How to Prevent Catastrophic Failure in Production ML Systems

Martin Goodson
Chief Scientist/CEO (Evolution AI)

EvolutionAI

Who am I?

Four types of data leakage

Data leakage: when a machine learning model uses information that it shouldn’t have access too

1. Leaking test data into training data

Article Topic Classifier

Science

http://www.dailymail.co.uk/sciencetech/article-5559683/Incredible-atlas-reveals-speed-people-moving-urban-areas.html

https://www.independent.co.uk/news/science/spacex-crew-dragon-iss-docking-capsule-space-station-a8805381.html

Health

https://www.independent.co.uk/news/health/ovarian-cancer-new-blood-test-rare-tumours-biophysical-society-a8803186.html

Article Topic Classifier

Class	Test Precision	Test Recall
Technology	**0.97**	**0.99**
News	0.85	0.81
Showbiz	0.82	0.80
Sport	0.72	0.74

AMAZING PERFORMANCE!

http://www.dailymail.co.uk/sciencetech/article-5559683/Incredible-atlas-reveals-speed-people-moving-urban-areas.html

http://www.dailymail.co.uk/sciencetech/article-5572947/Stunning-satellite-images-reveal-planets-largest-cities-mesmerising-detail.html http://www.dailymail.co.uk/sciencetech/article-5559683/Incredible-atlas-reveals-speed-people-moving-urban-areas.html

http://www.dailymail.co.uk/sciencetech/article-5572947/Stunning-satellite-images-reveal-planets-largest-cities-mesmerising-detail.html

Training Data

http://www.dailymail.co.uk/sciencetech/article-5559683/Incredible-atlas-reveals-speed-people-moving-urban-areas.html

Test Data

http://www.dailymail.co.uk/sciencetech/article-5572947/Stunning-satellite-images-reveal-planets-largest-cities-mesmerising-detail.html

EvolutionAI

After segregating on publisher

Class	Test Precision	Test Recall
Technology	**0.55**	**0.51**
News	0.65	0.62
Showbiz	0.62	0.62
Sport	0.68	0.69

CIFAR image data base

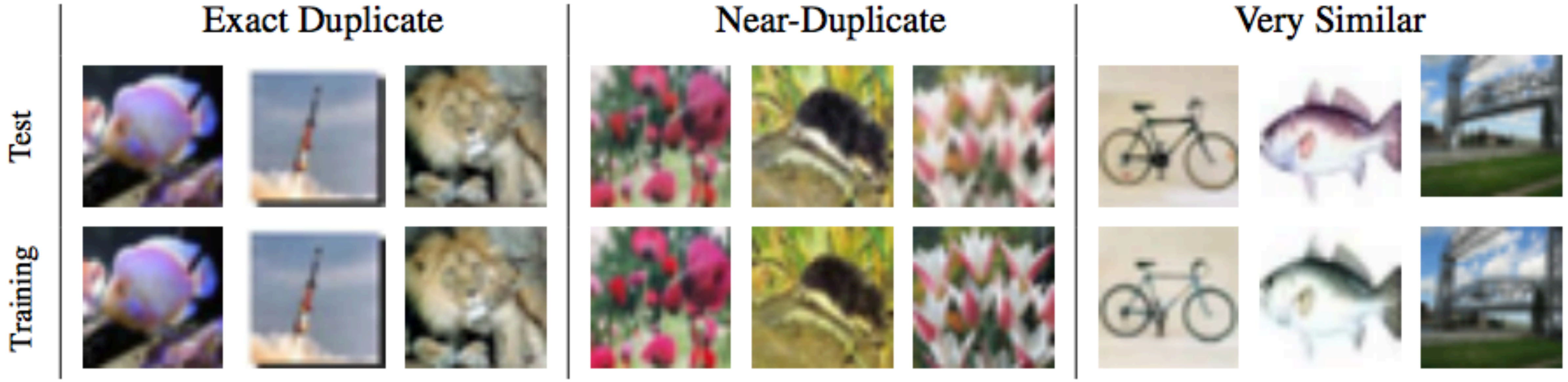


Figure 1: Examples for different types of duplicates between the CIFAR-100 test and training set. The top row shows images from the test set and the bottom row shows their nearest neighbors from the training set in a CNN feature space. Please see the main text for a description of the three categories of duplicates.

Bjorn Barz & Joachim Denzler. 2019

2. Leaking data temporally into training data

‘PROSSURG’

‘PROstate SURGery’

https://www.kaggle.com/wiki/Leakage/history/21889

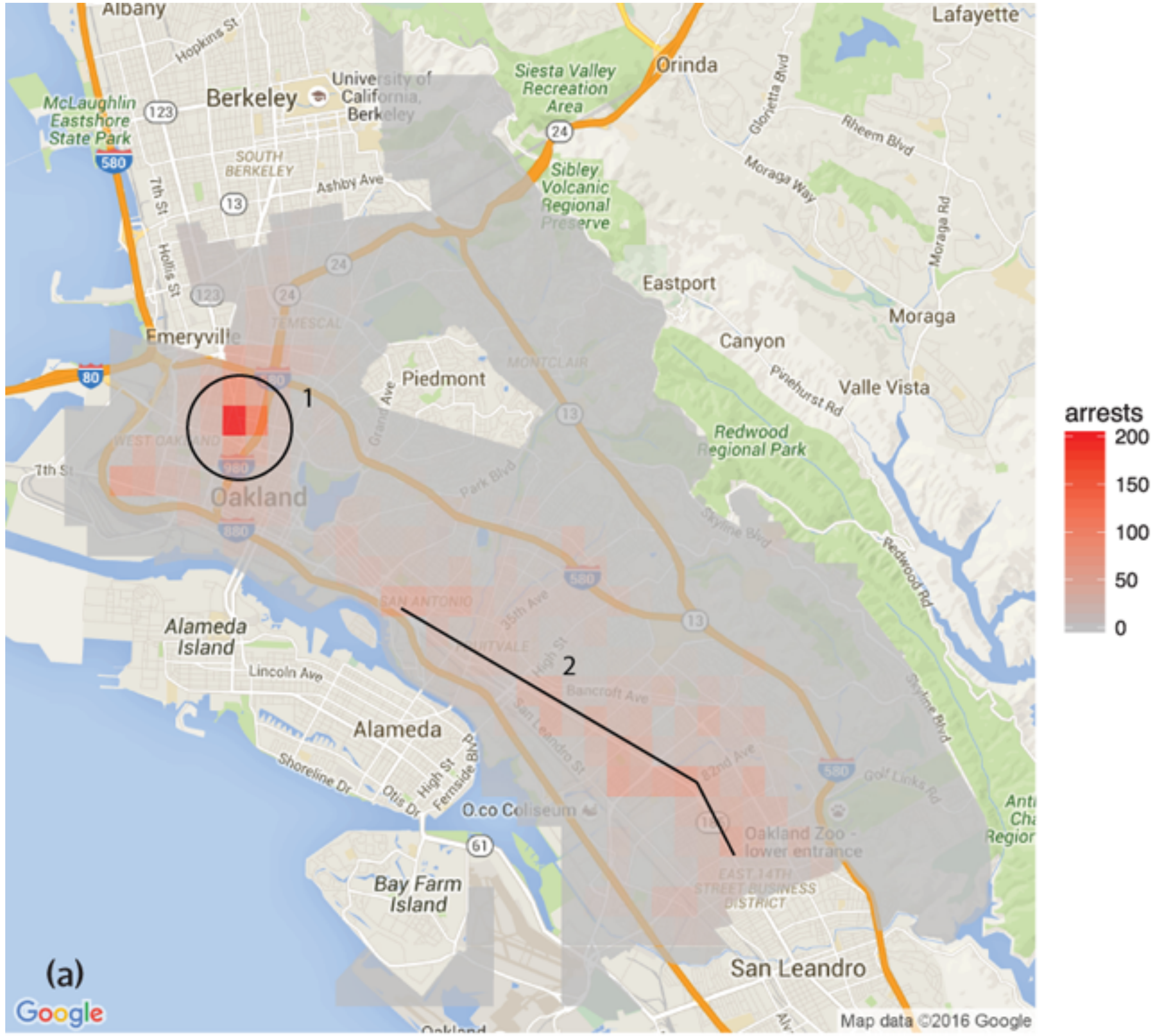
3. Leaking predictions into training data: *feedback loops*

