ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

Are Shortest Rationales the Best Explanations for Human Understanding?

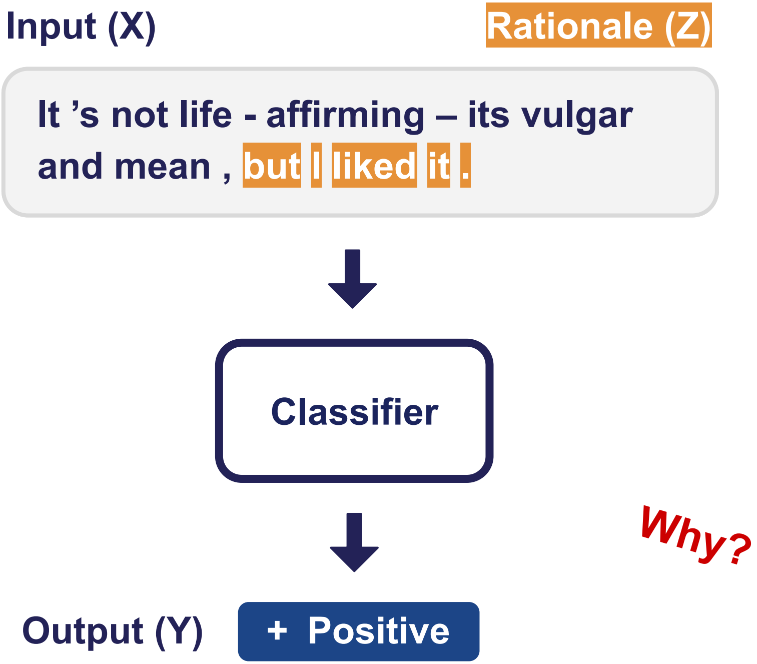
[✦]**Hua Shen**, [♦]Tongshuang Wu, [✦]Wenbo Guo, [✦]Ting-Hao ‘Kenneth’ Huang

[✦]The Pennsylvania State University, [♦]University of Washington

PennState
College of Information
Sciences and Technology

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Interpretation with rationales



Input (X)

Rationale (Z)

It ’s not life - affirming – its vulgar and mean , but I liked it .

Classifier

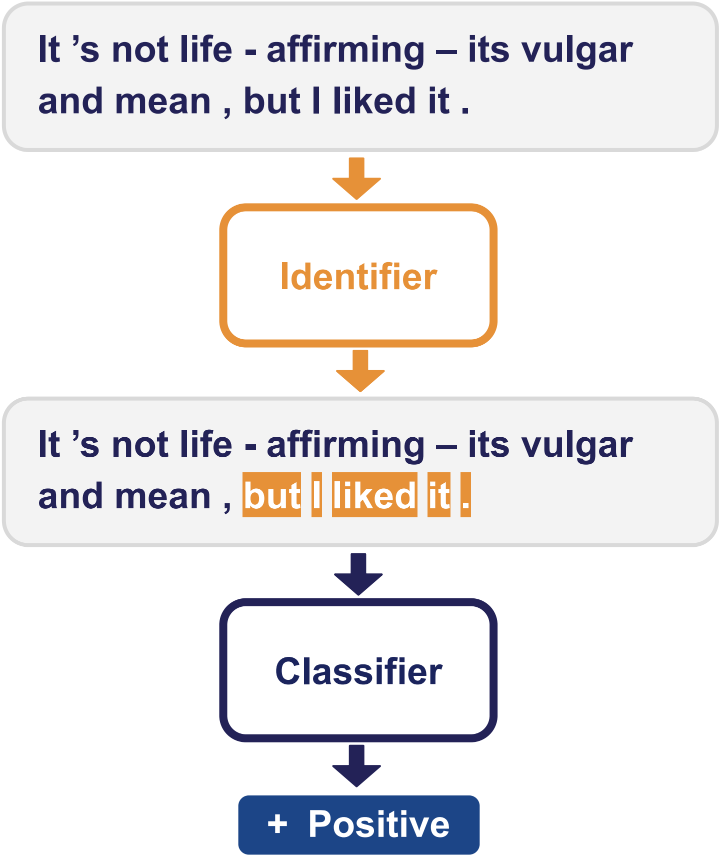
Why?

Output (Y)

+ Positive

Rationale: a sufficient **subset of input** text to **explain** the **model’s prediction**.

Self-explainable rationalizing methods



It ’s not life - affirming – its vulgar and mean , but I liked it .

