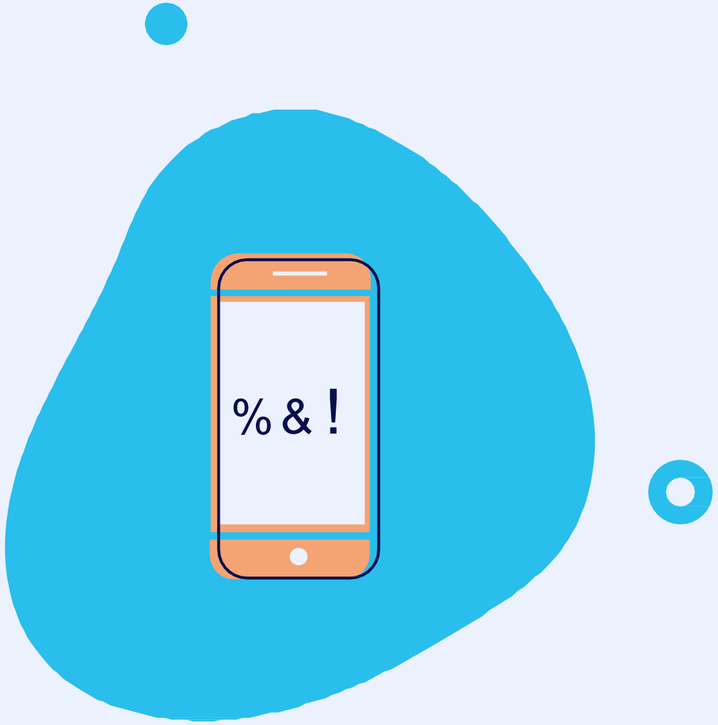
VALLEY AI 2020

EMAKIA

A system filtering out harassment on incoming social media data



Corinne David
corinne@emakia.org

"We weren't expecting any of the abuse and harassment and the ways that people have weaponized the platform."

Jack Dorsey
Founder & CEO of Twitter
January 19, 2019

OBJECTIVE:

Create a Safer Environment

→ WOMEN

Greta Thunberg

→ MINORS

Goo Hara

→ PERSONS of COLOR

Ilhan Omar

Christine Blasey Ford

→ DISADVANTAGE GROUPS

Tyler Clementi

Alexandria Ocasio-Cortez

Sulli

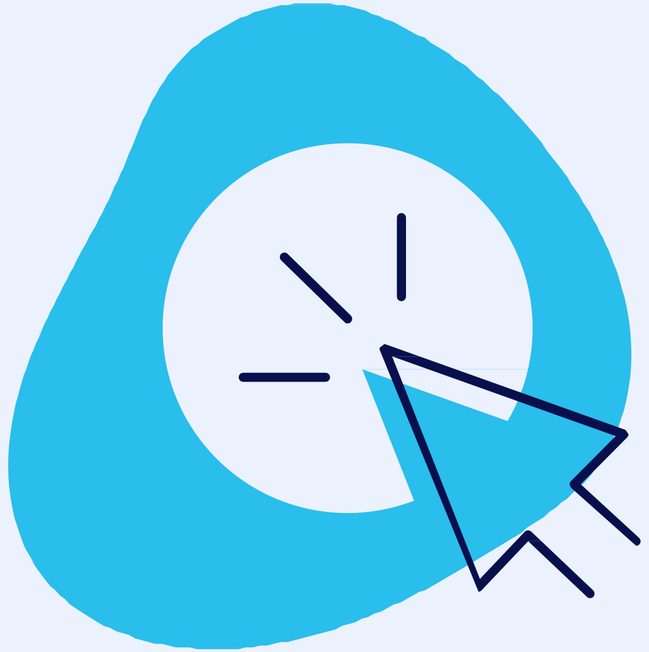
→ YOU?

