ODSC

Finding Rare Events in Text

Debanjana Banerjee
Senior Data Scientist, Walmart Global Tech

Sep 15th, 2021
Workshop,
Open Data Science Conference, India

Walmart Global Tech

Hello, fellow explorers!

Your Instructor for the day!!

Debanjana Banerjee

Senior Data Scientist,

Walmart Global Tech

My first workshop

- please be kind!

- Born and bred in the world of Statistics (I worship the old gods!)
- 4+ years in the industry
- In love with the diverse ML use cases retail has to offer

debanjana-banerjee

Our Agenda today

1. Introduction & Examples
2. Rare Events in Text: Defining Characteristics
3. Q&A
4. Text Mining
5. PU Learning
6. Q&A
7. Break
8. Semi Supervised Learners
9. iCASSTLe
10. Q&A

View slides for this workshop here:
github.com/debanjana-banerjee/Finding-Rare-Events-in-Text-ODSC-2021-
Available after the session!

debanjana-banerjee

Wands at the ready!

We will communicate on the Event X Ai Platform

Channel: **wed-debanjana-banerjee-finding-rare-events-in-text**

What do we need to do before we get started?

1. Create a folder named **FRET** on your Google Drive with sufficient space (+300MB)
2. Download **GloVe** file shared and load to folder **FRET** on your Google drive (this may take a while!)
3. Two data files will be shared shortly. We will load those onto **the same folder** on Google Drive

debanjana-banerjee

Finding Rare Events in Text – what is it all about?

Fraud Detection 01

Ethics Compliance 02

Consumer Safety 03

Social Media Compliance 04

Rare Events in Text

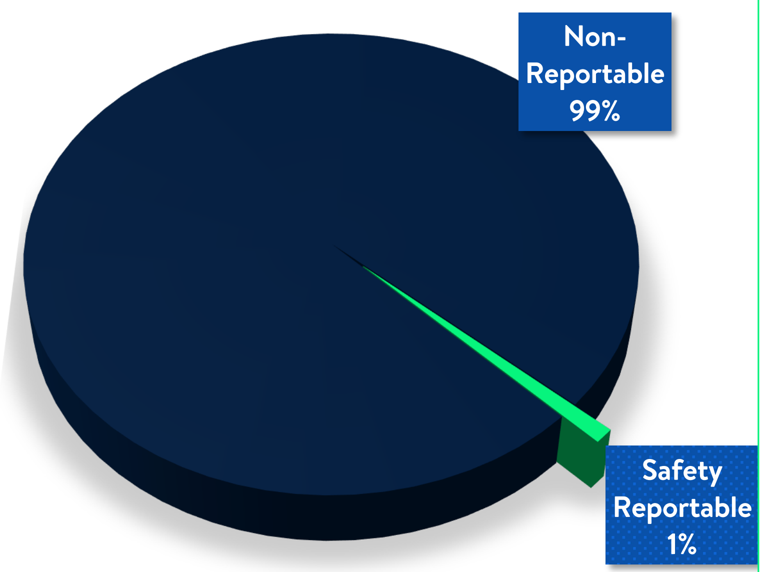
Rare Events!!

Global Tech

debanjana-banerjee

Rare Events in Consumer Safety

Customer Reviews



Non-Reportable Reviews

- Product Enquiry
- Compliments
- User guidance
- Discounts
- Sharing product info, etc.

Safety-Reportable Reviews

- Freshness of Produce
- Allergen Ingredients
- Safety guidelines on electronics
- Slippery tubs
- Brake safety on bikes, etc.

Let's look at a few examples!

Rare Events are not the same as Anomalies

Anomalies

- Generally, do not adhere to specification (i.e., cannot be described by a common theme)
- Variability in rate of imbalance is limited (rate of imbalance is always very high)
- Calls for unsupervised learning
- Sample dependent

Rare Events

- Adhere to certain specification (i.e., usually described by a common theme)
- Allows variability in rate of imbalance (depending on degree of rarity)
- May pertain to supervised or, unsupervised learning
- Sample independent

debanjana-banerjee

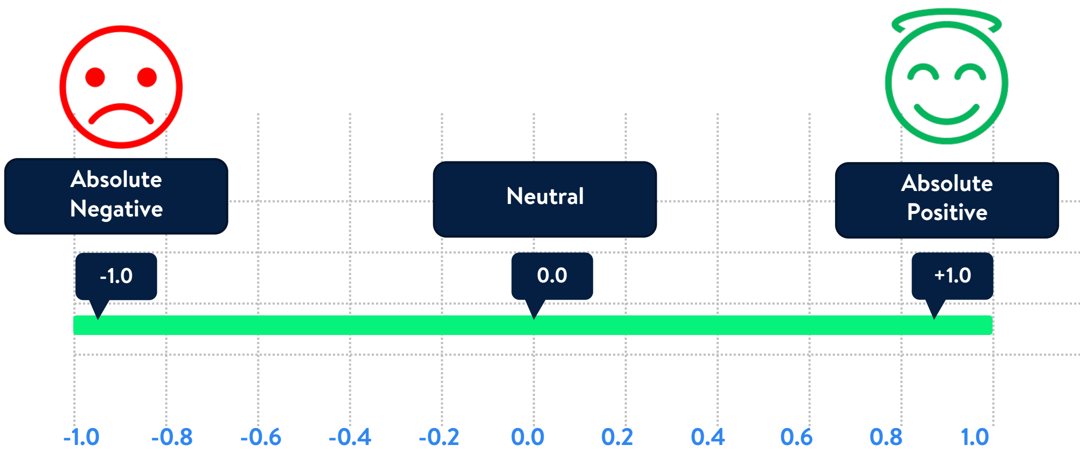
Rare Events in Text

Defining Characteristics

Global Tech

debanjana-banerjee

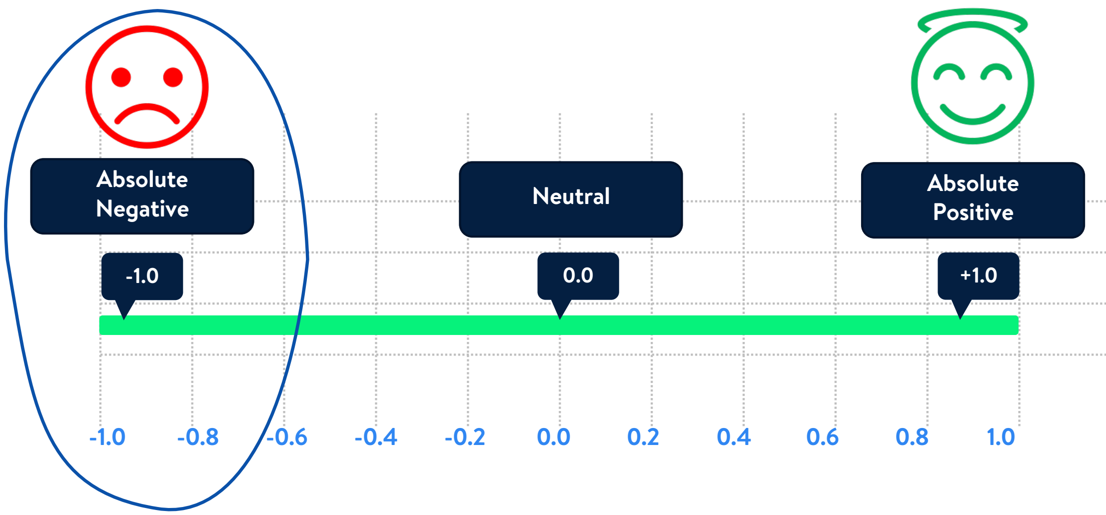
Sentiment Inclination



Depending on the specification, a rare event is likely to have inclination toward a polar sentiment

