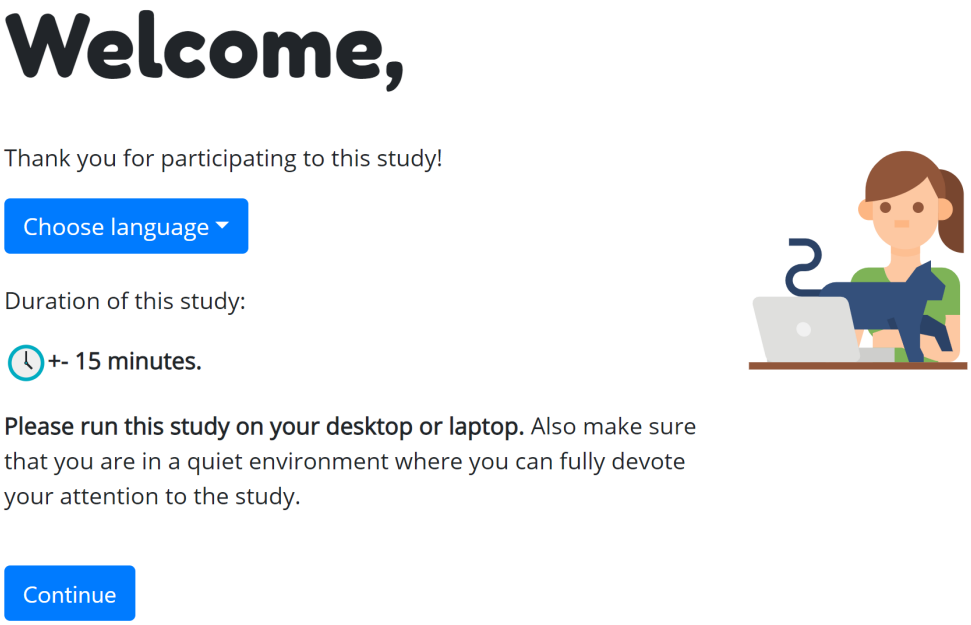
Motivate your choice

3. Were there problems or disagreements with either explanation methods that severely affected your trust in the AI?

4. Do you have any other comments?

Appendix D

Intro and informed consent



Who am I

I am Kenan Ekici and I am a Master IT student at the KU Leuven. You will be contributing to my Master's thesis through this study.

If you encounter problems or if you have questions, send an email to **kenan.ekici@student.kuleuven.be**

Back Continue

About

Title research: Illuminating the black box: Explaining machine learning algorithms that detect fake news on Twitter

I am researching **Explainable AI (XAI)** solutions in the context of COVID-19 fake news prediction on Twitter.

What is explainable AI?

The lack of transparency towards Artificial Intelligence models makes it hard for people without expert knowledge to understand how they work. This can be problematic. For instance, when we have to rely on these AI models. Explainable AI entails solutions that make it possible for the Artificial Intelligence model to explain its predictions.

Back Continue

Study progress

The study will progress as follows:

1. Explore the first interface and finish presented tasks.
2. Fill in questionnaire.
3. Explore the second interface and finish presented tasks.
4. Fill in questionnaire.
5. Fill in final questionnaire.
6. Done