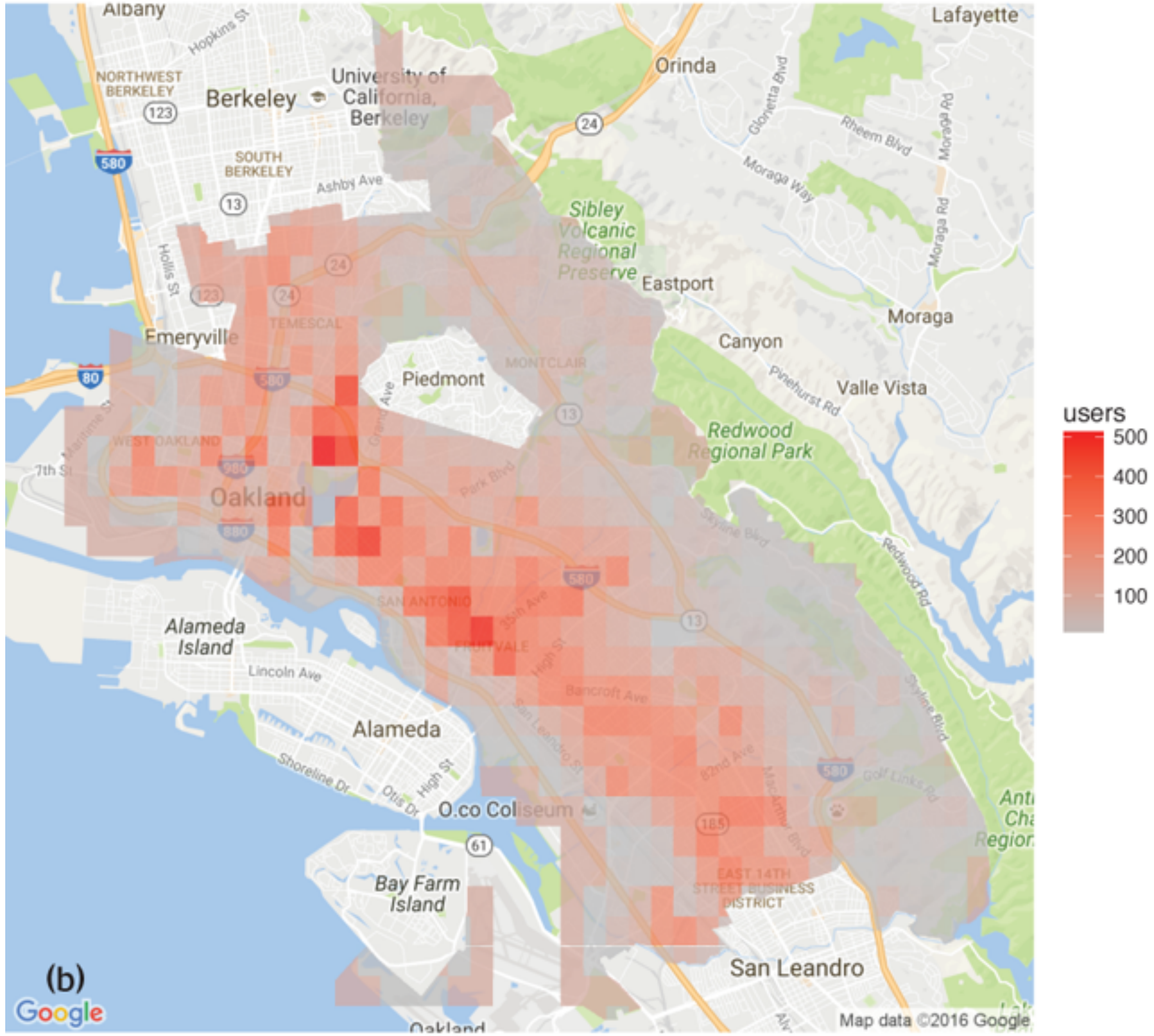
PREDPOL®

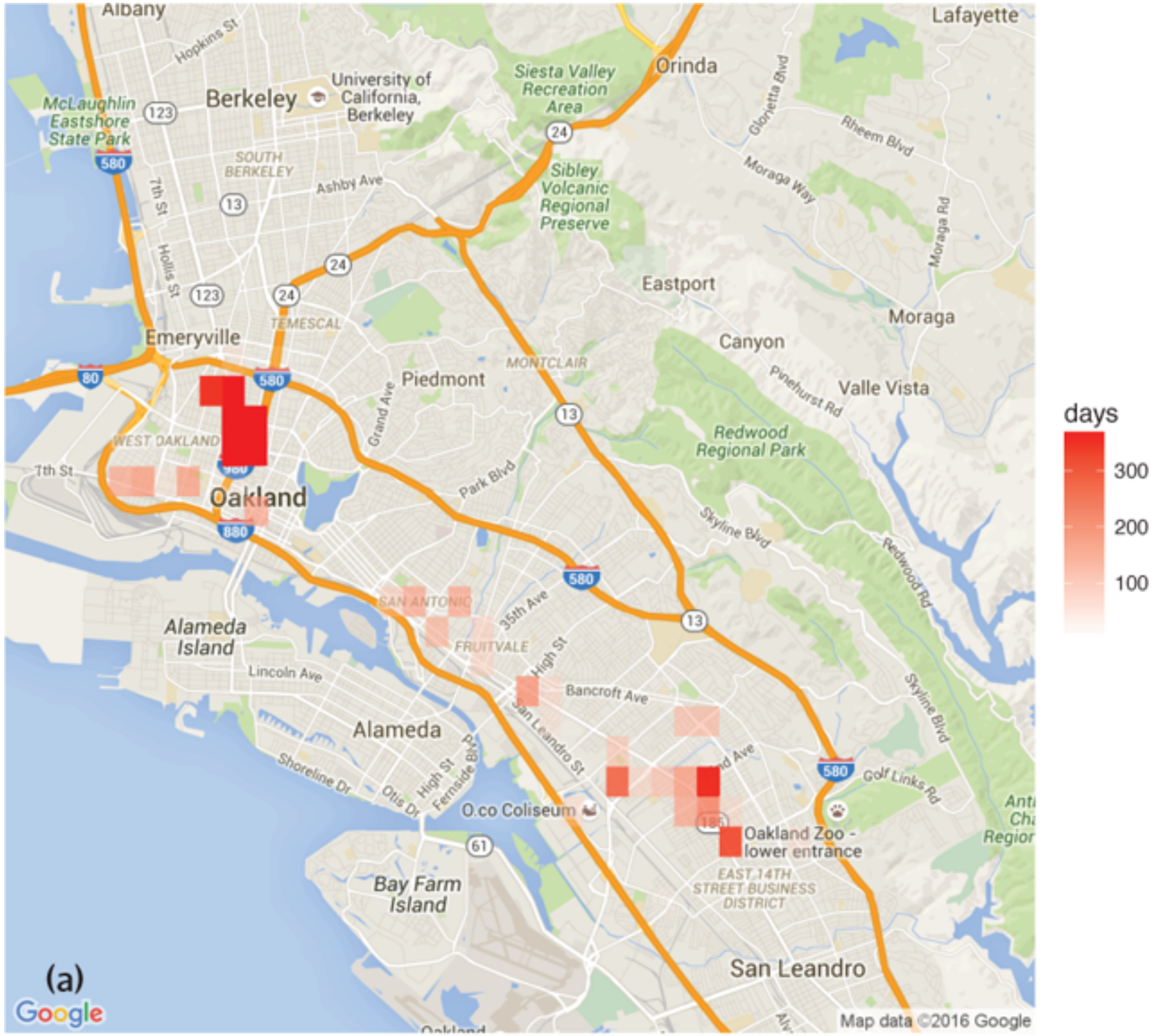
The Predictive Policing Company®

PredPol® uses artificial intelligence to help you prevent crime by predicting when and where crime is most likely to occur, allowing you to optimize patrol resources and measure effectiveness.

