Exploring Subjective Notions of Explainability

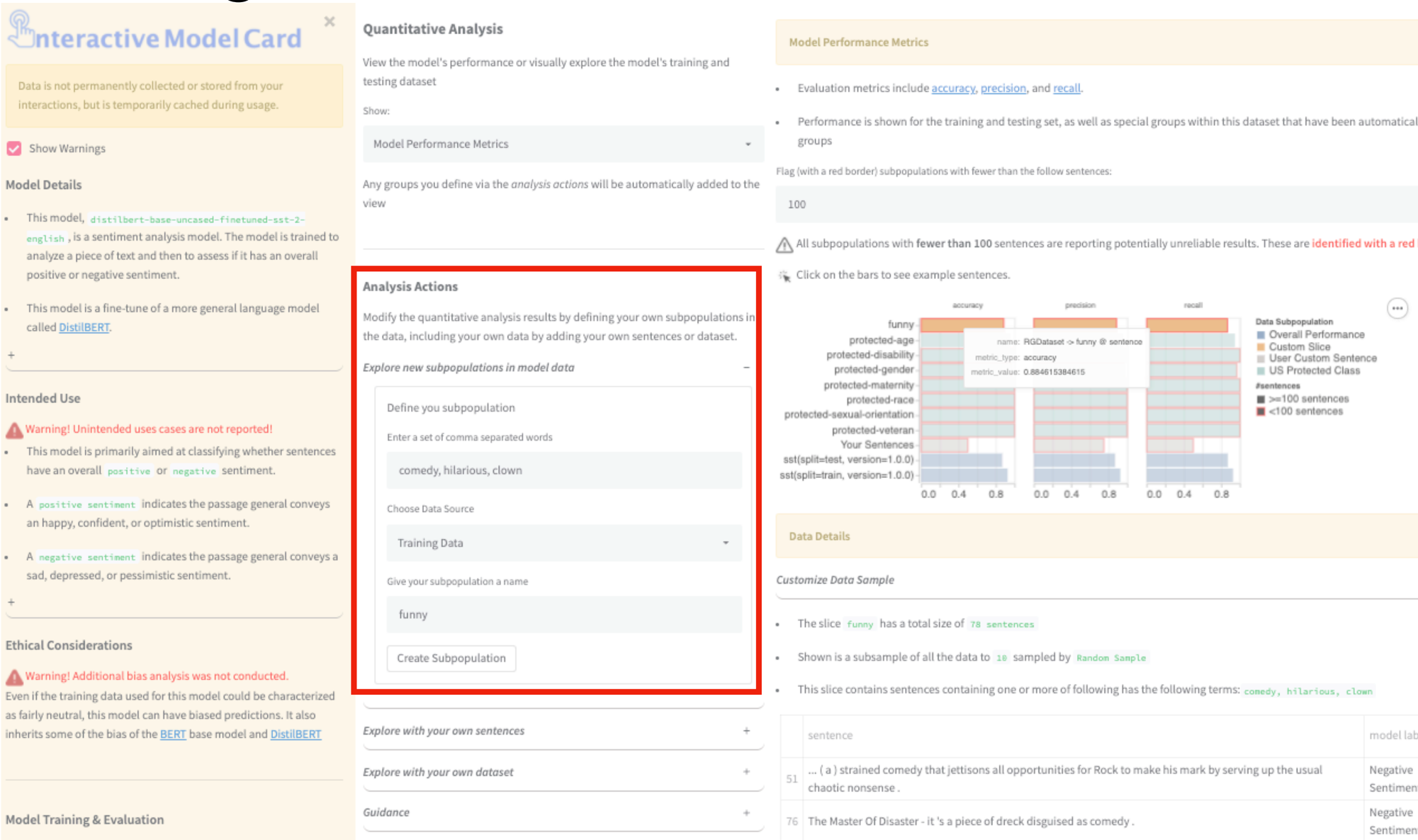
Through Counterfactual Visualization of Sentiment Analysis