Back Continue

Informed consent

- **The study will be conducted in English. I have an adequate profiency in English.**
- **I am familiar with Twitter and I am at least 18 years old.**
- I will consider taking the study serious and answering the questionnaires honestly.
- I understand what is expected of me during this research.
- I know that my participation may be associated to risks or discomforts: exposure to tweets containing misleading information.
- I or others can benefit from this research in the following ways: reflect on self-trust in Artificial Intelligence models, learn about Artificial Intelligence models and increase digital literacy.
- I understand that my participation to this study is voluntary. I have the right to stop participating at any time. I do not have to give a reason for this and I know that it will not have any negative repercussions for me. In the context of the GDPR the collected data will be processed under public interest as the legal basis. When you end your participation the data that were already collected can still legally be included in the research and do not need to be deleted by KU Leuven.
- I consider doing the study alone and do not use any other external tools, information or help. I can contact Kenan Ekici at kenan.ekici@student.kuleuven.be during the study in case I need help.
- The questionnaires are fully anonymous.
- The results of this study can be used for scientific goals and may be published. My name will not be published. The anonymity and confidentiality of the data will be protected in all stages of the research.
- For questions and for the execution of my rights (access to my data, rectification of the data, ...) after my participation I know that I can contact: kenan.ekici@student.kuleuven.be. More information with regard to privacy in research can be found at https://kuleuven.be/privacy/en/. With further questions about privacy issues I can contact the data protection officer: dpo@kuleuven.be
- In case of complaints or other concerns with regard to the ethical aspects of this research I can contact the Social and Societal Ethics Committee of KU Leuven: smec@kuleuven.be

Download informed consent

I have read and understood the information on this page and I have received an answer to all my questions regarding this research. Click the check box below to give your consent to participate. The study will proceed in English.

☐ I give my consent to participate.

Back Continue

About you

How old are you?

Enter age

How would you describe your own expertise in Artificial Intelligence? (0 = none, 6 = expert)

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6

How often do you use (visit) Twitter?

○ several times a day
○ about once a day
○ 3-5 days a week
○ 1-2 days a week
○ every few weeks
○ less often
○ never

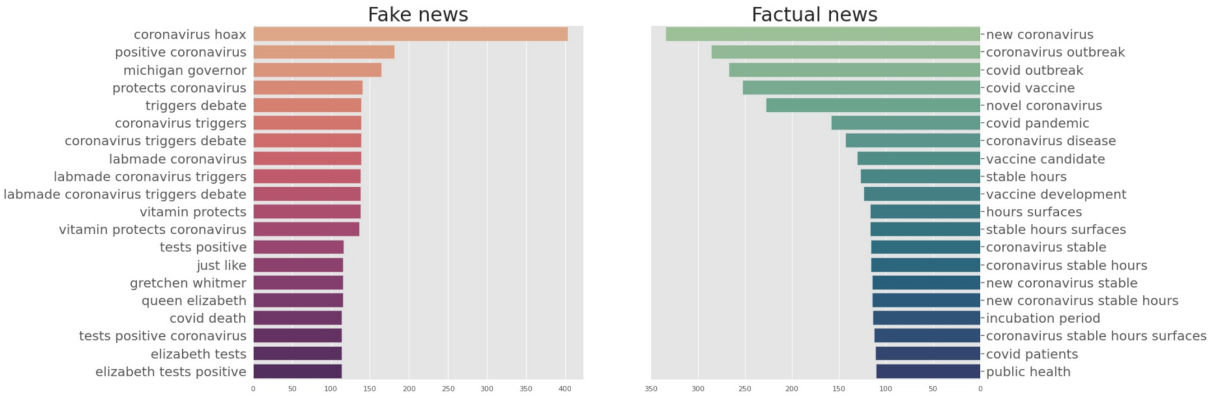
Back Start study

Appendix E

Extra reading material

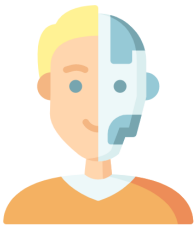
Claims

There are certain false and true claims about COVID-19 that are continously being spread, retweeted and reused in different contexts. The reason we know why they are false or true is because the claims were factchecked by community or professionals. Due to frequent occurence of these claims, it is possible for an algorithm to collect these claims and turn them into useful features. The AI can then learn from these features in order to predict fake or factual news. It will use both the absence and presence of certain claims to make this prediction. Some of these claims have more weight than others as they occur more frequently. Below, you can see a list of the most important claims for both fake and factual news. Note that the algorithm does not understand the claims nor is it entirely certain that they are indeed claims. It is possible that it collects a combination of words that are not necessarily claims, have anything to do with COVID-19, or even make sense.