Lum & Isaac, 2016


(a)
1
2
arrests
200
150
100
50
0
Albany
Berkeley
University of California, Berkeley
Emeryville
Piedmont
Oakland
Alameda Island
Alameda
Bay Farm Island
San Leandro
Orinda
Lafayette
Moraga
Canyon
Valle Vista
Eastport
O.co Coliseum
Oakland Zoo - lower entrance
Map data ©2016 Google
Google


(b)
users
500
400
300
200
100
Albany
Lafayette
Berkeley
University of California, Berkeley
Orinda
NORTHWEST BERKELEY
SOUTH BERKELEY
Sibley Volcanic Regional Preserve
Eastport
Moraga
Emeryville
Canyon
Piedmont
Valle Vista
Redwood Regional Park
WEST OAKLAND
Oakland
Alameda Island
Alameda
O.co Coliseum
Bay Farm Island
San Leandro
Google
Map data ©2016 Google



EvolutionAI

[In 2016] the Mesa Police Department in Maricopa County entered **a three-year contract with the predictive policing software company, PredPol,** which required the police department to provide local crime data.

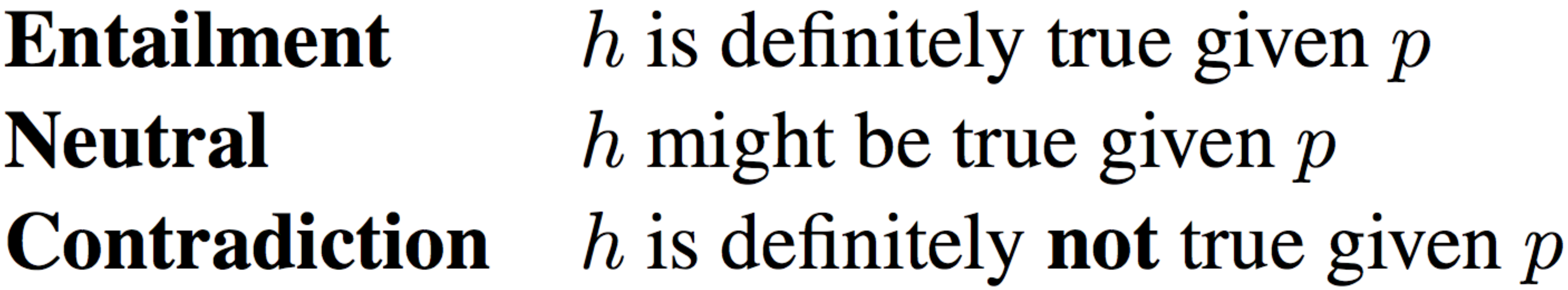
In 2011, the Department of Justice [documented Maricopa County Sheriff's Office's] pattern of discriminatory behavior between 2007 and 2011, including discriminatory policing against Latino residents; unlawful stops and arrests…

…**police data reflected the department's unlawful and racially biased practices**.

Richardson et al 2019

4. Leaking labels into training input data

Natural language inference

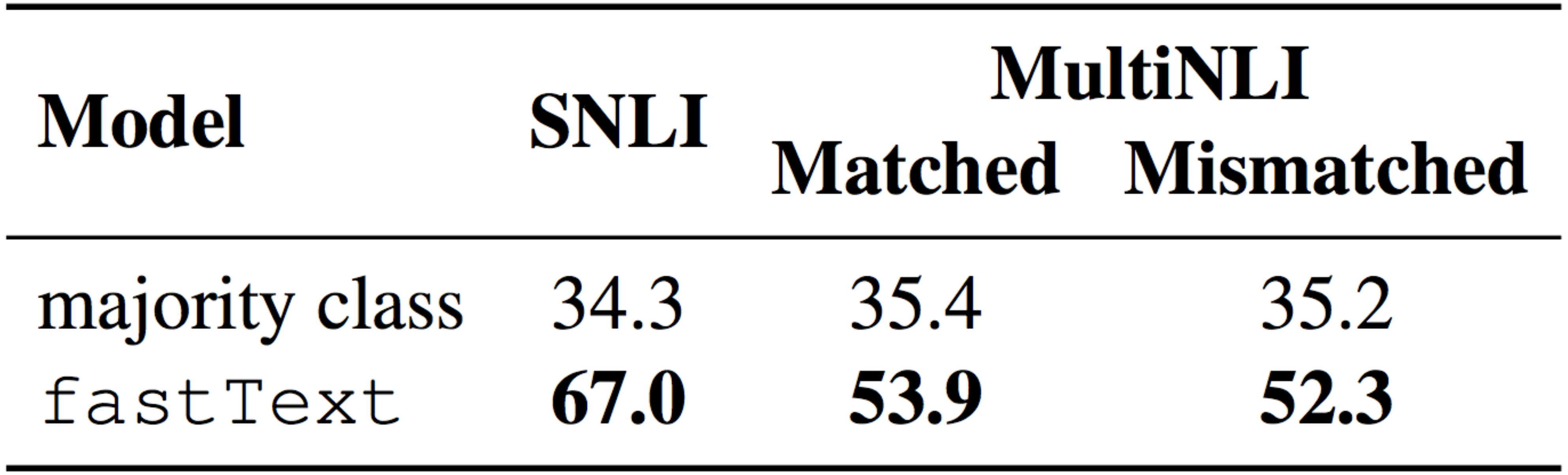


Entailment	h is definitely true given p
Neutral	h might be true given p
Contradiction	h is definitely **not** true given p

Natural language inference

Premise: A man inspects the uniform of a figure in some East Asian country.
Hypothesis: The man is sleeping.
Label: contradiction

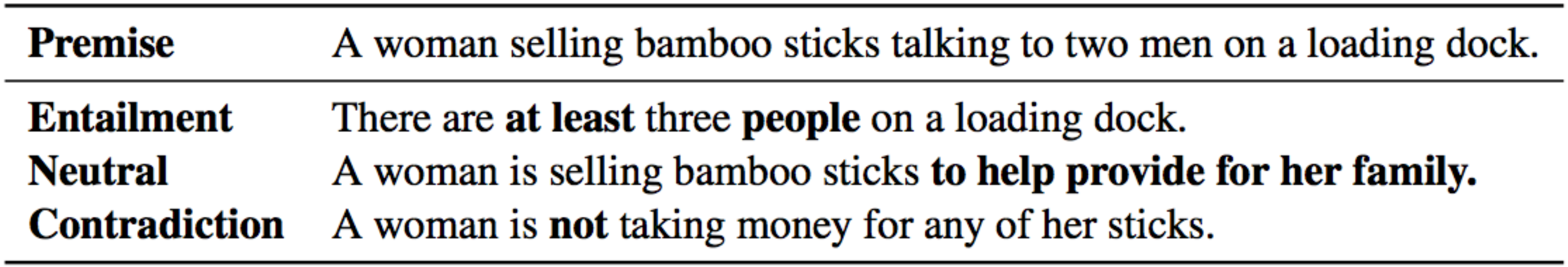
Natural language inference



Model	SNLI	MultiNLI	
		Matched	Mismatched
majority class	34.3	35.4	35.2
`fastText`	**67.0**	**53.9**	**52.3**