IT CAN BE DONE

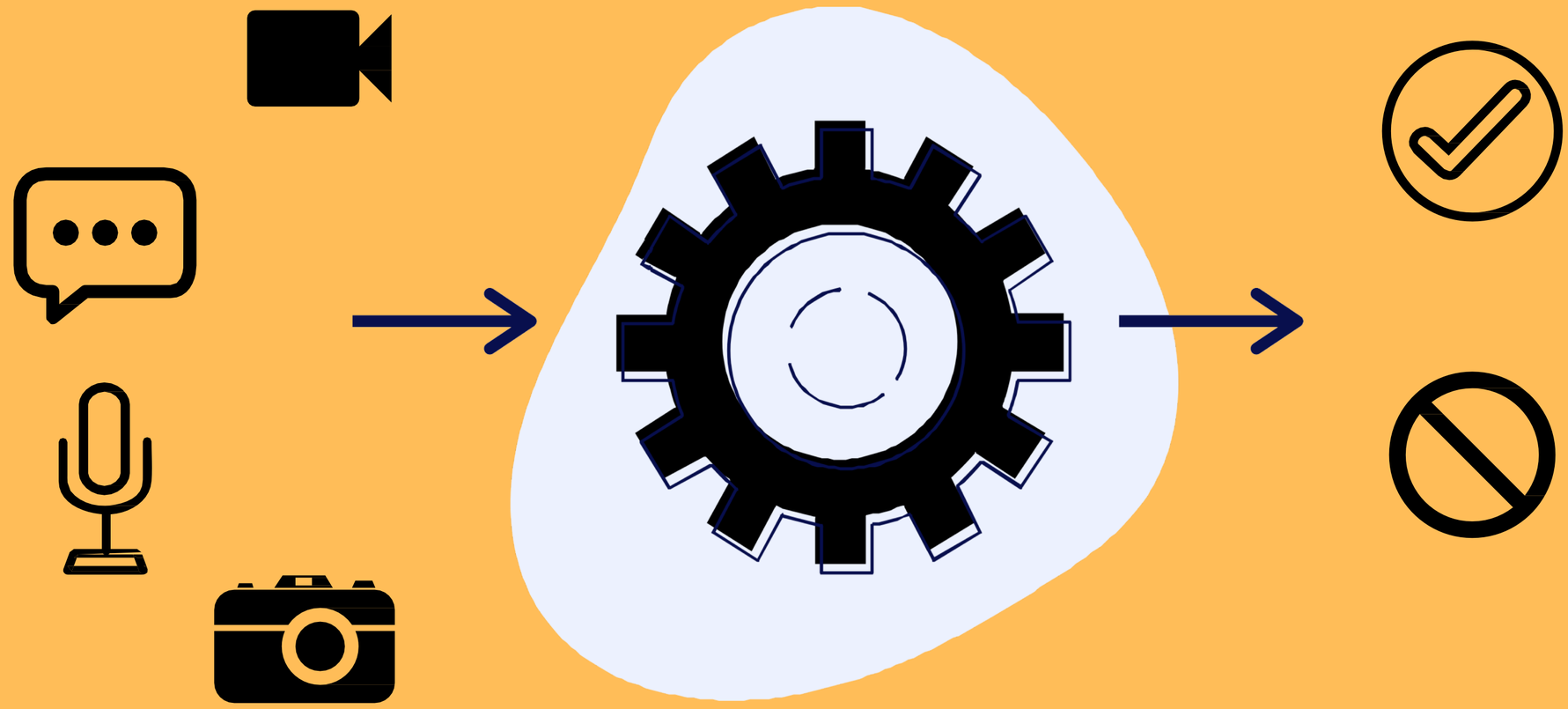
→ NYT & Google: Filter incoming data using text classifier to train a model on 16 million comments

→ Email Software uses text classifiers to determine whether incoming mail is sent to the inbox folder or the spam folder