Social bots?

Social bots are automated Twitter accounts that may be used in large volumes for malicious reasons, such as spreading fake news!



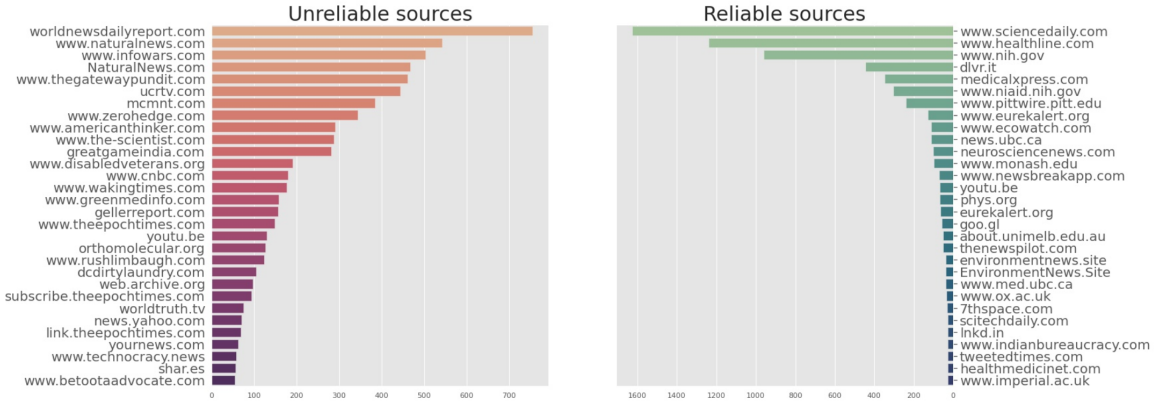
How does the AI know the difference?

There are several key indicators that the AI uses to distinguish between real users and social bots.

- Followers to friends ratio (#followers/#friends) — **Strongly** negatively correlated
- Likes to age ratio (#friends/#days active)
- Favourites count
- Statuses count
- Username length
- Tweet to age ratio (#tweets posted/#days active)
- Has url (yes/no)
- Has location (yes/no)

Sources

The source may be an important if not the most important element we look for in order consider if a claim is valid. Much like traditional media, there exist a set of news outlets on the web that may for most people be considered untrustworthy. It goes without saying that any tweet that uses these types of sources to back their claim could by default be considered misinformation or should at least be interpreted with care. However, even the sources that we consider trustworthy may be simply wrong. With the latter, we know at least that this is more of a rare occurrence.



How can the AI tell?

The AI has a one dimensional definition of a source. It uses two lists of collected sources, one that is more associated with misinformation and one that is known to contain more reliable sources of information. In the lists below, you may see some sources in the unreliable sources section that you might consider trustworthy. And that may of course be true. Unfortunately however, those sources (at time of collection) were frequently used to back up misinformation more than they were used to back up factual information.

Writing style

Tweets that contain misinformation may be written in a way to not only capture your attention, but also to persuade you. Studies such as [1] have shown that fake news is written similarly as "satire" and that fake news writers tend to use certain heuristics to try and persuade you of their false claims. The allowed maximum length of a Tweet means that writers of fake news may tend to use stop words and squeeze in many proper nouns as possible.

They may also use unnecessary amount of capital letters or hash tags in their Tweets to create a false sense of urgency or discomfort. It may also be worthwhile to consider whether the Tweet is written subjectively or objectively or carries a certain sentiment (angry vs. neutral).

How can the AI tell?

The AI considers some of the heuristics mentioned above during its prediction. The way it predicts whether a Tweet is written objectively or subjectively is by considering the amount of verbs and adjectives as subjectively written content tends to contain more adjectives and adverbs [2].