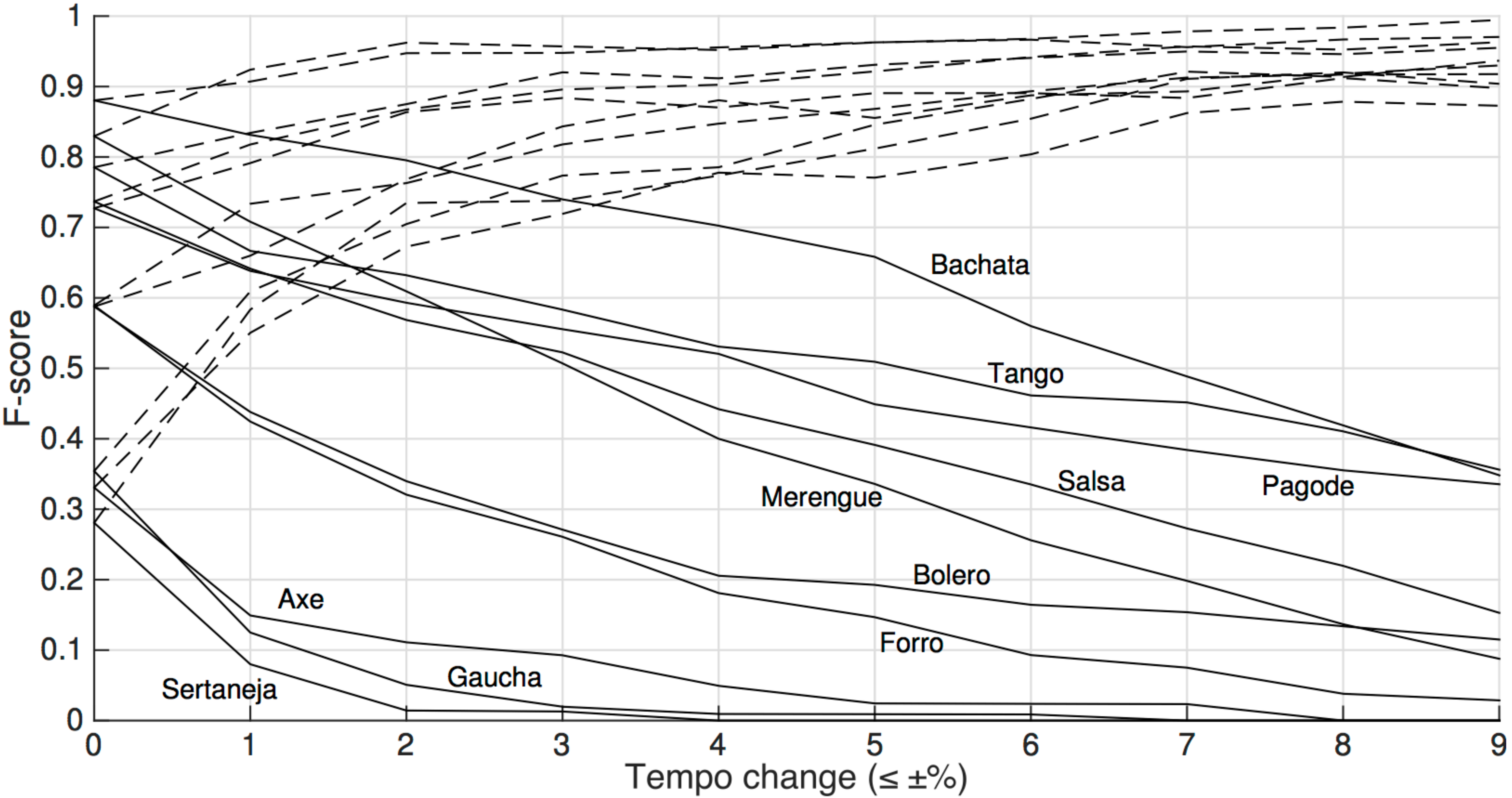
Gururangan et al. 2018

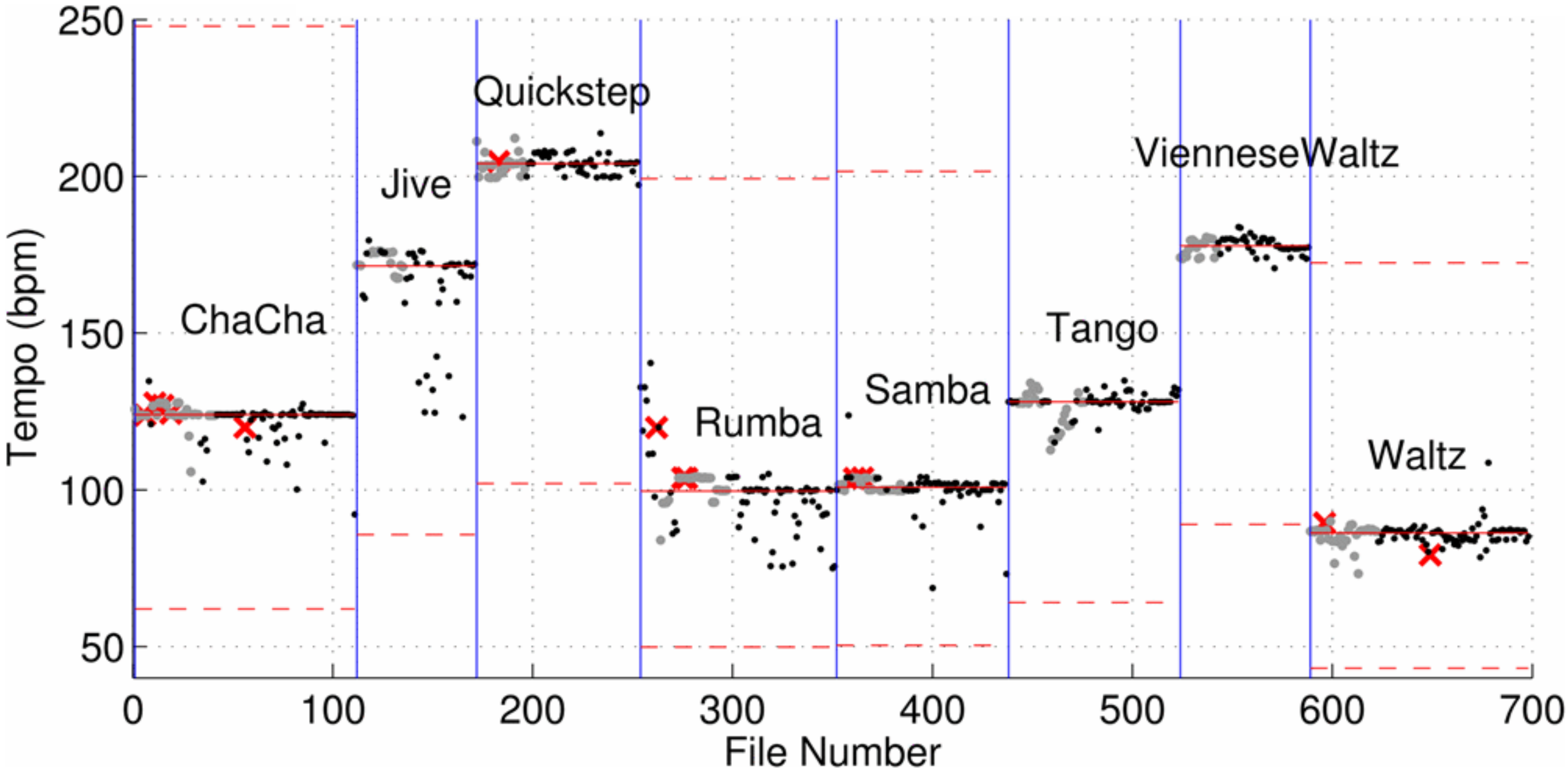
Natural language inference



Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

How widespread is this problem?