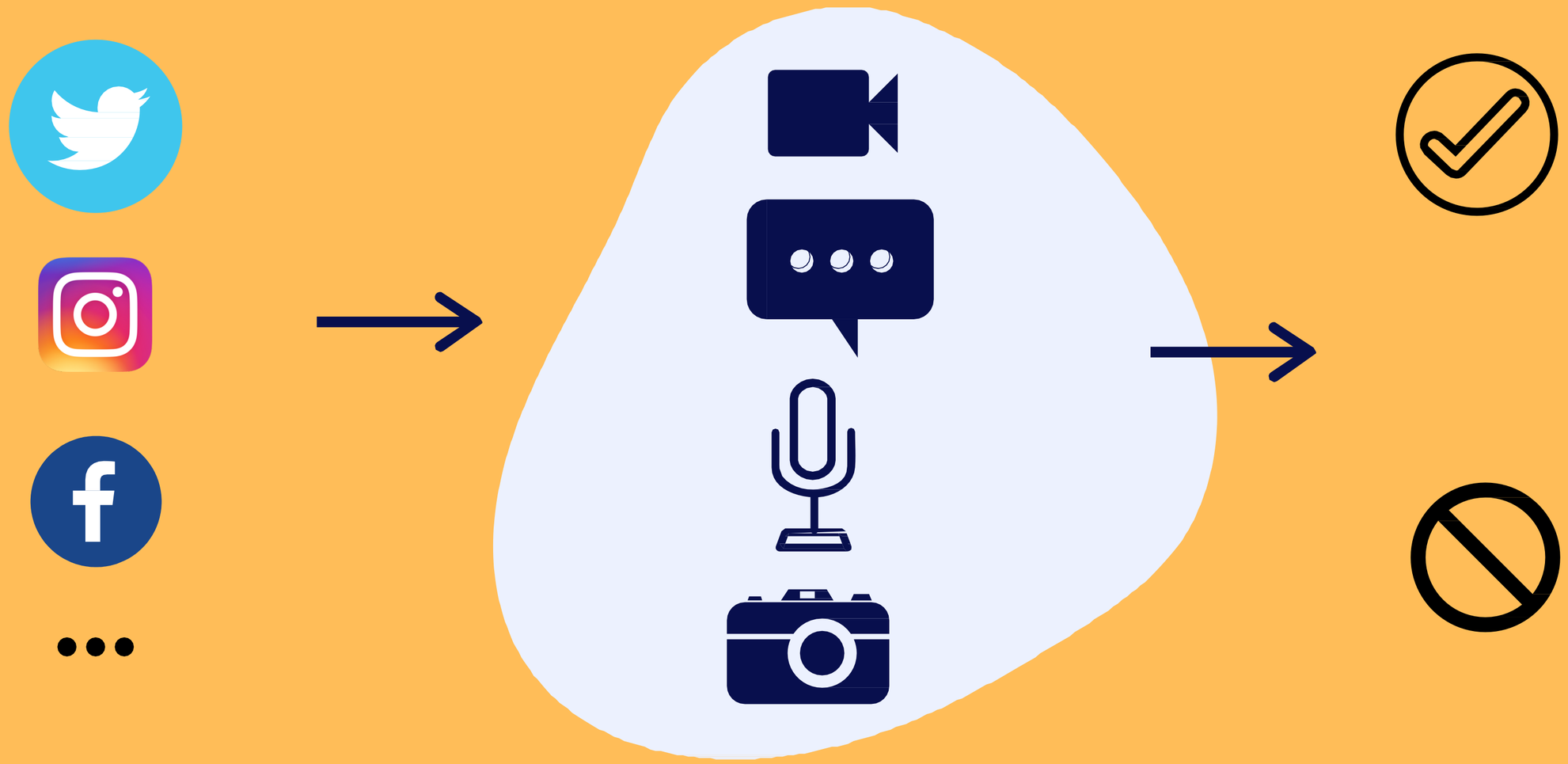
→ Discussion forums use text classifiers to determine if a comment is appropriate



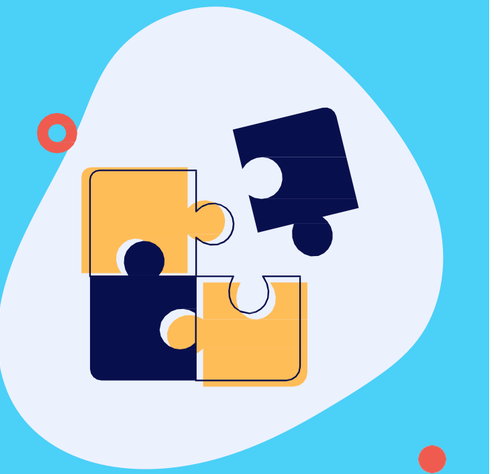
EMAKIA's multimodal solution



A closer look



THE SYSTEM



DATA

- Collect, label and clean data
  70 000 English Tweets,
  200 000 Italian labeled Tweets