References:

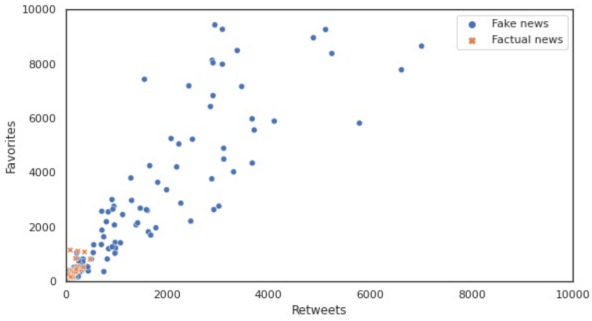
[1] This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News

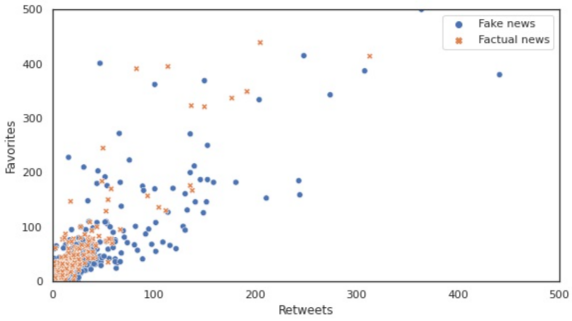
[2] Adjectives as indicators of subjectivity in documents

Social impact

Misinformed tweets are likely to become very popular as they are retweeted and favorited significantly more than tweets that contain reliable factual information. There are a variety of reasons why this may be the case. The tweet for example might be retweeted by someone with a large social outreach or importance.

Consider the plot on the left below. Notice in this case how tweets with thousands of retweets and likes are labeled as fake news. When we zoom in, it becomes harder to distinguish fake news. However, we see that factual news tweets simply do not receive the same amount of outreach.





Bibliography

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, PP:1–1, 09 2018.

[2] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1:e9, 12 2017.

[3] Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, and Preslav Nakov. Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society, 2020.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019.

[5] Thor Benson. Twitter Bots Are Spreading Massive Amounts of COVID-19 Misinformation, November 2020.

[6] Bettina Berendt, Rafael Hautekiet Peter Burger, Alexander Pleijter Jan Jagers, and Peter Van Aelst. Factrank: Developing automated claim detection for dutch-language fact-checkers. 11 2020.

[7] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. How to recommend?: User trust factors in movie recommender systems. pages 287–300, 03 2017.

[8] J. Brooke. Sus: A 'quick and dirty' usability scale. 1996.

[9] Cody Buntain and Jennifer Golbeck. Automatically identifying fake news in popular twitter threads. *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, Nov 2017.

[10] Alex Cao, Keshav Chintamani, Abhilash Pandya, and Richard Ellis. Nasa tlx: Software for assessing subjective mental workload. *Behavior research methods*, 41:113–7, 03 2009.

[11] Siming Chen, Lijing Lin, and Xiaoru Yuan. Social media visual analytics. *Computer Graphics Forum*, 36:563–587, 06 2017.

[12] GIOVANNI LUCA CIAMPAGLIA and FILIPPO MENCZER. These are the three types of bias that explain all the fake news, pseudoscience, and other junk in your News Feed, November 2020.

[13] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 191–198, New York, NY, USA, 2016. Association for Computing Machinery.

[14] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 963–972. International World Wide Web Conferences Steering Committee, 2017.

[15] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset, 2020.

[16] Dr. Dataman. *Explain Your Model with the SHAP Values*, 2019 (accessed January 23 , 2021).

[17] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 2016.

[18] Ryan Daws. Twitter's latest acquisition tackles fake news using AI, November 2020.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[20] ec.europa.eu. *Trustworthy AI*, 2020 (accessed October 28, 2020).

[21] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, Jun 2016.

[22] Vijaya Gadde and Kayvon Beykpour. Additional steps we're taking ahead of the 2020 US Election, November 2020.