F-score
Tempo change (≤ ±%)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3
4
5
6
7
8
9
Bachata
Tango
Salsa
Pagode
Merengue
Bolero
Forro
Axe
Gaucha
Sertaneja

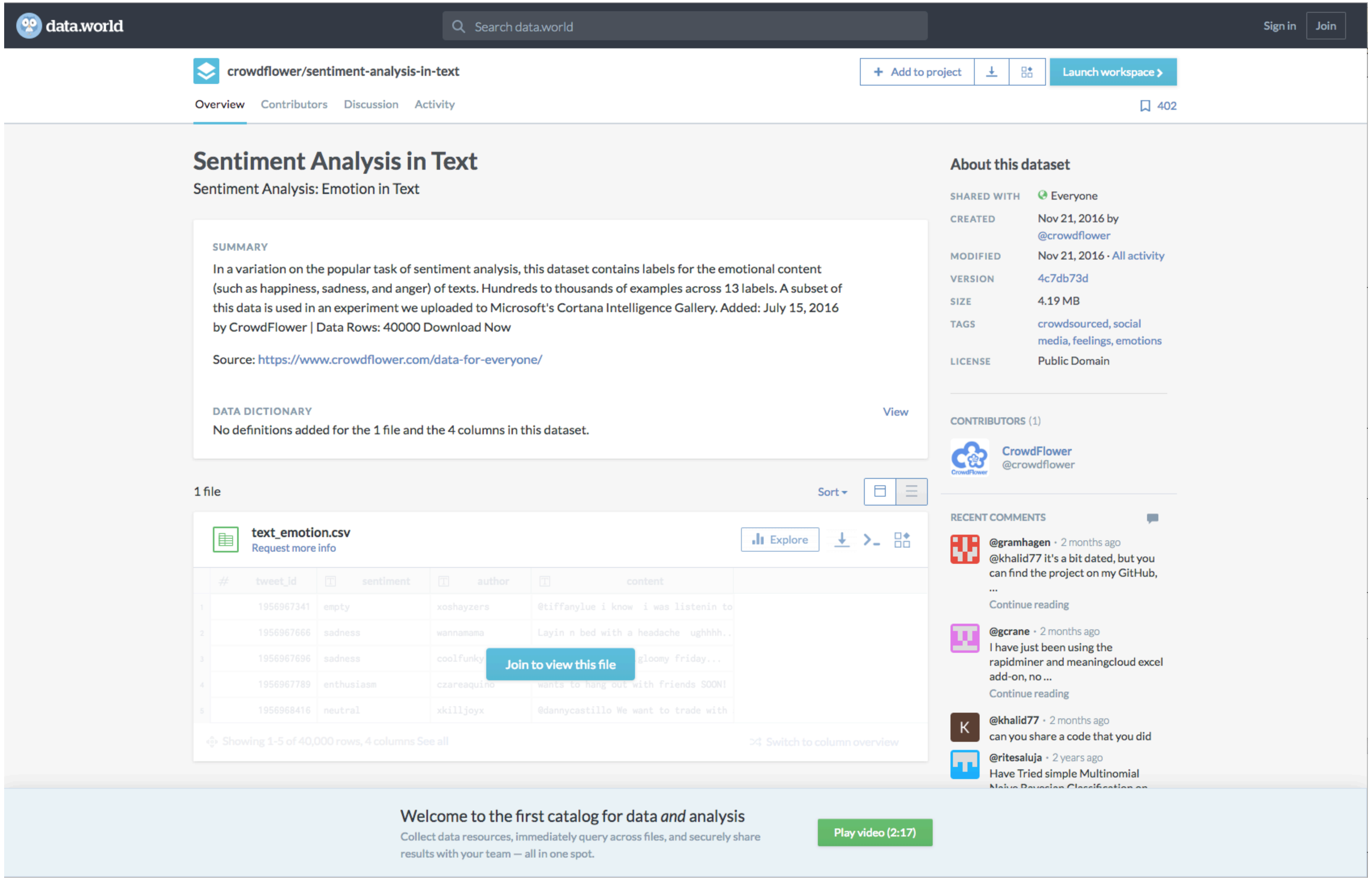


Tempo (bpm)
250
200
150
100
50
0
100
200
300
400
500
600
700
File Number
ChaCha
Jive
Quickstep
Rumba
Samba
Tango
VienneseWaltz
Waltz

‘… none of the evaluations in these many works is valid to produce conclusions with respect to recognizing genre…’

Sturm, 2013

A recent example that caused me some problems



crowdflower/sentiment-analysis-in-text

Sentiment Analysis in Text

Sentiment Analysis: Emotion in Text

SUMMARY

In a variation on the popular task of sentiment analysis, this dataset contains labels for the emotional content (such as happiness, sadness, and anger) of texts. Hundreds to thousands of examples across 13 labels. A subset of this data is used in an experiment we uploaded to Microsoft's Cortana Intelligence Gallery. Added: July 15, 2016 by CrowdFlower | Data Rows: 40000 Download Now

Source: https://www.crowdflower.com/data-for-everyone/

DATA DICTIONARY

No definitions added for the 1 file and the 4 columns in this dataset.

1 file

text_emotion.csv

#	tweet_id	sentiment	author	content
1	1956967341	empty	xoshayzers	@tiffanylue i know i was listenin to
2	1956967666	sadness	wannamama	Layin n bed with a headache ughhhh..
3	1956967696	sadness	coolfunky	gloomy friday...
4	1956967789	enthusiasm	czareaquino	wants to hang out with friends SOON!
5	1956968416	neutral	xkilljoyx	@dannycastillo We want to trade with

Showing 1-5 of 40,000 rows, 4 columns See all

About this dataset

SHARED WITH	Everyone
CREATED	Nov 21, 2016 by @crowdflower
MODIFIED	Nov 21, 2016 · All activity
VERSION	4c7db73d
SIZE	4.19 MB
TAGS	crowdsourced, social media, feelings, emotions
LICENSE	Public Domain

CONTRIBUTORS (1)

CrowdFlower
@crowdflower

RECENT COMMENTS

@gramhagen · 2 months ago
@khalid77 it's a bit dated, but you can find the project on my GitHub, ...
Continue reading