TRAIN MODEL

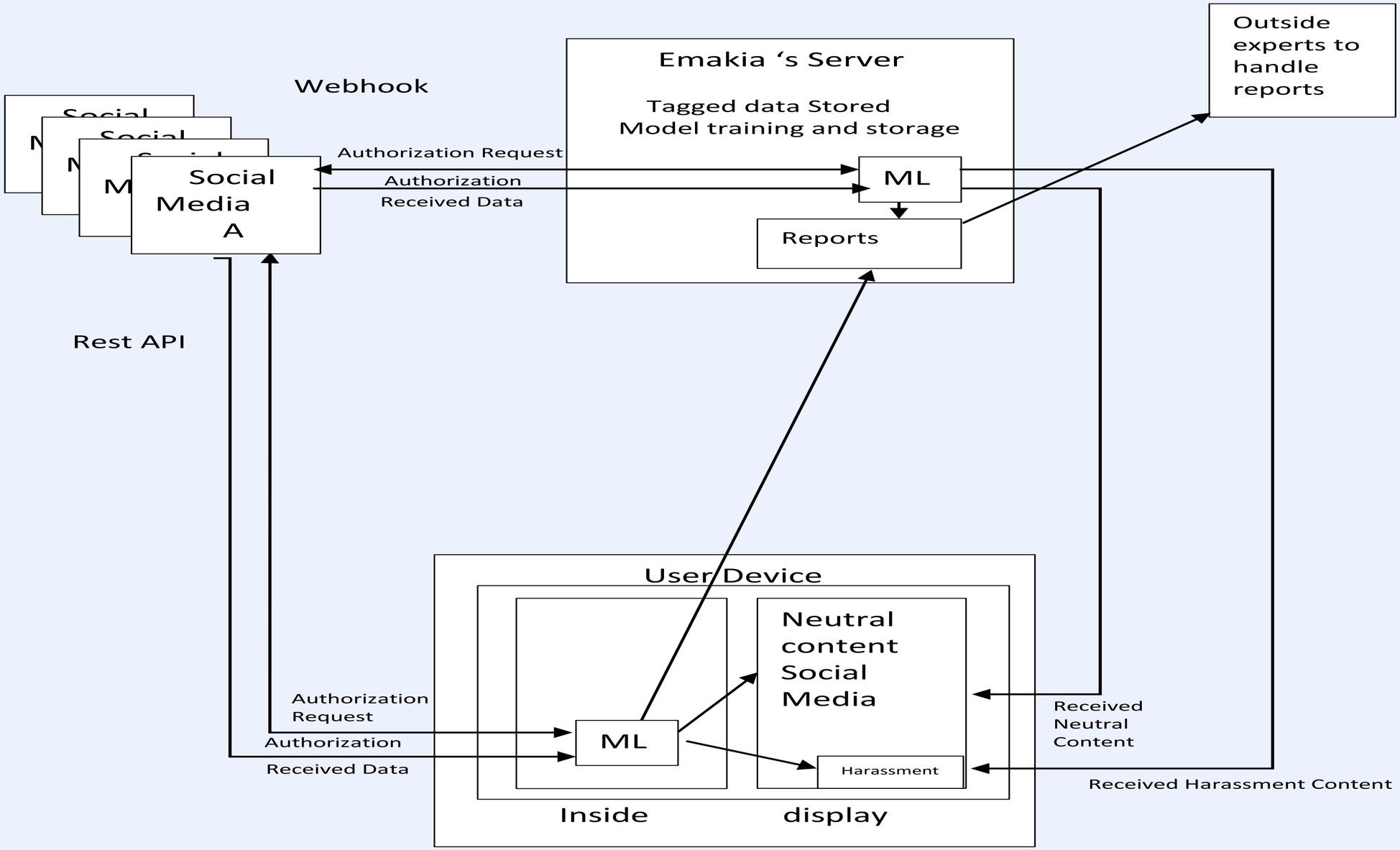
- Apple Core ML Classifiers
- Google AutoML Classifiers

ENAËLLE

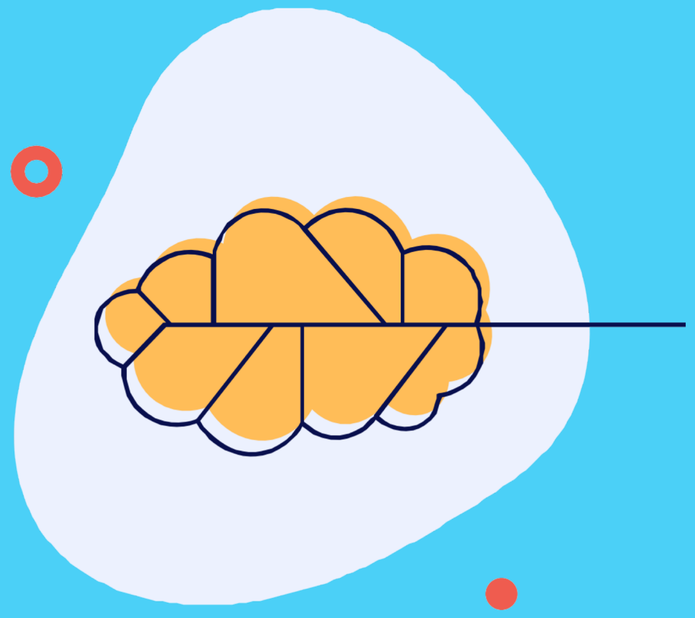
- Mobile Application running on iPhone or Android
- Detect harassment on real-time data with the model