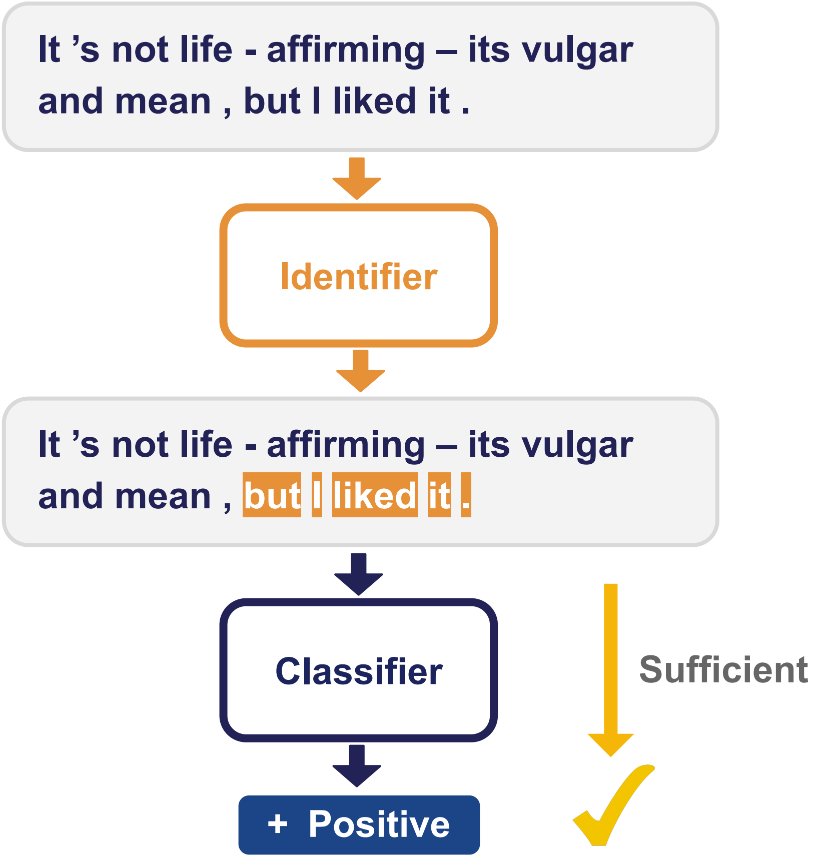
It ’s not life - affirming – its vulgar and mean , but I liked it .

What is a good rationale?

Sufficiency

Conciseness

What is a good rationale?



It ’s not life - affirming – its vulgar and mean , but I liked it .

Identifier

It ’s not life - affirming – its vulgar and mean , but I liked it .

Classifier

+ Positive

Sufficient

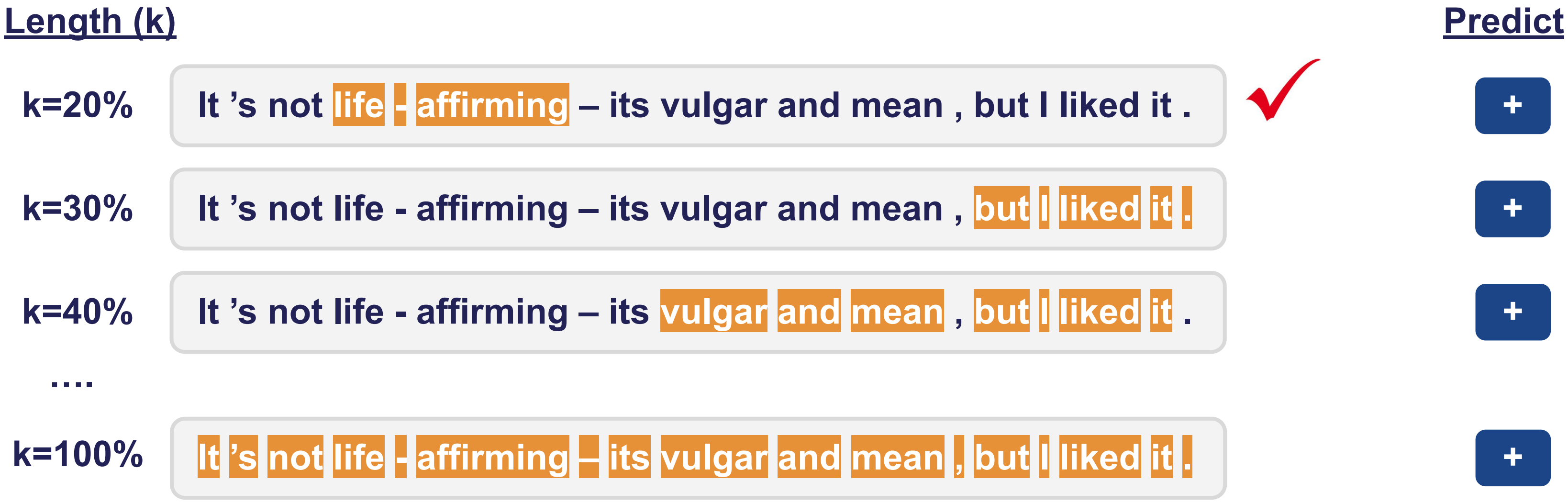
✓

Sufficiency

Sufficient subset of input to predict correct label

What is a good rationale?