debanjana-banerjee

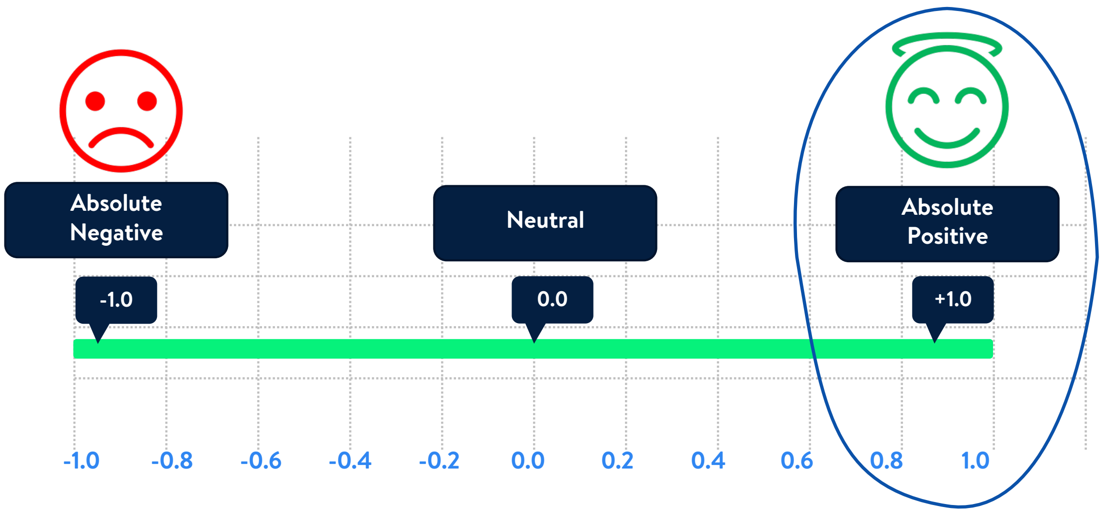
Sentiment Inclination



Depending on the specification, a rare event is likely to have inclination toward a polar sentiment

debanjana-banerjee

Sentiment Inclination



Depending on the specification, a rare event is likely to have inclination toward a polar sentiment

Sentiment Inclination

Depending on the specification, a rare event is likely to have inclination toward a polar sentiment

Safety-Reportable Reviews

"My salad was stale. The lettuce had blackened"

Sentiment Polarity: **-0.12**

Global Tech

debanjana-banerjee

Token Sensitivity

Review A

"Customer stated ABC microwave started smoking when she used it"

Review B

"Customer stated ABC microwave worked smoothly when she used it"

Token Sensitivity

Review A

"Customer stated ABC microwave started smoking when she used it"

Review B

"Customer stated ABC microwave worked smoothly when she used it"

Review C

"Customer stated he purchased the bike a year ago and the chains came off"

Review D

"Customer stated he purchased the bike a week ago and the chains came off"

Token Sensitivity

Review A

“Customer stated ABC microwave **started smoking** when she used it”

Review B

“Customer stated ABC microwave **worked smoothly** when she used it”

Review C

“Customer stated he purchased the bike **a year ago** and the chains came off”

Review D

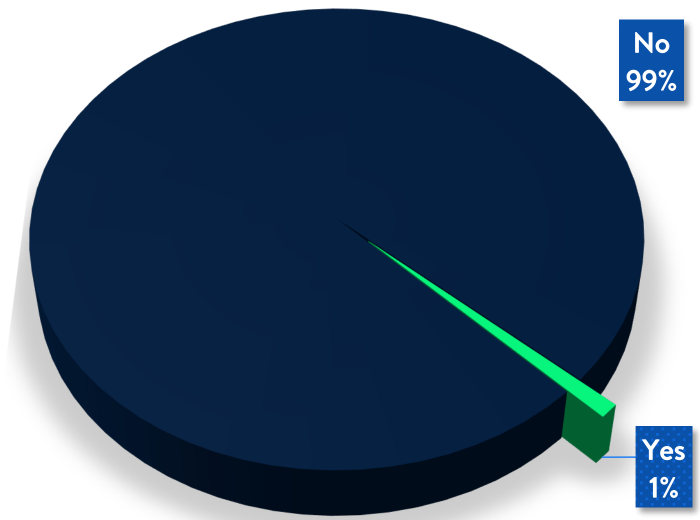
“Customer stated he purchased the **bike a week ago** and the chains came off”

For rare events in text, only certain tokens in the text dictate reportability or, non-reportability of the case

debanjana-banerjee

Data Availability

Is it a rare event?



- Rare Events are scarcely observed
  - Poor data availability

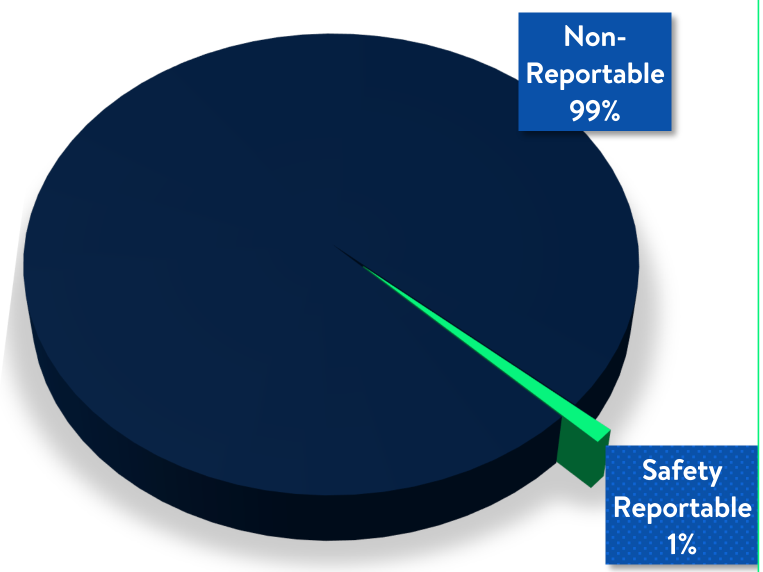
Quantity:
Number of available instances is low

Quality:
Available instances are not exhaustive for event specification

- Similar instances are repeated
- Not all sub-classes of the rare events are observed in the data

Data Availability

Consumer Safety



Non-Reportable Reviews

- Product Enquiry
- Compliments
- User guidance
- Discounts
- Sharing product info, etc.

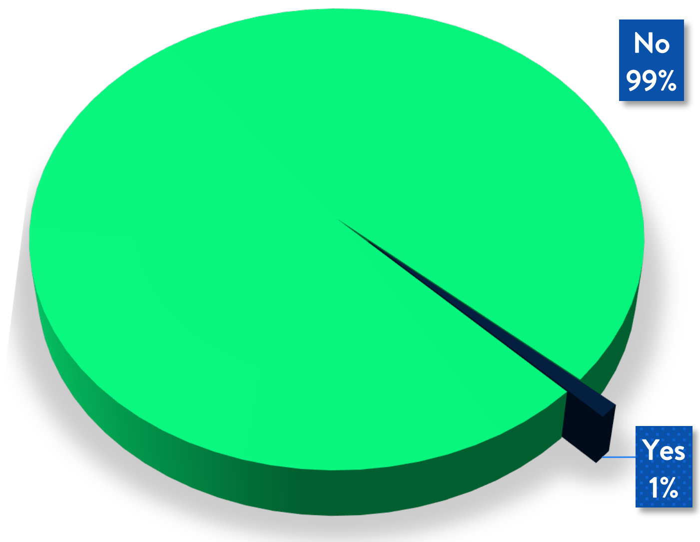
Safety-Reportable Reviews

- Freshness of Produce
- Allergen Ingredients
- Safety guidelines on electronics
- Slippery tubs
- Brake safety on bikes, etc

Of these, your data may show only a few kind!

No or, poor record of quality non-reportables

Is it a rare event?