REPORTS

- Harassment received
- History of the harassing senders, friends and followers

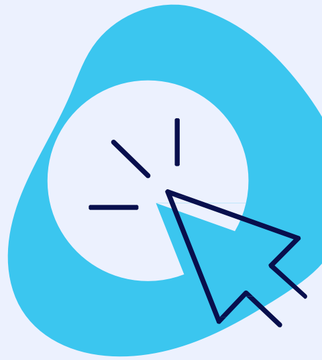

Outside experts to handle reports
Emakia ‘s Server
Tagged data Stored
Model training and storage
Webhook
Social Media A
Authorization Request
Authorization
Received Data
ML
Reports
Rest API
User Device
Neutral content Social Media
Authorization Request
Authorization
Received Data
ML
Harassment
Received Neutral Content
Received Harassment Content
Inside
display

What's running on the server?



- Collect, label and store data
- Validation of the data
- Train ML classifiers for each language with the labeled data
- Programs to search social media data for harassment data
- Process text, image, audio, and video
- Create reports of the detected harassment, fake news, deepfake

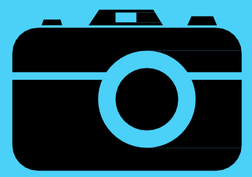
ENAËLLE

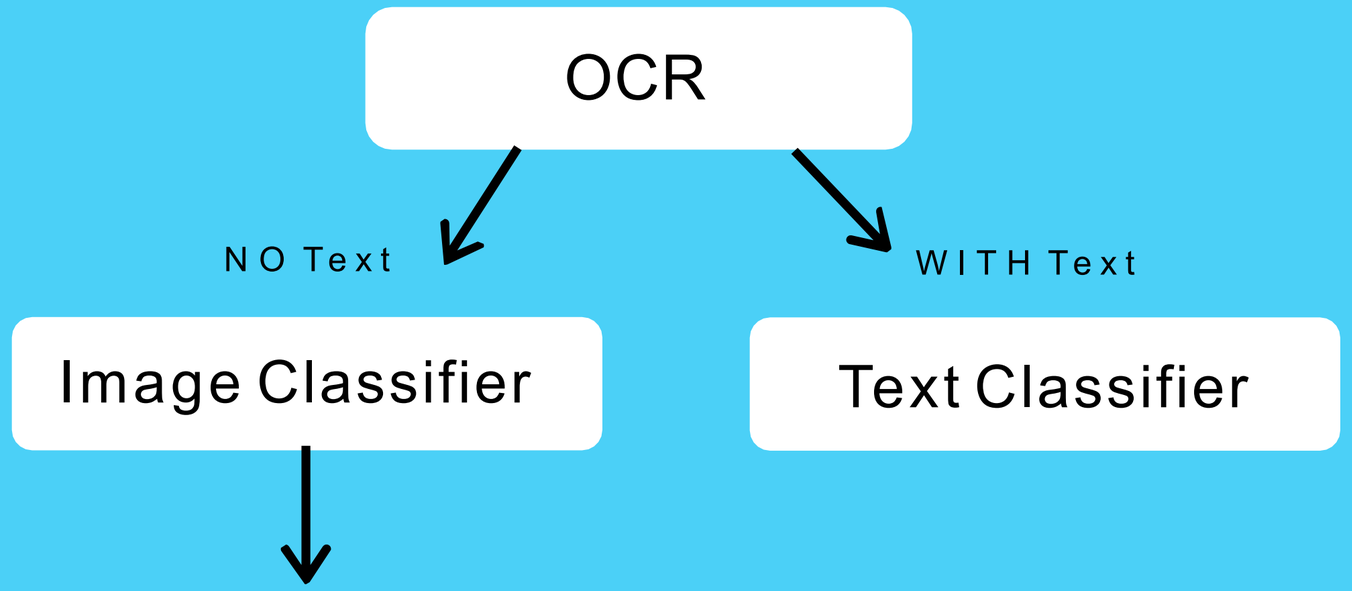


- Register an email of a buddy/mentor; Emakia sends reports to the user and the buddy system
- Portal to access the different social media platforms
- Provide the same functionality as the social media platform
- Neutral content of the social media platform is displayed
- Harassing content is accessible with a Tab bar