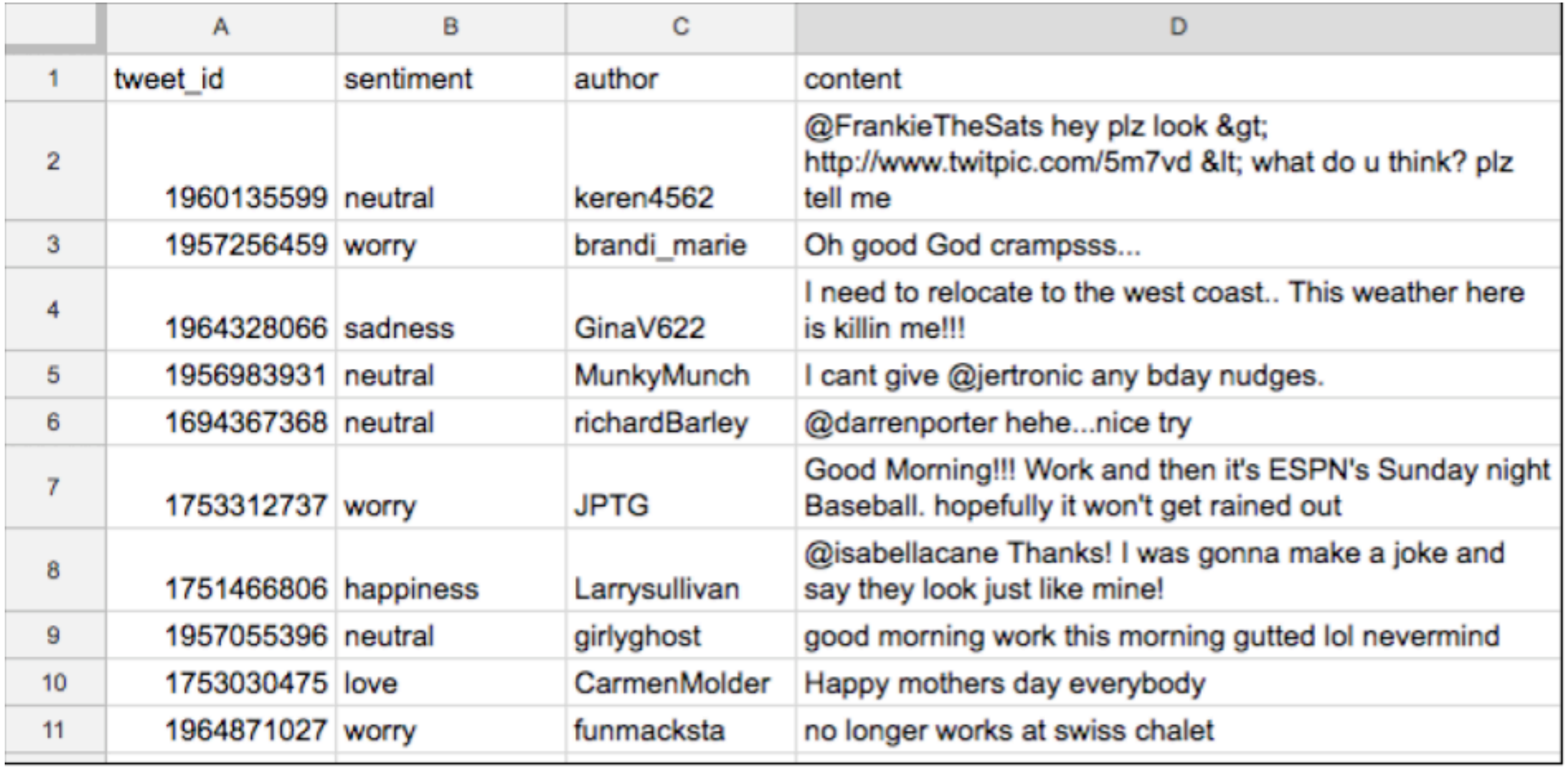
@gcrane · 2 months ago
I have just been using the rapidminer and meaningcloud excel add-on, no ...
Continue reading

@khalid77 · 2 months ago
can you share a code that you did

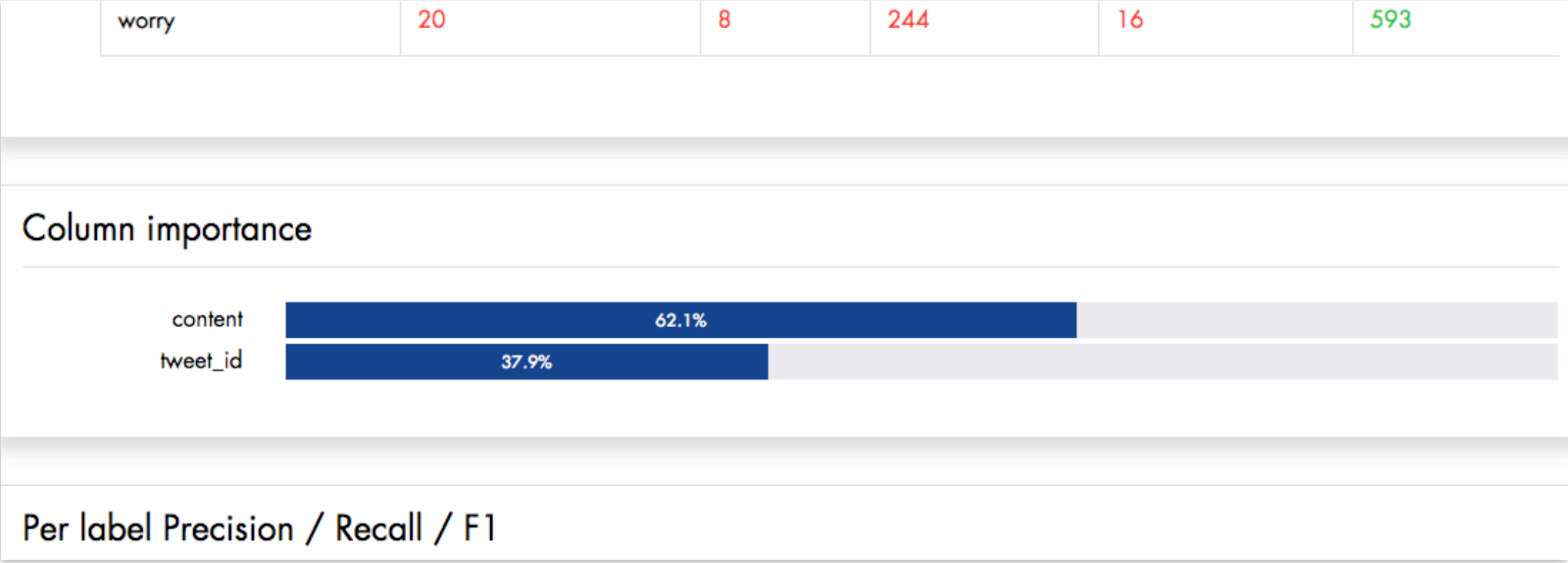
@ritesaluja · 2 years ago
Have Tried simple Multinomial Naive Bayesian Classification on

Welcome to the first catalog for data *and* analysis

Collect data resources, immediately query across files, and securely share results with your team — all in one spot.



	A	B	C	D
1	tweet_id	sentiment	author	content
2	1960135599	neutral	keren4562	@FrankieTheSats hey plz look &gt; http://www.twitpic.com/5m7vd &lt; what do u think? plz tell me
3	1957256459	worry	brandi_marie	Oh good God crampsss...
4	1964328066	sadness	GinaV622	I need to relocate to the west coast.. This weather here is killin me!!!
5	1956983931	neutral	MunkyMunch	I cant give @jertronic any bday nudges.
6	1694367368	neutral	richardBarley	@darrenporter hehe...nice try
7	1753312737	worry	JPTG	Good Morning!!! Work and then it's ESPN's Sunday night Baseball. hopefully it won't get rained out
8	1751466806	happiness	Larrysullivan	@isabellacane Thanks! I was gonna make a joke and say they look just like mine!
9	1957055396	neutral	girlyghost	good morning work this morning gutted lol nevermind
10	1753030475	love	CarmenMolder	Happy mothers day everybody
11	1964871027	worry	funmacksta	no longer works at swiss chalet