- **Quality of Non-Reportables is important**

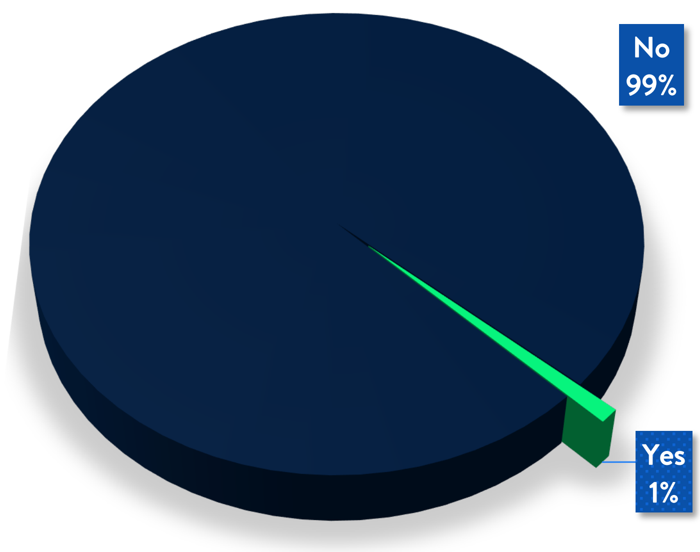
Non-Reportable Case A
Customer stated ABC microwave **worked smoothly** when she used it

Non-Reportable Case B
Customer stated he purchased the bike **a year ago** and the chains came off

debanjana-banerjee

Data Quality Issues in Rare Event Extraction

Is it a rare event?



Poor Quality and low quantity of Positives (Reportables)

Poor Quality or, no record of Negatives (Non-Reportables)

Positive Unlabeled Learning

debanjana-banerjee

Questions so far?

Please post on Event X Ai Platform

Channel: **wed-debanjana-banerjee-finding-rare-events-in-text**

Global Tech

debanjana-banerjee

Text Mining

Global Tech

debanjana-banerjee

Text Pre-Processing

- **Text Cleaning**
- **Lemmatization**
- **Numeric Representation of Text**

Please upload the following csv's to FRET folder in google drive

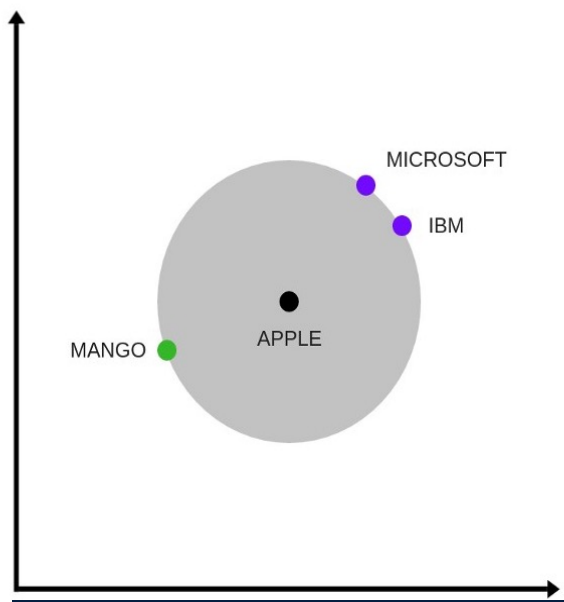
- **FRET_Test.csv**
- **FRET_Positive.csv**

Let's try this hands-on!

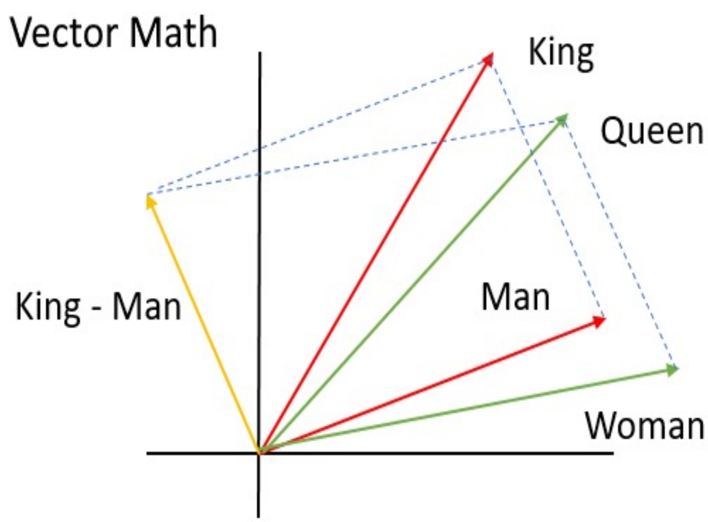
Notebook link is shared on **wed-debanjana-banerjee-finding-rare-events-in-text**

What are Text Embeddings?

Embeddings are numeric representation of meaningful words (or, phrases) such that their inter-relationships are preserved in the vector forms.



In the vector space above, the word 'Apple' is equidistant from the words 'Microsoft', 'IBM' and 'Mango', justifying two meanings of the word 'Apple'



In the vector space above, the distance between 'King' and 'Queen' is the same as that of 'Man' and 'Woman' capturing pairwise symmetry of words

GloVe

Global Tech

debanjana-banerjee

Training your own text embeddings using a corpus

- Context Window
- Term Co-occurrence Matrix

TCM

Vocab: child, got, rash, diaper, not, fit

Context Window : ∞

Original Text 1

- The child got a rash from the diapers

Original Text 2

- The diapers did not fit my child

Cleaned Text 1

- child got rash diaper

Cleaned Text 2

- diaper not fit child

	child	got	rash	diaper	not	fit
child	0	1	1	2	1	1
got	1	0	1	1	0	0
rash	1	1	0	1	0	0
diaper	2	1	1	0	1	1
not	1	0	0	1	0	1
fit	1	0	0	1	1	0

Matrix Factorization

Tokens

Tokens

$$\begin{array}{c|ccccc} & w_1 & w_2 & w_3 & w_4 & w_5 \\ \hline w_1 & y_{11} & y_{12} & y_{13} & y_{14} & y_{15} \\ w_2 & y_{21} & y_{22} & y_{23} & y_{24} & y_{25} \\ w_3 & y_{31} & y_{32} & y_{33} & y_{34} & y_{35} \\ w_4 & y_{41} & y_{42} & y_{43} & y_{44} & y_{45} \\ w_5 & y_{51} & y_{52} & y_{53} & y_{54} & y_{55} \end{array} \approx \underbrace{\begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \\ \beta_{41} & \beta_{42} \\ \beta_{51} & \beta_{52} \end{bmatrix}}_{\beta} \underbrace{\begin{bmatrix} x_{11} & x_{21} & x_{31} & x_{41} & x_{51} \\ x_{12} & x_{22} & x_{32} & x_{42} & x_{52} \end{bmatrix}}_{X^T}$$

Loadings

Factors

TCM

Matrix Factorization

Tokens

Tokens	w_1	w_2	w_3	w_4	w_5
w_1	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
w_2	y_{21}	y_{22}	y_{23}	y_{24}	y_{25}
w_3	y_{31}	y_{32}	y_{33}	y_{34}	y_{35}
w_4	y_{41}	y_{42}	y_{43}	y_{44}	y_{45}
w_5	y_{51}	y_{52}	y_{53}	y_{54}	y_{55}

TCM

$\approx$

Embedding for token w_1

Loadings

β_{11}	β_{12}
β_{21}	β_{22}
β_{31}	β_{32}
β_{41}	β_{42}
β_{51}	β_{52}

β

Factors

x_{11}	x_{21}	x_{31}	x_{41}	x_{51}
x_{12}	x_{22}	x_{32}	x_{42}	x_{52}

X^T

Matrix Factorization

Tokens

Tokens	w_1	w_2	w_3	w_4	w_5
w_1	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
w_2	y_{21}	y_{22}	y_{23}	y_{24}	y_{25}
w_3	y_{31}	y_{32}	y_{33}	y_{34}	y_{35}
w_4	y_{41}	y_{42}	y_{43}	y_{44}	y_{45}
w_5	y_{51}	y_{52}	y_{53}	y_{54}	y_{55}