Text Classification

Image Classification





1. APPLE:

Core ML Vision

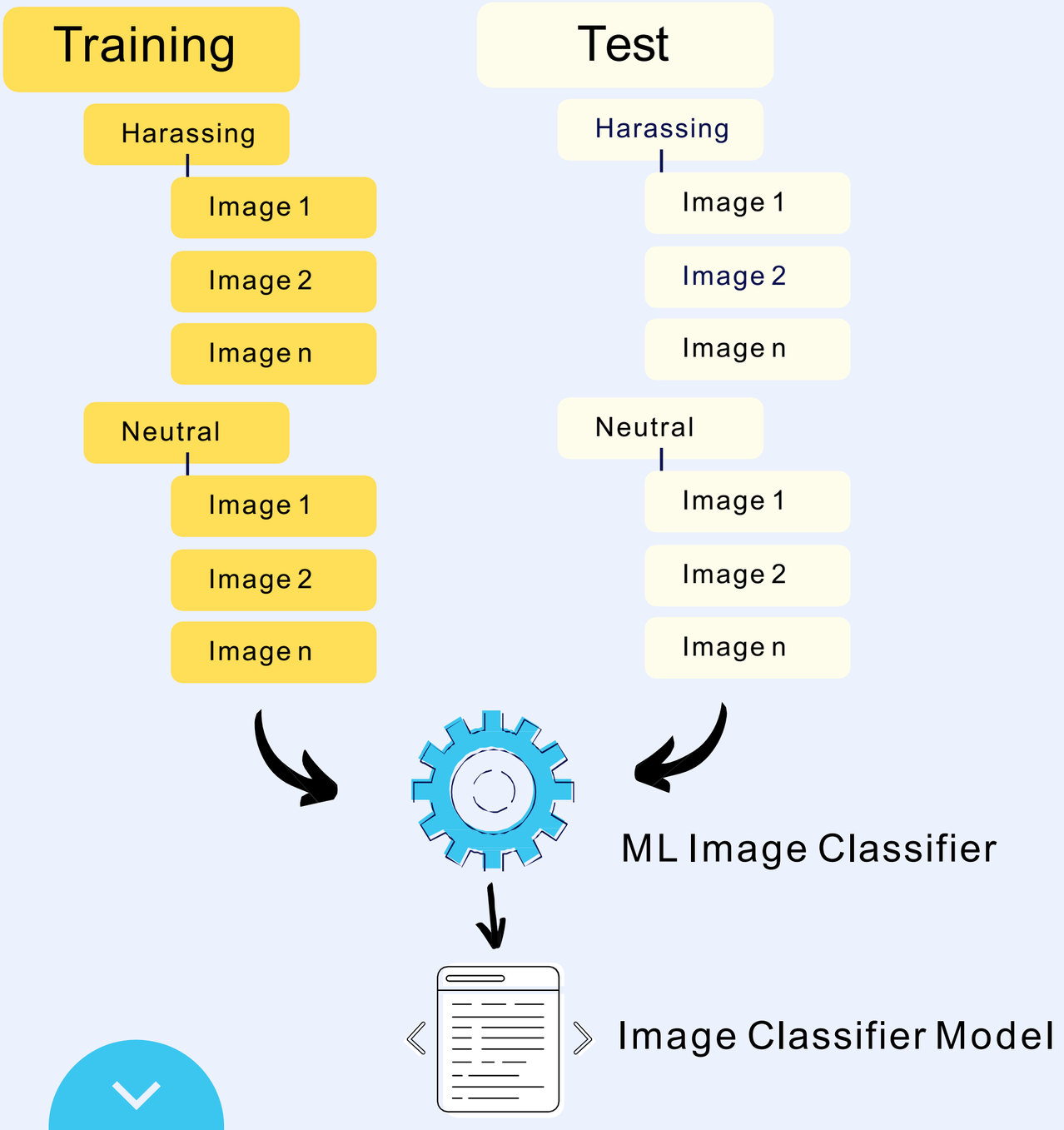
Apple Vision Framework

2. GOOGLE:

Auto ML Vision

Vision API

HOW TO FEED THE DATA?

