Cyberbullism Detection System

Course Final Project aa:2024/2025

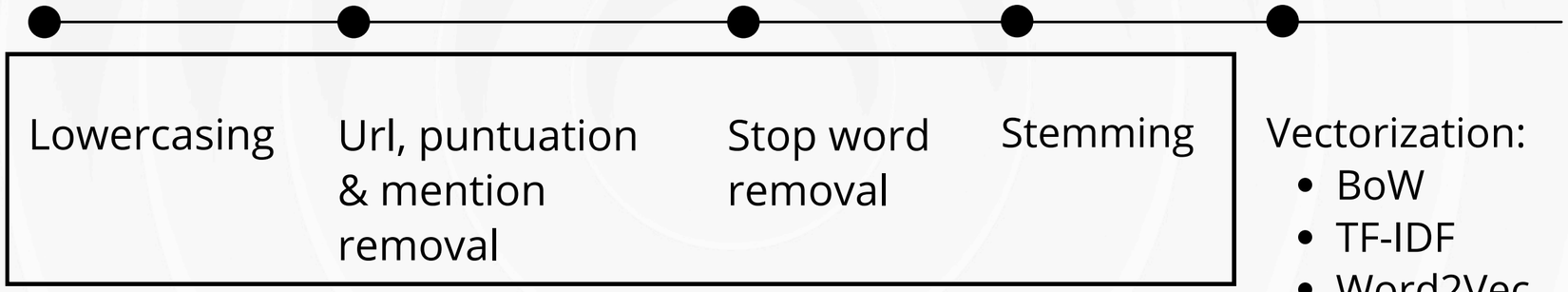
Segreto Mattia

Index

Dataset & Data Visualization

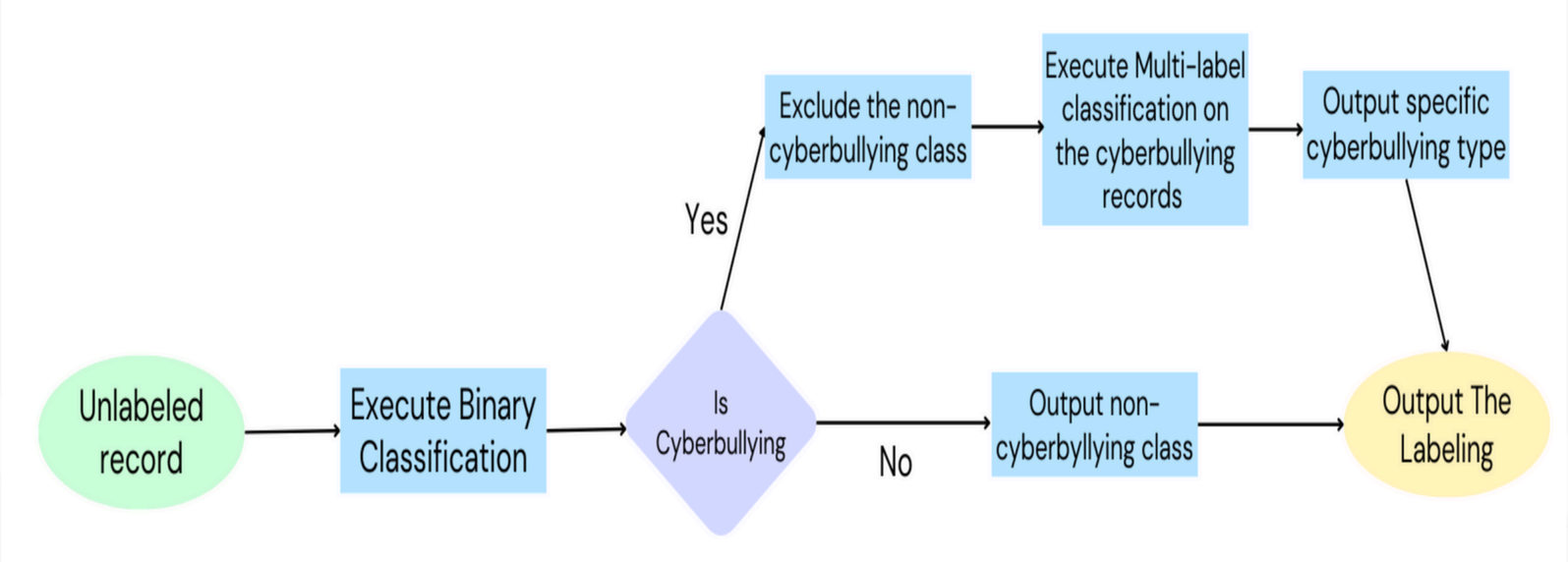
- Labaled Dataset
- Almost **50000** tweets
- Balanced Among Multiclass Label
- Original Feature:
  - **tweet_text**
  - **cyberbullying_type:** it’s the label used for **multiclass classification step**
- Created Feature:
  - **cyberbullying**: it’s a binary label created for **binary classification step**