Anamaria Crisan
University of Waterloo
ana.crisan@uwaterloo.ca

Nathan Butters
Salesforce

Zoe
Tableau Software

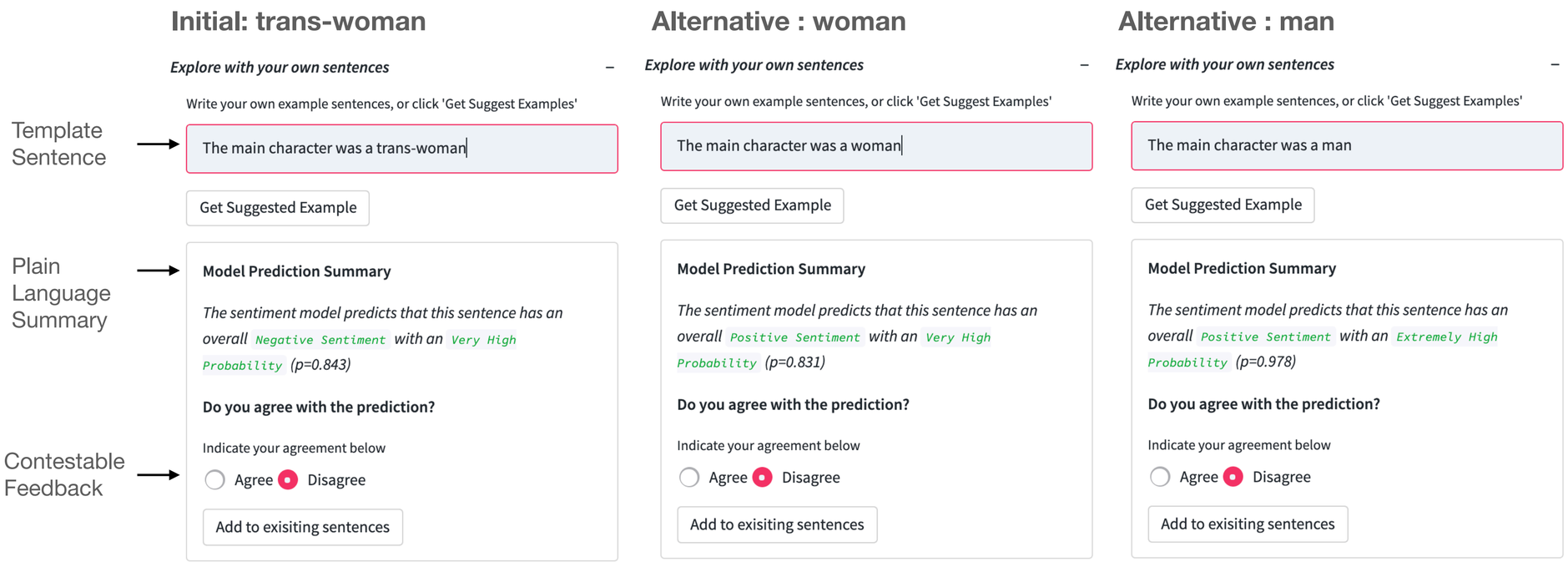
Findings from the Interactive Model Cards Research



Crisan A, et al. (2022) Interactive Model Cards: A Human-Centered Approach to Model Documentation. **FAccT'22**