Conciseness Rationale length to be as short as possible

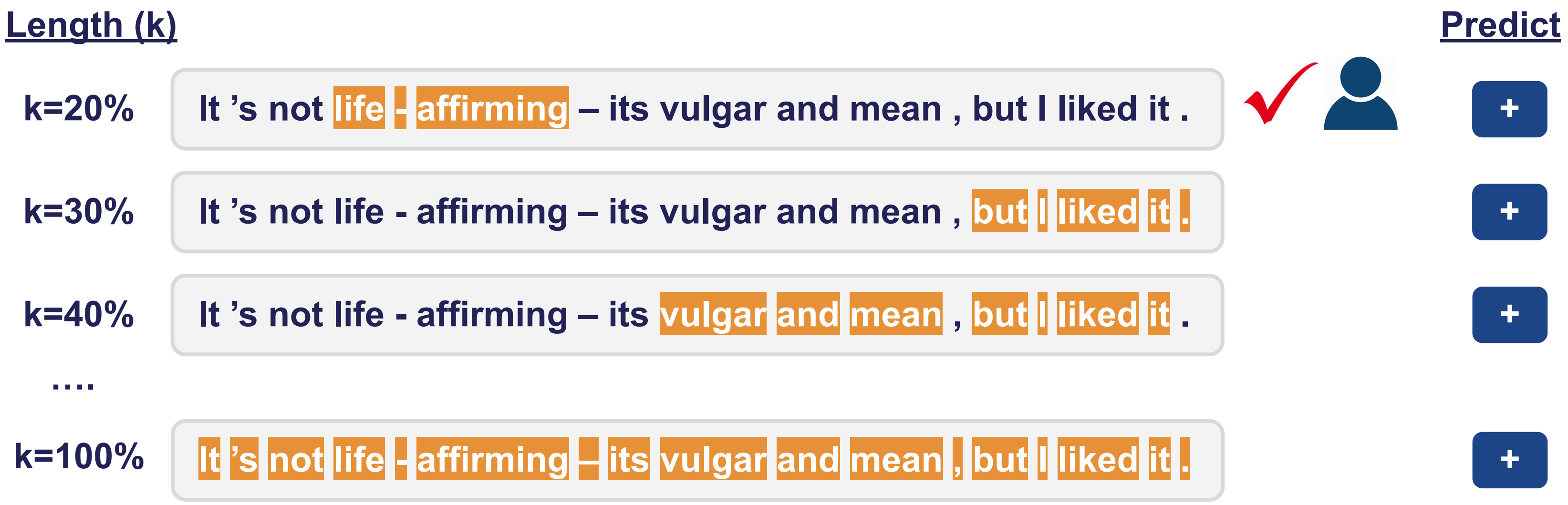


Length (k)			Predict
k=20%	It ’s not life - affirming – its vulgar and mean , but I liked it .	✓	+
k=30%	It ’s not life - affirming – its vulgar and mean , but I liked it .		+
k=40%	It ’s not life - affirming – its vulgar and mean , but I liked it .		+
….			
k=100%	It ’s not life - affirming – its vulgar and mean , but I liked it .		+

Conciseness: To be validated…

Conciseness

Yet to be validated by human studies !



Length (k)			Predict
k=20%	It ’s not life - affirming – its vulgar and mean , but I liked it .	✓	+
k=30%	It ’s not life - affirming – its vulgar and mean , but I liked it .		+
k=40%	It ’s not life - affirming – its vulgar and mean , but I liked it .		+
….			
k=100%	It ’s not life - affirming – its vulgar and mean , but I liked it .		+

Implicit Assumption: “shorter rationales are more intuitive to humans” ?

“Are Shortest Rationales the Best Explanations for Human Understanding?”

How does **rationale length affect human understanding**?

Core Statement

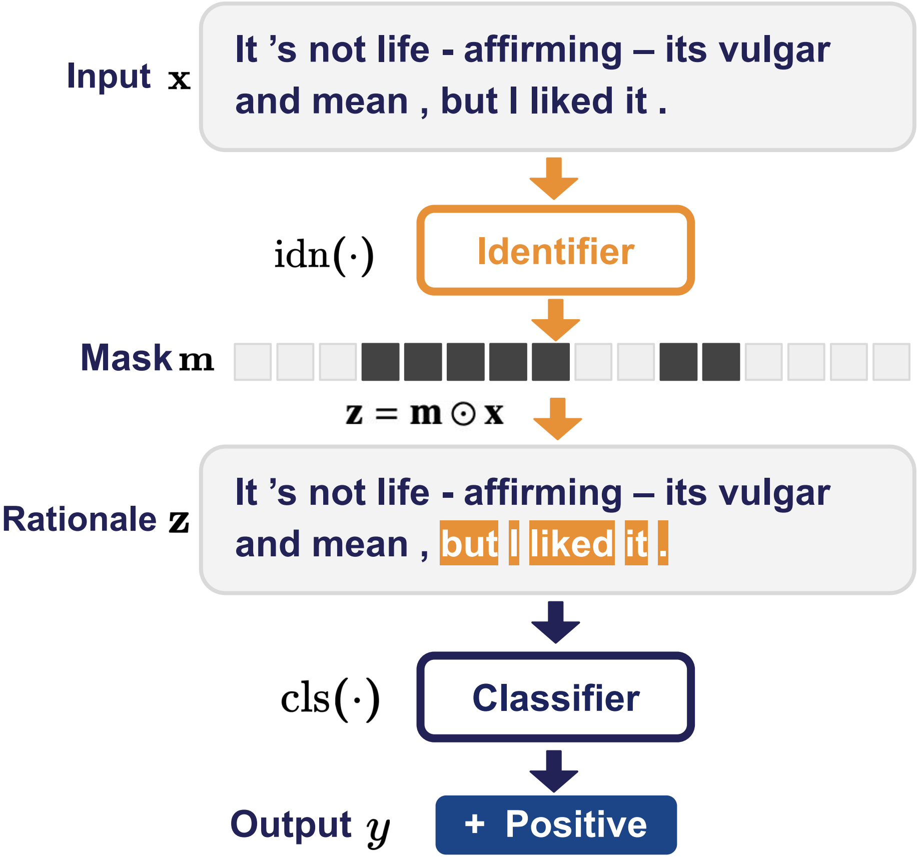
We find that ***shortest rationales*** *are* ***largely NOT the best for human understanding !***

Method

- We design a self-explaining model, **LimitedInk**, which allows users to **extract rationales at any target length**.

- We conduct **user studies** to investigate the **effect of rationale length** on **human understanding** using LimitedInk.

LimitedInk: A self-explaining model with rationale length control



Input $\mathbf{x}$

It ’s not life - affirming – its vulgar and mean , but I liked it .