TCM

$\approx$

Embedding for token w_1

Loadings

β_{11}	β_{12}
β_{21}	β_{22}
β_{31}	β_{32}
β_{41}	β_{42}
β_{51}	β_{52}

β

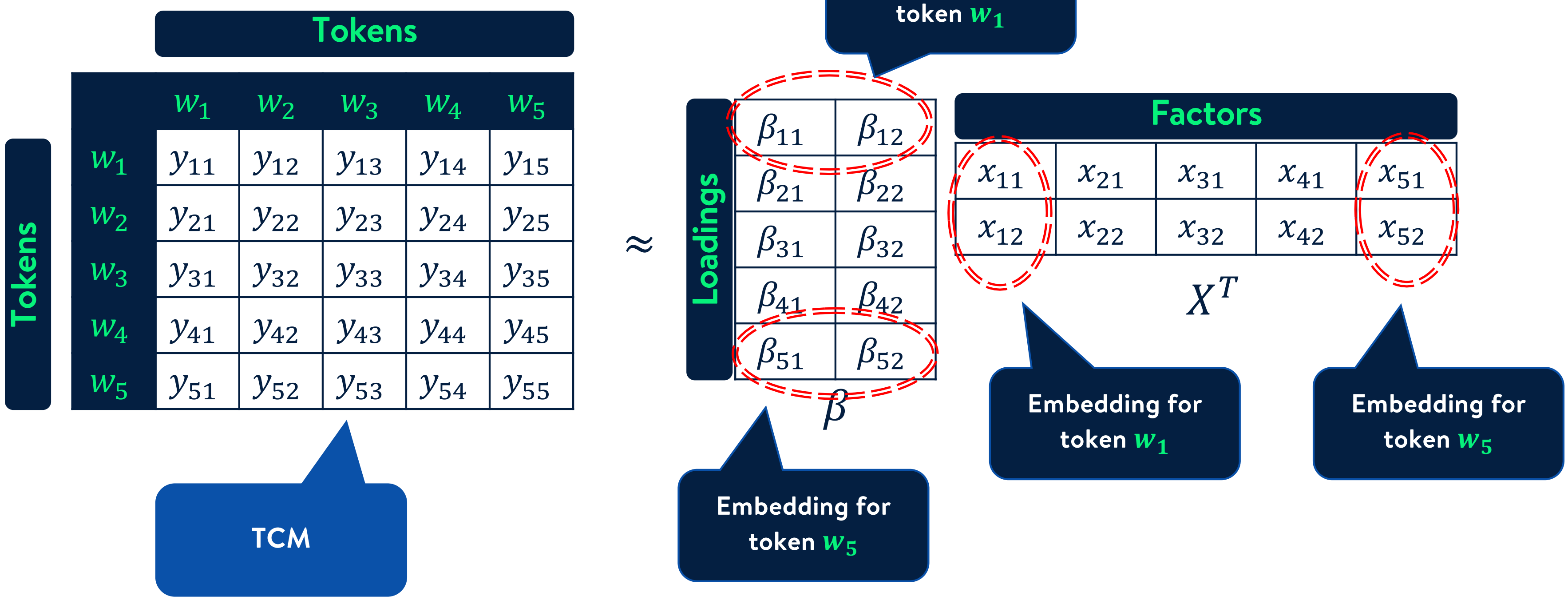
Embedding for token w_5

Factors

x_{11}	x_{21}	x_{31}	x_{41}	x_{51}
x_{12}	x_{22}	x_{32}	x_{42}	x_{52}

X^T

Matrix Factorization



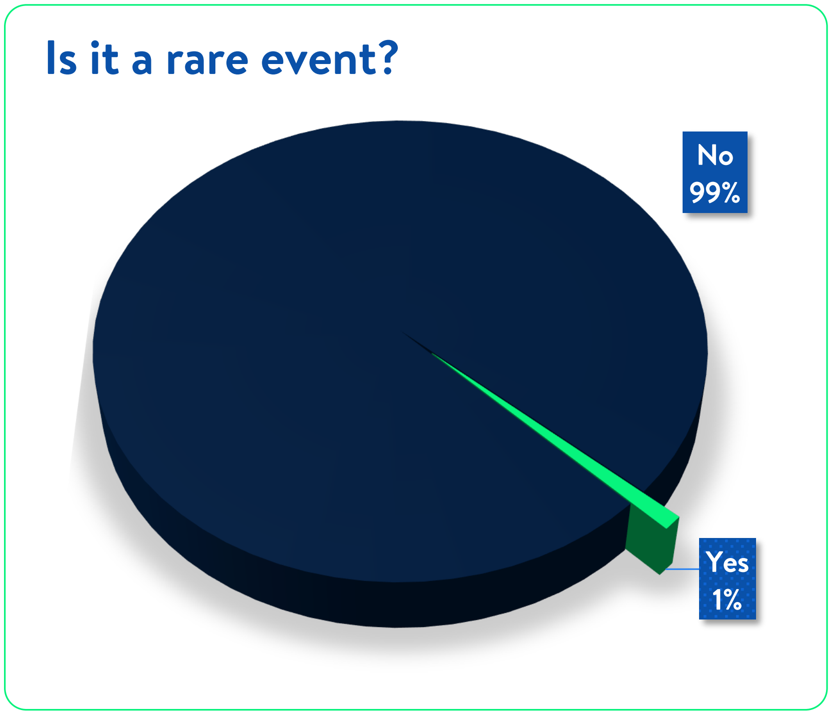
debanjana-banerjee

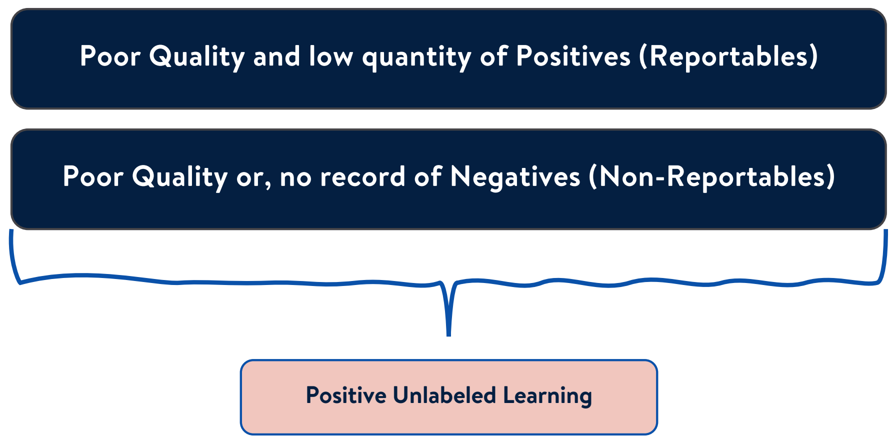
Positive Unlabeled (PU) Learning

Global Tech

debanjana-banerjee

Data Quality Issues in Rare Event Extraction

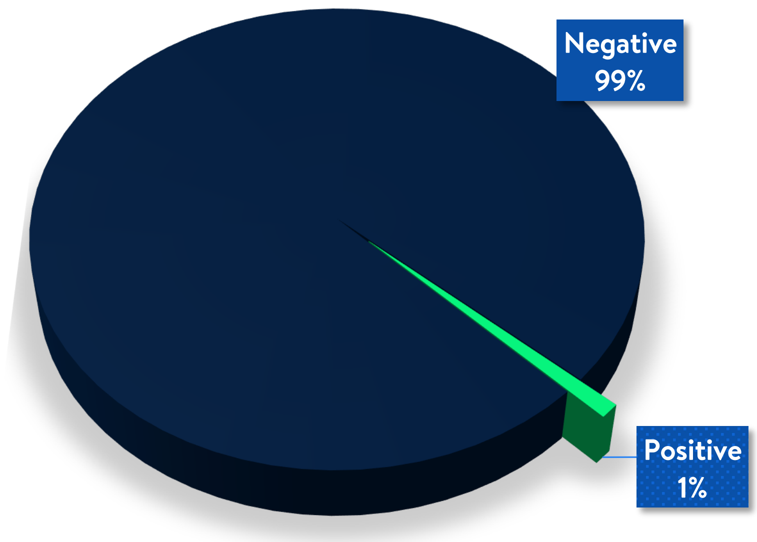




debanjana-banerjee

Rare Event Extraction as a Binary Classification technique

Is it a rare event of interest?