[23] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models, 2018.

[24] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, Jun. 2019.

[25] Michael Henderson, Ke Jiang, Martin Johnson, and Lance Porter. Measuring twitter use: Validating survey-based measures. *Social Science Computer Review*, page 089443931989624, 12 2019.

[26] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects, 2019.

[27] https://kpattorney.com. *Liability Self-Driving cars*, 2020 (accessed October 28, 2020).

[28] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter?: User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–27, 11 2019.

[29] Mark T. Keane and Barry Smyth. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In Ian Watson and Rosina Weber, editors, *Case-Based Reasoning Research and Development*, pages 163–178, Cham, 2020. Springer International Publishing.

[30] Josua Krause, Adam Perer, and Enrico Bertini. A user study on the effect of aggregating explanations for interpreting machine learning models. 08 2018.

[31] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. volume 2015, 03 2015.

[32] Onyi Lam. *Detecting subjectivity and tone with automated text analysis tools*, 2018 (accessed March 16, 2021).

[33] Susan Li. *COVID-19 fake news dataset*, 2020 (accessed October 28, 2020).

[34] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Apr 2020.

[35] Lion Gu, Vladimir Kropotov, and Fyodor Yarochkin. The fake news machine: How propagandists abuse the internet and manipulate the public, 2017.

[36] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[37] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2018.

[38] Sina Mohseni, Eric Ragan, and Xia Hu. Open issues in combating fake news: Interpretability as an opportunity, 2019.

[39] Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., ... Ragan, E. Trust evolution over time in explainable ai for fake news detection. *CHI20*, Apr 2020.

[40] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan 2020.

[41] Vincent Müller and Nick Bostrom. *Future Progress in Artificial Intelligence: A Survey of Expert Opinion*, volume 1, pages 553–571. 03 2016.

[42] Greg Nichols. More than half of Twitter's 'Reopen America' calls from bots, study finds, November 2020.

[43] Jakob Nielsen. *Thinking Aloud: The Number 1 Usability Tool*, 2012 (accessed March 16, 2021).

[44] Gordon Pennycook. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 06 2018.

[45] Caitlin Petre, Brooke Erin Duffy, and Emily Hund. "gaming the system": Platform paternalism and the politics of algorithmic visibility. *Social Media + Society*, 5(4):2056305119879995, 2019.

[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[47] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.

[48] J. Schnebly and S. Sengupta. Random forest twitter bot classifier. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0506–0512, 2019.

[49] Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. An exploratory study of covid-19 misinformation on twitter, 2020.

[50] Gautam Kishore Shahi and Durgesh Nandini. Fakecovid-a multilingual cross-domain fact check news dataset for covid-19.

[51] Elisa Shearer and Katerina Eva Matsa. News Use Across Social Media Platforms 2018, November 2020.

[52] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. 05 2019.

[53] Sina Mohseni and Niloofar Zarei and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems, 2018.

[54] Twitter. Transparency Reports, November 2020.

[55] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA, 2019. Association for Computing Machinery.

[56] Adam Waytz. The Psychology Behind Fake News, November 2020.

[57] Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. Xfake: Explainable fake news detector with visualizations. *The World Wide Web Conference on - WWW '19*, 2019.

[58] Jianlong Zhou, Syed Z. Arshad, Simon Luo, and Fang Chen. Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making. In Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler, editors, *16th IFIP Conference on Human-Computer Interaction (INTERACT)*, volume LNCS-10516 of *Human-Computer Interaction – INTERACT 2017*, pages 23–39, Bombay, India, September 2017. Springer International Publishing. Part 1: Security and Trust.

[59] Xinyi Zhou and Reza Zafarani. A survey of fake news. *ACM Computing Surveys*, 53(5):1–40, Oct 2020.

AFDEL
Straat nr bus
3000 LEUVEN, BEL
tel. + 32 16 00 0
fax + 32 16 00 0
www.kuleuven.be