$\text{idn}(\cdot)$ Identifier

Mask $\mathbf{m}$

$\mathbf{z} = \mathbf{m} \odot \mathbf{x}$

Rationale $\mathbf{z}$

It ’s not life - affirming – its vulgar and mean , but I liked it .

$\text{cls}(\cdot)$ Classifier

Output y

+ Positive

Optimization Objective

$$\min_{\theta_{\text{idn}},\theta_{\text{cls}}} \underbrace{\mathbb{E}_{\mathbf{z}\sim\mathbf{idn}(\mathbf{x})}\mathcal{L}(\mathbf{cls}(\mathbf{z}),y)}_{\text{sufficient prediction}} + \underbrace{\lambda\Omega(\mathbf{m})}_{\text{regularization}}$$

Sufficient rationale for correct predictions

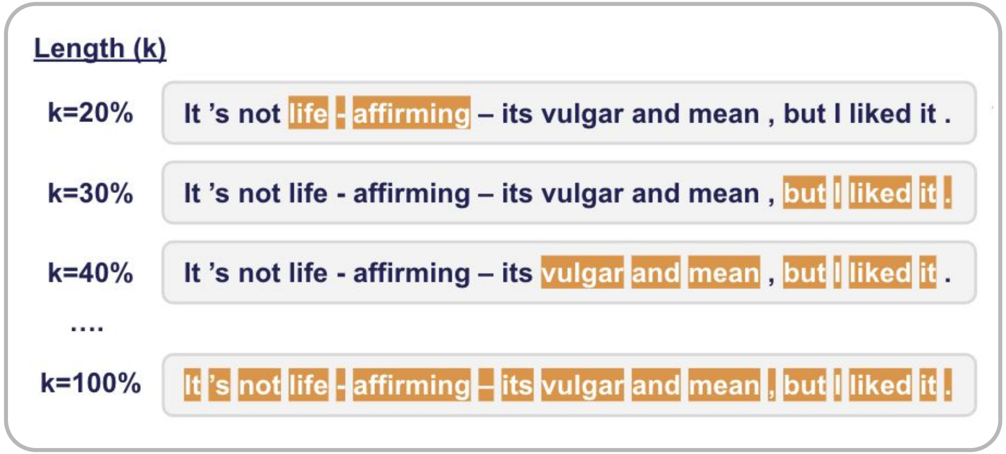
Regularization on masks properties

How to control rationale length in LimitedInk

Control Rationale Length

Gumbel-Softmax Sampling

Vector and Sort Regularization



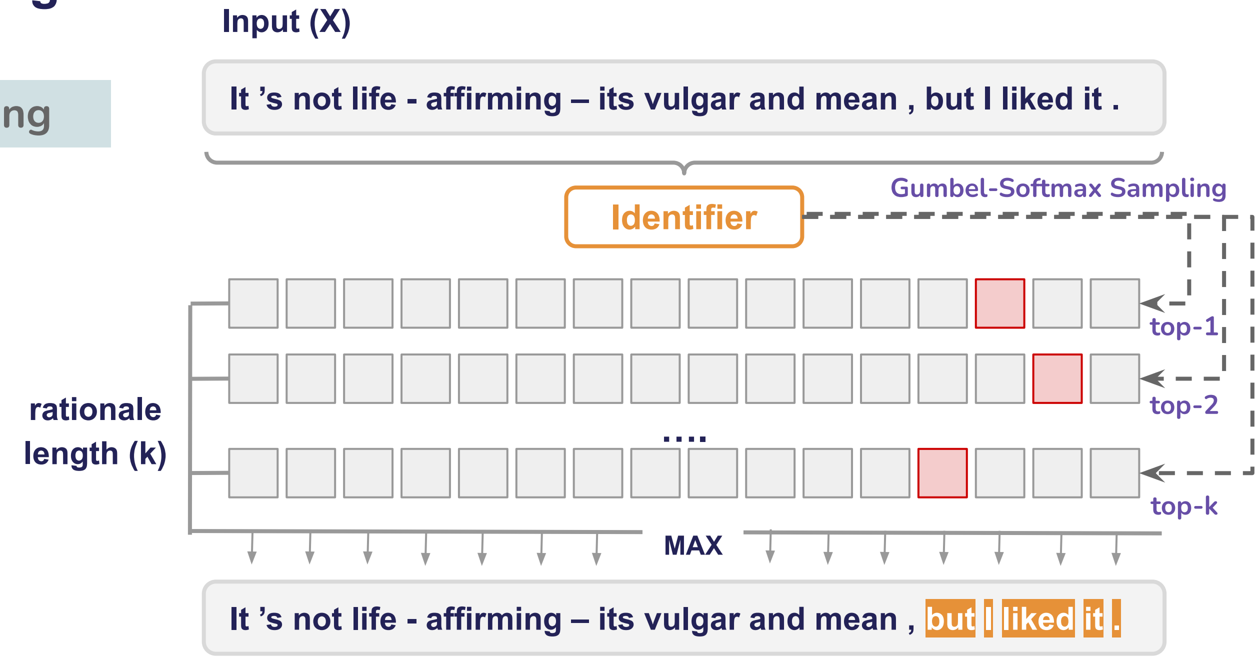
Length (k)

k=20%	It 's not life - affirming – its vulgar and mean , but I liked it .
k=30%	It 's not life - affirming – its vulgar and mean , but I liked it .
k=40%	It 's not life - affirming – its vulgar and mean , but I liked it .
....	
k=100%	It 's not life - affirming – its vulgar and mean , but I liked it .