Data Preprocessing & Vectorization



Preprocessing Phase

Classification Pipeline



Binary Classification

- **LogisticRegression** is the **best** model in this classification step
- There is **not** a significant statistic difference across **Bow** and **TF-IDF** vectorization.

Model Evaluation & Model Comparison Tables

Vectorizer	Classifier	Mean f1_macro	Mean Acc.	Mean Prec.	Mean Rec.	Mean F1
TF-IDF	LogisticRegression	0.726069	0.804865	0.954029	0.804667	0.872959
BoW	LogisticRegression	0.725971	0.813960	0.941532	0.828254	0.881224
TF-IDF	LinearSVM	0.723712	0.800409	0.956956	0.796364	0.869248
BoW	LinearSVM	0.723284	0.808796	0.944441	0.818757	0.877098
BoW	RandomForest	0.717003	0.782376	0.975631	0.757807	0.853000
TF-IDF	RandomForest	0.716418	0.788064			
W2V-1	LogisticRegression	0.688008	0.771394			
W2V-1	LinearSVM	0.685688	0.767620			
W2V-1	RandomForest	0.677147	0.810945			

Model	Victories
LogisticRegression + TF-IDF	7
LogisticRegression + BoW	7
LinearSVM + TF-IDF	3
LinearSVM + BoW	3
RandomForest + BoW	-1
RandomForest + TF-IDF	-1
LogisticRegression + W2V-1	-4
LinearSVM + W2V-1	-6
RandomForest + W2V-1	-8

Multiclass Classification

- All classifiers achieve good results on the evaluated metrics
- **RandomForest + TF-IDF** over perform w.r.t all the other classifiers

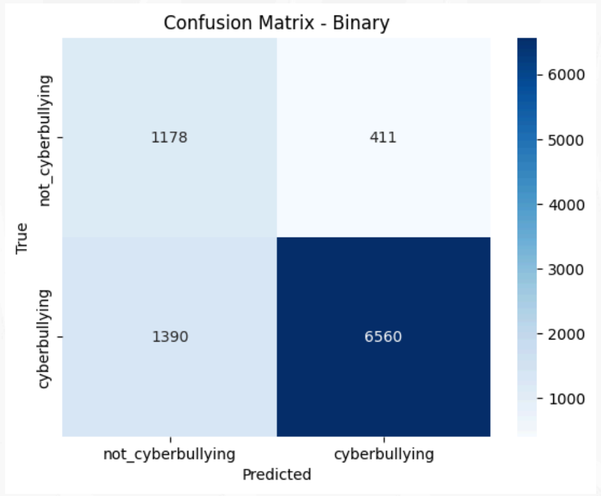
Model Evaluation & Model Comparison Tables

Vectorizer	Classifier	Mean F1	Mean Acc.	Mean Prec.	Mean Rec.
TF-IDF	RandomForest	0.932648	0.932258	0.934340	0.932075
BoW	LogisticRegression	0.928662	0.928075	0.931726	0.927980
BoW	LinearSVM	0.928478	0.927855	0.932123	0.927780
BoW	RandomForest	0.928417	0.928012	0.929886	0.927791
TF-IDF	LogisticRegression	0.927957	0.927383	0.930868	0.927272
TF-IDF	LinearSVM	0.927945	0.927477	0.930539	0.927351
W2V-1	RandomForest	0.892812			
W2V-1	LinearSVM	0.878282			
W2V-1	LogisticRegression	0.878051			