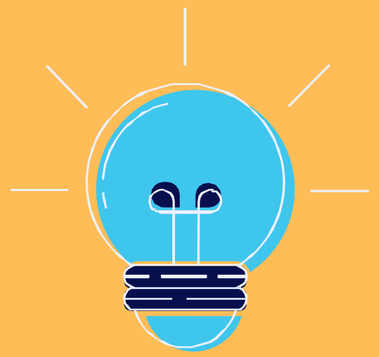


Core ML Audio

Core ML Video Apple Vision

AutoML Video

Audio & Video classification



Other applications

FAKE NEWS DETECTION

- Reference source with fact checking websites
- Labeled data - fake or true
- Train the model with labeled data and features

VIDEO CLASSIFIER FOR DEEPFAKE VIDEO

- Detect artifacts on the video
- Detect Face Swap
- Detect Lip Sync on the audio file of the video

THIRD PARTY DATA

→ NAACL_SRW_2016.csv (Waseem et al., 2016) “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter”

→ BullyingV3.0.zip (Xu, 2012) “Learning from bullying traces in social media”

→ Twitter-hate-speech-classifier-DFE-a845520.csv from data.world created on Nov. 21, 2016 by @crowdflowerdata

→ Labeled_data.csv (Waseem et al., 2016) “Understanding Abuse: A Typology of Abusive Language Detection Subtasks”

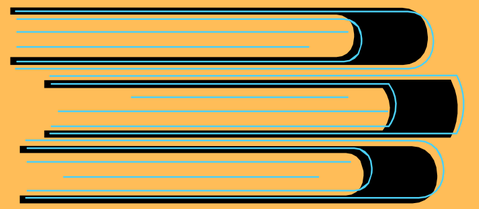
→ OnlineHarassmentDataset.csv (Golbeck et al., 2017). “A Large Labeled Corpus for Online Harassment Research”

OUR DATA

→ A Program Collecting Harassing Data uses the standard search Twitter API to obtain tweets with specific harassment terms unknown to the model

→ Using the bag-of-words adaptive filter and retraining the text classifier with content yet unknown to the model

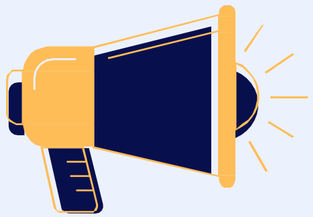
→ Italian data gathered by Claudia Zaghi



Bag of words resources

→ Hatebase_dict.csv provided by Hatebase, an online database of hate speech“Automatic Detection of Cyberbullying on Social Media” by Love Engman Master's Thesis in Computing Science

→ “Automatic Detection of Cyberbullying on Social Media” by Love Engman Master's Thesis in Computing Science



Validation of Labeled Data & Bag-of-Words

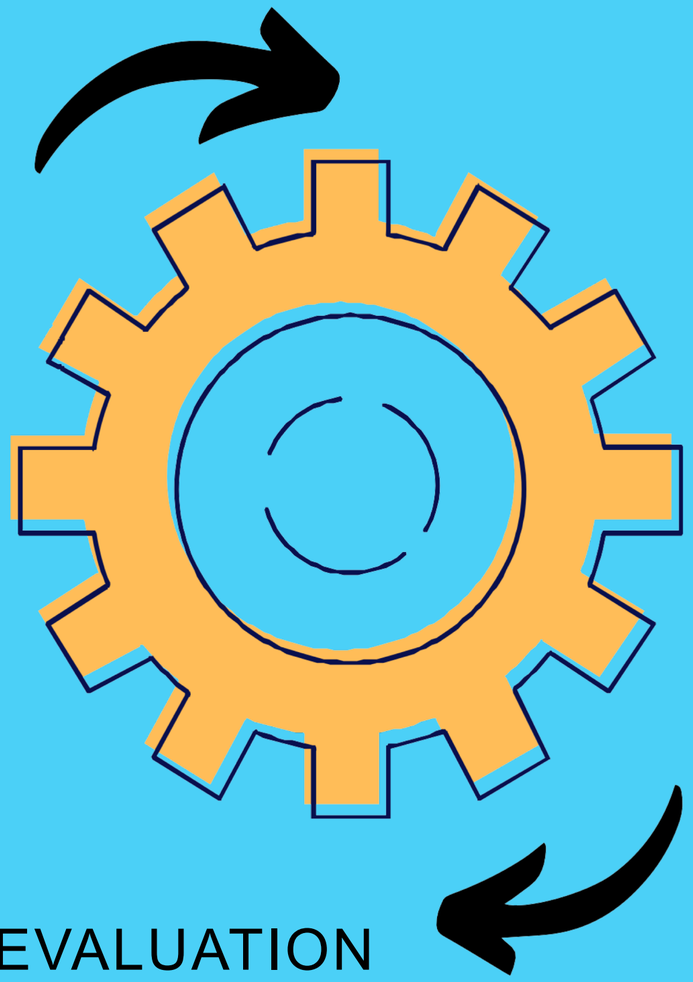
- Verify that the data are labeled correctly
- Evaluate if a sentence is harassing by comparing its content against the three bag-of-words sets
- Neutral label might have some terms from the moderate set or the double meaning set
- Harassing sentences usually contain several abusive words and moderate words
- Bag-of-words updated with new terms and the label is corrected if needed