How to control rationale length in LimitedInk

Control Rationale Length

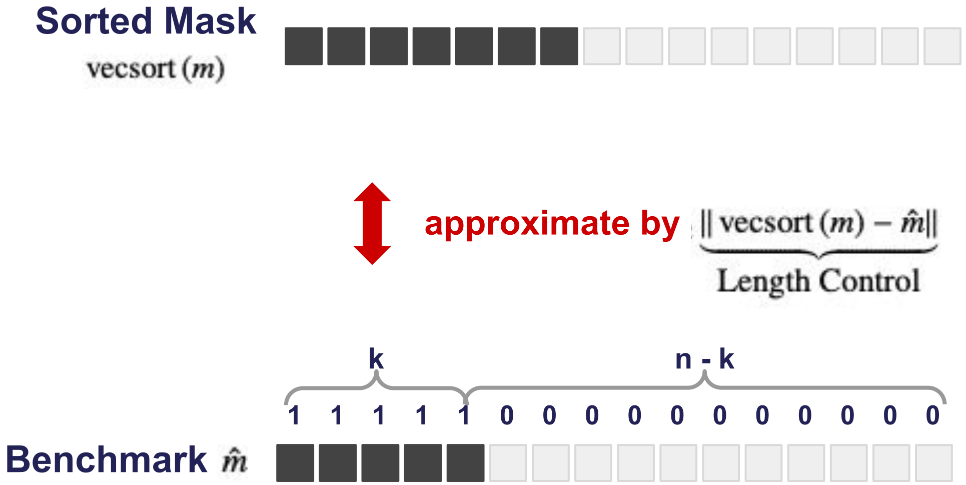
Gumbel-Softmax Sampling


Input (X)
It ’s not life - affirming – its vulgar and mean , but I liked it .
Identifier
Gumbel-Softmax Sampling
top-1
top-2
....
top-k
rationale length (k)
MAX
It ’s not life - affirming – its vulgar and mean , but I liked it .

How to control rationale length in LimitedInk

Control Rationale Length

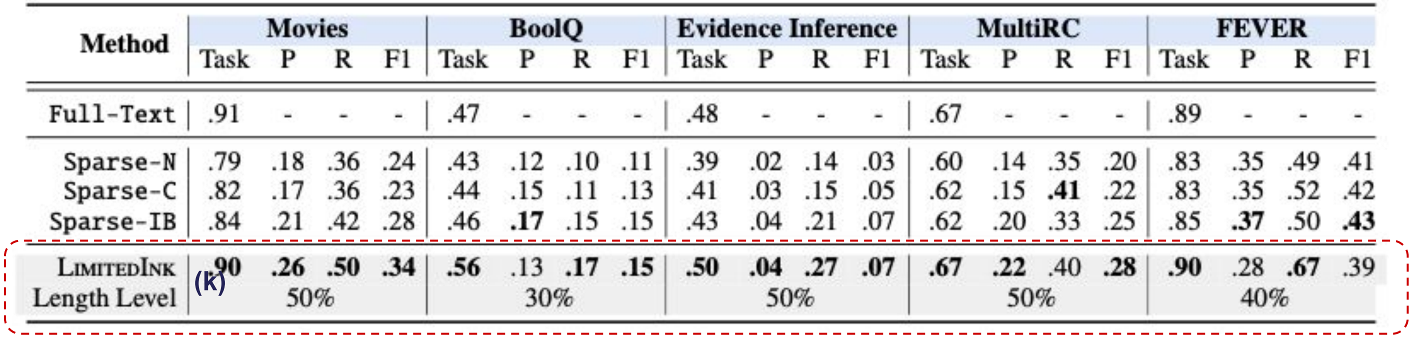
Vector and Sort Regularization



LimitedInk Performance

Evaluation Metrics

- **End-task classification: Task**, weighted average F1
- **Human-annotated rationale agreement: P**recision, **R**ecall, Token-level **F1**