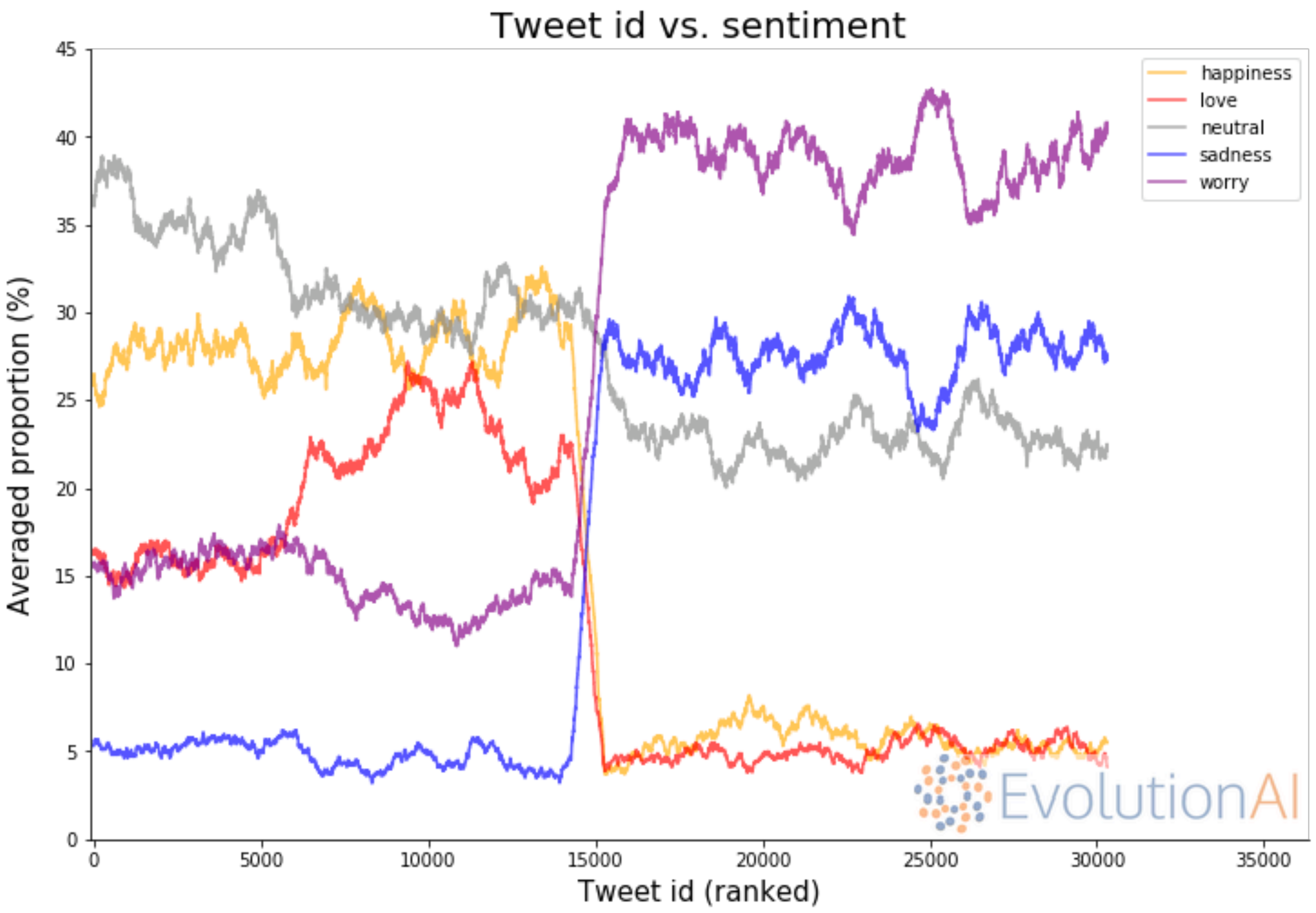
worry	20	8	244	16	593

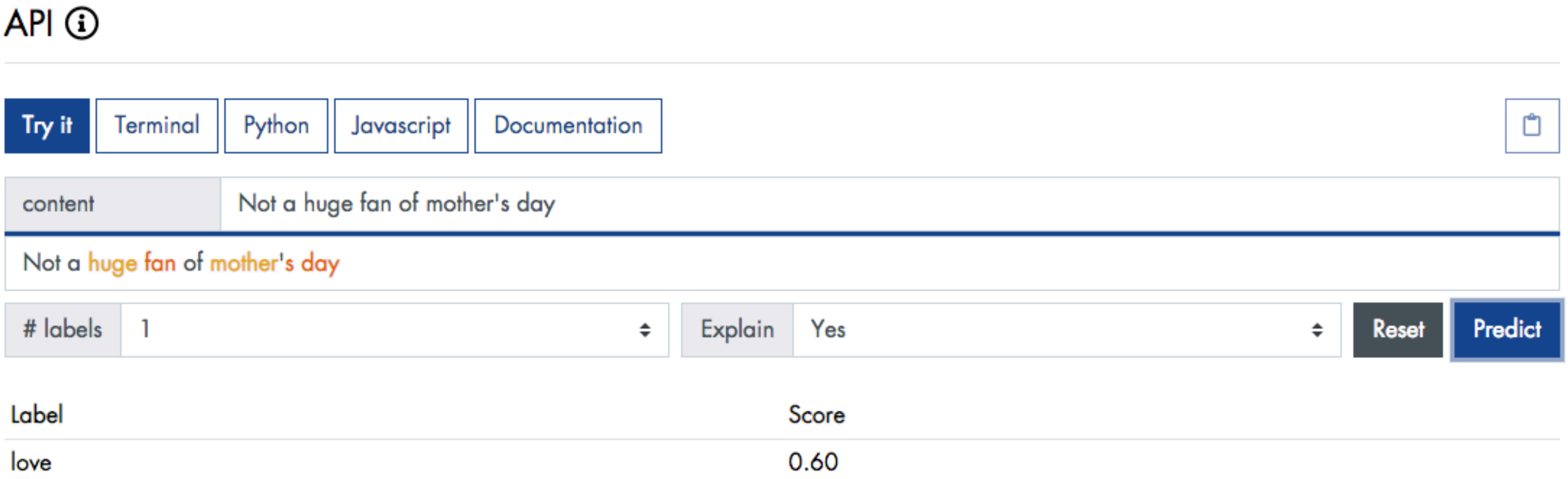
Column importance

content 62.1%

tweet_id 37.9%

Per label Precision / Recall / F1


Tweet id vs. sentiment
Averaged proportion (%)
0
5
10
15
20
25
30
35
40
45
Tweet id (ranked)
0
5000
10000
15000
20000
25000
30000
35000
happiness
love
neutral
sadness
worry
EvolutionAI