Core ML Model creation

1. DATA Labeled Tweets

63,000 English

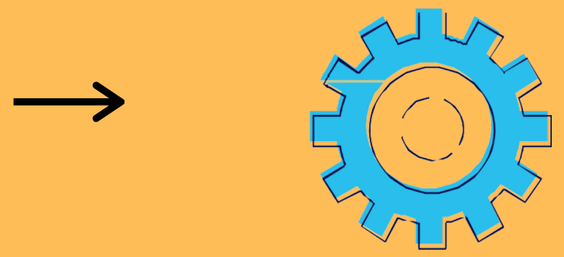
128,000 Italian



2. TRAINING

25 to 29 iterations

3. EVALUATION



accuracy on training set:

99.34% English

98.12 % Italian

Text Classifier Model

Test set

accuracy on test set:

90.21% English

88.56 % Italian

Text Classifier Training & Testing

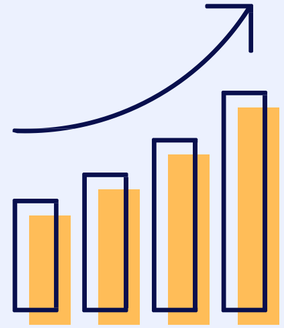
Automatic Feature extraction

entity recognition, script identification, tokenization, lemmatization,
parts-of-speech tagging, and language identification

Model definition with Max Entropy

Choice among Logistic Regression, Nearest Neighbor Classifier,
Support Vector Machines, Boosted Decision Trees, Random Forests.

Text Classifier results



English Data

True\Predict	Harassment	Neutral
Harassment	87 %	13 %
Neutral	2 %	98 %

Auto ML

True\Predict	Harassment	Neutral
Harassment	85.5 %	14.5%
Neutral	8 %	92 %

Core ML

Italian Data

True\Predict	Harassment	Neutral
Harassment	88 %	12 %
Neutral	5 %	95 %

Auto ML

True\Predict	Harassment	Neutral
Harassment	86.5 %	13.5 %
Neutral	9.5 %	90.5 %

Core ML

Testing on Real-Time Data

- Enaëlle
- Testing application
    - Collect tweets from a list of users
        - with a high incidence of harassing terms
        - with a different style of writing
    - The application filters and evaluates tweets from the list with classifiers

Improve the Accuracy

Core ML Text

Auto ML Text

Same results?

YES

NO

Label the content

Label the content

- BoW
- History of the senders
- Followers history
- Results of the models

REPORT

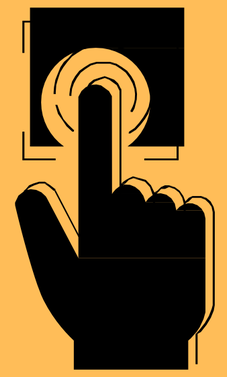
- The report analyzes
  - How the harassment spreads
  - What the harassment topics are
  - What are the emotional sentiment the abusive content carries
  - Geographic location of the senders
- The report is sent to the appropriate responder to resolve the issue
- Request the social media platform to remove the harassing content
- A follow-up checks done to see what action was taken
- Reports are stored
- The report and any follow-ups are sent to the users and their buddy-mentors
- The sender is automatically blocked

OUTGOING DATA

→ The ML classifier is not applied to the outgoing data

→ When the user composes a tweet, it is sent with no alteration

CUSTOM ML MODELS



→ Customized ML classifier models are applied to incoming data on mobile devices and on the Emakia server

→ The classifiers are tuned to the user's definition of what is harassing or not

→ This process is done by retraining the classifier with content specific to the user preferences

Questions?

CORINNE DAVID
Emakia
corinne@emakia.org