Rare Event of Interest: Consumer Safety

Negative

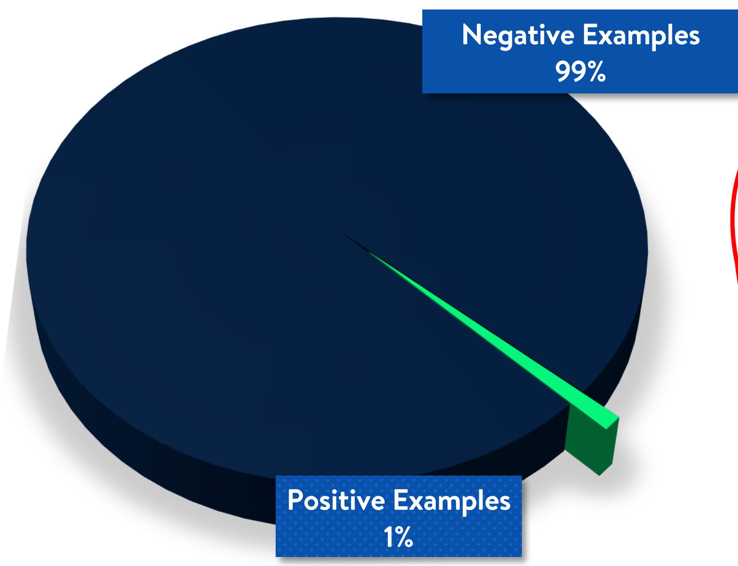
The text (review) is non-reportable
True > 99% of the times

Positive

The text (review) is safety-reportable
True < 1% of the times

Training Data in an ideal set-up

Ideal distribution of training data



Rare Event of Interest: Consumer Safety

Negative

The text (review) is non-reportable

True > 99% of the times

Positive

The text (review) is safety-reportable

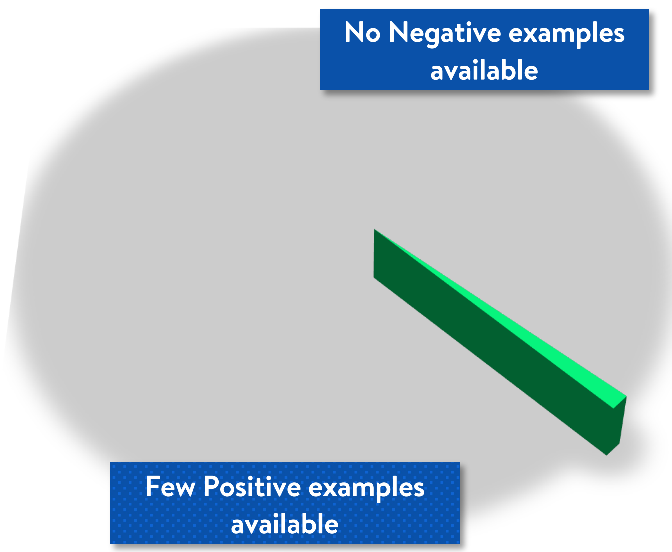
True < 1% of the times

True State

debanjana-banerjee

Training Data in a PU set-up (way more probable!)

Available distribution of training data



Rare Event of Interest: Consumer Safety

Negative

The text (review) is non-reportable

True > 99% of the times

Positive

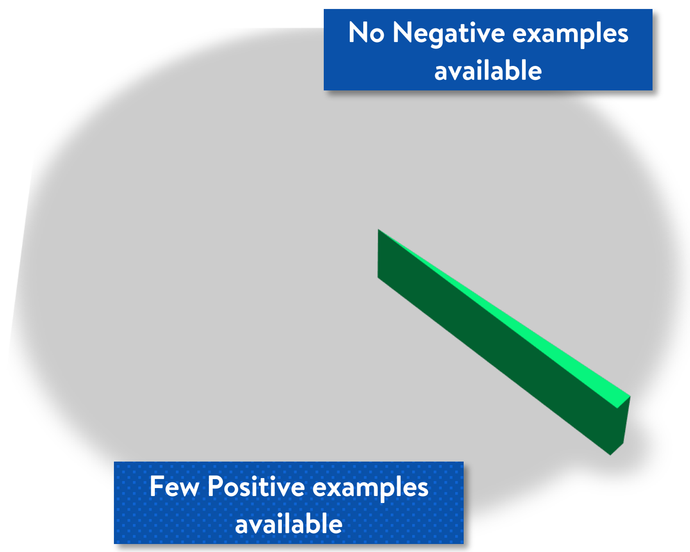
The text (review) is safety-reportable

True < 1% of the times

True State

Problem Formulation in an Imbalanced PU classifier

Available distribution of training data



- Binary Classification
- Two classes – Positive & Negative
- Class imbalance – Positive is a Rare Event
- Few examples available from Positive class
  - Data quality unknown
- No example available from Negative class
- Exact Rate of Imbalance unknown

Quick Break

Back in 10 mins

Global Tech

debanjana-banerjee

Semi Supervised Learners

When do we use semi-supervised learning?

- Limited Training Data
  - Usually labeled by experts
  - Hence, expensive
- Huge amount of unlabeled data
- Classes can be assumed to be distinctly separable
- Feature X is highly 'informative' about label Y

Is our use case eligible for SSL?

- **Limited Training Data**
  - Incomplete Training Data
  - Only positive examples available
- **Huge amt of unlabeled data**
  - Customer reviews database is huge

debanjana-banerjee

When do we use semi-supervised learning?

- Limited Training Data
  - Usually labeled by experts
  - Hence, expensive
- Huge amount of unlabeled data
- Classes can be assumed to be distinctly separable
- Feature X is highly ‘informative’ about label Y

Is our use case eligible for SSL?

- **Limited Training Data**
  - Incomplete Training Data
  - Only positive examples available
- **Huge amt of unlabeled data**
  - Customer reviews database is huge
- **We will assume**
  - **Classes are distinctly separable**
  - Text is enough to identify positive cases i.e., $\boldsymbol{X}$ **is informative about** $\boldsymbol{Y}$

Entropy Regularization in SSL

- Entropy is a measure of randomness (stability) in a random variable

debanjana-banerjee

Entropy Regularization in SSL

- Entropy is a measure of randomness (stability) in a random variable
- Entropy of our label Y $given$ X:

  $H(Y|X) = E(-\ln P(Y|X))$

debanjana-banerjee

Entropy Regularization in SSL

- Entropy is a measure of randomness (stability) in a random variable
- Entropy of our label Y *given* X:

  $$H(Y|X) = E(-\ln P(Y|X))$$

  Given the value of X, how stable is the value of Y

Entropy Regularization in SSL

- Entropy is a measure of randomness (stability) in a random variable
- Entropy of our label Y *given* X:

  $H(Y|X) = E(-\ln P(Y|X))$

  Given the value of X, how stable is the value of Y

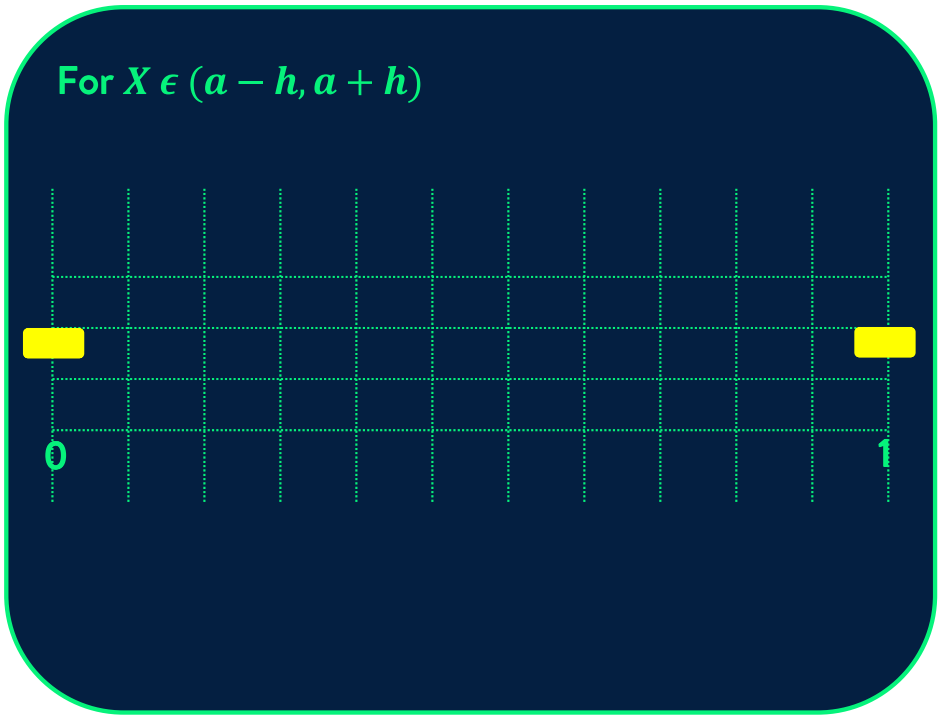
  Y can take only two values: **0 and 1**

Why regularize on Entropy?

- **SSL claims X is informative about Y**
  - i.e., X alone can act as a good predictor for Y
- **High $H(Y|X)$ indicates Y can have high variance for fixed values of X**
- **SSL uses unlabeled data to make predictions on Y**
  - This makes sense only if X actually has enough information about Y
  - i.e., **$H(Y|X)$ is low**

debanjana-banerjee

Entropy Regularization in SSL



Why regularize on Entropy?

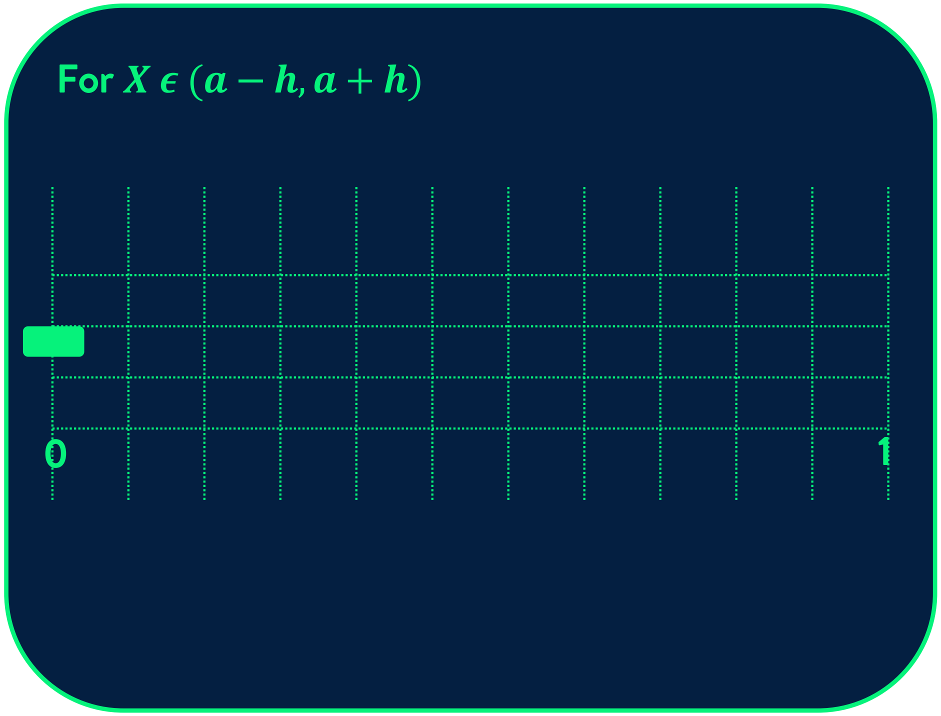
- **SSL claims X is informative about Y**
  - i.e., X alone can act as a good predictor for Y

- **High $H(Y|X)$ indicates Y can have high variance for fixed values of X**

- **SSL uses unlabeled data to make predictions on Y**
  - This makes sense only if X actually has enough information about Y
  - i.e., **$H(Y|X)$ is low**
  - **Y is stable for given ε-nbh of X**

debanjana-banerjee

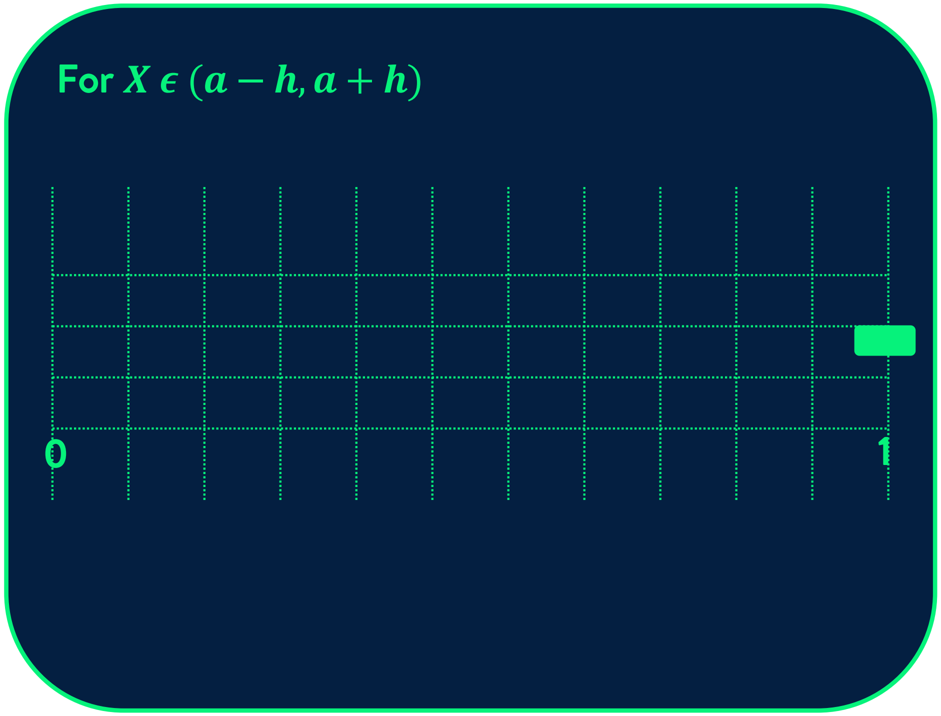
Entropy Regularization in SSL



Why regularize on Entropy?