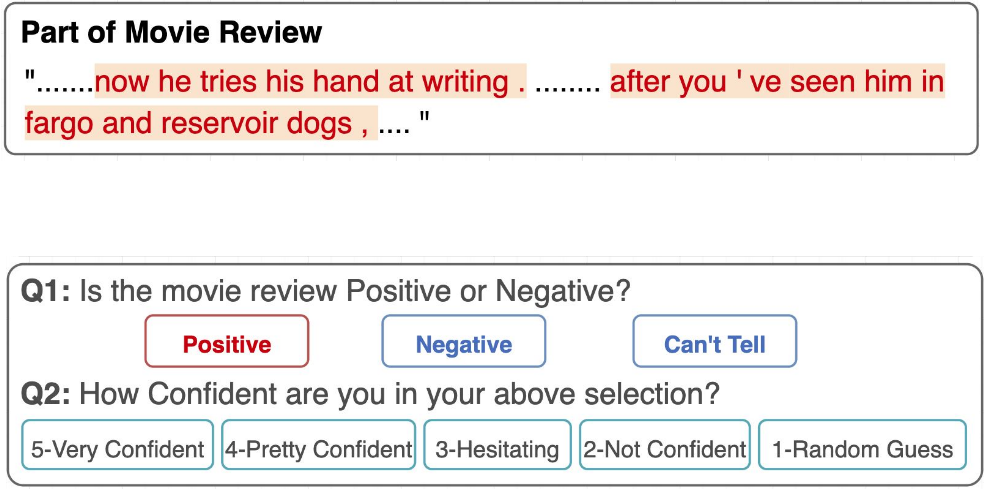


Method	Movies				BoolQ				Evidence Inference				MultiRC				FEVER			
	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1	Task	P	R	F1
Full-Text	.91	-	-	-	.47	-	-	-	.48	-	-	-	.67	-	-	-	.89	-	-	-
Sparse-N	.79	.18	.36	.24	.43	.12	.10	.11	.39	.02	.14	.03	.60	.14	.35	.20	.83	.35	.49	.41
Sparse-C	.82	.17	.36	.23	.44	.15	.11	.13	.41	.03	.15	.05	.62	.15	**.41**	.22	.83	.35	.52	.42
Sparse-IB	.84	.21	.42	.28	.46	**.17**	.15	.15	.43	.04	.21	.07	.62	.20	.33	.25	.85	**.37**	.50	**.43**
LimitedInk	**.90**	**.26**	**.50**	**.34**	**.56**	.13	**.17**	**.15**	**.50**	**.04**	**.27**	**.07**	**.67**	**.22**	.40	**.28**	**.90**	.28	**.67**	.39
Length Level (k)	50%				30%				50%				50%				40%			

Results

LimitedInk **performs compatible with three SOTA baselines** on the two common rationale metrics in five ERASER text classification benchmark datasets.

Human Study: description of dataset and human task



Part of Movie Review

"…….now he tries his hand at writing . …….. after you ' ve seen him in fargo and reservoir dogs , …. "

Q1: Is the movie review Positive or Negative?

Q2: How Confident are you in your above selection?

Key Components of the User Interface

Sentiment Analysis: we randomly sampled **100** reviews (correct prediction) from the **Movie review** test set

prediction

confidence

MTurk Workers

Human Study: preparing rationales

LimitedInk Rationales

k=10%

... not ....

k=20%

…….. life - affirming …..

k=30%

…….. but I liked it ….

k=40%

..vulgar and mean ….. but I liked it ……

k=50%

………… its vulgar and mean , but I liked it ....

≈

Random Baseline

k=10%

... affirming ....

k=20%

...... affirming – its ...

k=30%

.... – its vulgar and mean ....

k=40%

...... not life - affirming ... I liked it ....

k=50%

... life - affirming – its vulgar and mean , but ......

Human Study: effect of rationale length on human understanding

Review1 [k=10%]

........liked this movie......

Review2 [k=20%]

..obvious he a loser ...loser and not..

Review3 [k=30%]

.....the story starts to take off when his uncle dies.. not having a job......

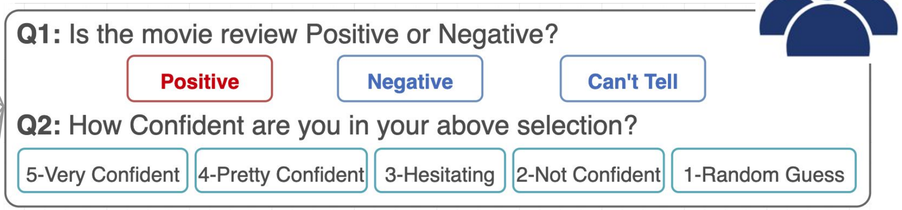
Review4 [k=40%]

little bit of casting was not done...extraordinary acting skills.....look drop dead gorgeous in any situation .....

Review5 [k=50%]

.... = = = = = ... this of course does n ' t mean its bad . ... arye cross is the stereotypica single male who falls in love . ...

Rationales in one Webpage



Q1: Is the movie review Positive or Negative?

Positive | Negative | Can't Tell

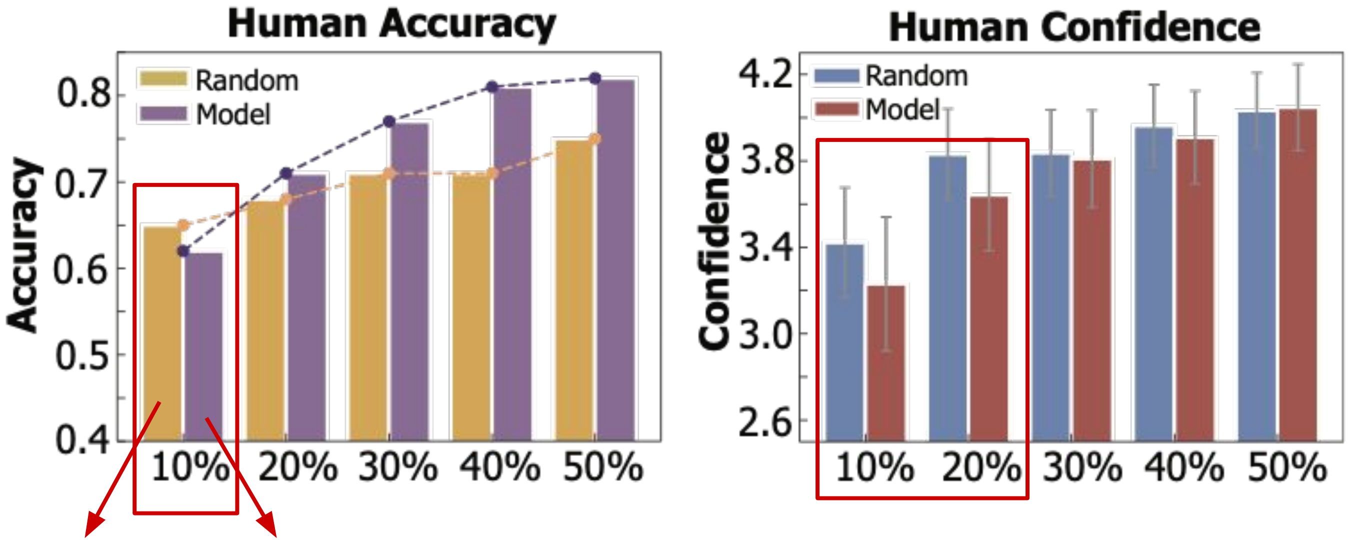
Q2: How Confident are you in your above selection?