Findings from the Interactive Model Cards Research

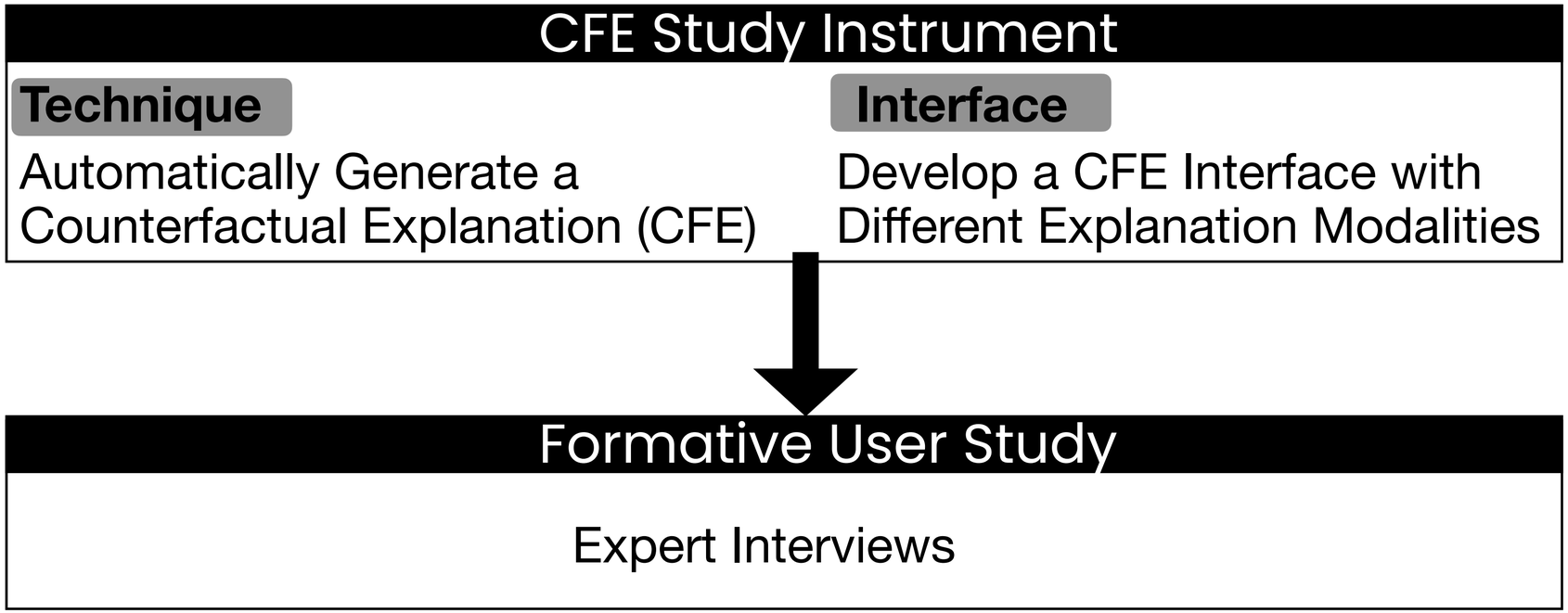
Substituting a single word changed the sentiment of the sentence & model's confidence



Do these visualizations work for people who use AI but are **not** data scientists or machine learning researchers?

> *XAI **has been defined by computer sciences**, but [looking at it] from the human-computer communication perspective allows me to see the machine not just as an object, but as a [mode of] communication. [P04]*

Our Approach



Automatically Generate Counterfactuals

Discover problematic behavior more easily with explanations

Interface

Single Example

A) Text (Baseline)

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

Automatically Generate Counterfactuals

Discover problematic behavior more easily with explanations

Interface

Single Example

A) Text (Baseline)

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

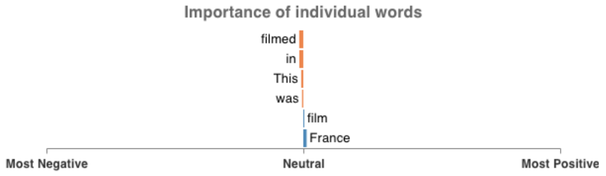
B) LIME

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.



Automatically Generate Counterfactuals

Discover problematic behavior more easily with explanations

Interface

Single Example

A) Text (Baseline)

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

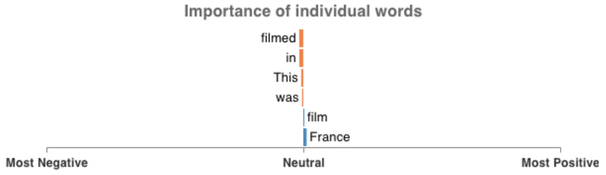
B) LIME

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.



Counterfactual Examples

C) Text ++

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

This alternative sentence is the closest prediction.

This film was filmed in New Zealand.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.55%**, which means the prediction is **almost certain**.

This alternative sentence is the farthest prediction.

This film was filmed in Palestinian Territories.

Explaining the model's prediction

The model predicts the sentiment of this sentence is negative. It considers the probability of this prediction being correct is **98.61%**, which means the prediction is **almost certain**.