Model	Victories
RandomForest + TF-IDF	8
LogisticRegression + BoW	4
RandomForest + BoW	2
LinearSVM + BoW	2
LinearSVM + TF-IDF	1
LogisticRegression + TF-IDF	1
RandomForest + W2V-1	-4
LinearSVM + W2V-1	-6
LogisticRegression + W2V-1	-8

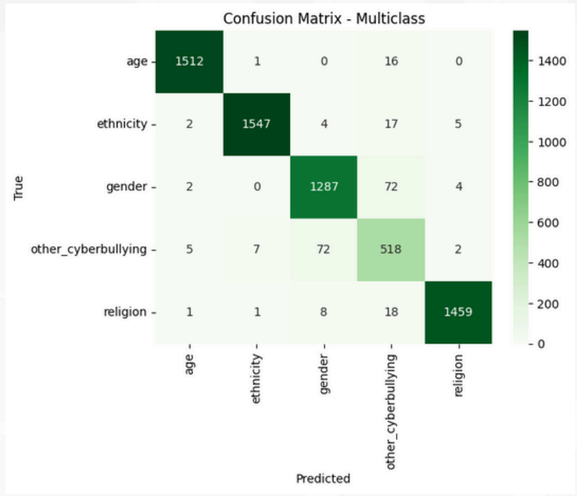
Pipeline Evaluation

Binary Stage:



LogisticRegression+BoW

Multiclass Stage:



Randomforest + TF-IDF

Explainability Module Overview

- Global Explainability
  - Global Classifier Explanation
  - Pattern Mining
- Local Explainability

Pattern Mining

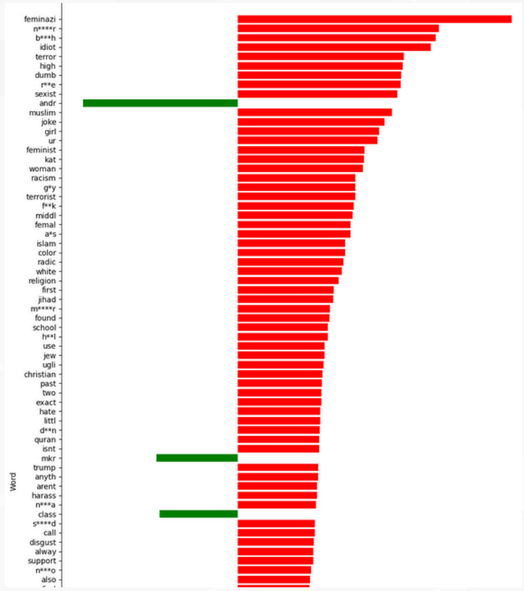
Closed and maximal itemset were analyzed in order to discover if the cyberbullyng phenomenon has **frequent pattern** that shold be useful to discover **interesting informations**

Class	Number of closed itemsets	Number of maximal itemsets
Multiclass Target		
age	147	25
ethnicity	131	25
religion	63	30
gender	62	20
other_cyberbullying	8	8
Binary Target		
cyberbullying	51	23
not_cyberbullying	5	5

Binary Model Global Explanation

LogisticRegression + BoW:

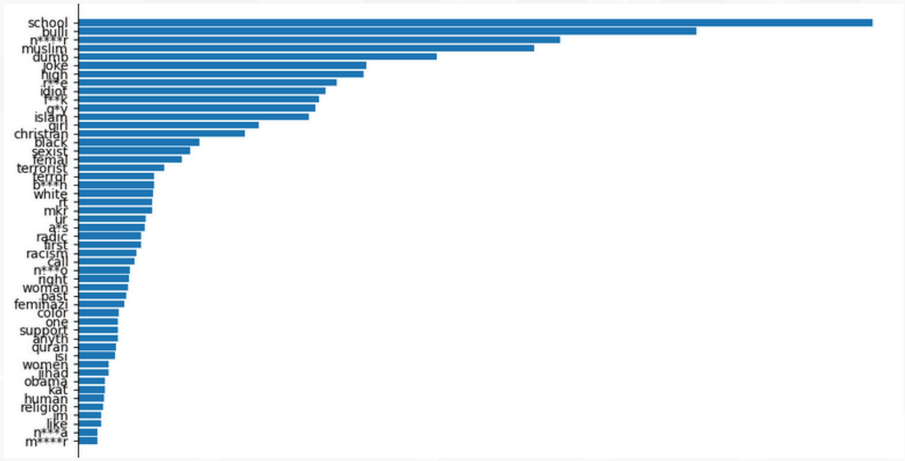
- Coefficient of Logistic Regression has been used(**.coef_**)
- The overall system is recognizing the **cyberbullying** class isted of distinguish the two classes.
- Module can mislead, global trend no.