5-Very Confident | 4-Pretty Confident | 3-Hesitating | 2-Not Confident | 1-Random Guess

We **strictly control** the workers' **participation**.

Therefore, participants **cannot see the same review repeatedly** to gradually see all the words.

Key Findings



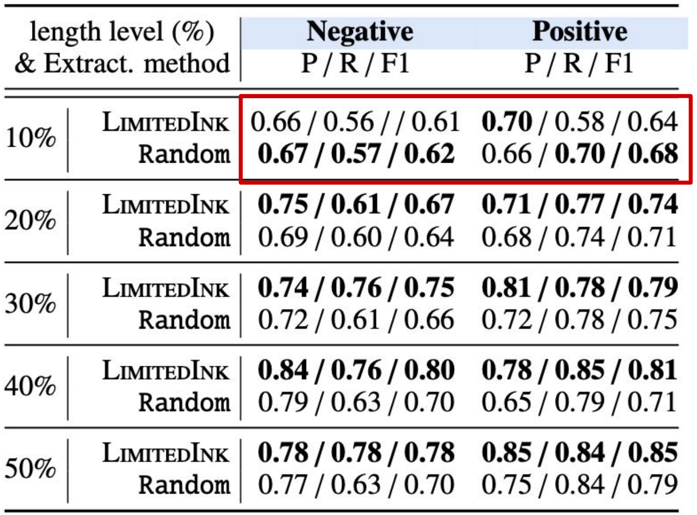
0.63

0.61

Human accuracy and confidence, at the shortest level (i.e., 10% length), are **lower than** the random baseline

The **shortest rationales** are **NOT the best** for human understanding.

Key Findings



length level (%) & Extract. method		Negative	Positive
		P / R / F1	P / R / F1
10%	LimitedInk	0.66 / 0.56 / / 0.61	**0.70** / 0.58 / 0.64
	Random	**0.67 / 0.57 / 0.62**	0.66 / **0.70 / 0.68**
20%	LimitedInk	**0.75 / 0.61 / 0.67**	**0.71 / 0.77 / 0.74**
	Random	0.69 / 0.60 / 0.64	0.68 / 0.74 / 0.71
30%	LimitedInk	**0.74 / 0.76 / 0.75**	**0.81 / 0.78 / 0.79**
	Random	0.72 / 0.61 / 0.66	0.72 / 0.78 / 0.75
40%	LimitedInk	**0.84 / 0.76 / 0.80**	**0.78 / 0.85 / 0.81**
	Random	0.79 / 0.63 / 0.70	0.65 / 0.79 / 0.71
50%	LimitedInk	**0.78 / 0.78 / 0.78**	**0.85 / 0.84 / 0.85**
	Random	0.77 / 0.63 / 0.70	0.75 / 0.84 / 0.79

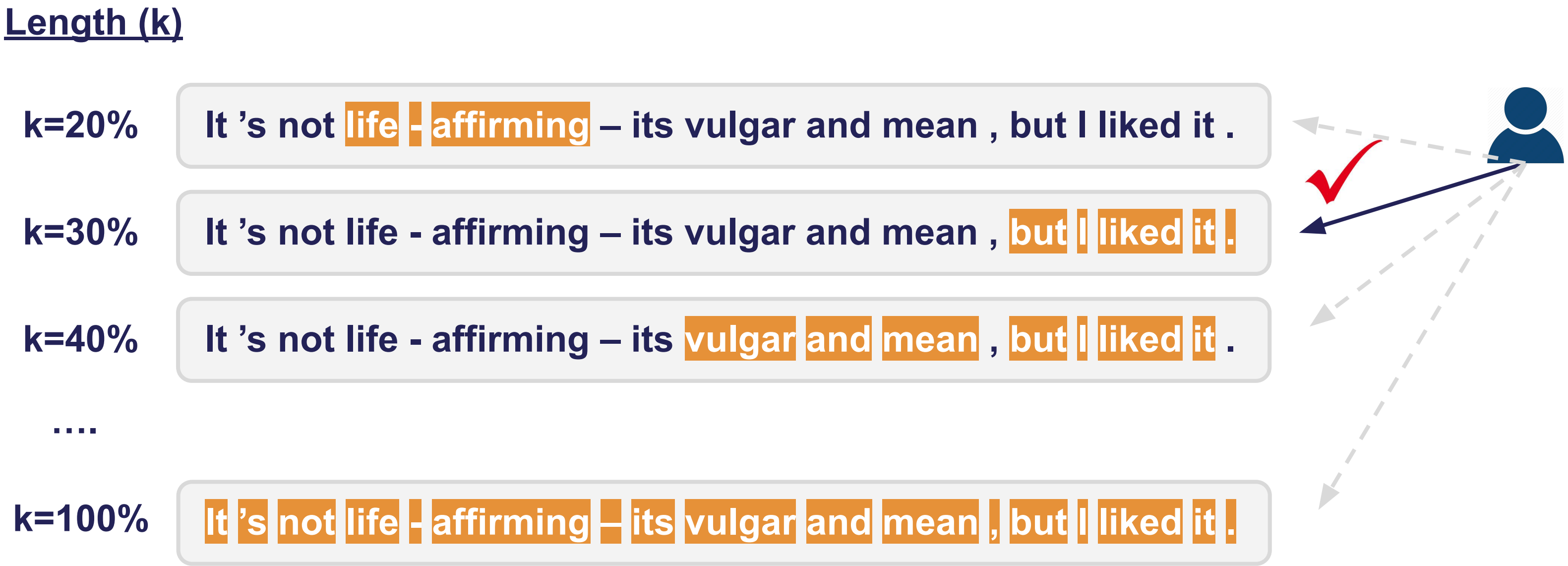
Human performance (i.e., Precision / Recall / F1 Score) on each category;

Again, the shortest rationales are NOT the most human-understandable.

