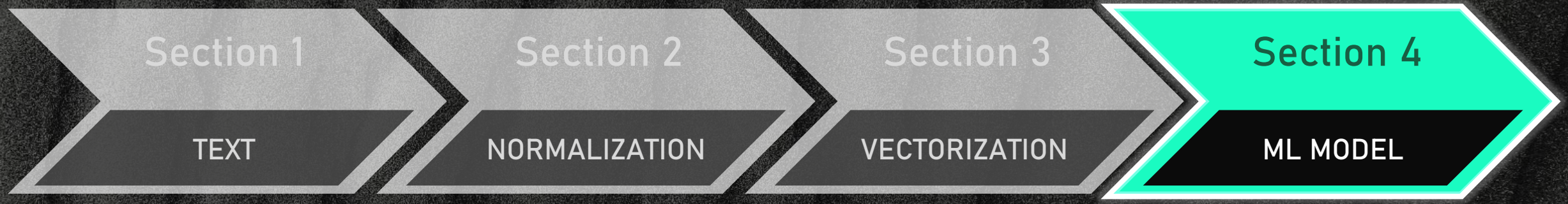
AIOutsider

POWERED BY

Section 4
Sentiment Analysis

Want more? Head over to www.AIOutsider.com

Section Overview



- Text Mining & NLP
- Sentiment Analysis
- Google Colab
- Dataset Overview
- Descriptive Statistics

- Text Normalization
- Features Cleaning
- Tokenization
- Stemming
- Lemmatization

- Text Representation
- Negative/Positive
- Bag-of-Words
- TF-IDF

- Logistic Regression
- Model Training
- Model Evaluation
- Model Prediction

AIOutsider.com

NEED A MODEL?

SENTIMENT STEP BY STEP

TWEET

TOKEN

VECTOR

SENTIMENT

« @AIOutsider Hellooooo, I'm happy!!!! »

[« @user », « hello », « I », « am », « happy »]

[.55 , 0 , .74 , 0 , 0 , .18 , .41 , 0 , 0 , .11]

The Sentiment is Positive!