Multiclass Model Global Explanation

RandomForest + TF-IDF:

- built in method has been used (**.feature_importances_**)
- most frequent terms can be inuitively associated to **harmful content** and specific categories

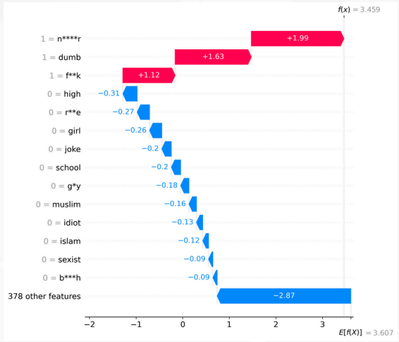


Binary Model Local Explanation

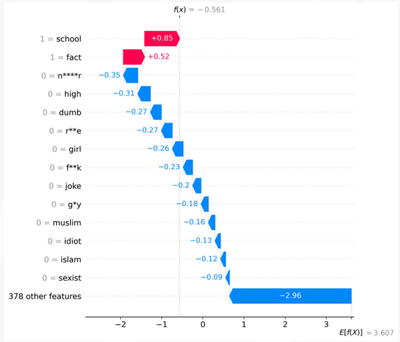
LogisticRegression + BoW:

- **Shap** has been used as post hoc methods
- **positive** coefficients lead towards **cyberbullying**
- **negative** coefficients lead toward **not_cyberbullying**

Cyberbullying:
0.9695 as SHAP prediction



Not_cyberbullying:
0.3633 as SHAP prediction



Multiclass Model Local Explanation

RandomForest + TF-IDF:

- **TreeInterpreter** has been used as post hoc methods for **tree-based** mode
- **positive** coefficients lead towards **predicted label**
- **negative** coefficients lead toward **other label**

feature	contribution	tfidf_value
n****r	0.321969	0.588248
dumb	0.264371	0.596447
f**k	0.171966	0.546091
a*s	-0.014309	0.000000
bulli	0.013783	0.000000
school	0.012158	0.000000
black	-0.006779	0.000000
muslim	0.006439	0.000000
r**e	0.005373	0.000000
joke	0.005160	0.000000
high	0.005145	0.000000
idiot	0.004514	0.000000
obama	-0.004471	0.000000
g*y	0.004391	0.000000

feature	contribution	tfidf_value
sexist	0.435919	0.434186
rt	0.122834	0.324170
im	0.074218	0.316331
woman	0.060055	0.405737
joke	-0.027372	0.000000
r**e	-0.025950	0.000000
christian	0.024148	0.000000
g*y	-0.023147	0.000000
bulli	0.019080	0.000000
right	0.018494	0.000000
school	0.015308	0.000000
hate	-0.013947	0.400914
n****r	0.012817	0.000000
femal	-0.011987	0.000000

References

- Wang, J., Fu, K., & Lu, C. T. (2020). ***"SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection.* 2020 IEEE International Conference on Big Data".** https://ieeexplore.ieee.org/document/9378065
- Ahmadinejad, M., Shahriar, N., & Fan, L(2023). **"Self-Training for Cyberbully Detection: Achieving High Accuracy with a Balanced Multi-Class Dataset".** https://www.proquest.com/openview/2e6b484d78e3a1fe0486ec1217dd574c/1?cbl=18750&diss=y&pq-origsite=gscholar
- Sharma, P., Mirzan, S. R., Bhandari, A., Pimpley, A., Eswaran, A., Srinivasan, S., & Shao, L. (2020). ***"Evaluating Tree Explanation Methods for Anomaly Reasoning: A Case Study of SHAP TreeExplainer and TreeInterpreter"***. https://arxiv.org/abs/2010.06734

Thanks for the attention