Take-away Message

Shortest rationales are largely NOT the best for human understanding

Discussion: *Rethink how to define a good rationale?*



Length (k)

k=20% It ’s not life - affirming – its vulgar and mean , but I liked it .

k=30% It ’s not life - affirming – its vulgar and mean , but I liked it .

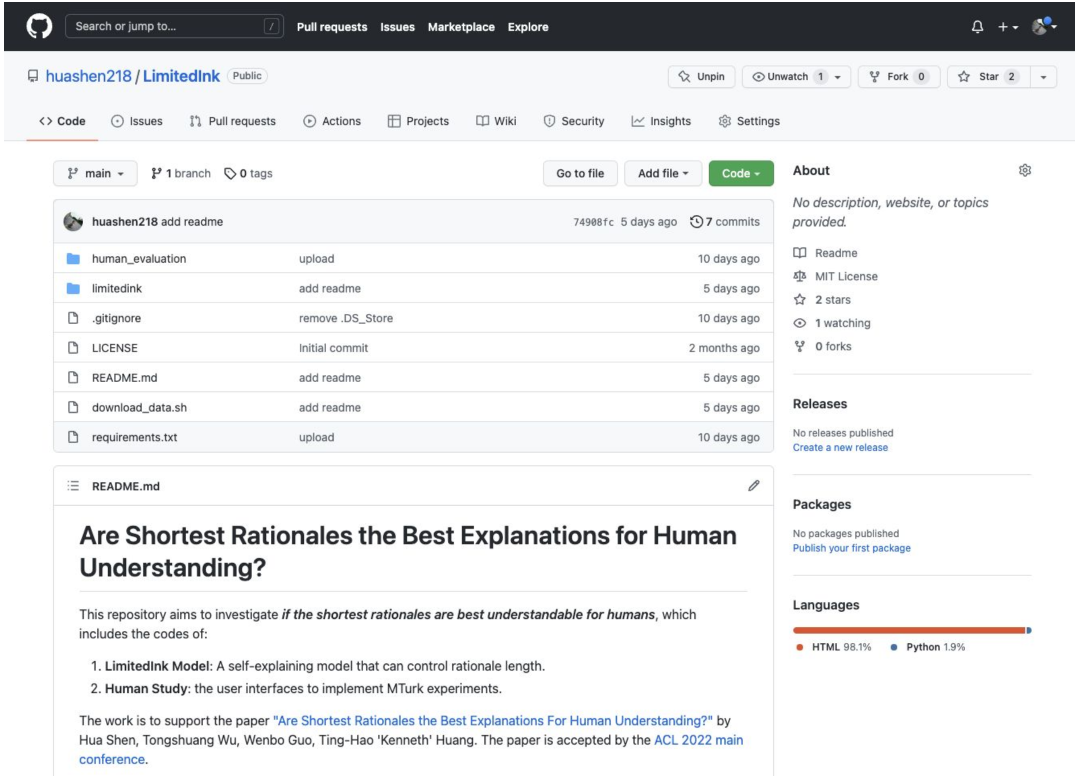
k=40% It ’s not life - affirming – its vulgar and mean , but I liked it .

….

k=100% It ’s not life - affirming – its vulgar and mean , but I liked it .

★ Learn to find the **right balance** between the **rationale length** and **model accuracy**.

Github: https://github.com/ huashen218/LimitedInk.git



huashen218 add readme 74908fc 5 days ago 7 commits

human_evaluation	upload	10 days ago
limitedink	add readme	5 days ago
.gitignore	remove .DS_Store	10 days ago
LICENSE	Initial commit	2 months ago
README.md	add readme	5 days ago
download_data.sh	add readme	5 days ago
requirements.txt	upload	10 days ago

README.md

Are Shortest Rationales the Best Explanations for Human Understanding?

This repository aims to investigate ***if the shortest rationales are best understandable for humans***, which includes the codes of:

1. **LimitedInk Model**: A self-explaining model that can control rationale length.
2. **Human Study**: the user interfaces to implement MTurk experiments.

The work is to support the paper "Are Shortest Rationales the Best Explanations For Human Understanding?" by Hua Shen, Tongshuang Wu, Wenbo Guo, Ting-Hao 'Kenneth' Huang. The paper is accepted by the ACL 2022 main conference.

About

No description, website, or topics provided.

Readme
MIT License
2 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● HTML 98.1% ● Python 1.9%

Check out our open-source **code** of **LimitedInk** and **Human Study** at Github!

Hua Shen

huashen218@psu.edu

@SarahHShen1

Tongshuang (Sherry) Wu

wtshuang@cs.washington.edu

@tongshuangwu

Wenbo Guo

wzg13@psu.edu

@WenboGuo4

Ting-Hao ‘Kenneth’ Huang

txh710@psu.edu

@windx0303

Thank you!