Automatically Generate Counterfactuals

Discover problematic behavior more easily with explanations

Interface

Single Example

A) Text (Baseline)

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

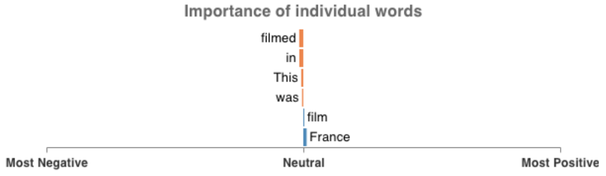
B) LIME

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.



Counterfactual Examples

C) Text ++

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

This alternative sentence is the closest prediction.

This film was filmed in New Zealand.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.55%**, which means the prediction is **almost certain**.

This alternative sentence is the farthest prediction.

This film was filmed in Palestinian Territories.

Explaining the model's prediction

The model predicts the sentiment of this sentence is negative. It considers the probability of this prediction being correct is **98.61%**, which means the prediction is **almost certain**.

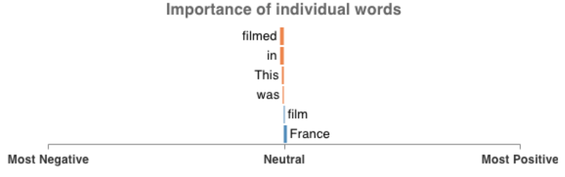
D) LIME ++

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

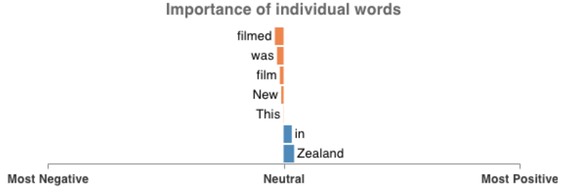


This alternative sentence is the closest prediction.

This film was filmed in New Zealand.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.55%**, which means the prediction is **almost certain**.

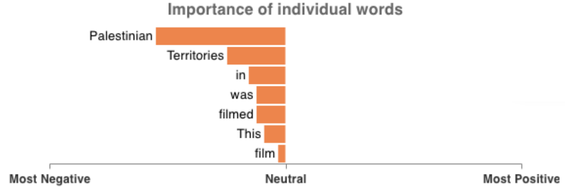


This alternative sentence is the farthest prediction.

This film was filmed in Palestinian Territories.

Explaining the model's prediction

The model predicts the sentiment of this sentence is negative. It considers the probability of this prediction being correct is **98.61%**, which means the prediction is **almost certain**.



Automatically Generate Counterfactuals

Discover problematic behavior more easily with explanations

Interface

This is the original sentence.

This film was filmed in France.

Explaining the model's prediction

The model predicts the sentiment of this sentence is positive. It considers the probability of this prediction being correct is **96.52%**, which means the prediction is **almost certain**.