- **SSL claims X is informative about Y**
  - i.e., X alone can act as a good predictor for Y
- **High $H(Y|X)$ indicates Y can have high variance for fixed values of X**
- **SSL uses unlabeled data to make predictions on Y**
  - This makes sense only if X actually has enough information about Y
  - i.e., **$H(Y|X)$ is low**
  - **Y is stable for given ε-nbh of X**

Entropy Regularization in SSL

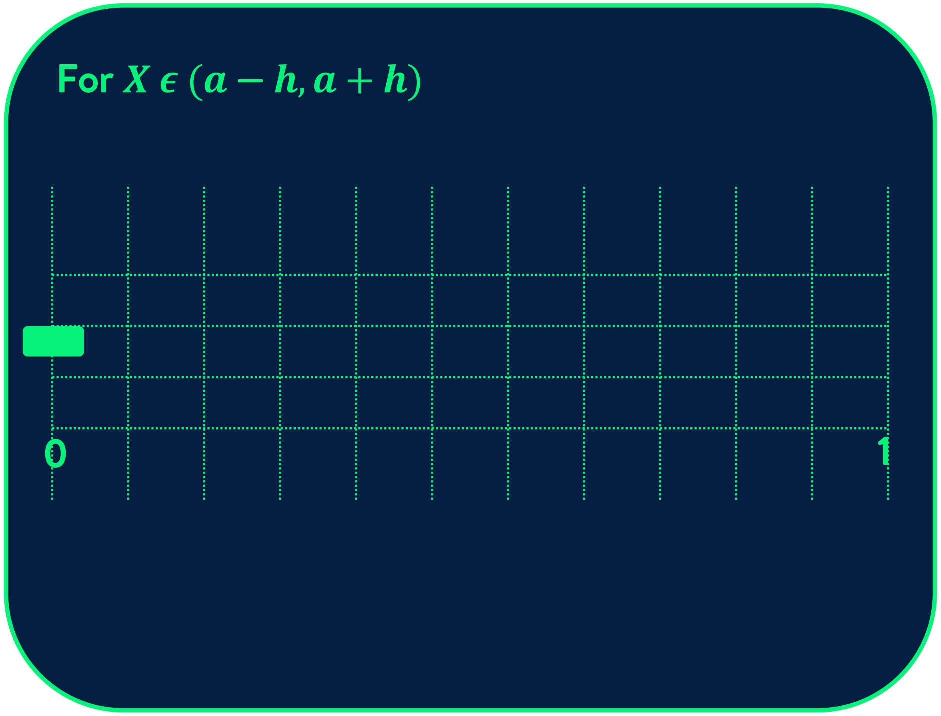


Why regularize on Entropy?

- **SSL claims X is informative about Y**
  - i.e., X alone can act as a good predictor for Y
- **High $H(Y|X)$ indicates Y can have high variance for fixed values of X**
- **SSL uses unlabeled data to make predictions on Y**
  - This makes sense only if X actually has enough information about Y
  - i.e., **$H(Y|X)$ is low**
  - **Y is stable for given ε-nbh of X**

debanjana-banerjee

Entropy Regularization in SSL



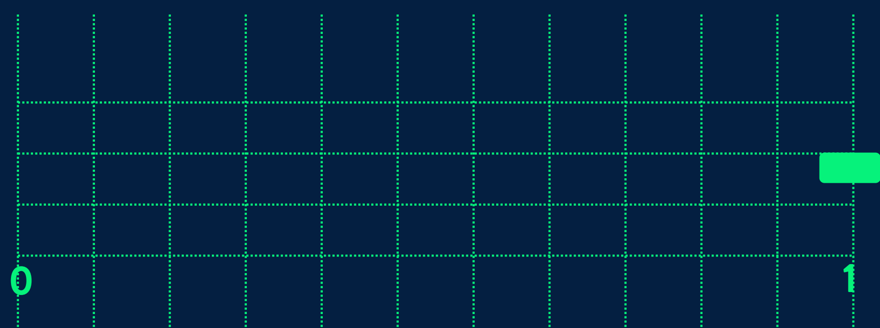
Why regularize on Entropy?

- **SSL claims X is informative about Y**
  - i.e., X alone can act as a good predictor for Y
- **High $H(Y|X)$ indicates Y can have high variance for fixed values of X**
- **SSL uses unlabeled data to make predictions on Y**
  - This makes sense only if X actually has enough information about Y
  - i.e., **$H(Y|X)$ is low**
  - **Y is stable for given ε-nbh of X**

debanjana-banerjee

Entropy Regularization in SSL

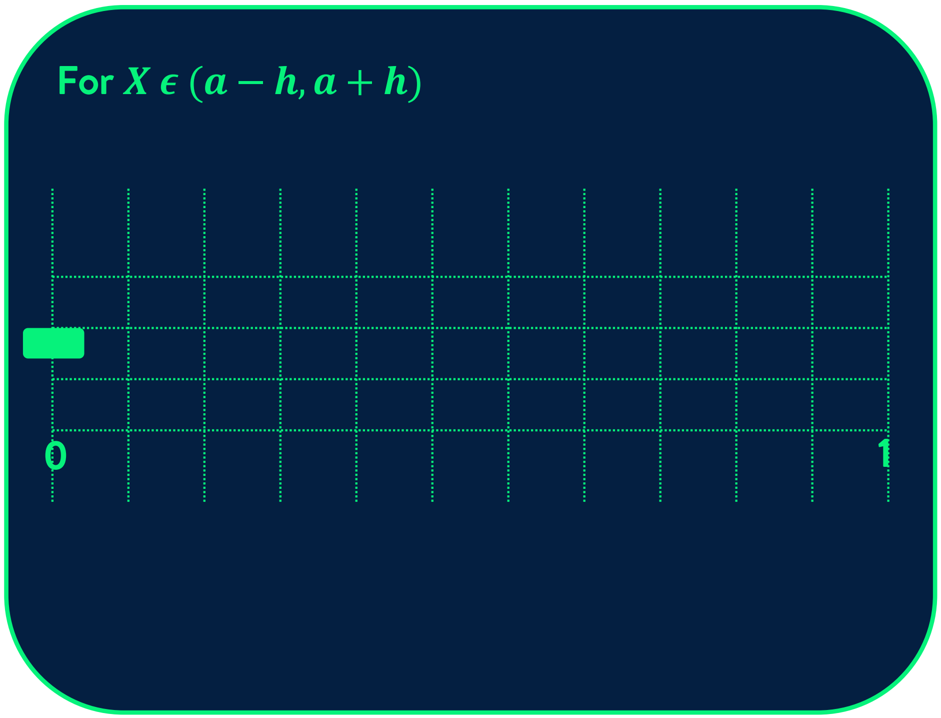
For $X \epsilon (a-h, a+h)$



Why regularize on Entropy?

- **SSL claims X is informative about Y**
  - i.e., X alone can act as a good predictor for Y
- **High $H(Y|X)$ indicates Y can have high variance for fixed values of X**
- **SSL uses unlabeled data to make predictions on Y**
  - This makes sense only if X actually has enough information about Y
  - i.e., **$H(Y|X)$ is low**
  - **Y is stable for given ε-nbh of X**

Entropy Regularization in SSL

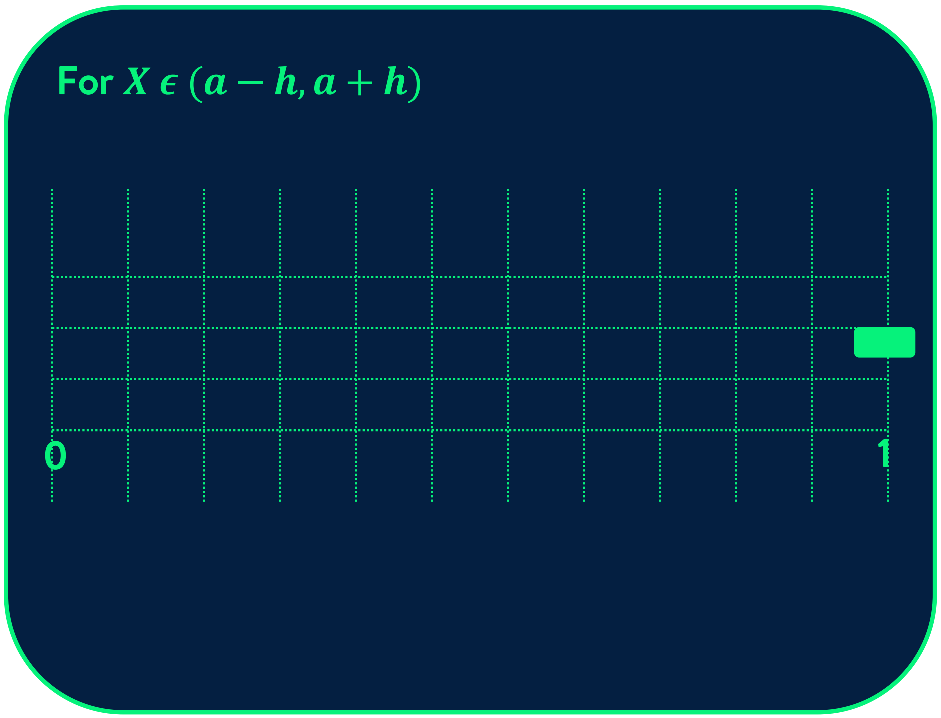


Why regularize on Entropy?