content: Not a huge fan of mother's day

Not a huge fan of mother's day

labels: 1

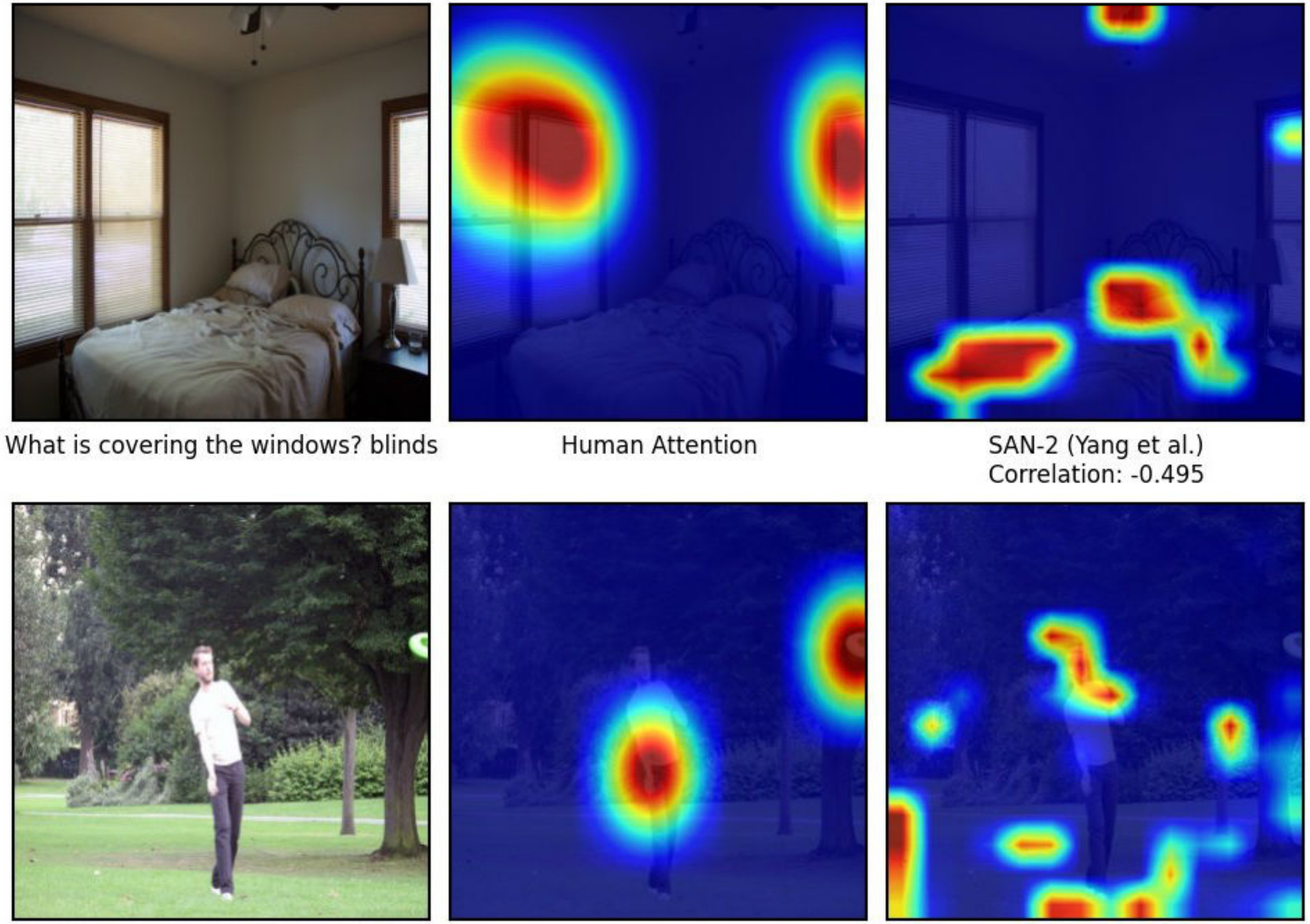
Explain: Yes

Label	Score
love	0.60

How can you be sure you got any of this right?

1. Understand the decision-making basis of your model

Feature	Coefficient
PROSSURG	**0.983**
PSA _NGML	0.003
PCA3 _NGML	0.005



What is covering the windows? blinds

Human Attention

SAN-2 (Yang et al.)
Correlation: -0.495

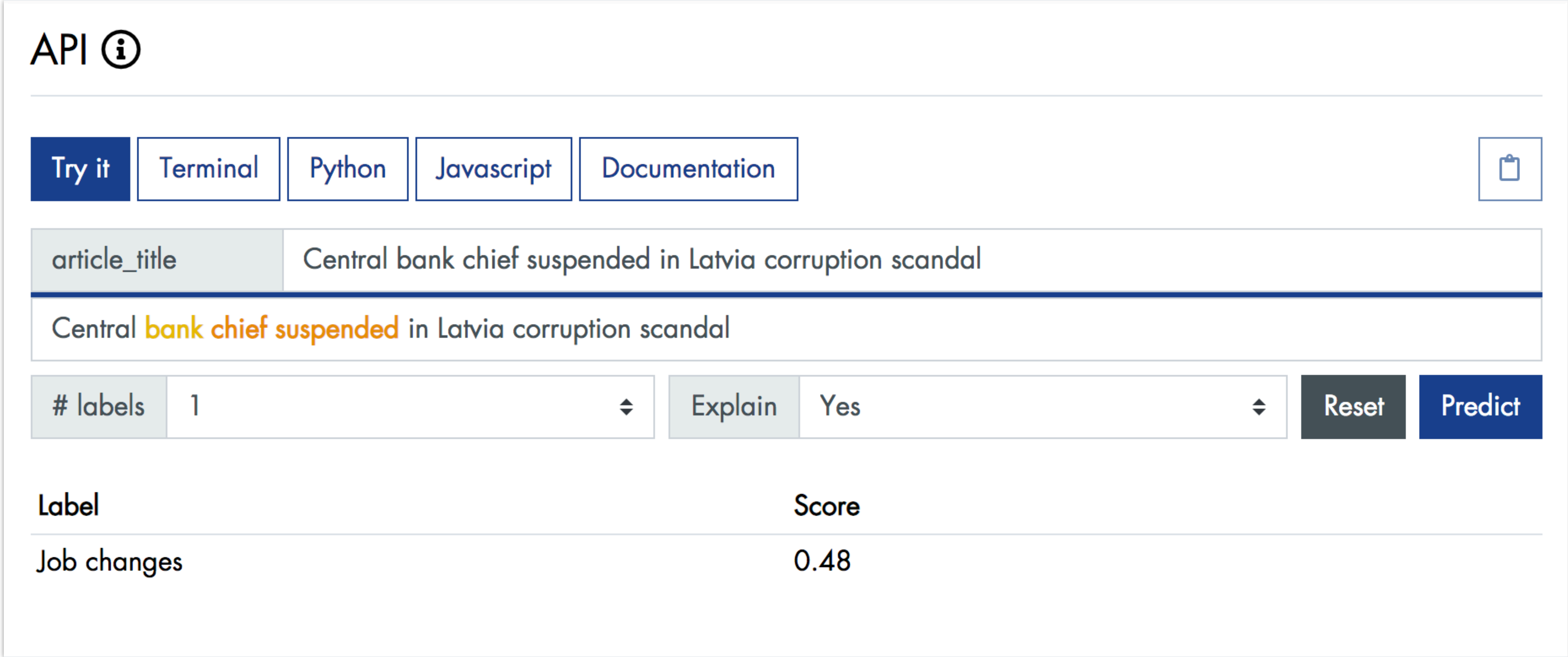
What is the man doing? playing frisbee

Human Attention

SAN-2 (Yang et al.)
Correlation: -0.060

Das et al. 2017

Explainability in NLP



article_title: Central bank chief suspended in Latvia corruption scandal

Central **bank chief suspended** in Latvia corruption scandal

labels: 1

Explain: Yes

Label	Score
Job changes	0.48

2. Test in a real-world setting as early as possible

References

Das et al., Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? Comput Vis Image Underst, 2017 .

Sturm, Kereliuk, and Pikrakis, “A closer look at deep learning neural networks with low-level spectral periodicity features,” in Proc. CIP 2014.

Sturm, B. The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval. J. New Music Res, 2014

To predict and serve? Lum & Isaac. Significance (2016)

The Encyclopedia of Weapons of World War II. Chris Bishop, Sterling Publishing Company, Inc., 2002

https://www.kaggle.com/wiki/Leakage/history/21889

Bjorn Barz & Joachim Denzler. Do we train on test data? Purging CIFAR of near-duplicates, 2019.

Gururangan, S et al. Annotation Artifacts in Natural Language Inference Data, 2018

Richardson, R et al. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. New York University Law Review, 2019

EvolutionAI

Get in touch:
Martin@evolution.ai