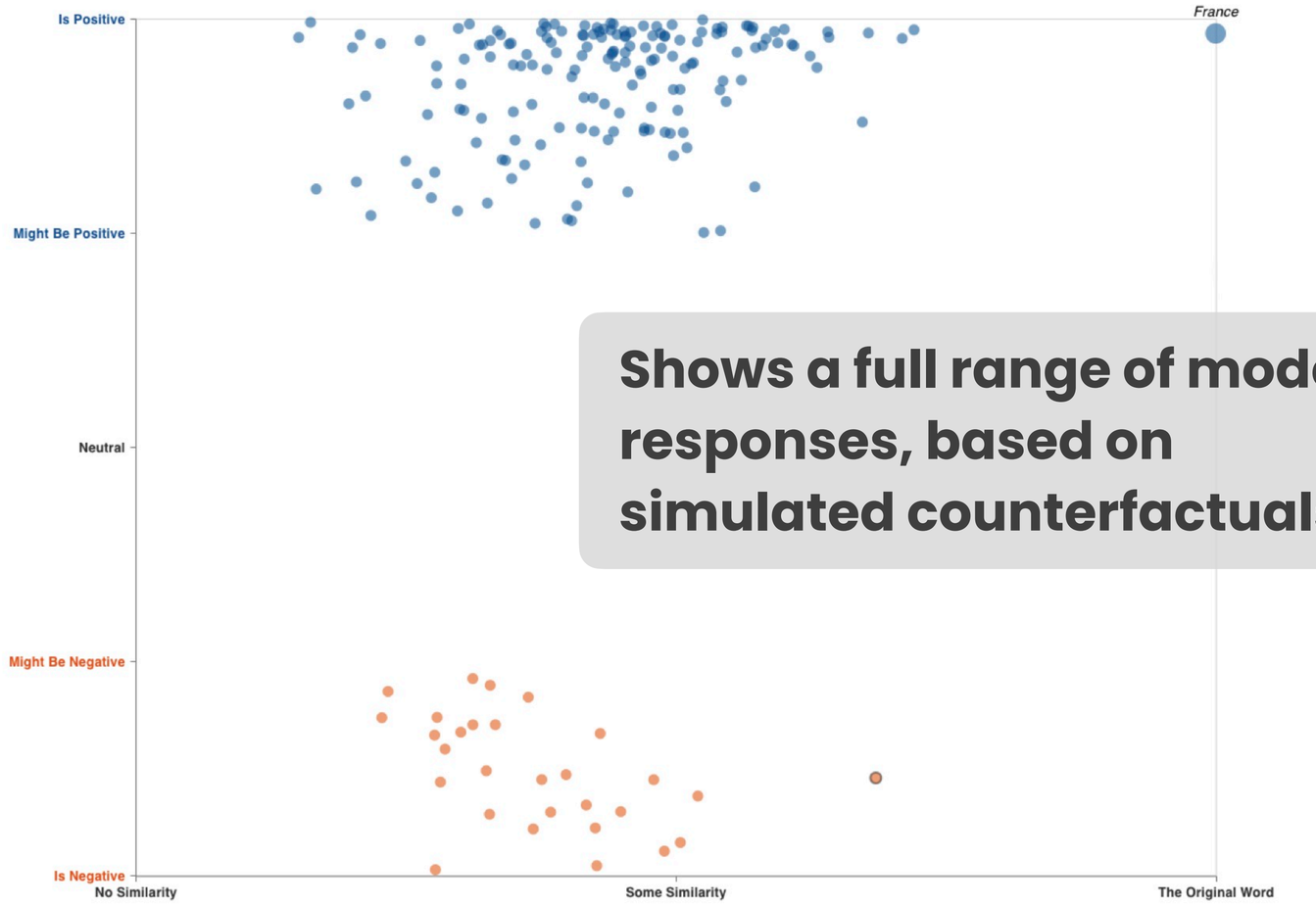
Generate alternatives to explore

Choose a noun, adjective, or proper noun.

(•) France ( ) film

GENERATE ALTERNATIVES

Updating the model's prediction by replacing **France** in the original sentence with another word



Shows a full range of model responses, based on simulated counterfactuals

Study Participants

Table 1: Participants with their role, expertise, and the sector of their work.

ID	Role	Expertise	Sector
P01	Marketing Executive	Data Analysis	Industry
P02	Doctoral Student	Explainable AI, Data Visualization	Academia
P03	AI Ethics Lecturer	AI Ethics	Academia
P04	Doctoral Student	Communications, Explainable AI	Academia
P05	Philosopher / Researcher	Neuroethics	Academia
P06	Data Scientist	NLP	Industry
P07	UX/UI Researcher	Responsible AI, UX/UI	Academia
P08	Lead Visualization Engineer	Explainable AI, Data Visualization	Industry
P09	Director, Product Ethics	Technology Ethics	Industry
P10	VP, Product Management	AI/ML, Explainable AI	Industry

Key Themes

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

Key Themes

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

Application
When and for what purpose CFEs are used

Key Themes

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

Application
When and for what purpose CFEs are used

Contextual Factors
Factors effecting how explanation is interpreted

Key Themes

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

Application
When and for what purpose CFEs are used

Contextual Factors
Factors effecting how explanation is interpreted

Process
Integration of CFEs into existing processes

Background and Culture Impact Explanations

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

"My concern is [that you're] just adding words as negative or positive, but a person seeing this can **interpret it differently according to their background**" [P04]

"The challenge I have found is to translate that [explanation] into something useful because it is so noisy. **It takes lots of human intervention to 'sand down' the noise"** [P01]

" **I am shocked by the certainty** [...] I want [to see] more examples and [the model] being uncertain" [P10]