- **SSL claims X is informative about Y**
  - i.e., X alone can act as a good predictor for Y
- **High $H(Y|X)$ indicates Y can have high variance for fixed values of X**
- **SSL uses unlabeled data to make predictions on Y**
  - This makes sense only if X actually has enough information about Y
  - i.e., **$H(Y|X)$ is low**
  - **Y is stable for given ε-nbh of X**

debanjana-banerjee

Entropy Regularization in SSL

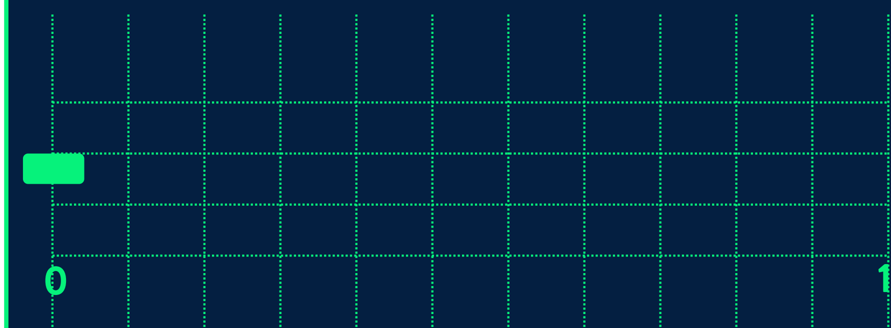


Why regularize on Entropy?

- **SSL claims X is informative about Y**
  - i.e., X alone can act as a good predictor for Y
- **High $H(Y|X)$ indicates Y can have high variance for fixed values of X**
- **SSL uses unlabeled data to make predictions on Y**
  - This makes sense only if X actually has enough information about Y
  - i.e., **$H(Y|X)$ is low**
  - **Y is stable for given ε-nbh of X**

Entropy Regularization in SSL

For $X \epsilon (a-h, a+h)$



Why regularize on Entropy?

- **SSL claims X is informative about Y**
    - i.e., X alone can act as a good predictor for Y
- **High $H(Y|X)$ indicates Y can have high variance for fixed values of X**
- **SSL uses unlabeled data to make predictions on Y**
    - This makes sense only if X actually has enough information about Y
    - i.e., **$H(Y|X)$ is low**
    - **Y is stable for given ε-nbh of X**

debanjana-banerjee

iCASSTLe

Global Tech

debanjana-banerjee

Algorithm Overview

Component Scores

Get component scores dictating degree of reportability
(positive association with the rare event of interest)

RScore Formulation

Combine component scores to form a single measure of reportability
(positive association with the rare event of interest)

Stage I Classification

Perform filtering based on derived thresholds on RScore to obtain representative training data

Stage II Classification

Perform Semi Supervised Classification using the enhanced training data to obtain final positive cases

debanjana-banerjee

Component Scores for Reportability & RScore Formulation

Sentiment Score

Keyword Score

Similarity Score

Similarity with labeled positives

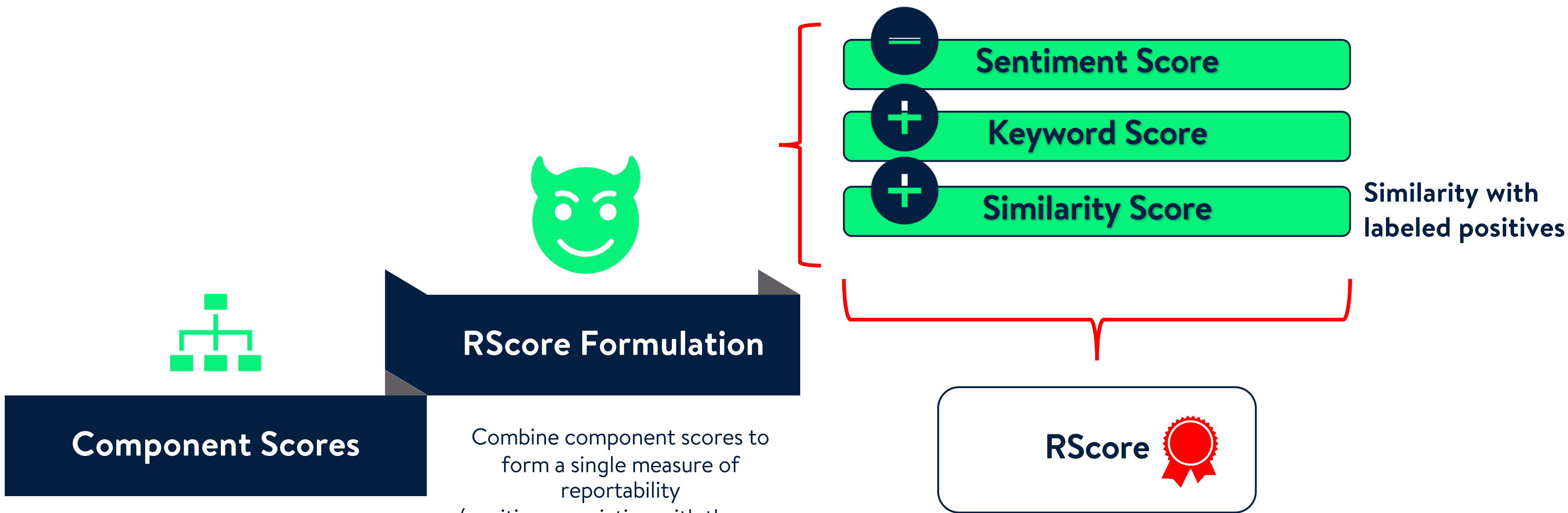
RScore Formulation

Combine component scores to form a single measure of reportability
(positive association with the rare event of interest)

Component Scores

Get component scores dictating degree of reportability
(positive association with the rare event of interest)

Component Scores for Reportability & RScore Formulation



Combine component scores to form a single measure of reportability (positive association with the rare event of interest)

Get component scores dictating degree of reportability (positive association with the rare event of interest)

debanjana-banerjee

Stage I Classification: Obtaining Training Negatives

Global Threshold for RScore

Q_{LR}^{K1} : $K_{1\ th}$ quantile of the Rscore values for labeled positive examples

Local Threshold for RScore

Q_{U}^{K1} : $K_{2\ th}$ quantile of the Rscore values for unlabeled cases

Anything not classified as Stage I Positive is labeled as Stage I Negative

Stage I Classification

Perform filtering based on derived thresholds on RScore to obtain representative training data

The j^{th} unlabeled case is classified as Stage I Positive iff

$$RScore_j > min(Q_{LR}^{K1}, Q_{U}^{K2})$$

debanjana-banerjee

Stage II Classification: SSL

Stage II Classification

Perform Semi Supervised Classification using the enhanced training data to obtain final positive cases

Labeled Data for Stage II (SSL)

- Original labeled Positives
- Top $K\%$ of positives + negatives obtained in **Stage I** (ranked by RScore)

Unlabeled Data for Stage II (SSL)

All original test cases except top $K\%$ of positives + negatives obtained in **Stage I** (ranked by RScore)

Quantity to Minimize: $L + \lambda H(Y|X)$

Entropy of Y given X

Loss Function

Regularization Constant

debanjana-banerjee

Q&A

Please post on Event X Ai Platform

Channel: **wed-debanjana-banerjee-finding-rare-events-in-text**

Global Tech

debanjana-banerjee

Thank You!

debanjanabanerjee1993@gmail.com

github.com/debanjana-banerjee

linkedin.com/in/debanjana-banerjee

Appendix

Global Tech

debanjana-banerjee