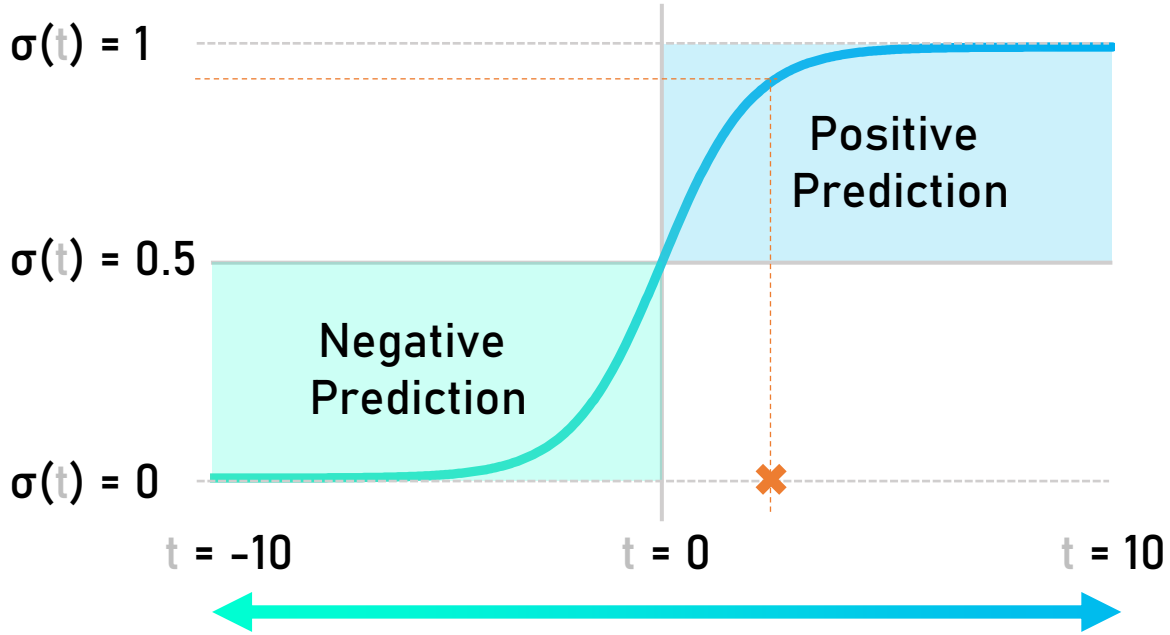
AIOutsider.com

LOGISTIC REGRESSION

WHAT IS LOGISTIC REGRESSION

Definition

Use of logistic function to model a binary dependant variable



Logistic Function

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

$$t = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

Linear combination

where β are learned parameters
and x are explanatory input variables

PREDICT WITH LOGISTIC REGRESSION

Tweet Corpus	Vectors	t	σ(t)	Sentiment
TWEET 1	[1350 , 2214] (x_1 x_2)	-2.97	0.0487	NEGATIVE
TWEET 2				
TWEET 3				
TWEET 4	[3561 , 2145]	4.77	0.9915	POSITIVE
TWEET 5				

$$t = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

Learned parameters

$$\beta_0 = -0.4783 , \beta_1 = 0.0034 , \beta_2 = -0.0032$$

AIOutsider.com

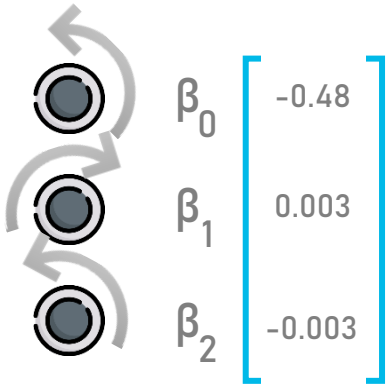
MODEL TRAINING

PARAMETERS ESTIMATION

PHASE 1: TRAINING

$$t = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

	x_1	x_2
TWEET 1	1350	554
TWEET 2	254	2214
TWEET 3	2156	148
TWEET 4	245	1876
TWEET 5	445	1644
TWEET 6	5350	224
TWEET 7	4250	974
TWEET 8	4762	248
TWEET 9	125	1999
TWEET 10	788	8456



β_0	-0.48
β_1	0.003
β_2	-0.003

Predictions σ(t)	Actual Sentiments
+	+
-	-
+	+
-	-
-	-
+	+
+	+
+	+
-	-
-	-

Tweet Corpus

Predictions σ(t)

Actual Sentiments

TRAIN/TEST SPLIT

TWEET 1	-
TWEET 2	-
TWEET 3	-
TWEET 4	+
TWEET 5	-
TWEET 6	+
TWEET 7	-
TWEET 8	+
TWEET 9	+
TWEET 10	-

DATASET

TRAINING

TWEET 1	-
TWEET 3	-
TWEET 4	+
TWEET 6	+
TWEET 5	-
TWEET 9	+
TWEET 10	-
TWEET 7	-

Learned optimal parameters

$\beta_0 = -0.4783$, $\beta_1 = 0.0034$, $\beta_2 = -0.0032$

TESTING

TWEET 2	
TWEET 8	

Model Predictions

-	=	-
-	≠	+

Real Sentiments

AIOutsider.com

MODEL PERFORMANCE

ACCURACY MEASURE

TWEET 1

« Sad? Certainly not! »

- Prediction
+ Actual Sentiment

17.000 Tweets of Test Set

Actual \ Prediction	-	+
-	TRUE NEGATIVE 8476	FALSE POSITIVE 154
+	FALSE NEGATIVE 385	TRUE POSITIVE 7985

Accuracy

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} = 96.8\%$$

FOLLOW US!

AIOutsider.com

twitter.com/AIOutsider_

Disclaimer