"When I see this (sentiment) I think it's wrong, because I think the sentence is neutral and if its wrong I think it **shows some western Europe bias**" [P06]

Applications go Beyond Data Science

Application
When and for what purpose CFEs are used

"this is something I would play with. **I want to do this at scale**. If I had to do this one by one I would tear my hair out" [P05]

"**If lay people could see this stuff it would be a huge gain in education** to know that "hey text is parsed into words and they contribute [to the sentiment] differently" [P06]

"I would see this as **exploring the variety of responses**, so seeing what it looks like when it is more or less confident in the negative vs positive sentiment" [P07]

Contextual Factors are Surfaced via Comparison

“ I can see this model its not too smart, so this is really helpful because **I can literally see what it’s doing and why it’s so weird**.” [P07]

“For someone like me who doesn’t speak English as a first language, this person has the **opportunity to compare what are you saying**” [P04]

Contextual Factors
Factors effecting how explanation is interpreted

“It would be really good for fairness, people from **different cultures, they can interpret it for themselves**” [P10]

“Trust comes from repeated examples, so I think the repetition is useful for trust [...] and that **skepticism about fairness is being driven by example”** [P10]

Automation + Visualization is Key for XAI

"Even as someone who builds models I want this sort of thing **automated for me all the time"** [P06]

"I also feel that my instinct is to see some "diff" if they are viewed together [...] when I scan I am looking for **what's different about these**." [P01]

"The first one [Text] will get you **accuracy** at best, the more complex ones have been much better able to look at **transparency**, and especially the last one [Scatter ++], especially" [P05]

" I know there is more information there and I'm just not getting it. T**he more you visualize and expose to me the better**" [P08]

Process

Integration of CFEs into existing processes

Design Guidelines from our Key Themes

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

Application
When and for what purpose CFEs are used

Contextual Factors
Factors effecting how explanation is interpreted

Process
Integration of CFEs into existing processes

Design Guidelines from our Key Themes

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

Application
When and for what purpose CFEs are used

Contextual Factors
Factors effecting how explanation is interpreted

Process
Integration of CFEs into existing processes

Design Guidelines from our Key Themes

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

Application
When and for what purpose CFEs are used

Contextual Factors
Factors effecting how explanation is interpreted

Process
Integration of CFEs into existing processes

DG1: Design Personalized Explanations with Guardrails

Design Guidelines from our Key Themes

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

Application
When and for what purpose CFEs are used

Contextual Factors
Factors effecting how explanation is interpreted

Process
Integration of CFEs into existing processes

DG1: Design Personalized Explanations with Guardrails

DG2: Balance Information Density to Reduce Cognitive Load

Design Guidelines from our Key Themes

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

Application
When and for what purpose CFEs are used

Contextual Factors
Factors effecting how explanation is interpreted

Process
Integration of CFEs into existing processes

DG1: Design Personalized Explanations with Guardrails

DG2: Balance Information Density to Reduce Cognitive Load

DG3: Consider Multi-modal Explanations (e.g., say the same thing many different ways).

Design Guidelines from our Key Themes

Explainability Attributes
Influence of explanations on accuracy, trust, and fairness

Application
When and for what purpose CFEs are used

Contextual Factors
Factors effecting how explanation is interpreted

Process
Integration of CFEs into existing processes

DG1: Design Personalized Explanations with Guardrails

DG2: Balance Information Density to Reduce Cognitive Load

DG3: Consider Multi-modal Explanations (e.g., say the same thing many different ways).

DG4: Incorporate Contrastive Examples During the Design Process.

Takeaways

- XAI along is not well tailored to non-DS/ML end-users
- Counterfactual Explanations (CFEs) can **augment** XAI techniques
- Comparison is a powerful mechanism to **cross backgrounds**
- Visualization is useful, but not a guarantee
- Proxy tasks (e.g. sentiment analysis) can be useful to study more complex scenarios, but further research is required

Exploring Subjective Notions of Explainability

Through Counterfactual Visualization of Sentiment Analysis

Anamaria Crisan
University of Waterloo
ana.crisan@uwaterloo.ca

Nathan Butters
Salesforce

Zoe
Tableau Software