ADVANCED CONTENT MODERATOR

Comment

SOCIAL MEDIA COMMENT CLASSIFIER

Presented by Igor Hufnagel

SUMMARY

OVERVIEW

Introduction to Content Moderation

DATA

Dataset Collection, Preprocessing Techniques and Undersampling for Optimal Training.

APPLICATION

Large Language Model (LLMs), Machine Learning (ML) and Web App

CONCLUSION

Discussion and Future Projects

CONTENT MODERATION

- The significance in digital platforms.
- There are many challenges faced in content moderation, such as diverse communication forms and potential misinterpretations.
- It is important to have accurate content assessments for user satisfaction and platform safety.

DATA

DATASETS

- Hate Comment Dataset
- Suicide Watch Dataset
- Sexually Explicit Comments Dataset
- Cyberbullying Dataset
- Twitter Sentiment Analysis Dataset

TECHNIQUES

- Cleaning
- Tokenization
- Stop Words Removal
- Stemming
- Lemmatization
- Undersampling

TESTING

- Linear Regression Model
- Random Forest
- **Support Vector Classifier (SVC)**
- Flask App

RESULTS

Accuracy: 92%

Precision: 92%

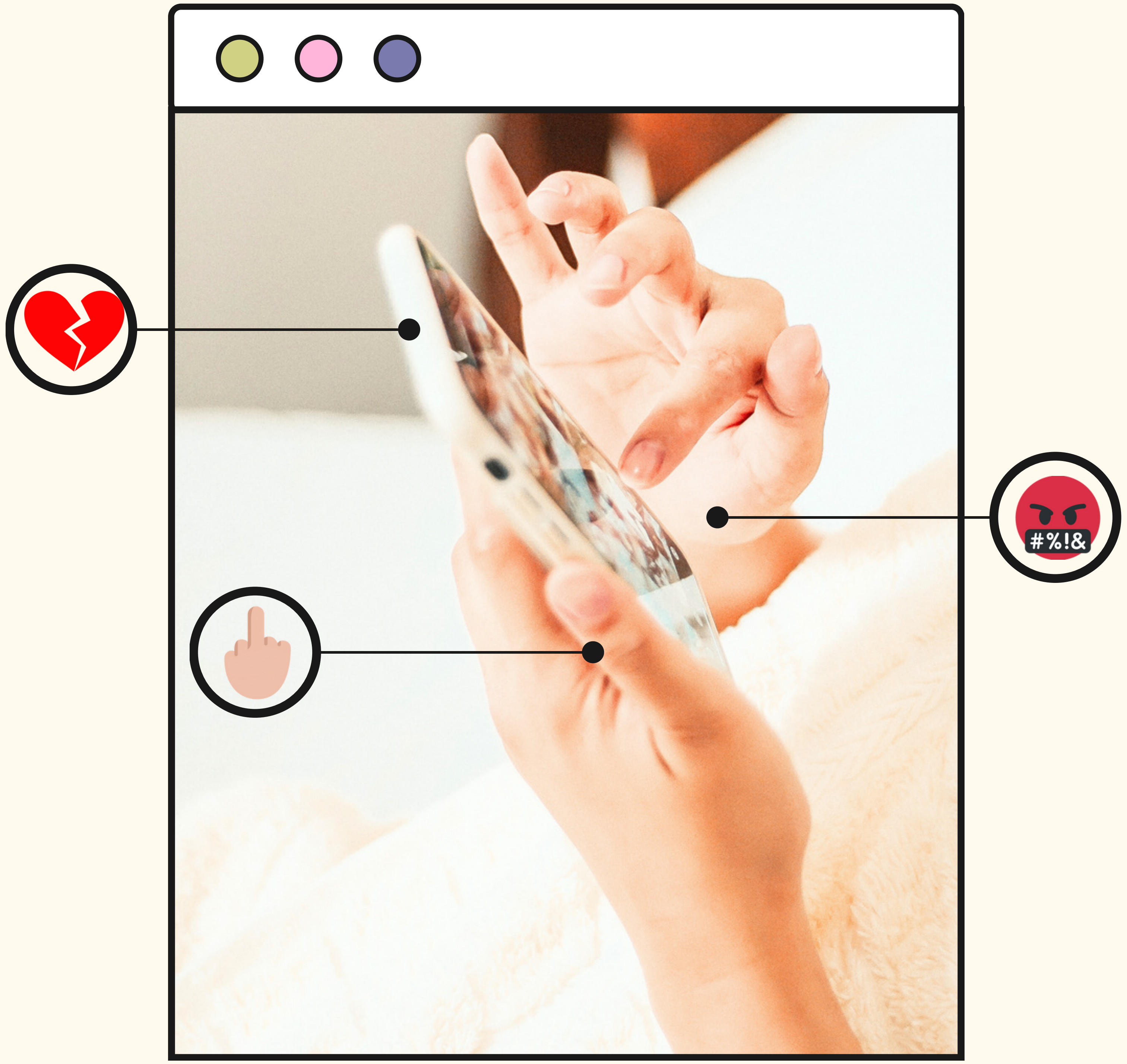
Recall: 92%

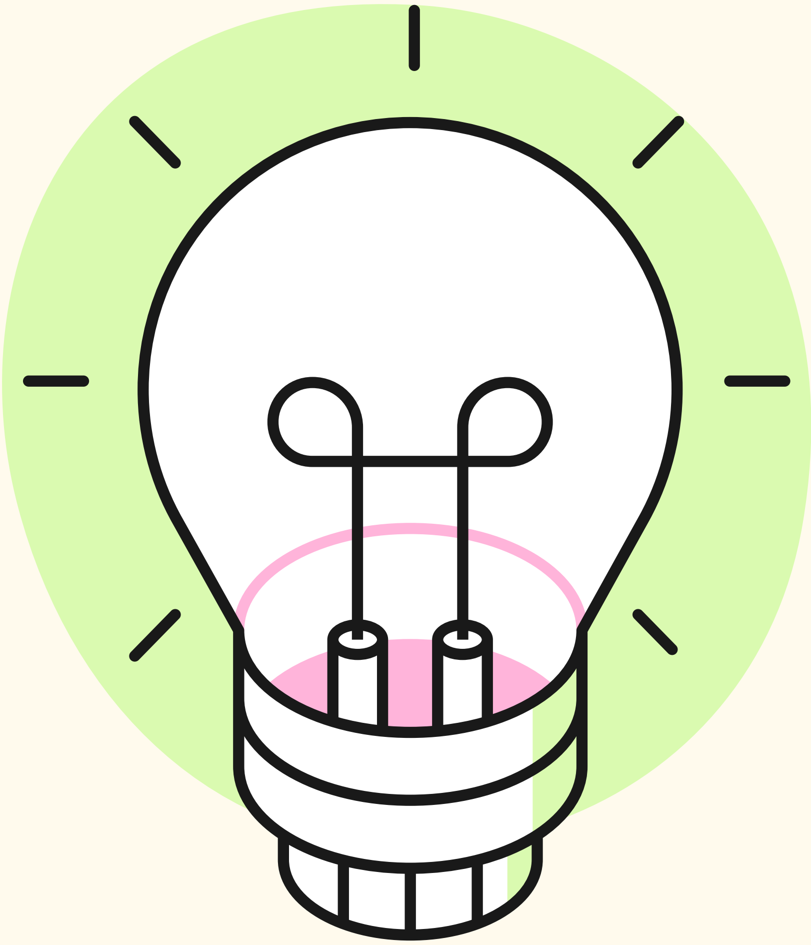
- Accuracy tells how many times the ML model was correct overall.
- Precision is how good the model is at predicting a specific category.
- Recall tells you how many times the model was able to detect a specific category.

COMMENT CLASSIFICATOR

5 (Five) Categories:

- General Guidelines
- Suicide
- Sexually Explicit
- Hate
- Bullying





CONCLUSION AND FUTURE PROJECTS

- Datasets
- Multilingual Content Moderation
- Contextual Analysis
- Real-Time Moderation using Social Media APIs
- Enhanced Content Moderator Interface

THANK YOU

Github

Linkedin